NBER WORKING PAPER SERIES

RACIAL INEQUALITY IN THE 21ST CENTURY:
THE DECLINING SIGNIFICANCE OF DISCRIMINATION

Roland G. Fryer, Jr

Working Paper 16256
http://www.nber.org/papers/w16256

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
August 2010

I am enormously grateful to Lawrence Katz, Steven Levitt, Derek Neal, William Julius Wilson and numerous other colleagues whose ideas and collaborative work fill this chapter. Vilsa E. Curto and Meghan L. Howard provided truly exceptional research assistance. Support from the Education Innovation Laboratory at Harvard University (EdLabs), is gratefully acknowledged. Correspondence can be addressed to the author at: Department of Economics, Harvard University, 1805 Cambridge Street, Cambridge MA, 02138. The usual caveat applies. The views expressed herein are those of the author and do not necessarily reflect the views of the National Bureau of Economic Research.

Racial Inequality in the 21st Century: The Declining Significance of Discrimination
Roland G. Fryer, Jr
NBER Working Paper No. 16256
August 2010
JEL No. I20,J01,J15,J71

ABSTRACT

There are large and important differences between blacks and whites in nearly every facet of life - earnings, unemployment, incarceration, health, and so on. This chapter contains three themes. First, relative to the 20th century, the significance of discrimination as an explanation for racial inequality across economic and social indicators has declined. Racial differences in social and economic outcomes are greatly reduced when one accounts for educational achievement; therefore, the new challenge is to understand the obstacles undermining the development of skill in black and Hispanic children in primary and secondary school. Second, analyzing ten large datasets that include children ranging in age from eight months old to seventeen years old, I demonstrate that the racial achievement gap is remarkably robust across time, samples, and particular assessments used. The gap does not exist in the first year of life, but black students fall behind quickly thereafter and observables cannot explain differences between racial groups after kindergarten. Third, we provide a brief history of efforts to close the achievement gap. There are several programs -- various early childhood interventions, more flexibility and stricter accountability for schools, data-driven instruction, smaller class sizes, certain student incentives, and bonuses for effective teachers to teach in high-need schools, which have a positive return on investment, but they cannot close the achievement gap in isolation. More promising are results from a handful of high-performing charter schools, which combine many of the investments above in a comprehensive framework and provide an "existence proof" -- demonstrating that a few simple investments can dramatically increase the achievement of even the poorest minority students. The challenge for the future is to take these examples to scale.

Roland G. Fryer, Jr
Department of Economics
Harvard University
Littauer Center 208
Cambridge, MA 02138
and NBER
rfryer@fas.harvard.edu

An online appendix is available at:
http://www.nber.org/data-appendix/w16256

"In the 21st Century, the best anti-poverty program around is a world-class education."

President Barack Obama, State of the Union Address (January 27, 2010)

1 Introduction

Racial inequality is an American tradition. Relative to whites, blacks earn twenty-four percent less, live five fewer years, and are six times more likely to be incarcerated on a given day. Hispanics earn twenty-five percent less than whites and are three times more likely to incarcerated.[1] At the end of the 1990s, there were one-third more black men under the jurisdiction of the corrections system than there were enrolled in colleges or universities (Ziedenberg and Schiraldi, 2002). While the majority of barometers of economic and social progress have increased substantially since the passing of the civil rights act, large disparities between racial groups have been and continue to be an everyday part of American life.

Understanding the causes of current racial inequality is a subject of intense debate. A wide variety of explanations have been put forth, which range from genetics (Jensen, 1973; Rushton, 1995) to personal and institutional discrimination (Darity and Mason, 1998; Pager, 2007; Krieger and Sidney, 1996) to the cultural backwardness of minority groups (Reuter, 1945; Shukla, 1971). Renowned sociologist William Julius Wilson argues that a potent interaction between poverty and racial discrimination can explain current disparities (Wilson, 2010).

Decomposing the share of inequality attributable to these explanations is exceedingly difficult, as experiments (field, quasi-, or natural) or other means of credible identification are rarely available.[2] Even in cases where experiments are used (i.e., audit studies), it is unclear precisely what is being measured (Heckman, 1998). The lack of success in convincingly identifying root causes of racial inequality has often reduced the debate to a competition of "name that residual" – arbitrarily assigning identity to unexplained differences between racial groups in economic outcomes after accounting for a set of confounding factors. The residuals are often interpreted as "discrimination," "culture," "genetics," and so on. Gaining a better understanding of the root causes of racial inequality is of tremendous importance for social policy, and the purpose of this chapter.

This chapter contains three themes. First, relative to the 20th century, the significance of discrimination as an explanation for racial inequality across economic and social indicators has declined. Racial differences in social and economic outcomes are greatly reduced when one accounts for educational achievement; therefore, the new challenge is to understand the obstacles undermining the achievement of black and Hispanic children in primary and secondary school. Second, analyzing ten large datasets that include children ranging in age from eight months old to seventeen years old, we demonstrate that the racial achievement gap is remarkably robust across time, samples, and particular assessments used. The gap does not exist in the first year of life, but black students fall behind quickly thereafter and observables cannot explain differences between racial groups after kindergarten.

Third, we provide a brief history of efforts to close the achievement gap. There are several programs – various early childhood interventions, more flexibility and stricter accountability for

[1] The Hispanic-white life expectancy gap actually favors Hispanics in the United States. This is often referred to as the "Hispanic Paradox" (Franzini, Ribble, and Keddie, 2001).

[2] List (2005), which examines whether social preferences impact outcomes in the actual market through field experiments in the sportscard market, is a notable exception.

schools, data-driven instruction, smaller class sizes, certain student incentives, and bonuses for effective teachers to teach in high-need schools, which have a positive return on investment, but they cannot close the achievement gap in isolation.[3] More promising are results from a handful of high-performing charter schools, which combine many of the investments above in a comprehensive model and provide a powerful "existence proof" – demonstrating that a few simple investments can dramatically increase the achievement of even the poorest minority students.

An important set of questions is: (1) whether one can boil the success of these charter schools down to a form that can be taken to scale in traditional public schools; (2) whether we can create a competitive market in which only high-quality schools can thrive; and (3) whether alternative reforms can be developed to eliminate achievement gaps. Closing the racial achievement gap has the potential to substantially reduce or eliminate many of the social ills that have plagued minority communities for centuries.

2 The Declining Significance of Discrimination

One of the most important developments in the study of racial inequality has been the quantification of the importance of pre-market skills in explaining differences in labor market outcomes between blacks and whites (Neal and Johnson, 1996; O'Neill, 1990). Using the National Longitudinal Survey of Youth 1979 (NLSY79), a nationally representative sample of 12,686 individuals aged 14 to 22 in 1979, Neal and Johnson (1996) find that educational achievement among 15- to 18-year-olds explains all of the black-white gap in wages among young women and 70 percent of the gap among men. Accounting for pre-market skills also eliminates the Hispanic-white gap. Important critiques such as racial bias in the achievement measure (Darity and Mason, 1998; Jencks, 1998), labor market dropouts, or the potential that forward-looking minorities underinvest in human capital because they anticipate discrimination in the market cannot explain the stark results.[4]

We begin by replicating the seminal work of Neal and Johnson (1996) and extending their work in four directions. First, the most recent cohort of NLSY79 is between 42 and 44 years old (15 years older than in the original analysis), which provides a better representation of the lifetime gap. Second, we perform a similar analysis with the National Longitudinal Survey of Youth 1997 cohort (NLSY97). Third, we extend the set of outcomes to include unemployment, incarceration, and measures of physical health. Fourth, we investigate the importance of pre-market skills among graduates of thirty-four elite colleges and universities in the College and Beyond database, 1976 cohort.

To understand the importance of academic achievement in explaining life outcomes, we follow the lead of Neal and Johnson (1996) and estimate least squares models of the form:

$$outcome_i = \sum_R \beta_R R_i + \Gamma X_i + \varepsilon_i, \tag{1}$$

where i indexes individuals, X_i denotes a set of control variables, and R_i is a full set of racial

[3] For details on the treatment effects of these programs, see Jacob and Ludwig (2008), Guskey and Gates (1985), and Fryer (2010).

[4] Lang and Manove (2006) show that including years of schooling in the Neal and Johnson (1996) specification causes the gap to increase - arguing that when one controls for AFQT performance, blacks have higher educational attainment than whites and that the labor market discriminates against blacks by not financially rewarding them for their greater education.

identifiers.

Table 1 presents racial disparities in wage and unemployment for men and women, separately.[5] The odd-numbered columns present racial differences on our set of outcomes controlling only for age. The even-numbered columns add controls for the Armed Forces Qualifying Test (AFQT) – a measure of educational achievement that has been shown to be racially unbiased (Wigdor and Green, 1991) – and its square. Black men earn 39.4 percent less than white men; black women earn 13.1 percent less than white women. Accounting for educational achievement drastically reduces these inequalities – 39.4 percent to 10.9 percent for black men and 13.1 percent *lower* than whites to 12.7 percent *higher* for black women.[6] An eleven percent difference between white and black men with similar educational achievement is a large and important number, but a small fraction of the original gap. Hispanic men earn 14.8 percent less than whites in the raw data – 62 percent less than the raw black-white gap – which reduces to 3.9 percent more than whites when we account for AFQT. The latter is not statistically significant. Hispanic women earn six percent less than white women (not significant) without accounting for achievement. Adding controls for AFQT, Hispanic women earn sixteen percent *more* than comparable white women and these differences are statistically significant.

Labor force participation follows a similar pattern. Black men are more than twice as likely to be unemployed in the raw data and thirty percent more likely after controlling for AFQT. For women, these differences are 3.8 and 2.9 times more likely, respectively. Hispanic-white differences in unemployment with and without controlling for AFQT are strikingly similar to black-white gaps.

Table 2 replicates Table 1 using the NLSY97.[7] The NLSY97 includes 8,984 youths between the ages of 12 and 16 at the beginning of 1997; these individuals are 21 to 27 years old in 2006-2007, the most recent years for which wage measures are available. In this sample, black men earn 17.9 percent less than white men and black women earn 15.3 percent less than white women. When we account for educational achievement, racial differences in wages measured in the NLSY97 are strikingly similar to those measured in NLSY79 – 10.9 percent for black men and 4.4 percent for black women. The raw gaps, however, are much smaller in the NLSY97, which could be due either to the younger age of the workers and a steeper trajectory for white males (Farber and Gibbons, 1996) or to real gains made by blacks in recent years. After adjusting for age, Hispanic men earn 6.5 percent less than white men and Hispanic women earn 5.7 percent less than white women, but accounting for AFQT eliminates the Hispanic-white gap for both men and women.

Black men in the NLSY97 are almost three times as likely to be unemployed, which reduces to twice as likely when we account for educational achievement. Black women are roughly two and a half times more likely to be unemployed than white women, but controlling for AFQT reduces this gap to seventy-five percent more likely. Hispanic men are twenty-five percent more likely to be unemployed in the raw data, but when we control for AFQT, this difference is eliminated. Hispanic women are fifty percent more likely than white women to be unemployed and this too is eliminated by controlling for AFQT. Similar to the NLSY79, controlling for AFQT has less of an impact on racial differences in unemployment than on wages.

Table 3 employs a Neal and Johnson specification on two social outcomes: incarceration and physical health. The NLSY79 asks the "type of residence" in which the respondent is living during

[5] Summary statistics for NLSY79 are displayed, by race, in Appendix Table 1.

[6] This may be due, in part, to differential selection out of the labor market between black and white women. See Neal (2005) for a detailed account of this.

[7] Summary statistics for NLSY97 are displayed, by race, in Appendix Table 2.

each administration of the survey, which allows us to construct a measure of whether the individual was ever incarcerated when the survey was administered across all years of the sample.[8] The NLSY97 asks individuals if they have been sentenced to jail, an adult corrections institution, or a juvenile corrections institution in the past year for each yearly follow-up survey of participants. In 2006, the NLSY79 included a 12-Item Short Form Health Survey (SF-12) for all individuals over age 40. The SF-12 consists of twelve self-reported health questions ranging from whether the respondent's health limits him from climbing several flights of stairs to how often the respondent has felt calm and peaceful in the past four weeks. The responses to these questions are combined to create physical and mental component summary scores.

Adjusting for age, black males are about three and a half times and Hispanics are about two and a half times more likely to have ever been incarcerated when surveyed.[9] Controlling for AFQT, this is reduced to about eighty percent more likely for blacks and fifty percent more likely for Hispanics. Again, the racial differences in incarceration after controlling for achievement is a large and important number that deserves considerable attention in current discussions of racial inequality in the United States. Yet, the importance of educational achievement in the teenage years in explaining racial differences is no less striking.

The final two columns of Table 3 display estimates from similar regression equations for the SF-12 physical health measure, which has been standardized to have a mean of zero and standard deviation of one for ease of interpretation. Without accounting for achievement, there is a black-white disparity of 0.15 standard deviations in self-reported physical health for men and 0.23 standard deviations for women. For Hispanics, the differences are -0.140 for men and 0.030 for women. Accounting for educational achievement eliminates the gap for men and cuts the gap in half for black women [-0.111 (0.076)]. The remaining difference for black women is not statistically significant. Hispanic women report better health than white women with or without accounting for AFQT.

Extending Neal and Johnson (1996) further, we turn our attention to the College and Beyond (C&B) Database, which contains data on 93,660 full-time students who entered thirty-four elite colleges and universities in the fall of 1951, 1976, or 1989. We focus on the cohort from 1976.[10] The C&B data contain information drawn from students' applications and transcripts, Scholastic Aptitude Test (SAT) and the American College Test (ACT) scores, standardized college admissions exams that are designed to assess a student's readiness for college, as well as information on family demographics and socioeconomic status in their teenage years.[11] The C&B database also includes responses to a survey administered in 1995 or 1996 to all three cohorts that provides detailed

[8] Lochner and Moretti (2004) use a similar approach to determine incarceration rates, using type of residence in Census data and in the NLSY79.

[9] We focus on the estimates from NLSY79 because we have many more years of observations for these individuals than for those in the NLSY97, which gives us a more accurate picture of incarceration.

[10] There are two reasons for this. First, the 1976 College & Beyond cohort can be reasonably compared to the NLSY79 cohort because they are all born within a seven-year period. Second, there are issues with using either the 1951 or the 1989 data. The 1951 cohort presents issues of selection bias - black students who entered top colleges in this year were too few in number and those who did were likely to be incredibly motivated and intelligent students, in comparison to both their non-college-going black peers and their white classmates. The 1989 cohort is problematic because the available wage data for that cohort was obtained when that cohort was still quite young. Wage variance is likely to increase a great deal beyond the levels observed in the available wage data. Additionally, some individuals who have high expected earnings were pursuing graduate degrees at the time wage data were gathered, artificially depressing their observed wages.

[11] Ninety-two percent of the sample has valid SAT scores.

information on post-college labor market outcomes. Wage data were collected when the respondents were approximately 38 years old, and reported as a series of ranges. We assigned individuals the midpoint value of their reported income range as their annual income.[12] The response rate to the 1996 survey was approximately 80 percent. Appendix Table 3 contains summary statistics used in our analysis.

Table 4 presents racial disparities in income for men and women from the 1976 cohort of the C&B Database.[13] The odd-numbered columns present raw racial differences. The even-numbered columns add controls for performance on the SAT and its square.[14] Black men from this sample earn 27.3 percent less than white men, but when we account for educational achievement, the gap shrinks to 15.2 percent. Black women earn more than white women by 18.6 percent, which increases to an advantage of 28.6 percent when accounting for SAT scores. There are no differences in income between Hispanics and whites with or without accounting for achievement. Hispanic men earn 3.8 percent less than similarly aged white men (not statistically significant) and one percent less when one accounts for pre-college scores.

In developing countries, eradicating poverty takes a large and diverse set of strategies: battling disease, fighting corruption, building schools, providing clean water, and so on (Schultz and Strauss, 2008). In the United States, important progress toward racial equality can be made if one ensures that black and white children obtain the same skills. This is an enormous improvement over the battles for basic access and equality that were fought in the 20th century, but we must now work to close the racial achievement gaps in education – high-quality education is the new civil rights battleground.[15]

3 Basic Facts About Racial Differences in Achievement Before Kids Enter School

We begin our exploration of the racial achievement gap with data on mental function in the first year of life. This approach has two virtues. First, nine months is one of the earliest ages at which one can reliably test cognitive achievement in infants. Second, data on the first year of life provide us with a rare opportunity to potentially understand whether genetics is an important factor in explaining racial differences later in life.[16]

[12] Individuals in the wage range "less than $1000" are excluded from the analysis as they cannot have made this wage as full-time workers and therefore should not be compared to the rest of the sample.

[13] A measure of current unemployment for the individuals surveyed was also created. However, only 39 out of 19,257 with valid answers as to employment status could be classified as unemployed, making an analysis of unemployment by race infeasible. Although 1,876 reported that they were not currently working for reasons other than retirement, the vast majority of these individuals were out of the labor force rather than unemployed. More details on this variable can be found in the data appendix.

[14] The SAT is presently called the SAT Reasoning Test and the letters "SAT" no longer stand for anything. At the time these SAT scores were gathered, however, the test was officially called the "Scholastic Aptitude Test" and was believed to function as a valid intelligence test. The test also had a substantially different format and included a different range of question types.

[15] This argument requires an important leap of faith. We have demonstrated that educational achievement is correlated with better economic and social outcomes, but we have not proven that this relationship is causal. We will come back to this in the conclusion.

[16] Some scholars have argued that the combination of high heritability of innate ability (typically above 0.6 for adults, but somewhat lower for children, e.g., Neisser et al. (1996) or Plomin et al. (2000), and persistent racial gaps in test scores is evidence of genetic differences across races (Jensen, 1973, 1998; Rushton and Jensen, 2005).

There are only two datasets that are both nationally representative and contain assessments of mental function before the first year of life. The first is the U.S. Collaborative Perinatal Project (CPP) (Bayley, 1965), which includes over 31,000 women who gave birth in twelve medical centers between 1959 and 1965. The second dataset is the Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), a nationally representative sample with measures of mental functioning (a shortened version of the Bayley Scale of Infant Development) for over 10,000 children aged one and under. Summary statistics for the variables we use in our core specifications are displayed by race in Appendix Tables 4 (CPP) and 5 (ECLS-B).

Figures 1 and 2 plot the density of mental test scores by race at various ages in the ECLS-B and CPP data sets, respectively.[17] In Figure 1, the test score distributions on the Bayley Scale at age nine months for children of different races are visually indistinguishable. By age two, the white distribution has demonstrably shifted to the right. At age four, the cognitive score is separated into two components: literacy (which measures early language and literacy skills) and math (which measures early mathematics skills and math readiness). Gaps in literacy are similar to disparities at age two; early math skills differences are more pronounced. Figure 2 shows a similar pattern using the CPP data. At age eight months, all races look similar. By age four, whites are far ahead of blacks and Hispanics and these differences continue to grow over time. Figures 1 and 2 make one of the key points of this section: the commonly observed racial achievement gap only emerges after the first year of life.

To get a better sense of the magnitude (and standard errors) of the change from nine months to seven years old, we estimate least squares models of the following form:

$$outcome_{i,a} = \sum_{R} \beta_R R_i + \Gamma X_i + \varepsilon_{i,a} \tag{2}$$

where i indexes individuals, a indexes age in years, and R_i corresponds to the racial group to which an individual belongs. The vector X_i captures a wide range of possible control variables including demographics, home and prenatal environment; $\varepsilon_{i,a}$ is an error term. The variables in the ECLS-B and CPP datasets are similar, but with some important differences.[18] In the ECLS-B dataset, demographic variables include the gender of the child, the age of the child at the time of assessment (in months), and the region of the country in which the child lives. Home environment variables include a single socioeconomic status measure (by quintile), the mother's age, the number of siblings, and the family structure (child lives with: "two biological parents," "one biological parent,"and so on). There is also a "parent as teacher" variable included in the home environment variables. The"parent as teacher" score is coded based on interviewer observations of parent-child interactions in a structured problem-solving environment and is based on the Nursing Child Assessment Teaching Scale (NCATS). Our set of prenatal environment controls include: the birthweight of the child (in 1000-gram ranges), the amount premature that the child was born (in 7-day ranges), and a set of dummy variables representing whether the child was a single birth, a twin, or one in a birth of three or more.

In the CPP dataset, demographic variables include the age of the child at the time of assessment

As Nisbett (1998) and Phillips et al. (1998) argue, however, the fact that blacks, whites, and Asians grow up in systematically different physical and social environments makes it difficult to draw strong, causal, genetically-based conclusions.

[17] This analysis is a replication and extension of Bayley (1965) and Fryer and Levitt (2004).

[18] For more information on the coding of these variables, see the data appendix.

(in months) and the gender of the child. Our set of home environment variables provides rich proxies of the environment in which children were reared. The set of home variables includes: parental education (both mother's and father's, which have been transformed to dichotomous variables ranging from "high school dropout" to "college degree or more"), parental occupation (a set of mutually exclusive and collectively exhaustive dummy variables: "no occupation,""professional occupation,"or "non-professional occupation"), household income during the first three months of pregnancy (in $500 ranges), mother's age, number of siblings, and each mother's reaction to and interactions with the child, which are assessed by the interviewer (we indicate whether a mother is indifferent, accepting, attentive, over-caring, or if she behaves in another manner). The set of prenatal environment controls for the CPP is the same as the set of prenatal environment controls in the ECLS-B dataset. Also included in the analysis of both datasets is interviewer fixed effects, which adjust for any mean differences in scoring of the test across interviewers.[19] It is important to stress that a causal interpretation of the coefficients on the covariates is likely to be inappropriate; we view these particular variables as proxies for a broader set of environmental and behavioral factors.

The coefficients on the race variables across the first three waves of ECLS-B and CPP datasets are presented in Table 5. The omitted race category is non-Hispanic white, so the other race coefficients are relative to that omitted group. Each column reflects a different regression and potentially a different dataset. The odd-numbered columns have no controls. The even-numbered columns control for interviewer fixed effects, age at which the test was administered, the gender of the child, region, socioeconomic status, variables to proxy for a child's home environment (family structure, mother's age, number of siblings, and parent-as-teacher measure) and prenatal condition (birth weight, premature birth, and multiple births).[20] Even-numbered columns for CPP data omit region and the parent-as-teacher measure, which are unique to ECLS-B.[21]

In infancy, blacks lag whites by 0.077 (0.031) standard deviations in the raw ECLS-B data. Hispanics and Asians also slightly trail whites by 0.025 (0.029) and 0.027 (0.040), respectively. Adding our set of controls eliminates these trivial differences. The patterns in the CPP data are strikingly similar. Yet, raw gaps of almost 0.4 standard deviations between blacks and whites are present on the test of mental function in the ECLS-B at age two. Even after including extensive controls, a black-white gap of 0.219 (0.036) standard deviations remains. Hispanics look similar to blacks. Asians lag whites by a smaller margin than blacks or Hispanics in the raw data but after including controls they are the worst-performing ethnic group. By age four, a large test score gap has emerged for blacks and Hispanics in both datasets – but especially in the CPP. In the raw CPP data, blacks lag whites by almost 0.8 standard deviations and Hispanics fare even worse. The inclusion of controls reduces the gap to roughly 0.3 standard deviations for blacks and 0.5 standard deviations for Hispanics. In the ECLS-B, black math scores trail white scores by 0.337 (0.032) in

[19] In ECLS, each of the 13 regions was staffed by one field supervisor and between 14 and 19 interviewers, for a total of 256 field staff (243 interviewers), who conducted an average of 42 child assessments each. The number of interviews per interviewer ranges from 1 to 156. Almost all interviewers assessed children from different races (Bethel et al., 2004). There are 184 interviewers in CPP for eight-month-olds, 305 for four-year-olds, and 217 for seven-year-olds. In the CPP, there are many interviewers for whom virtually all of the children assessed were of the same race.

[20] Because the age at which the test is taken is such an important determinant of test performance, we include separate indicators for months of age in our specification.

[21] It should also be noted that in the CPP dataset, there is not a single SES measure, but the set of variables including parental education, parental occupation, and family income provides a rich proxy for socioeconomic status.

the raw data and trail by 0.130 (0.036) with controls. Black-white differences in literacy are -0.195 (0.031) without controls and 0.020 (0.035) with controls. The identical estimates for Hispanics are -0.311 (0.029) and -0.174 (0.034) in math; -0.293 (0.028) and -0.103 (0.033) in literacy. Asians are the highest-performing ethnic group in both subjects on the age four tests. Racial disparities at age seven, available only in CPP, are generally similar to those at age four.

There are at least three possible explanations for the emergence of racial differences with age. The first is that the skills tested in one-year-olds are not the same as those required of older children, and there are innate racial differences only in the skills that are acquired later. For instance, an infant scores high if she babbles expressively or looks around to find the source of the noise when a bell rings, while older children are tested directly on verbal skills and puzzle-solving ability. Despite these clear differences in the particular tasks undertaken, the outcomes of these early and subsequent tests are correlated by about 0.30, suggesting that they are, to some degree, measuring a persistent aspect of a child's ability.[22] Also relevant is the fact that the Bayley Scales of Infant Development (BSID) score is nearly as highly correlated with measures of parental IQ as childhood aptitude tests.

Racial differences in rates of development are a second possible explanation for the patterns in our data. If black infants mature earlier than whites, then black performance on early tests may be artificially inflated relative to their long-term levels. On the other hand, if blacks are less likely to be cognitively stimulated at home or more likely to be reared in environments that Shonkoff (2006) would label as characterized by "toxic stress," disruptions in brain development may occur which may significantly retard cognitive growth.

A third possible explanation for the emerging pattern of racial gaps is that the relative importance of genes and environmental factors in determining test outcomes varies over time. In contrast to the first two explanations mentioned above, under this interpretation, the measured differences in test scores are real, and the challenge is to construct a model that can explain the racial divergence in test scores with age.

To better understand the third explanation, Fryer and Levitt (forthcoming-b) provide two statistical models that are consistent with the data presented above. Here we provide a brief overview of the models and their predictions.

The first parameter of interest is the correlation between test scores early on and later in life. Fryer and Levitt (forthcoming-b) assign a value of 0.30 to that correlation. The measured correlation between test scores early and late in life and parental test scores is also necessary for the analysis. Based on prior research (e.g., Yeates et al., 1983), we take these two correlations as 0.36 and 0.39, respectively.[23] The estimated black-white test score gap at young ages is taken as 0.077 based on our findings in ECLS-B, compared to a gap of 0.78 at later ages based on our findings in CPP.

The primary puzzle raised by our results is the following: how does one explain small racial gaps on the BSID test scores administered at ages 8 to 12 months and large racial gaps in tests of mental ability later in life, despite the fact that these two test scores are reasonably highly correlated with one another ($\rho = 0.3$), and both test scores are similarly correlated with parental

[22] Nonetheless, Lewis and McGurk (1972) are pessimistic about the generalizability of these infant test scores. Work focusing on infant attention and habituation is also predictive of future test scores (e.g., Bornstein and Sigman, 1986; McCall and Carriger, 1993), but unfortunately our data do not include such information.

[23] It is important to note that substantial uncertainty underlies these correlations, which are based on a small number of studies carried out on a non-representative sample.

test scores ($\rho \geq 0.3$)?

The Basic Building Blocks

Let θ_a denote the measured test score of an individual at age a. We assume that test scores are influenced by an individual's genetic make-up (G) and his environment (E_a) at age a. The simplest version of the canonical model of genes and environment takes the following form:

$$\theta_a = \alpha_a G + \beta_a E_a + \varepsilon_a \tag{3}$$

In this model, the individual's genetic endowment is fixed over time, but environmental factors vary and their influence may vary. θ_a, G, and E_a are all normalized into standard deviation units. Initially we will assume that G, E_a, and ϵ_a are uncorrelated for an individual at any point in time (this assumption will be relaxed below), and that E_a and the error terms for an individual at different ages are also uncorrelated.[24] There will, however, be a positive correlation between an individual's genetic endowment G and the genetic endowment of his or her mother (which we denote G_m). We will further assume, in accord with the simplest models of genetic transmission, that the correlation between G and G_m is 0.50.[25]

We are interested in matching two different aspects of the data: (1) correlations between test scores, and (2) racial test score gaps at different ages. The test score correlations of interest are those of an individual at the age of one (for which we use the subscript b for baby) and later in childhood (denoted with subscript c).

Under the assumptions above, these correlations are as follows:

$$corr(\theta_b, \theta_m) = 0.5\alpha_b\alpha_m = 0.36 \tag{4}$$

$$corr(\theta_c, \theta_m) = 0.5\alpha_c\alpha_m = 0.39 \tag{5}$$

$$corr(\theta_b, \theta_c) = \alpha_b\alpha_c = 0.30 \tag{6}$$

where the 0.5 in the first two equations reflects the assumed genetic correlation between mother and child, and the values 0.36, 0.39, and 0.30 are our best estimates of the empirical values of these correlations based on past research cited above.

The racial test score gaps in this model are given by:

$$\Delta\theta_b = \alpha_b\Delta G + \beta_b\Delta E_b = 0.077 \tag{7}$$

$$\Delta\theta_c = \alpha_c\Delta G + \beta_c\Delta E_c = 0.854 \tag{8}$$

where the symbol Δ in front of a variable signifies the mean racial gap between blacks and whites for that variable. The values 0.077 and 0.854 represent our estimates of the black-white test score gap at ages nine months and seven years from Table 5.[26] For Hispanics, these differences are 0.025

[24] Allowing for an individual's environment to be positively correlated at different points in time causes this simple model to show even greater divergence from what is observed in the data. We relax the assumption that environment is not correlated across ages for an individual when we introduce a correlation between parental test scores and the child's environment below.

[25] As noted below, factors such as assortative mating can cause that correlation to be higher.

[26] Note that the racial gap at age seven is based on earlier CPP data. The evidence suggests that racial gaps have diminished over time (Dickens and Flynn, 2006). Thus, a value of 0.854 in equation (7) may be too large. The only implication this has for solving our model is to reduce the black-white differences in environment that are necessary

and 0.846, respectively.

Solving Equations (4)-(6), this simple model yields a value of 1.87 for α_m^2. Under the assumptions of the model, however, the squared value of the coefficients α and β represent the share of the variance in the measured test score explained by genetic and environmental factors, respectively, meaning that α_m^2 is bounded at one. Thus, this simple model is not consistent with the observed correlations in the data. The correlation between child and mother test scores observed in the data is too large relative to the correlation between the child's own test scores at different ages.

Consequently, we consider two extensions to this simple model that can reproduce these correlations in the data: assortative mating and allowing for a mother's test score to influence the child's environment.[27]

Assortative mating

If women with high G mate with men who also have high G, then the parent child $corr(G, G_m)$ is likely to exceed 0.50. Assuming a value of $\alpha_m^2 = 0.80$, which is consistent with prior research, the necessary $corr(G, G_m)$ to solve the system of equations above is roughly 0.76, which requires the correlation between parents on G to be around 0.50, not far from the 0.45 value reported for that coefficient in a literature review (Jensen, 1978).[28] With that degree of assortative mating, the other parameters that emerge from the model are $\alpha_b = 0.53$ and $\alpha_c = 0.57$. Using these values of α_b and α_c, it is possible to generate the observed racial gaps in (7) and (8). If we assume as an upper bound that environments for black and Hispanic babies are the same as those for white babies (i.e., $\Delta E_b = 0$) in Equation (7), then the implied racial gap in G is a modest 0.145 standard deviations for blacks and 0.04 for Hispanics[29].

To fit Equation (8) requires $\beta_c \Delta E_c = 0.77$. If $\beta_c = 0.77$ (implying that environmental factors explain about half of the variance in test scores), then a one standard deviation gap in environment between black and white children and a 1.14 standard deviation gap between Hispanic and white children would be needed to generate the observed childhood racial test score gap[30]. If environmental factors explain less of the variance, a larger racial gap in environment would be needed. Taking a simple non-weighted average across environmental proxies available in the ECLS yields a 1.2 standard deviation gap between blacks and whites[31].

to close the model. We use the raw racial gaps in this analysis, rather than the estimates controlling for covariates, because our goal in this section is to decompose the differences into those driven by genes versus environments. Many of the covariates included in our specifications could be operating through either of those channels.

[27] A third class of models that we explored has multiple dimensions of intelligence (e.g., lower-order and higher-order thinking) that are weighted differently by tests administered to babies versus older children. We have not been able to make such a model consistent with the observed correlations without introducing either assortative mating or allowing the mother's test score to influence the child's environment.

[28] The correlation of 0.5 can be derived as follows. Let $G = 0.5G(M) + 0.5G(F)$. Taking the correlation of both sides with respect to $G(M)$ and assuming unit variance, $corr(G, G(M)) = 0.76$ only if $corr(G(M), G(F)) = 0.5$.

[29] Allowing black babies to have worse environments makes the implied racial gap in G even smaller.

[30] Estimates from Fryer and Levitt (2004) on racial differences in achievement when black, white, Asian, and Hispanic students enter kindergarten, along with the assortative mating model above, imply that even smaller differences in environment explain later test scores.

[31] Fryer and Levitt (2004) find a 0.75 standard deviation difference between blacks and whites in socioeconomic status, a 0.83 standard deviation gap in the number of children's books in the home, a 1.30 standard deviation difference in female-headed households, a 1.51 standard deviation difference in whether or not one feels safe in their neighborhood, a 1.5 standard deviation difference in the percentage of kids in their school who participate in the free lunch program, and a 1.31 difference in the amount of loitering reported around the school by non-students. All estimates are derived by taking the difference in the mean of a variable between blacks and whites and dividing by

Allowing parental test scores to influence the child's environment

A second class of model consistent with our empirical findings is one in which the child's environment is influenced by the parent's test score, as in Dickens and Flynn (2001). One example of such a model would be

$$\theta_a = \alpha_a G + \beta_a E_a(\theta_m, \tilde{E_a}) + \varepsilon_a \tag{9}$$

where Equation (9) differs from the original Equation (3) by allowing the child's environment to be a function of the mother's test score, as well as factors $\tilde{E_a}$ that are uncorrelated with the mother's test score. In addition, we relax the earlier assumption that the environments an individual experiences as a baby and as a child are uncorrelated. We do not, however, allow for assortative mating in this model. Under these assumptions, Equation (9) produces the following three equations for our three key test score correlations

$$corr(\theta_b, \theta_m) = 0.5\alpha_b\alpha_m + \beta_b cov(E_b, \theta_m) = 0.36 \tag{10}$$

$$corr(\theta_c, \theta_m) = 0.5\alpha_c\alpha_m + \beta_c cov(E_c, \theta_m) = 0.39 \tag{11}$$

$$corr(\theta_b, \theta_c) = \alpha_b\alpha_c + \beta_b\beta_c cov(E_b, E_c) = 0.30 \tag{12}$$

Allowing parental ability to influence the child's environment introduces extra degrees of freedom; indeed, this model is so flexible that it can match the data both under the assumption of very small and large racial differences in G (e.g., $\Delta G \leq 1$ standard deviation). In order for our findings to be consistent with small racial differences in G, the importance of environmental factors must start low and grow sharply with age. In the most extreme case (where environment has no influence early in life: $\beta_b = 0$), solving Equations (10) and (12) implies $\alpha_b = 0.80$ and $\alpha_c = 0.37$. If $\beta_c = 0.77$ (as in the assortative mating model discussed above), then a correlation of 0.29 between the mother's test score and the child's environment is necessary to solve Equation (11). The mean racial gap in G implied by Equation (7) is 0.096 standard deviations. To match the test score gap for children requires a mean racial difference in environmental factors of approximately one standard deviation.

A model in which parents' scores influence their offspring's environment is, however, equally consistent with mean racial gaps in G of one standard deviation. For this to occur, G must exert little influence on the baby's test score, but be an important determinant of the test scores of children. Take the most extreme case in which G has no influence on the baby's score (i.e., $\alpha_b = 0$). If genetic factors are not directly determining the baby's test outcomes, then environmental factors must be important. Assuming $\beta_b = 0.80$, Equation (10) implies a correlation between the mother's test score and the baby's environment of 0.45. If we assume that the correlation between the baby's environment and the child's environment is 0.70, then Equation (12) implies a value of $\beta_c = 0.54$. If we maintain the earlier assumption of $\alpha_m^2 = 0.80$, as well as a correlation between the mother's test score and the child's environment of 0.32, then a value of $\alpha_c = 0.49$ is required to close the model. If there is a racial gap of one standard deviation in G, then Equations (7) and (8) imply 0.096 and 0.67 standard deviation racial gaps in environment factors for babies and children, respectively, to fit our data.

Putting the pieces together, the above analysis shows that the simplest genetic models are not consistent with the evidence presented on racial differences in the cognitive ability of infants. These

the standard deviation for whites. The socioeconomic composite measure contains parental income, education, and occupation.

inconsistencies can be resolved in two ways: incorporating assortative mating or allowing parental ability to affect the offspring's environment. With assortative mating, our data imply a minimal racial gap in intelligence (0.11 standard deviations as an upper bound), but a large racial gap in environmental factors. When parent's ability influences the child's environment, our results can be made consistent with almost any value for a racial gap in G (from roughly zero to a full standard deviation), depending on the other assumptions that are made. Thus, despite stark empirical findings, our data cannot resolve these difficult questions – much depends on the underlying model.

4 Interventions to Foster Human Capital Before Children Enter School

In the past five decades there have been many attempts to close the racial achievement gap before kids enter school.[32] Table 6 provides an overview of twenty well-known programs, the ages they serve, and their treatment effects (in the cases in which they have been credibly evaluated).

Perhaps the most famous early intervention program for children involved 64 students in Ypsilanti, Michigan, who attended the Perry Preschool program in 1962. The program consisted of a 2.5-hour daily preschool program and weekly home visits by teachers, and targeted children from disadvantaged socioeconomic backgrounds with IQ scores in the range of 70-85. An active learning curriculum - High/Scope - was used in the preschool program in order to support both the cognitive and non-cognitive development of the children over the course of two years beginning when the children were three years old. Schweinhart, Barnes, and Weikart (1993) find that students in the Perry Preschool program had higher test scores between the ages of 5 and 27, 21 percent less grade retention or special services required, 21 percent higher graduation rates, and half the number of lifetime arrests in comparison to children in the control group. Considering the financial benefits that are associated with the positive outcomes of the Perry Preschool, Heckman et al. (2009) estimated that the rate of return on the program is between 7 and 10 percent, passing a cost-benefit analysis.

Another important intervention, which was initiated three years after the Perry Preschool program is Head Start. Head Start is a preschool program funded by federal matching grants that is designed to serve 3- to 5-year-old children living at or below the federal poverty level.[33] The program varies across states in terms of the scope of services provided, with some centers providing full-day programs and others only half-day. In 2007, Head Start served over 900,000 children at an average annual cost of about $7,300 per child.

Evaluations of Head Start have often been difficult to perform due to the non-random nature of enrollment in the program. Currie and Thomas (1995) use a national sample of children and compare children who attended a Head Start program with siblings who did not attend Head Start, based on the assumption that examining effects within the family unit will reduce selection bias. They find that those children who attended Head Start scored higher on preschool vocabulary tests but that for black students, these gains were lost by age ten. Using the same analysis method with updated data, Garces, Thomas, and Currie (2002) find several positive outcomes associated with Head Start attendance. They conclude that there is a positive effect from Head Start on the

[32] See Carneiro and Heckman (2003) for a nice review of policies to foster human capital.

[33] Local Head Start agencies are able to extend coverage to those meeting other eligibility criteria, such as those with disabilities and those whose families report income between 100 and 130 percent of the federal poverty level.

probability of attending college and - for whites - the probability of graduating from high school. For black children, Head Start led to a lower likelihood of being arrested or charged with a crime later in life.

Puma et al. (2005), in response to the 1998 reauthorization of Head Start, conduct an evaluation using randomized admission into Head Start.[34] The impact of being offered admission into Head Start for three and four year olds is 0.10 to 0.34 standard deviations in the areas of early language and literacy. For 3-year-olds, there were also small positive effects in the social-emotional domain (0.13 to 0.18 standard deviations) and on overall health status (0.12 standard deviations). Yet, by the time the children who received Head Start services have completed first grade, almost all of the positive impact on initial school readiness has faded. The only remaining impacts in the cognitive domain are a 0.08 standard deviation increase in oral comprehension for 3-year-old participants and a 0.09 standard deviation increase in receptive vocabulary for the 4-year-old cohort (Puma et al., 2010).[35]

A third, and categorically different, program is the Nurse Family Partnership. Through this program, low-income first-time mothers receive home visits from a registered nurse beginning early in the pregnancy that continue until the child is two years old – a total of fifty visits over the first two years. The program aims to encourage preventive health practices, reduce risky health behaviors, foster positive parenting practices, and improve the economic self-sufficiency of the family. In a study of the program in Denver in 1994-95, Olds et al. (2002) find that those children whose mothers had received home visits from nurses (but not those who received home visits from paraprofessionals) were less likely to display language delays and had superior mental development at age two. In a long-term evaluation of the program, Olds et al. (1998) find that children born to women who received nurse home visits during their pregnancy between 1978 and 1980 have fewer juvenile arrests, convictions, and violations of probation by age fifteen than those whose mothers did not receive treatment.

Other early childhood interventions – many based on the early success of the Perry Preschool, Head Start, and the Nurse Family Partnership – include the Abecedarian Project, the Early Training Project, the Infant Health and Development Program, the Milwaukee Project, and Tulsa's universal pre-kindergarten program. The Abecedarian Project provided full-time, high-quality center-based childcare services for four cohorts of children from low-income families from infancy through age five between 1971 and 1977. Campbell and Ramey (1994) find that at age twelve, those children who were randomly assigned to the project scored 5 points higher on the Wechsler Intelligence Scale and 5-7 points higher on various subscales of the Woodcock-Johnson Psycho-Educational Battery achievement test. The Early Training Project provided children from low-income homes with summertime experiences and weekly home visits during the three summers before entering first grade in an attempt to improve the children's school readiness. Gray and Klaus (1970) report that children who received these intervention services maintained higher Stanford-Binet IQ scores (2-5 points) at the end of fourth grade. The Infant Health and Development Program specifically targeted families with low birthweight, preterm infants and provided them with weekly home visits during the child's first year and biweekly visits through age three, as well as enhanced early childhood

[34] Students not chosen by lottery to participate in Head Start were not precluded from attending other high-quality early childhood centers. Roughly ninety percent of the treatment sample and forty-three percent of the control sample attended center-based care.

[35] The Early Head Start program, established in 1995 to provide community-based supplemental services to low-income families with infants and toddlers, has similar effects (Administration for Children and Families, 2006).

educational care and bimonthly parent group meetings. Brooks-Gunn, Liaw, and Klebanov (1992) report that this program had positive effects on language development at the end of first grade, with participant children scoring 0.09 standard deviations higher on receptive vocabulary and 0.08 standard deviations higher on oral comprehension. The Milwaukee Project targeted newborns born to women with IQs lower than 80; mothers received education, vocational rehabilitation, and child care training while their children received high-quality educational programming and three balanced meals daily at "infant stimulation centers" for seven hours a day, five days a week until the children were six years old. Garber (1988) finds that this program resulted in an increase of 23 points on the Stanford-Binet IQ test at age six for treatment children compared to control children.

Unlike the other programs described, Tulsa's preschool program is open to all 4-year-old children. It is a basic preschool program that has high standards for teacher qualification (a college degree and early childhood certification are both required) and a comparatively high rate of penetration (63 percent of eligible children are served). Gormley et al. (2005) use a birthday cutoff regression discontinuity design to evaluate the program and find that participation improves scores on the Woodcock-Johnson achievement test significantly (from 0.38 to 0.79 standard deviations).

Beyond these highly effective programs, Table 6 demonstrates that there is large variance in the effectiveness of well-known early childhood programs. The Parents as Teachers Program, for instance, shows mixed and generally insignificant effects on initial measures of cognitive development (Wagner and Clayton, 1999). In an evaluation of the Houston Parent-Child Development Centers, Andrews et al. (1982) find no significant impact on children's cognitive skills at age one and mixed impacts on cognitive development at age two. Even so, the typical early childhood intervention passes a simple cost-benefit analysis.[36]

There are two potentially important caveats going forward. First, most of the programs are built on the insights gained from Perry and Head Start, yet what we know about infant development in the past five decades has increased dramatically. For example, psychologists used to assume that there was a relatively equal degree of early attachment across children but they now acknowledge that there is a great deal of variance in the stability of early attachment (Thompson, 2000). Tying new programs to the lessons learned from previously successful programs while incorporating new insights from biology and developmental psychology is both the challenge and opportunity going forward.

Second, and more important for our purposes here, even the most successful early interventions cannot close the achievement gap in isolation. If we truly want to eliminate the racial achievement gap, early interventions may or may not be necessary but the evidence forces one to conclude that they are not sufficient.

5 The Racial Achievement Gap in Kindergarten through 12th Grade

As we have seen, children begin life on equal footing, but important differences emerge by age two and their paths quickly diverge. In this section, we describe basic facts about the racial achievement

[36] Researchers consider a variety of outcomes in determining the monetary value of the benefits of such programs, including the program's impact on need for special education services, grade retention, incarceration rates, and wages. Heckman et al. (2009) estimate that the long-term return on investment of the Perry Preschool program is between seven and ten percent.

gap from the time children enter kindergarten to the time they exit high school. Horace Mann famously argued that schools were “the great equalizer,” designed to eliminate differences between children that are present when they enter school because of different background characteristics. As this section will show, if anything, schools currently tend to exacerbate group differences.

Basic Facts about Racial Differences in Educational Achievement Using ECLS-K

The Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K) is a nationally representative sample of over 20,000 children entering kindergarten in 1998. Information on these children has been gathered at six separate points in time. The full sample was interviewed in the fall and spring of kindergarten, and the spring of first, third, fifth, and eighth grades. Roughly 1,000 schools are included in the sample, with an average of more than twenty children per school in the study. As a consequence, it is possible to conduct within-school or even within-teacher analyses.

A wide range of data is gathered on the children in the study, which is described in detail at the ECLS website http://nces.ed.gov/ecls. We utilize just a small subset of the available information in our baseline specifications, the most important of which are cognitive assessments administered in kindergarten, first, third, fifth, and eighth grades. The tests were developed especially for the ECLS, but are based on existing instruments including Children’s Cognitive Battery (CCB); Peabody Individual Assessment Test - Revised (PIAT-R); Peabody Picture Vocabulary Test-3 (PPVT-3); Primary Test of Cognitive Skills (PTCS); and Woodcock-Johnson Psycho-Educational Battery - Revised (WJ-R). The questions are administered orally through spring of first grade, as it is not assumed that they know how to read until then. Students who are missing data on test scores, race, or gender are dropped from our sample. Summary statistics for the variables we use in our core specifications are displayed by race in Appendix Table 6.

Table 7 presents a series of estimates of the racial test score gap in math (Panel A) and reading (Panel B) for the tests taken over the first nine years of school. Similar to our analysis of younger children in the previous section, the specifications estimated are least squares regressions of the form:

$$outcome_{i,g} = \sum_{R} \beta_R R_i + \Gamma X_i + \varepsilon_{i,g} \tag{13}$$

where $outcome_{i,g}$ denotes an individual i’s test score in grade g and X_i represents an array of student-level social and economic variables describing each student’s environment. The variable R_i is a full set of race dummies included in the regression, with non-Hispanic white as the omitted category. In all instances, we use sampling weights provided in the dataset.

The vector X_i contains a parsimonious set of controls – the most important of which is a composite measure of socio-economic status constructed by the researchers conducting the ECLS survey. The components used in the SES measure are parental education, parental occupational status, and household income. Other variables included as controls are gender, child’s age at the time of enrollment in kindergarten, WIC participation (a nutrition program aimed at relatively low income mothers and children), mother’s age at first birth, birth weight, and the number of children’s books in the home.[37] When there are multiple observations of social and economic variables (SES, number of books in the home, and so on), for all specifications, we only include the value recorded in the fall kindergarten survey.[38] While this particular set of covariates might seem idiosyncratic,

[37] A more detailed description of each of the variables used is provided in the data appendix.

[38] Including all the values of these variables from each survey or only those in the relevant years does not alter the

Fryer and Levitt (2004) have shown that results one obtains with this small set of variables mirror the findings when they include an exhaustive set of over 100 controls. Again, we stress that a causal interpretation is unwarranted; we view these variables as proxies for a broader set of environmental and behavioral factors. The odd-numbered columns of Table 7 present the differences in means, not including any covariates. The even-numbered columns mirror the main specification in Fryer and Levitt (2004).

The raw black-white gap in math when kids enter school is 0.393 (0.029), shown in column one of Panel A. Adding our set of controls decreases this difference to 0.100 (0.035). By fifth grade, Asians continue to outperform other racial groups and Hispanics have gained ground relative to whites, but blacks have lost significant ground. The black-white achievement gap in fifth grade is 0.539 (0.033) standard deviations without controls and 0.304 (0.048) with controls. Disparities in eighth grade look similar, but a peculiar aspect of ECLS-K (very similar tests from kindergarten through eighth grade with different weights on the components of the test) masks potentially important differences between groups. If one restricts attention on the eighth grade exam to subsections of the test which are not mastered by everyone (eliminating the counting and shapes subsection, for example), a large racial gap emerges. Specifically, blacks are trailing whites by 0.961 (0.055) in the raw data and 0.422 (0.093) with the inclusion of controls.

The black-white test score gap grows, on average, roughly 0.60 standard deviations in the raw data and 0.30 when we include controls between the fall of kindergarten and spring of eighth grade. The table also illustrates that the control variables included in the specification shrink the gap a roughly constant amount of approximately 0.30 standard deviations regardless of the year of testing. In other words, although blacks systematically differ from whites on these background characteristics, the impact of these variables on test scores is remarkably stable over time. Whatever factor is causing blacks to lose ground is likely operating through a different channel.[39]

In contrast to blacks, Hispanics gain substantial ground relative to whites, despite the fact that they are plagued with many of the social problems that exist among blacks – low socioeconomic status, inferior schools, and so on. One explanation for Hispanic convergence is increases in English proficiency, though we have little direct evidence on this question.[40] Calling into question that hypothesis is the fact that after controlling for other factors Hispanics do not test particularly poorly on reading, even upon school entry. Controlling for whether or not English is spoken in the home does little to affect the initial gap or the trajectory of Hispanics.[41] The large advantage enjoyed by Asians in the first two years of school is maintained. We also observe striking losses by girls relative to boys in math – over two-tenths of a standard deviation over the four-year period – which is consistent with other research (Becker and Forsyth, 1994; Fryer and Levitt, forthcoming-a).

Panel B of Table 7 is identical to Panel A, but estimates racial differences in reading scores rather than math achievement. After adding our controls, black children score very similarly to whites in reading in the fall of kindergarten. As in math, however, blacks lose substantial ground relative to

results.

[39] The results above are not likely a consequence of the particular testing instrument used. If one substitutes the teachers' assessment of the student's ability as the dependent variable, virtually identical results emerge. Results are available from the author upon request.

[40] Hispanics seem to increase their position relative to whites in states where English proficiency is known to be a problem (Arizona, California, and Texas).

[41] One interesting caveat: Hispanics are also less likely to participate in preschool, which could explain their poor initial scores and positive trajectory. However, including controls for the type of program/care children have prior to entering kindergarten does nothing to explain why Hispanics gain ground.

other racial groups over the first nine years of school. The coefficient on the indicator variable black is 0.009 standard deviations above whites in the fall of kindergarten and 0.246 standard deviations below whites in the spring of fifth grade, or a loss of over 0.25 standard deviations for the typical black child relative to the typical white child. In eighth grade, the gap seems to shrink to 0.168 (0.051), but accounting for the fact that a large fraction of students master the most basic parts of the exam left over from the early elementary years gives a raw gap of 0.918 (0.060) and 0.284 (0.090) with controls. The impact of covariates – explaining about 0.2 to 0.25 of a standard deviation gap between blacks and whites across most grades – is slightly smaller than in the math regressions. Hispanics experience a much smaller gap relative to whites, and it does not grow over time. The early edge enjoyed by Asians diminishes by third grade.

One potential explanation of such large racial achievement gaps, even after accounting for differences in the schools that racial minorities attend, is the possibility that they are assigned inferior teachers within schools. If whites and Asians are more likely to be in advanced classes with more skilled teachers then this sorting could exacerbate differences and explain the divergence over time. Moreover, with such an intense focus on teacher quality as a remedy for racial achievement gaps, it useful to understand whether and the extent to which gaps exist when minorities and non-minorities have the same teacher. This analysis is possible in ECLS-K – the data contain, on average, 3.3 students per teacher within each year of data collection (note that because the ECLS surveys subsamples within each classroom, this does not reflect the true student-teacher ratios in these classrooms).

Table 8 estimates the racial achievement gap in math and reading over the first nine years of school including teacher fixed effects. For each grade, there are two columns. The first column estimates racial differences with school fixed effects on a sample of students for whom we have valid information on their teacher. This restriction reduces the sample approximately one percent from the original sample in Table 7. Across all grades and both subjects, accounting for sorting into classrooms has very little marginal impact on the racial achievement gap beyond including school fixed effects. The average gain in standard deviations from including teacher fixed effects is only about 0.014. The minimum marginal gain from including the teacher controls is 0.006 and the maximum difference is 0.072; however, in several cases the gap is not actually reduced by including teacher fixed effects. There are two important takeaways. First, differential sorting within schools does not seem to be an important contributor to the racial achievement gap. Second, although much has been made of the importance of teacher quality in eliminating racial disparities (Levin and Quinn, 2003; Barton, 2003), the above analysis suggests that racial gaps among students with the same teacher are stark.

In an effort to uncover the factors that are associated with the divergent trajectories of blacks and whites, Table 9 explores the sensitivity of these "losing ground" estimates across a wide variety of subsamples of the data. We report only the coefficients on the black indicator variable and associated standard errors in the table. The top row of the table presents the baseline results using a full sample and our parsimonious set of controls (corresponding to Tables 7 and 8). For the eighth grade scores, we restrict the test to components that are not mastered by all students.[42] In that specification, blacks lose an average of 0.356 (0.047) standard deviations in math and 0.483 (0.060) in reading relative to whites over the first nine years of school.

Surprisingly, blacks lose similar amounts of ground across many subsets of the data, including

[42] Using the full eighth grade test reduces the magnitude of losing ground by roughly half, but the general patterns are the same.

by sex, location type, and whether or not a student attends private schools. The results vary quite a bit across the racial composition of schools, quintiles of the socioeconomic status distribution, and by family structure. Blacks in schools with greater than fifty percent blacks lose substantially more ground in math than do blacks in greater than fifty percent white schools. In reading, their divergence follows similar paths. The top three SES quintiles lose more ground than the lower two quintiles in both math and reading, but the differences are particularly stark in reading. The two largest losing ground coefficients in the table are for the fourth and fifth quintile of SES in reading. Black students in these categories lose ground at an alarming rate - roughly 0.6 standard deviations over 9 years. This latter result could be related to the fact that, in the ECLS-K, a host of variables which are broad proxies for parenting practices differ between blacks and whites. For instance, black college graduates have the same number of children's books for their kids as white high school graduates. A similar phenomenon emerges with respect to family structure; the most ground is lost, relative to whites, by black students who have both biological parents. Investigating within-race regressions, Fryer and Levitt (2004) show that the partial correlation between SES and test scores are about half the magnitude for blacks relative to whites. In other words, there is something that higher income buys whites that is not fully realized among blacks. The limitation of this argument is including these variables as controls does not substantially alter the divergence in black-white achievement over the first nine years of school. This issue is beyond the scope of this chapter but deserves further exploration.

We conclude our analysis of ECLS-K by investigating racial achievement gaps on questions assessing specific skills in kindergarten and eighth grade. Table 10 contains unadjusted means on questions tested in each subsample of the test. The entries in the table are means of probabilities that students have mastered the material in that subtest. Math sections include: counting, numbers, and shapes; relative size; ordinality and sequence; adding and subtracting; multiplying and dividing; place value; rate and measurement; fractions; and area and volume. Reading sections include: letter recognition, beginning sounds, ending sounds, sight words, words in context, literal inference, extrapolation, evaluation, nonfiction evaluation, and complex syntax evaluation. In kindergarten, the test excluded fractions and area and volume (in math) as well as nonfiction evaluation and complex syntax evaluation (in reading).

All students enter kindergarten with a basic understanding of counting, numbers, and shapes. Black students have a probability of 0.896 (0.184) of having mastered this material and the corresponding probability for whites is 0.964 (0.102). Whites outpace blacks on all other dimensions. Hispanics are also outpaced by whites on all dimensions, while Asians actually fare better than whites on all dimensions. By eighth grade, students have essentially mastered six out of the nine areas tested in math, and six out of the ten in reading. Interestingly, on every dimension where there is room for growth, whites outpace blacks—and by roughly a constant amount. Blacks only begin to close the gap after white students have demonstrated mastery of a specific area and therefore can improve no more. While it is possible that this implies that blacks will master the same material as whites but on a longer timeline, there is a more disconcerting possibility - as skills become more difficult, a non-trivial fraction of black students may never master the skills. If these skills are inputs into future subject matter, then this could lead to an increasing black-white achievement gap. The same may apply to Hispanic children, although they are closer to closing the gap with white students than blacks are.

In summary, using the ECLS-K – a recent and remarkably rich nationally representative dataset of students from the beginning of kindergarten through their eighth grade year – we demonstrate an important and remarkably robust racial achievement gap that seems to grow as children age. Blacks underperform whites in the same schools, the same classrooms, and on every aspect of each cognitive assessment. Hispanics follow a similar, though less stark, pattern.

Basic Facts about Racial Differences in Educational Achievement Using CNLSY79

Having exhausted possibilities in the ECLS-K, we now turn to the Children of the National Longitudinal Survey of Youth 1979 (CNLSY79). The CNLSY79 is a survey of children born to NLSY79 female respondents that began in 1986. The children of these female respondents are estimated to represent over 90 percent of all the children ever to be born to this cohort of women. As of 2006, a total of 11,466 children have been identified as having been born to the original 6,283 NLSY79 female respondents, mostly during years in which they were interviewed. In addition to all the mother's information from the NLSY79, the child survey includes assessments of each child as well as additional demographic and development information collected from either the mother or child. The CNLSY79 includes the Home Observation for Measurement of Environment (HOME), an inventory of measures related to the quality of the home environment, as well as three subtests from the full Peabody Individual Achievement Test (PIAT) battery: the Mathematics, Reading Recognition, and Reading Comprehension assessments. We use the Mathematics and Reading Recognition assessments for our analysis.[43]

Most children for whom these assessments are available are between the ages of five and fourteen. Administration of the PIAT Mathematics assessment is relatively straightforward. Children enter the assessment at an age-appropriate item (although this is not essential to the scoring) and establish a "basal" by attaining five consecutive correct responses. If no basal is achieved then a basal of "1" is assigned. A "ceiling" is reached when five of seven items are answered incorrectly. The non-normalized raw score is equivalent to the ceiling item minus the number of incorrect responses between the basal and the ceiling scores. The PIAT Reading Recognition subtest measures word recognition and pronunciation ability, essential components of reading achievement. Children read a word silently, then say it aloud. PIAT Reading Recognition contains 84 items, each with four options, which increase in difficulty from preschool to high school levels. Skills assessed include matching letters, naming names, and reading single words aloud. Appendix Table 7 contains summary statistics for variables used in our analysis.

To our knowledge, the CNLSY is the only large nationally representative sample that contains achievement tests both for mothers and their children, allowing one to control for maternal academic achievement in investigating racial disparities in achievement. Beyond the simple transmission of any genetic component of achievement, more educated mothers are more likely to spend time with their children engaging in achievement-enhancing activities such as reading, using academically stimulating toys, encouraging young children to learn the alphabet and numbers, and so on (Klebanov, 1994).

Tables 11 and 12 provide estimates of the racial achievement gap, by age, for children between the ages of five and fourteen.[44] Table 11 provides estimates for elementary school ages and Table 12 provides similar estimates for middle school aged childen. Both tables contain two panels: Panel A presents results for math achievement and Panel B presents results for reading achievement. The

[43] Results from analysis of the Reading Comprehension assessment are qualitatively very similar to results from using the Reading Recognition assessment and are available from the author upon request.

[44] This corresponds, roughly, to kindergarten entry through ninth grade. To avoid complications due to potential differences in grade retention by race, we analyze CNLSY data by age.

first column under each age presents raw racial differences (and includes dummies for the child's age in months and for the year in which the assessment was administered). The second column adds controls for race, gender, free lunch status, special education status, whether the child attends a private school, family income, the HOME inventory, mother's standardized AFQT score, and dummies for the mother's birth year. Most important of these controls, and unique relative to other datasets, is maternal AFQT.

Two interesting observations emerge. First, gaps in reading are large and positive for blacks relative to whites for children under the age of seven. At age five, blacks are 0.174 (0.042) standard deviations behind whites. Controlling for maternal IQ, blacks are 0.395 (0.045) standard deviations *ahead* of whites. The black advantage, after controlling for maternal AFQT, steadily decreases as children age. At age fourteen, blacks are one-quarter standard deviation behind whites even after controlling for maternal achievement – a loss of roughly 0.650 standard deviations in ten years.

A second potentially important observation is that, in general, the importance of maternal achievement is remarkably constant over time. Independent of the raw data, maternal achievement demonstrably shifts the black coefficient roughly 0.4 to 0.5 standard deviations relative to whites. At age five, the raw difference between blacks and whites is -0.579 (0.040) in math and -0.174 (0.042) in reading. Accounting for maternal AFQT, these differences are -0.147 (0.046) and 0.395 (0.045) – a 0.432 standard deviation shift in math and 0.569 shift in reading. At age fourteen, maternal achievement explains 0.531 standard deviations in math and 0.446 in reading despite the fact that the raw gaps on both tests increased substantially. The stability of the magnitudes in the shift of the gap once one controls for maternal AFQT suggests that whatever is causing blacks to lose ground relative to whites is operating through a different channel.

Basic Facts about Racial Differences in Achievement Using District Administrative Files

Thus far we have concentrated on nationally representative samples because of their obvious advantages. Yet, using the restricted-use version of ECLS-K, we discovered that some large urban areas with significant numbers of chronically underperforming schools may not be adequately represented. For instance, New York City contains roughly 3.84 percent of black school children, but is only 1.46 percent of the ECLS-K Sample. Chicago has 2.42 percent of the population of black students and is only 1.13 percent of the ECLS-K sample. Ideally, sample weights would correct for this imbalance, but if schools with particular characteristics (i.e., predominantly minority and chronically poor performing) are not sampled or refuse to participate for any reason, weights will not necessarily compensate for this imbalance.

To understand the impact of this potential sampling problem, we collected administrative data from four representative urban school districts: Chicago, Dallas, New York City, and Washington, DC. The richness of the data varies by city, but all data sets include information on student race, gender, free lunch eligibility, behavioral incidents, attendance, matriculation with course grades, whether a student is an English Language Learner (ELL), and special education status. The data also include a student's first and last names, birth date, and address. We use address data to link every student to their census block group and impute the average income of that block group to every student who lives there. In Dallas and New York we are able to link students to their classroom teachers. New York City administrative files also contain teacher value-added data for teachers in grades four through eight and question-level data for each student's state assessment.

The main outcome variable in these data is an achievement assessment unique to each city. In May of every school year, students in Dallas public elementary schools take the Texas Assessment of

Knowledge and Skills (TAKS) if they are in grades three through eight. New York City administers mathematics and English Language Arts tests, developed by McGraw-Hill, in the winter for students in third through eighth grade. In Washington, DC, the DC Comprehensive Assessment System (DC-CAS) is administered each April to students in grades three through eight and ten. All Chicago students in grades three through eight take the Illinois Standards Achievement Test (ISAT). See the Data Appendix for more details on each assessment.

One drawback of using school district administrative files is that individual-level controls only include a mutually exclusive and collectively exhaustive set of race dummies, indicators for free lunch eligibility, special education status, and whether a student is an ELL student. A student is income-eligible for free lunch if her family income is below 130 percent of the federal poverty guidelines, or categorically eligible if (1) the student's household receives assistance under the Food Stamp Program, the Food Distribution Program on Indian Reservations (FDPIR), or the Temporary Assistance for Needy Families Program (TANF); (2) the student was enrolled in Head Start on the basis of meeting that program's low-income criteria; (3) the student is homeless; (4) the student is a migrant child; or (5) the student is a runaway child receiving assistance from a program under the Runaway and Homeless Youth Act and is identified by the local educational liaison. Determination of special education and ELL status varies by district. For example, in Washington, DC, special education status is determined through a series of observations, interviews, reviews of report cards and administration of tests. In Dallas, any student who reports that his or her home language is not English is administered a test and ELL status is based on the student's score. Appendix Tables 8 - 11 provide summary statistics used in our analysis in Chicago, Dallas, New York, and Washington, DC, respectively.

Table 13 presents estimates of the racial achievement gap in math (Panel A) and reading (Panel B) for New York City, Washington, DC, Dallas, and Chicago using the standard least squares specification employed thus far. Each city contains three columns. The first column reports the raw racial gap with no controls. The second column adds a small set of individual controls available in the administrative files in each district and the final column under each city includes school fixed effects.

In NYC, blacks trail whites by 0.696 (0.024) standard deviations, Hispanics trail whites by 0.615 (0.023), and Asians outpace whites by 0.266 (0.022) in the raw data. Adding sex, free lunch status, ELL status, special education status, age (including quadratic and cubic terms), and income quintiles reduces these gaps to 0.536 (0.020) for blacks and 0.335 (0.018) for Hispanics. Asians continue to outperform other racial groups. Including school fixed effects further suppresses racial differences for blacks and Hispanics – yielding gaps of 0.346 (0.005) and 0.197 (0.005), respectively. The Asian gap increases modestly with the inclusion of school fixed effects.

Dallas follows a pattern similar to NYC – there is a black-white gap of 0.690 (0.124) in the raw data which decreases to 0.678 (0.108) with the inclusion of controls, and 0.528 (0.031) with school fixed effects. Asians and Hispanics in Dallas follow a similar pattern to that documented in NYC. Both Chicago and Washington, DC, have raw racial gaps that hover around one standard deviation for blacks and 0.75 for Hispanics. Accounting for differences in school assignment reduces the black-white gaps to 0.657 (0.029) in DC and 0.522 (0.011) in Chicago – roughly half of the original gaps. Asians continue to outpace all racial groups in Chicago and are on par with whites in Washington, DC.

Panel B of Table 13 estimates racial differences in reading achievement across our four cities. Similar to the results presented earlier using nationally representative samples, racial gaps on reading assessments are smaller than those on math assessments. In NYC, the raw gap is 0.634 (0.025) and the gap is 0.285 (0.005) with controls and school fixed effects. Dallas contains gaps of similar magnitude to those in NYC and adding school fixed effects has little effect on racial disparities. Chicago and Washington, DC, trail the other cities in the raw gaps – 0.846 (0.046) and 1.163 (0.073) respectively – but these differences are drastically reduced after accounting for the fact that blacks and whites attend different schools. The Chicago gap, with school fixed effects, is 0.381 (0.012) (45 percent of the original gap) and the corresponding gap in DC is 0.599 (0.030). These gaps are strikingly similar in magnitude to racial differences in national samples such as ELCS-K and CNLSY79, suggesting that biased sampling is not a first-order problem.

Thus far, we have concentrated on average achievement across grades three through eight in NYC, Chicago, and DC, and grades three through five in Dallas. Our analysis of ECLS suggests that racial gaps increase over time. Krueger and Whitmore (2001) and Phillips, Crouse, and Ralph (1998) also find that the black-white achievement gap widens as children get older, which they attribute to the differential quality of schools attended by black and white students. Figure 3 plots the raw black-white achievement gap in math (Panel A) and reading (Panel B) for all grades available in each city. In math, DC shows a remarkable increase in the gap as children age – increasing from 0.990 (0.077) in third grade to 1.424 (0.174) in eighth grade. The gap in NYC also increased with age, but much less dramatically. Racial disparities in Chicago are essentially flat across grade levels, and, if anything, racial differences decrease in Dallas. A similar pattern is observed in reading: the gap in DC is increasing over time whereas the gap in other cites is relatively flat. The racial achievement gap in reading in DC is roughly double that in any other city. Figure 4 provides similar data for Hispanics. Hispanics follow a similar, but less consistent, pattern as blacks.

In NYC and Dallas, we were able to obtain data on classroom assignments that allow us to estimate models with teacher fixed effects. In elementary school, we assign the student's main classroom teacher. In middle schools we assign teachers according to subject: for math (resp. ELA) assessment scores, we compare students with the same math (resp. ELA) teacher. In Dallas, there are 1,950 distinct teachers in the sample, with an average of 14 students per teacher. In New York City, there are 16,398 ELA teachers and 16,069 math teachers, with an average of about 25 students per teacher (note that in grades three through five, the vast majority of students have the same teacher for both ELA and math, so the actual number of distinct teachers in the dataset is 20,064.)

Table 14 supplements our analysis by including teacher fixed effects in NYC (Panel A) and Dallas (Panel B) for both math and reading. Each city contains four columns, two for math and two for reading. For comparison, the odd-numbered columns are identical to the school fixed effects specifications in Table 13, but estimated on a sample of students for which we have valid information on their classroom teacher. This restricted sample is 92 percent of the original for NYC and 99 percent of the original for Dallas. The even-numbered columns contain teacher fixed effects. Consistent with the analysis in ECLS-K, accounting for sorting into classrooms has a modest marginal effect on the racial achievement gap beyond the inclusion of school fixed effects. The percent reduction in the black coefficient in NYC is 20.0 percent in math and 25.0 percent in reading. In Dallas, these reductions are 0.9 percent and 3.0 percent, respectively.

Table 15 concludes our analysis of our school district administrative files by investigating the source of the racial achievement gap in NYC across particular skills tested. The math section of the NYC state assessment is divided into five strands: number sense and operations, algebra, geometry, measurement, and statistics and probability. ELA exams are divided into three standards for grades three through eight: (1) information and understanding; (2) literary response and expression; and (3) critical analysis and evaluation. The information and understanding questions measure a student's ability to gather information from spoken language and written text and to transmit knowledge orally and textually. Literary response and expression refers to a student's ability to make connections to a diverse set of texts and to speak and write for creative expression. Critical analysis and evaluation measures how well a student can examine an idea or argument and create a coherent opinion in response. There is no clear pattern in the emphasizing or deemphasizing of particular topics between third and eighth grades. The ELA exams focus more heavily on information and understanding and literary response and expression than on critical analysis and evaluation across all years tested. The math exams focus heavily on number sense until eighth grade, when the focus shifts to algebra and geometry. There are also segments of geometry in fifth grade and statistics and probability in seventh grade.

The most striking observation about Table 15 is how remarkably robust the racial achievement gap in NYC is across grade levels and sets of skills tested. There are substantial racial gaps on every skill at every grade level. The disparities in reading achievement are roughly half as large as the disparities in math.

Putting the pieces together, there are four insights gleaned from our analysis in this section. First, racial achievement gaps using district administrative files, which contain all students in a school district, are similar in magnitude to those estimated using national samples. Second, the evidence as to whether gaps increase over time is mixed. Washington, DC, provides the clearest evidence that black and white paths diverge in school. Patterns from other cities are less clear. Third, school fixed effects explain roughly fifty percent of the gap; adding teacher fixed effects explains about twenty-three percent more in NYC and only about two percent more in Dallas. Fourth, and perhaps most troubling, black students are behind on every aspect of the achievement tests at every grade.

6 The Racial Achievement Gap in High School

We conclude our descriptive analysis of the racial achievement gap with high school-aged students using the National Education Longitudinal Survey (NELS).[45] The NELS consists of a nationally representative group of students who were in eighth grade in 1988 when the baseline survey and achievement test data were collected. Students were resurveyed in 1990 at the end of their tenth grade year and again in 1992 at the anticipated end of their high school career. All three waves consist of data from a student questionnaire, achievement tests, a school principal questionnaire, and teacher questionnaires; 1990 and 1992 follow-ups also include a dropout questionnaire, the baseline and 1992 follow-up also surveyed parents, and the 1992 follow-up contains student transcript information. NELS contains 24,599 students, in 2,963 schools and 5,351 math, science, English, and history classrooms initially surveyed in the baseline year. Eighty-two percent of these students completed a survey in each of the first three rounds.

[45] Similar results are obtained from the National Longitudinal Survey of Adolescent Health (Add Health) – a nationally representative sample of over 90,000 students in grades six through twelve. We chose NELS because it contains tests on four subject areas. Add Health only contains the results from the Peabody Picture Vocabulary Test. Results from Add Health are available from the author upon request.

The primary outcomes in the NELS data are four exams: math, reading comprehension, science, and social studies (history/citizenship/government). In the base year (eighth grade), all students took the same set of tests, but in order to avoid problematic "ceiling" and "floor" effects in the follow-up testing (tenth and twelfth grades for most participants) students were given test forms tailored to their performance in the previous test administration. There were two reading test forms and three math test forms; science and social studies tests remained the same for all students. Test scores were determined using Item Response Theory (IRT) scoring, which allowed the difficulty of the test taken by each student to be taken into account in order to estimate the score a student would have achieved for any arbitrary set of test items. Appendix Table 12 provides descriptive statistics.

Table 16 provides estimates of the racial achievement gap in high school across four subjects. For each grade, we estimate four empirical models. We begin with raw racial differences, which are displayed in the first column under each grade. Then, we add controls for race, gender, age (linear, quadratic, and cubic terms), family income, and dummies for parents' levels of education. The third empirical model includes school fixed effects and the fourth includes teacher fixed effects. The raw black-white gap in eighth grade math is 0.754 (0.025) standard deviations. Adding controls reduces the gap to 0.526 (0.021), and adding school fixed effects reduces the gap further to 0.400 (0.021), which is similar to the eighth grade disparities reported in ECLS. Including teacher fixed effects reduces the gap to 0.343 (0.031) standard deviations. In 10th and 12th grade, black-white disparities range from 0.734 (0.038) in the raw data to 0.288 (0.060) with teacher fixed effects in 10th grade, and 0.778 (0.045) to 0.581 (0.089) in 12th grade. Hispanics follow a similar trend, but the achievement gaps are nearly 40 percent smaller. In the raw data, Asians are the highest-performing ethnic group in eighth through twelfth grades. Including teacher fixed effects, however, complicates the story. Asians are 0.127 standard deviations ahead of whites in eighth grade. This gap diminishes over time and, by twelfth grade, Asian students trail whites when they have the same teachers.

Panels B, C, and D of Table 16, which estimate racial achievement gaps in English, history, and science, respectively, all show magnitudes and trends similar to those documented above in math. Averaging across subjects, the black-white gap in eighth grade is roughly 0.7 standard deviations. An identical calculation for Hispanics yields a gap of just under 0.6 standard deviations. Asians are ahead in math and on par with whites in all other subjects. In twelfth grade, black students significantly trail whites in science and math (0.911 (0.041) and 0.778 (0.045) standard deviations, respectively) and slightly less so in history and English. Hispanics and Asians demonstrate patterns in twelfth grade that are very similar to their patterns in eighth grade.

To close our analytic pipeline from nine months old to high school graduation, we investigate racial differences in high school graduation or GED within five years of their freshman year in high school [not shown in tabular form]. In the raw data, blacks are twice as likely as whites to not graduate from high school or receive a GED within five years of entering high school. Accounting for math and reading achievement scores in eighth grade explains all of the racial gap in graduation rates. Hispanics are 2.2 times more likely not to graduate and these differences are reduced to thirty percent more likely after including eighth grade achievement.

We learn four points from NELS. First, achievement gaps continue their slow divergence in the high school years. Second, gaps are as large in science and history as they are in subjects that are tested more often, such as math and reading. Third, similarly as in the preceding analysis, a substantial racial achievement gap exists after accounting for teacher fixed effects. Fourth, the well-documented disparities in graduation rates can be explained by eighth grade test scores. The last result is particularly striking.

7 Interventions to Foster Human Capital in School-Aged Children

In an effort to increase achievement and narrow differences between racial groups, school districts have become laboratories of innovative reforms, including smaller schools and classrooms (Nye et al., 1995; Krueger, 1999), mandatory summer school (Jacob and Lefgren, 2004), merit pay for principals, teachers, and students (Podgursky and Springer, 2007; Fryer, 2010), after-school programs (Lauer et al., 2006), budget, curricula, and assessment reorganization (Borman et. al., 2007), policies to lower the barrier to teaching via alternative paths to accreditation (Decker, Mayer, and Glaserman, 2004; Kane, Rockoff, and Staiger, 2008), single-sex education (Shapka and Keating, 2003), data-driven instruction (Datnow, Park, and Kennedy, 2008), ending social promotion (Greene and Winters, 2006), mayoral/state control of schools (Wong and Shen, 2002, 2005; Henig and Rich, 2004), instructional coaching (Knight, 2009), local school councils (Easton et al., 1993), reallocating per-pupil spending (Marlow, 2000; Guryan 2001), providing more culturally sensitive curricula (Protheroe and Barsdate, 1991; Thernstrom, 1992; Banks, 2001, 2006), renovated and more technologically savvy classrooms (Rouse and Krueger, 2004; Goolsbee and Guryan, 2006), professional development for teachers and other key staff (Boyd et al., 2008; Rockoff, 2008), and getting parents to be more involved (Domina, 2005).

The evidence on the efficacy of these investments is mixed. Despite their intuitive appeal, school choice, summer remediation programs, and certain mentoring programs show no effect on achievement (Krueger and Zhu, 2002; Walker and Vilella-Velez, 1992; Bernstein et al., 2009). Financial incentives for students, smaller class sizes, and bonuses for teachers in hard-to-staff schools show small to modest gains that pass a cost-benefit analysis (Fryer, 2010; Schanzenbach, 2007; Jacob and Ludwig, 2008). It is imperative to note: these programs have not been able to substantially reduce the achievement gap even in the most reform-minded school systems.

Even more aggressive strategies that place disadvantaged students in better schools through busing (Angrist and Lang, 2004) or significantly alter the neighborhoods in which they live (Jacob, 2004; Kling, Liebman, and Katz, 2007; Sanbonmatsu et al., 2006; Turney et al., 2006) have left the racial achievement gap essentially unchanged.

Table 17 describes seventeen additional interventions designed to increase achievement in public schools.[46] The first column lists the program name, the second column reports the grades treated, and the third column provides a brief description of each intervention. The final two columns provide information on the magnitude of the reported effect and a reference. The bulk of the evidence finds little to no effect of these interventions. Three programs seem to break this mold: Mastery Learning, Success for All, and self-affirmation essay writing. Mastery learning is a group-based, teacher-paced instructional model that is based on the idea that students must attain a level of mastery on a particular objective before moving on to a new objective. Guskey and Gates (1985) perform a meta-analysis of thirty-five studies on this instructional strategy and find that the average achievement effect size from mastery learning programs was 0.78 standard deviations. The

[46] This list was generated by typing in "school-aged interventions" into Google Scholar, National Bureau of Economic Research, and JSTOR. From the (much larger) original list, we narrowed our focus to those programs that contained credible identification.

effect sizes from within individual studies, however, ranged from 0.02 to 1.70 and varied significantly depending on the age of the students and the subject tested (Guskey and Gates, 1985).

Success for All is a school-level elementary school intervention that focuses on improving literacy outcomes for all students in order to improve overall student achievement that is currently used in 1,200 schools across the country (Borman et al., 2007). The program is designed to identify and address deficiencies in reading skills at a young age using a variety of instruction strategies, ranging from cooperative learning to data-driven instruction. Borman et al. (2007) use a cluster randomized trial design to evaluate the impacts of the Success for All model on student achievement. Forty-one schools from eleven states volunteered and were randomly assigned to either the treatment or control groups. Borman et al. (2007) find that Success for All increased student achievement by 0.36 standard deviations on phonemic awareness, 0.24 standard deviations on word identification, and 0.21 standard deviations on passage comprehension.

The self-affirmation essay writing intervention was intended specifically to improve the academic achievement of minorities by reducing the impact of stereotype threat. Seventh grade students were randomly assigned to either a treatment or control group. Both groups were given structured writing assignments three to five times over the course of two school years, but the treatment group was instructed to write about their personal values and why they were important, while the control group was given neutral essay topics. Cohen et al. (2009) find that for black students, this intervention increased GPA by 0.24 points and that the impact was even greater for low-achieving black students (0.41 GPA points). They also find that the program reduced the probability of being placed in remedial classes or being retained in a grade for low-achieving black students. It is unclear what the general equilibrium effects of such psychological interventions are.

Despite trillions spent, there is not one urban school district that has ever closed the racial achievement gap. Figures 5 and 6 show the achievement gap in percentage of students proficient for their grade level across eleven major US cities who participate in the National Assessment of Educational Progress (NAEP) – a nationally representative set of assessments administered every two years to fourth, eighth, and twelfth graders that cover various subject areas, including mathematics and reading.[47]

In every city there are large racial differences. In the Trial Urban District Assessment, among fourth graders, 43.2 percent of whites, 12 percent of blacks, and 16 percent of Hispanics are proficient in reading. In math, these numbers are 50.9, 14, and 20.9, respectively. Similarly, among eighth graders, 40.4 percent of whites, 10.6 percent of blacks, and 13.2 percent of Hispanics score proficient in reading. Math scores exhibit similarly marked racial differences. Washington, DC, has the largest achievement gap of participating cities in NAEP; there is a roughly seventy percent difference between blacks and whites on both subjects and both grade levels. At the other end of the spectrum, Cleveland has the smallest achievement gap – less than seventeen percentage points separate racial groups. Unfortunately, Cleveland's success in closing the achievement gap is mainly due to the dismal performance of whites in the school district and not due to increased performance of black students. Remarkably, there is very little variance in the achievement of minority students across

[47] Individual schools are first selected for participation in NAEP in order to ensure that the assessments are nationally representative, and then students are randomly selected from within those schools. Both schools and students have the option to not participate in the assessments. Tests are given in multiple subject areas in a given school in one sitting, with different students taking different assessments. Assessments are conducted between the last week of January and the first week in March every year. The same assessment is given to all students within a subject and a grade during a given administration.

NAEP districts. There is not one school district in NAEP in which more than twenty-one percent of black students are proficient in reading or math.

The lack of progress has fed into a long-standing and rancorous debate among scholars, policymakers, and practitioners as to whether schools alone can close the achievement gap, or whether the issues children bring to school as a result of being reared in poverty are too much for even the best educators to overcome. Proponents of the school-centered approach refer to anecdotes of excellence in particular schools or examples of other countries where poor children in superior schools outperform average Americans (Chenoweth, 2007). Advocates of the community-focused approach argue that teachers and school administrators are dealing with issues that actually originate outside the classroom, citing research that shows racial and socioeconomic achievement gaps are formed before children ever enter school (Fryer and Levitt, 2004; 2006) and that one-third to one-half of the gap can be explained by family-environment indicators (Phillips et al., 1998; Fryer and Levitt, 2004).[48] In this scenario, combating poverty and related social ills directly and having more constructive out-of-school time may lead to better and more focused instruction in school. Indeed, Coleman et al. (1966), in their famous report on equality of educational opportunity, argue that schools alone cannot solve the problem of chronic underachievement in urban schools.

The Harlem Children's Zone (HCZ), a 97-block area in central Harlem, New York, that combines reform-minded charter schools with a web of community services designed to ensure the social environment outside of school is positive and supportive for children from birth to college graduation, provides an extremely rare opportunity to understand whether communities, schools, or a combination of the two are the main drivers of student achievement.

Dobbie and Fryer (2009) use two separate statistical strategies to estimate the causal impact of attending the charter schools in the HCZ. First, they exploit the fact that HCZ charter schools are required to select students by lottery when the number of applicants exceeds the number of available slots for admission. In this scenario, the treatment group is composed of students who are lottery winners and the control group consists of students who are lottery losers. The second identification strategy explored in Dobbie and Fryer (2009) uses the interaction between a student's home address and her cohort year as an instrumental variable. This approach takes advantage of two important features of the HCZ charter schools: (1) anyone is eligible to enroll in HCZ's schools, but only students living inside the Zone are actively recruited by HCZ staff; and (2) there are cohorts of children that are ineligible due to the timing of the schools' opening and their age. Both statistical approaches lead to the same result: HCZ charter schools are effective at increasing the achievement of the poorest minority children.

Figures 7A and 7B provide a visual representation of the basic results from Dobbie and Fryer (2009). Figure 7A plots yearly, raw, mean state math test scores, from fourth to eighth grade, for four subgroups: lottery winners, lottery losers, white students in New York City public schools and black students in New York City public schools. Lottery winners are comprised of students who either won the lottery or who had a sibling who is already enrolled in the HCZ Promise Academy. Lottery losers are individuals who lost the lottery and did not have a sibling already enrolled. These

[48] The debate over communities or schools often seems to treat these approaches as mutually exclusive, evaluating policies that change one aspect of the schools or a student's learning environment. This approach is potentially informative on the various partial derivatives of the educational production function but is uninformative on the net effect of many simultaneous changes. The educational production function may, for example, exhibit either positive or negative interactions with respect to various reforms. Smaller classes and more time-on-task matter more (or less) if the student has good teachers; good teachers may matter more (or less) if the student has a good out-of-school environment, and so on.

represent "Intent-to-Treat"(ITT) estimates.

In fourth and fifth grade, before they enter the middle school, math test scores for lottery winners, losers, and the typical black student in New York City are virtually identical, and roughly 0.75 standard deviations behind the typical white student.[49] Lottery winners have a modest increase in sixth grade, followed by a more substantial increase in seventh grade and even larger gains by their eighth-grade year.

The "Treatment-on-Treated" (TOT) estimate, which is the effect of actually attending the HCZ charter school, is depicted in Panel B of Figure 7. The TOT results follow a similar pattern, showing remarkable convergence between children in the middle school and the average white student in New York City. After three years of "treatment," HCZ Promise Academy students have nearly closed the achievement gap in math – they are behind their white counterparts by 0.121 standard deviations (p-value = 0.113). If one adjusts for gender and free lunch, the typical eighth grader enrolled in the HCZ middle school outscores the typical white eighth grader in New York City public schools by 0.087 standard deviations, though the difference is not statistically significant (p-value = 0.238).

Figure 8A plots yearly state ELA test scores, from fourth to eighth grade. Treatment and control designations are identical to those in Figure 7A. In fourth and fifth grades, before they enter the middle school, ELA scores for lottery winners, losers, and the typical black student in NYC are not statistically different, and are roughly 0.65 standard deviations behind the typical white student.[50] Lottery winners and losers have very similar ELA scores from fourth through seventh grade. In eighth grade, HCZ charter students distance themselves from the control group. These results are statistically meaningful, but much less so than the math results. The TOT estimate, depicted in Panel B of Figure 8, follows an identical pattern with marginally larger differences between enrolled middle-school students and the control group. Adjusting for gender and free lunch pushes the results in the expected direction.[51]

7.1 What do the Results from HCZ Tell Us About Interventions to Close the Achievement Gap?

There are seven pieces of evidence that, taken together, suggest schools alone can dramatically increase the achievement of the poorest minority students – other community and broader investments may not be necessary. First, Dobbie and Fryer (2009) find no correlation between participation in community programs and academic achievement. Second, the IV strategy described above compares children inside the Zone's boundaries relative to other children in the Zone who were ineligible for the lottery, so the estimates are purged of the community bundle. Recall that IV estimates are larger than the lottery estimates, however, suggesting that communities alone are not the answer. Third, Dobbie and Fryer (2009) report that children inside the Zone garnered the

[49] This is similar in magnitude to the math racial achievement gap in nationally representative samples [0.082 in Fryer and Levitt (2006) and 0.763 in Campbell, Hombo, and Mazzeo (2000)].

[50] This is smaller than the reading racial achievement gap in some nationally representative samples [0.771 in Fryer and Levitt (2006) and 0.960 in Campbell, Hombo, and Mazzeo (2000)].

[51] Interventions in education often have larger impacts on math scores compared to reading or ELA scores (Decker, Mayer, and Glazerman, 2004; Rockoff, 2004; Jacob, 2005). This may be because it is relatively easier to teach math skills, or because reading skills are more likely to be learned outside of school. Another explanation is that language and vocabulary skills may develop early in life, making it difficult to impact reading scores in adolescence (Hart and Risley, 1995; Nelson, 2000).

same benefit from the schools as those outside the Zone, suggesting that proximity to the community programs is unimportant. Fourth, siblings of HCZ students who are in regular public schools, but likely have better-than-average access and information about HCZ community programs, have marginally lower absence rates but their achievement is unchanged (Dobbie and Fryer, 2009).

The final three pieces of evidence are taken from interventions outside of HCZ. The Moving to Opportunity experiment, which relocated individuals from high-poverty to low-poverty neighborhoods while keeping the quality of schools roughly constant, showed small positive results for girls and negative results for boys (Sanbonmatsu et al., 2006; Kling, Liebman, and Katz, 2007). This suggests that a better community, as measured by poverty rate, does not significantly raise test scores if school quality remains essentially unchanged.

Sixth, SEED charter schools – the only urban boarding school in America – which changes a student's home environment from Sunday evening to Friday afternoon by placing students in a dormitory living environment staffed by a full residential faculty that provides students with life skills instruction, homework help, and tutoring as needed, shows very small impacts on achievement (Curto, Fryer, and Howard, 2010).

The last pieces of evidence stem from the rise of a new literature on the impact of charter schools on achievement. While the bulk of the evidence finds only modest success (Hanushek et al., 2005; Hoxby and Rockoff, 2004; Hoxby and Murarka, 2009), there are growing examples of success that is similar to that achieved in HCZ – without community or broader investments. The Knowledge is Power Program (KIPP) is the nation's largest network of charter schools. Anecdotally, they perform at least as well as students from the HCZ on New York state assessments.[52] Angrist et al. (2010) perform the first quasi-experimental analysis of a KIPP school, finding large impacts on achievement. The magnitude of the gains are strikingly similar to those in HCZ. Figure 9 plots the reduced form effect of attending KIPP in Lynn, Massachusetts. Similar to the results of KIPP, Abdulkadiroglu et al. (2009) find that students enrolled in oversubscribed Boston charter schools with organized lottery files gain about 0.17 standard deviations per year in ELA and 0.53 standard deviations per year in math.[53]

8 Conclusion

In 1908, W.E.B Dubois famously noted that "the problem of the 20th century is the problem of the color line." America has undergone drastic changes in 102 years. The problem of the 21st century is the problem of the skill gap. As this chapter attempts to make clear, eliminating the racial skill gap will likely have important impacts on income inequality, unemployment, incarceration, health, and other important social and economic indices. The problem, to date, is that we do not know how to close the achievement gap.

Yet, there is room for considerable optimism. A key difference between what we know now and what we knew even two years ago lies in a series of "existence proofs" in which poor black and Hispanic students score on par with more affluent white students. That is, we now know

[52] On the New York state assessments in the 2008-09 school year, KIPP charter schools had student pass rates that were at least as high as those at the HCZ Promise Academy. This information can be accessed through the New York State Report Cards at https://www.nystart.gov/publicweb/CharterSchool.do?year=2008.

[53] However, the typical middle school applicant in Abdulkadiroglu et al. (2009) starts 0.286 and 0.348 standard deviations higher in fourth grade math and reading than the typical Boston student, and the typical high school applicant starts 0.380 standard deviations higher on both eighth grade math and reading tests.

that with some combination of investments, high achievement is possible for all students. That is an important step forward. Of course, there are many questions as to how one can use these examples to direct interventions that have the potential to close the achievement gap writ large.[54] An economist's solution might be to create a market for gap-closing schools with high-powered incentives for entrepreneurs to enter. The government's role would not be to facilitate the daily workings of the schools; it would simply fund those schools that close the achievement gap and withhold funds from those that do not. The non-gap-closing schools would go out of business and would be replaced by others that are more capable. In a rough sense, this is what is happening in Louisiana post-Hurricane Katrina, what cities such as Boston claim to do, and what reform-minded school leaders such as Chancellor Joel Klein in New York City have been trying accomplish within the constraints of the public system.

A second, potentially more politically expedient, way forward is to try and understand what makes some schools productive and others not. Hoxby and Murarka (2009) and Abdulkadiroglu et al. (2009) show that there is substantial variance in the treatment effect of charter schools – even though all are free from most constraints of the public system and the vast majority do not have staffs under collective bargaining agreements. Investigating this variance and its causes could reveal important clues about measures that could be taken to close the racial achievement gap.

Independent of how we get there, closing the racial achievement gap is the most important civil rights battle of the twenty-first century.

References

[1] Abdulkadiroglu, Atila, Joshua Angrist, Susan Dynarski, Thomas J. Kane, and Parag Pathak (2009), "Accountability and Flexibility in Public Schools: Evidence from Boston's Charters and Pilots", Working paper no. 15549 (NBER, Cambridge, MA).

[2] Administration for Children and Families (2006), "Preliminary Findings from the Early Head Start Prekindergarten Followup", U.S. Department of Health and Human Services Report, Washington, DC.

[3] Andrews, Susan Ring, Janet Berstein Blumenthal, Dale L. Johnson, Alfred J. Kahn, Carol J. Ferguson, Thomas M. Lancaster, Paul E. Malone, and Doris B. Wallace (1982), "The Skills of Mothering: A Study of Parent Child Development Centers", Monographs of the Society for Research in Child Development 47(6): 1-83.

[4] Angrist, Joshua D. and Kevin Lang (2004), "Does School Integration Generate Peer Effects? Evidence from Boston's Metco Program", The American Economic Review 94(5): 1613-1634.

[5] Angrist, Joshua D., Susan M. Dynarski, Thomas J. Kane, Parag A. Pathak, and Christopher R. Walters (2010), "Who Benefits from KIPP?", Working paper no. 15740 (NBER, Cambridge, MA).

[6] Banks, James A. (2001), "Approaches to Multicultural Curriculum Reform", in: James A. Banks and Cherry A.M. Banks, eds., Multicultural Education: Issues and Perspectives, 4th Edition (John Wiley & Sons, Inc., New York).

[54] See Curto, Fryer, and Howard (forthcoming) for more discussion on caveats to taking strategies from charter schools to scale.

[7] Banks, James A. (2006), Cultural Diversity and Education: Foundations, Curriculum, and Teaching (Pearson Education, Inc., Boston, MA).

[8] Barton, Paul E. (2003), “Parsing the Achievement Gap: Baselines for Tracking Progress”, Policy Information Report (Educational Testing Service Policy Information Report, Princeton, NJ).

[9] Bayley, Nancy (1965), “Comparisons of Mental and Motor Test Scores for Ages 1 to 15 Months by Sex, Birth Order, Race, Geographical Location, and Education of Parents”, Child Development, 36: 379-411.

[10] Becker, Douglas F. and Robert A. Forsyth (1994), “Gender Differences in Mathematics Problem Solving and Science: A Longitudinal Analysis”, International Journal of Educational Research 21(4): 407-416.

[11] Bernstein, Lawrence, Catherine Dun Rappaport, Lauren Olsho, Dana Hunt, and Marjorie Levin et al. (2009), “Impact Evaluation of the U.S. Department of Education’s Student Mentoring Program: Final Report”, U.S. Department of Education, Institute of Education Sciences, Washington, DC.

[12] Bethel, James, James L. Green, Graham Kalton, and Christine Nord (2004), “Early Childhood Longitudinal Study, Birth Cohort (ECLS–B), Sampling. Volume 2 of the ECLS-B Methodology Report for the 9-Month Data Collection, 2001–02”, U.S. Department of Education, NCES, Washington, DC.

[13] Bloom, Dan, Alissa Gardenhire-Crooks, and Conrad Mandsager (2009), “Reengaging High School Dropouts: Early Results of the National Guard Youth ChalleNGe Program Evaluation”, MDRC Report, New York.

[14] Borman, Geoffrey D., Robert E. Slavin, Alan C.K. Cheung, Anne M. Chamberlain, Nancy A. Madden, and Bette Chambers (2007), “Final Reading Outcomes of the National Randomized Field Trial of Success for All”, American Educational Research Journal 44(3): 701-731.

[15] Bornstein, Marc H. and Marian D. Sigman (1986), “Continuity in Mental Development from Infancy”, Child Development 57(2): 251-274.

[16] Boyd, Donald, Pamela Grossman, Hamilton Lankford, Susanna Loeb, and James Wyckoff (2008), “Teacher Preparation and Student Achievement”, Working paper no. 14314 (NBER, Cambridge, MA).

[17] Brooks-Gunn, Jeanne, Fong-ruey Liaw, and Pamela Kato Klebanov (1992), “Effects of early intervention on cognitive function of low birth weight preterm infants”, Journal of Pediatrics 120(3): 350-359.

[18] Campbell, Frances A. and Craig T. Ramey (1994), “Cognitive and School Outcomes for High-Risk African-American Students at Middle Adolescence: Positive Effects of Early Intervention”, American Educational Research Journal 32(4): 743-772.

[19] Campbell, Jay R., Catherine M. Hombo, John Mazzeo (2000), “NAEP 1999 Trends in Academic Progress: Three Decades of Student Performance”, U.S. Department of Education, NCES, Washington, DC.

[20] Carneiro, Pedro and James Heckman (2003), “Human Capital Policy”,Working paper no. 9495 (NBER, Cambridge, MA).

[21] Chenoweth, Karin (2007), “It’s Being Done”: Academic Success in Unexpected Schools (Harvard University Press, Cambridge, MA).

[22] Cohen, Geoffrey L., Julio Garcia, Valerie Purdie-Vaughns, Nancy Apfel, and Patricia Brzutoski (2009), “Recursive Processes in Self-Affirmation: Intervening to Close the Minority Achievement Gap”, Science 324(5925): 400-403.

[23] Coleman, James S., Ernest Q. Campbell, Carol J. Hobson, James McPartland, Alexander M. Mood, Frederic D. Weinfeld, and Robert L. York (1966), “Equality of Educational Opportunity”, U.S. Department of Health, Education, and Welfare, Office of Education, Washington, DC.

[24] Congressional Record, No. 11, p. H417 (daily ed. Jan. 27, 2010) (statement of The President).

[25] Cook, Thomas D., Farah-Naaz Habib, Meredith Phillips, Richard A. Settersten, Shobha C. Shagle, Serdar M. Degirmencioglu (1999), “Comer’s School Development Program in Prince George’s County, Maryland: A Theory-Based Evaluation”, American Educational Research Journal 36(3): 543-597.

[26] Corrin, William, Marie-Andree Somers, James J. Kemple, Elizabeth Nelson, Susan Sepanik, et al. (2009), “The Enhanced Reading Opportunities Study: Findings from the Second Year of Implementation”, U.S. Department of Education, Institute of Education Sciences, Washington, DC.

[27] Currie, Janet and Duncan Thomas (1995), “Does Head Start Make a Difference?”American Economic Review 85(3): 341-364.

[28] Curto, Vilsa E., Roland G. Fryer, and Meghan L. Howard (2010), “It May Not Take a Village: Increasing Achievement among the Poor”, Unpublished paper (Harvard University).

[29] Darity, Jr., William A. and Patrick L. Mason (1998), “Evidence on Discrimination in Employment: Codes of Color, Codes of Gender”, Journal of Economic Perspectives 12(2): 63-90.

[30] Datnow, Amanda, Vicki Park, and Brianna Kennedy (2008), “Acting on Data: How Urban High Schools Use Data to Improve Instruction”, Center on Educational Governance, USC Rossier School of Education, Los Angeles.

[31] Decker, Paul, Daniel Mayer, and Steven Glazerman (2004), “The effects of Teach for America on students: findings from a national evaluation”, Mathematica Policy Research, Inc. Report, Princeton, NJ.

[32] Dee, Thomas (2009), “Conditional Cash Penalties in Education: Evidence from the Learnfare Experiment”, Working paper no. 15126 (NBER, Cambridge, MA).

[33] Dickens, William T. and James R. Flynn (2001), “Heritability Estimates Versus Large Environmental Effects: The IQ Paradox Resolved”, Psychological Review 108(2): 346-369.

[34] Dickens, William T. and James R. Flynn (2006), “Black Americans Reduce the Racial IQ Gap: Evidence from Standardization Samples”, Psychological Science 17(10): 913-920.

[35] Dobbie, Will and Roland G. Fryer, Jr. (2009), "Are High Quality Schools Enough to Close the Achievement Gap? Evidence from a Social Experiment in Harlem", Working paper no. 15473 (NBER, Cambridge, MA).

[36] Domina, Thurston (2005), "Leveling the Home Advantage: Assessing the Effectiveness of Parental Involvement in Elementary School", Sociology of Education 78(3): 233-249.

[37] Easton, John Q., Susan Leigh Flinspach, Carla O'Connor, Mark Paul, Jesse Qualls, and Susan P. Ryan (1993), "Local School Council Governance: The Third Year of Chicago School Reform", Chicago Panel on Public School Policy and Finance, Chicago, IL.

[38] Farber, Henry S. and Robert Gibbons (1996), "Learning and Wage Dynamics", Quarterly Journal of Economics 111(4): 1007-1047.

[39] Franzini, L., J.C. Ribble, and A.M. Keddie (2001),"Understanding the Hispanic Paradox", Ethnicity and Disease 11: 496-518.

[40] Fryer, Roland 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): 447-464.

[41] Fryer, Roland G. and Steven D. Levitt (2006), "The Black-White Test Score Gap Through Third Grade", American Law and Economics Review 8(2): 249-281.

[42] Fryer, Roland G. and Steven D. Levitt (forthcoming-a), "An Empirical Analysis of the Gender Gap in Mathematics", American Economic Journal: Applied Economics.

[43] Fryer, Roland G. and Steven D. Levitt (forthcoming-b), "Testing for Racial Differences in the Mental Ability of Young Children", American Economic Review.

[44] Fryer, Roland G. (2010) "Financial Incentives and Student Achievement: Evidence from Randomized Trials", Unpublished paper (Harvard University).

[45] Garber, Howard L. (1988), "The Milwaukee Project: Preventing Mental Retardation in Children at Risk", National Institute of Handicapped Research Report, Washington, DC.

[46] Garces, Eliana, Duncan Thomas, and Janet Currie (2002), "Longer-Term Effects of Head Start", American Economic Review 92(4): 999-1012.

[47] Garet, Michael S., Stephanie Cronen, Marian Eaton, Anja Kurki, Meredith Ludwig, Wehmah Jones, Kazuaki Uekawa, Audrey Falk, Howard Bloom, Fred Doolittle, Pei Zhu, Laura Sztenjnberg, and Marsha Silverberg (2008), "The Impact of Two Professional Development Interventions on Early Reading Instruction and Achievement", U.S. Department of Education, Institute of Education Sciences, Washington, DC.

[48] Goolsbee, Austan and Jonathan Guryan (2006), "The Impact of Internet Subsidies in Public Schools", Review of Economics and Statistics 88(2): 336-347.

[49] Gormley, Jr., William T., Ted Gayer, Deborah Phillips, and Brittany Dawson (2005), "The Effects of Universal Pre-K on Cognitive Development", Developmental Psychology 41(6): 872-884.

[50] Gray, Susan W. and Rupert A. Klaus (1970), "The Early Training Project: A Seventh-Year Report", Child Development 41: 909-924.

[51] Greene, Jay P. and Marcus A. Winters (2006), "Getting Ahead by Staying Behind: An Evaluation of Florida's Program to End Social Promotion", Education Next 6(2): 65-69.

[52] Guryan, Jonathan (2001), “Does Money Matter? Regression-Discontinuity Estimates from Education Finance Reform in Massachusetts”, Working paper no. 8269 (NBER, Cambridge, MA).

[53] Guskey, Thomas R. and Sally L. Gates (1985), “A Synthesis of Research on Group-Based Mastery Learning Programs”, American Educational Research Association Presentation, Chicago, IL.

[54] Hanushek, Eric A., John Kain, Steven Rivkin, and Gregory Branch (2005), “Charter School Quality and Parental Decision Making with School Choice”, Working paper no. 11252, (NBER, Cambridge, MA).

[55] Hart, Betty and Todd R. Risley (1995), Meaningful Differences in the Everyday Experience of Young American Children (Brookes, Baltimore, MD).

[56] Hawkins, J. David, Rick Kosterman, Richard F. Catalano, Karl G. Hill, and Robert D. Abbott (2008), “Effects of Social Development Intervention in Childhood Fifteen Years Later”, Archives of Pediatrics & Adolescent Medicine 162(12): 1133-1141.

[57] Heckman, James J., Seong Hyeok Moon, Rodrigo Pinto, Peter A. Savelyev, and Adam Yavitz (2009), “The Rate of Return to the High/Scope Perry Preschool Program”, Working paper no. 15471 (NBER, Cambridge, MA).

[58] Heckman, James J. (1998), “Detecting Discrimination”, Journal of Economic Perspectives 12(2): 101-116.

[59] Heckman, James J. (1999), “Policies to Foster Human Capital”, Working paper no. 7288 (NBER, Cambridge, MA).

[60] Henig, Jeffrey R. and Wilbur C. Rich (2004), Mayors in the Middle: Politics, Race, and Mayoral Control of Urban Schools (Princeton University Press, Princeton, NJ).

[61] Hoxby, Caroline M. and Sonali Murarka (2009), “Charter Schools in New York City: Who Enrolls and How They Affect Their Students’ Achievement”, Working paper no. 14852 (NBER, Cambridge, MA).

[62] Hoxby, Caroline M. and Jonah E. Rockoff (2004), “The Impact of Charter Schools on Student Achievement”, Unpublished paper (Harvard University).

[63] Jacob, Brian A. and Lars Lefgren (2004), “Remedial Education and Student Achievement: A Regression-Discontinuity Analysis”, Review of Economics and Statistics 86(1): 226-244.

[64] Jacob, Brian A. and Jens Ludwig (2008), “Improving Educational Outcomes for Poor Children”, Working paper no. 14550 (NBER, Cambridge, MA).

[65] Jacob, Brian A. (2004), “Public Housing, Housing Vouchers, and Student Achievement: Evidence from Public Housing Demolitions in Chicago”, American Economic Review 94(1): 233-258.

[66] Jacob, Brian A. (2005), “Accountability, incentives and behavior: the impact of high-stakes testing in the Chicago Public Schools”, Journal of Public Economics 89: 761-796.

[67] James-Burdumy, Susanne, Wendy Mansfield, John Deke, Nancy Carey, Julieta Lugo-Gil, Alan Hershey, Aaron Douglas, Russell Gersten, Rebecca Newman-Gonchar, Joseph Dimino, Bonnie Faddis, and Audrey Pendleton (2009), “Effectiveness of Selected Reading Comprehen-

sion Interventions: Impacts on a First Cohort of Fifth-Grade Students", U.S. Department of Education, Institute of Education Sciences, Washington, DC.

[68] Jencks, Christopher (1998), "Racial Bias in Testing", in: Christopher Jencks and Meredith Phillips, eds., The Black-White Test Score Gap (The Brookings Institution Press, Washington, DC) pp. 55-85.

[69] Jensen, Arthur R. (1973), Educability and Group Differences (The Free Press, New York).

[70] Jensen, Arthur R. (1978), "Genetic and Behavioral Effects of Nonrandom Mating", in: Clyde E. Noble, ed., Human Variation: Biogenetics of Age, Race, and Sex (Academic Press, New York).

[71] Jensen, Arthur R. (1998), The G Factor: The Science of Mental Ability (Praeger, Westport, CT).

[72] Kane, Thomas J., Jonah E. Rockoff, and Douglas O. Staiger (2008), "What Does Certification Tell Us About Teacher Effectiveness? Evidence from New York City", Working paper no. 12155 (NBER, Cambridge, MA).

[73] Kemple, James J. (2008), "CareerAcademies: Long-Term Impacts on Labor Market Outcomes, Educational Attainment, and Transitions to Adulthood", MDRC Report, New York.

[74] Kemple, James J., Corinne M. Herlihy, and Thomas J. Smith (2005), "Making Progress Toward Graduation: Evidence from the Talent Development High School Model", MDRC Report, New York.

[75] Klebanov, Pamelo Kato (1994), "Does Neighborhood and Family Poverty Affect Mothers' Parenting, Mental Health, and Social Support?" Journal of Marriage and Family 56(2): 441-455.

[76] Kling, Jeffrey R., Jeffrey B. Liebman, and Lawrence F. Katz (2007), "Experimental Analysis of Neighborhood Effects", Econometrica 75(1): 83-119.

[77] Knight, Jim, ed. (2009), Coaching: Approaches and Perspectives (Corwin Press, Thousand Oaks, CA).

[78] Krieger, Nancy and Stephen Sidney (1996), "Racial Discrimination and Blood Pressure: The CARDIA Study of Young Black and White Adults", American Journal of Public Health 86(10): 1370-1378.

[79] Krueger, Alan B. (1999), "Experimental Estimates of Education Production Functions", Quarterly Journal of Economics 114(2): 497-532.

[80] Krueger, Alan B. and Diane Whitmore (2001), "Would Smaller Classes Help Close the Black White Achievement Gap?", Working paper no. 451 (Industrial Relations Section, Princeton University).

[81] Krueger, Alan B. and Pei Zhu (2002), "Another Look at the New York City School Voucher Experiment", Working paper no. 9418 (NBER, Cambridge, MA).

[82] Lally, J. Ronald, Peter L. Mangione, and Alice S. Honig (1987), "The Syracuse University Family Development Research Program: Long-Range Impact of an Early Intervention with Low-Income Children and Their Families", Center for Child & Family Studies, Far West Laboratory for Educational Research & Development, San Francisco, CA.

[83] Lang, Kevin and Michael Manove (2006), “Education and Labor-Market Discrimination”, Working paper no. 12257 (NBER, Cambridge, MA).

[84] Lauer, Patricia A., Motoko Akiba, Stephanie B. Wilkerson, Helen S. Apthorp, David Snow, and Mya L. Martin-Glenn (2006), “Out-of-School-Time Programs: A Meta-Analysis of Effects for At-Risk Students”, Review of Educational Research 76(2): 275-313.

[85] Levin, Jessica and Meredith Quinn (2003), “Missed Opportunities: How We Keep High-Quality Teachers Out of Urban Classrooms”, Unpublished paper (The New Teacher Project).

[86] Lewis, Michael and Harry McGurk (1972), “Evaluation of Infant Intelligence”, Science 178 (December 15): 1174-1177.

[87] List, John A. (2005), “The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions”, Working paper no. 11616 (NBER, Cambridge, MA).

[88] Lochner, Lance and Enrico Moretti (2004), “The Effect of Education on Crime: Evidence from Prison Inmates, Arrests, and Self-Reports”, American Economic Review 94(1): 155-189.

[89] Marlow, Michael L. (2000), “Spending, School Structure, and Public Education Quality: Evidence from California”, Economics of Education Review 19(1): 89-106.

[90] McCall, Robert B. and Michael S. Carriger (1993), “A Meta-Analysis of Infant Habituation and Recognition Memory Performance as Predictors of Later IQ”, Child Development 64(1): 57-79.

[91] Morrow-Howell Nancy, Melissa Jonson-Reid, Stacey McCrary, YungSoo Lee, and Ed Spitznagel (2009), “Evaluation of Experience Corps: Student Reading Outcomes”, Unpublished paper (Center for Social Development, George Warren Brown School of Social Work, Washington University, St. Louis, MO).

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

[93] Neal, Derek (2005), “Why Has Black-White Skill Convergence Stopped?”, Working paper no. 11090 (NBER, Cambridge, MA).

[94] Neisser, Ulric, Gwyneth Boodoo, Thomas J. Bouchard, Jr., A. Wade Boykin, Nathan Brody, Stephen J. Ceci, Diane F. Halpern, John C. Loehlin, Robert Perloff, Robert J. Sternberg, and Susana Urbina (1996), “Intelligence: Knowns and Unknowns”, American Psychologist 51(2): 77-101.

[95] Nelson, Charles A. (2000), “The Neurobiological Bases of Early Intervention”, in: Jack P. Shonkoff and Samuel J. Meisels, eds., Handbook of Early Childhood Intervention (Cambridge University Press, New York).

[96] Nisbett, Richard E. (1998), “Race, Genetics, and IQ”, in: Christopher Jencks and Meredith Phillips, eds., The Black-White Test Score Gap (The Brookings Institution Press, Washington, DC) pp. 86-102.

[97] Niswander, K.R. and M. Gordon (1972), The women and their pregnancies: the Collaborative Perinatal Study of the National Institute of Neurological Diseases and Stroke (US Government Print Office, Washington, DC).

[98] Nord, Christine, Carol Andreassen, Laura Branden, Rick Dulaney, Brad Edwards, Anne Elmore, Kristin Denton Flanagan, Philip Fletcher, Jim Green, Richard Hilpert, et al. (2004), "Early Childhood Longitudinal Study, Birth Cohort (ECLS-B), User's Manual for the ECLS-B Nine-Month Public- Use Data File and Electronic Code Book", U.S. Department of Education, NCES, Washington, DC.

[99] Nye, K.E. (1995), The Effect of School Size and the Interaction of School Size and Class Type on Selective Student Achievement Measures in Tennessee Elementary Schools, Unpublished doctoral dissertation (University of Tennessee, Knoxville, TN).

[100] Olds, David, Charles R. Henderson, Robert Cole, John Eckenrode, Harriet Kitzman, Dennis Luckey, Lisa Pettitt, Kimberly Sidora, Pamela Morris, and Jane Powers (1998), "Long-term Effects of Nurse Home Visitation on Children's Criminal and Antisocial Behavior", Journal of the American Medical Association 280(14): 1238-1244.

[101] Olds, David L., JoAnn Robinson, Ruth O'Brien, Dennis W. Luckey, Lisa M. Pettitt, Charles R. Henderson, Rosanna K. Ng, Karen L. Sheff, Jon Korfmacher, Susan Hiatt, and Ayelet Talmi (2002), "Home Visiting by Paraprofessionals and by Nurses: A Randomized, Controlled Trial", Pediatrics 110(3): 486-496.

[102] O'Neill, June (1990), "The Role of Human Capital in Earnings Differences Between Black and White Men", Journal of Economic Perspectives 4(4): 25-45.

[103] Pager, Devah (2007), "The Use of Field Experiments for Studies of Employment Discrimination: Contributions, Critiques, and Directions for the Future", Annals of the American Academy of Political and Social Science 609(1): 104-133.

[104] Phillips, Meredith, Jeanne Brooks-Gunn, Greg J. Duncan, Pamela Klebanov, and Jonathan Crane (1998), "Family Background, Parenting Practices, and the Black-White Test Score Gap", in: Christopher Jencks and Meredith Phillips, eds., The Black-White Test Score Gap (The Brookings Institution Press, Washington, DC) pp. 103-147.

[105] Phillips, Meredith, James Crouse, and John Ralph (1998), "Does the Black-White Test Score Gap Widen After Children Enter School?", in: Christopher Jencks and Meredith Phillips, eds., The Black-White Test Score Gap (The Brookings Institution Press, Washington, DC) pp. 229-272.

[106] Plomin, Robert, John C. DeFries, Gerald E. McClearn, and Peter McGuffin (2000), Behavioral Genetics (Worth, New York).

[107] Podgursky, Michael J. and Matthew G. Springer (2007), "Teacher Performance Pay: A Review", Journal of Policy Analysis and Management 26(4): 909-949.

[108] Protheroe, Nancy J. and Kelly J. Barsdate (1991), "Culturally Sensitive Instruction and Student Learning", Educational Research Center, Arlington, VA.

[109] Puma, Michael, Stephen Bell, Ronna Cook, Camilla Heid, and Michael Lopez, et al. (2005), "Head Start Impact Study: First Year Findings", U.S. Department of Health and Human Services, Washington, DC.

[110] Puma, Michael, Stephen Bell, Ronna Cook, Camilla Heid, et al. (2010), "Head Start Impact Study: Final Report", U.S. Department of Health and Human Services, Washington, DC.

[111] Reuter, E.B. (1945), "Racial Theory", American Journal of Sociology 50(6): 452-461.

[112] Rock, Donald A. and Jackson Stenner (2004), "Assessment Issues in the Testing of Children at School Entry", The Future of Children 15(1): 15-34.

[113] Rockoff, Jonah E. (2004), "The Impact of Individual Teachers on Student Achievement: Evidence from Panel Data", American Economic Review 94(2): 247-252.

[114] Rockoff, Jonah E. (2008), "Does Mentoring Reduce Turnover and Improve Skills of New Employees? Evidence from Teachers in New York City", Working paper no. 13868 (NBER, Cambridge, MA).

[115] Rouse, Cecilia E. and Alan B. Krueger (2004), "Putting Computerized Instruction to the Test: A Randomized Evaluation of a 'Scientifically Based' Reading Program", Economics of Education Review 23(4): 323-338.

[116] Rushton, J. Philippe and Arthur Jensen (2005), "Thirty Years of Research on Race Differences in Cognitive Ability", Psychology, Public Policy, and Law 11(2): 235-294.

[117] Rushton, J. Philippe (1995), "Race and crime: international data for 1989-1990", Psychological Reports 76(1): 307-12.

[118] Sanbonmatsu, Lisa, Jeffrey R. Kling, Greg J. Duncan, and Jeanne Brooks-Gunn (2006), "Neighborhoods and Academic Achievement: Results from the Moving to Opportunity Experiment", The Journal of Human Resources 41(4): 649-691.

[119] Schanzenbach, Diane Whitmore (2007), "What Have Researchers Learned from Project STAR?" Brookings Papers on Education Policy 2006/07: 205- 228.

[120] Schultz, T. Paul and John Strauss (2008), Handbook of Development Economics, Volume 4 (North-Holland, Amsterdam and New York).

[121] Schweinhart, Lawrence J., Helen V. Barnes, and David P. Weikart (1993), Significant Benefits: The High/Scope Perry Preschool Study Through Age 27 (High Scope Press, Ypsilanti, MI).

[122] Shapka, Jennifer D. and Daniel P. Keating (2003), "Effects of a Girls-Only Curriculum During Adolescence: Performance, Persistence, and Engagement in Mathematics and Science", American Educational Research Journal 40(4): 929-960.

[123] Shonkoff, Jack P. (2006), "A Promising Opportunity for Developmental and Behavioral Pediatrics at the Interface of Neuroscience, Psychology, and Social Policy: Remarks on Receiving the 2005 C. Anderson Aldrich Award", Pediatrics 118: 2187-2191.

[124] Shukla, S. (1971), "Priorities in Educational Policy", Economic and Political Weekly 6(30/32): 1649-1651, 1653-1654.

[125] Taggart, Robert (1995), Quantum Opportunity Program Opportunities (Industrialization Center of America, Philadelphia, PA).

[126] Thernstrom, Abigail (1992), "The Drive for Racially Inclusive Schools", Annals of the American Academy of Political and Social Science 523: 131-143.

[127] Thompson, Ross A. (2000), "The Legacy of Early Achievements", Child Development 71(1): 145-152.

[128] Turney, Kristin, Kathryn Edin, Susan Clampet-Lundquist, Jeffrey R. Kling, and Greg J. Duncan (2006), "Neighborhood Effects on Barriers to Employment: Results from a Randomized Housing Mobility Experiment in Baltimore", Brookings-Wharton Papers on Urban Affairs 2006: 137-187.

[129] Wagner, Mary M. and Serena L. Clayton (1999), "The Parents as Teachers Program: Results from Two Demonstrations", The Future of Children 9(1): 91-115.

[130] Walker, Gary and Frances Vilella-Velez (1992), "Anatomy of a Demonstration: The Summer Training and Education Program (STEP) from Pilot through Replication and Postprogram Impacts", Public/Private Ventures, Philadelphia, PA.

[131] Wigdor, Alexandra K. and Bert F. Green (1991), Performance Assessment for the Workplace, Volume 1 (National Academies Press, Washington, DC).

[132] Wilson, William Julius (2010), More than Just Race: Being Black and Poor in the Inner City (Issues of Our Time) (W.W. Norton & Company, New York).

[133] Wong, Kenneth L. and Francis X. Shen (2005), "When Mayors Lead Urban Schools: Assessing the Effects of Takeover", in: William G. Howell, ed., Beseiged: School Boards and the Future of Education Politics (The Brookings Institution Press, Washington, DC) pp. 81-101.

[134] Yeates, Keith Owen, David MacPhee, Frances A. Campbell, and Craig T. Ramey (1983), "Maternal IQ and Home Environment as Determinants of Early Childhood Intellectual Competence: A Developmental Analysis", Developmental Psychology 19(5): 731-739.

[135] Ziedenberg, Jason and Vincent Schiraldi (2002), "Cellblocks or Classrooms?: The Funding of Higher Education and Corrections and Its Impact on African American Men", Unpublished paper (Justice Policy Institute).

9 Appendix: Data Description

A. National Longitudinal Survey of Youth 1979 (NLSY79)

The National Longitudinal Survey of Youth, 1979 Cohort (NLSY79) is a panel data set with data from 12,686 individuals born between 1957 and 1964 who were first surveyed in 1979 when they were between the ages of 14 and 22. The survey consists of a nationally representative cross-section sample as well as a supplemental over-sample of blacks, Hispanics, and low-income whites. In our analysis, we include only and nationally representative cross-section and the over-samples of blacks and Hispanics. We drop 2,923 people from the military and low-income white oversamples and 4 more who have invalid birth years (before 1957 or after 1964). The 5,386 individuals who were born before 1962 are also not included in our analysis.

AFQT Score

The Armed Forces Qualification Test (AFQT) is a subset of four tests given as part of the Armed Services Vocational Aptitude Battery (ASVAB). AFQT scores as reported in the 1981 survey year are used. Scores for an individual were considered missing if problems were reported, if the procedures for the test were altered, or if no scores are reported (either valid or invalid skip) on the relevant ASVAB subtests.

The AFQT score is the sum of the arithmetic reasoning score, the mathematics knowledge score, and two times the verbal composite score. This composite score is then standardized by year of birth (in order to account for natural score differences arising because of differences in age when the test was taken) and then across the whole sample, excluding those with missing AFQT scores.

The variable AFQT^2 is simply constructed by squaring the standardized AFQT score.

Age

In order to determine an individual's age, we use the person's year of birth. The birth year given in 1981 (the year participants took the AFQT) is used if available; otherwise the year of birth given at the beginning of the data collection in 1979 is used. Those who report birth years earlier than 1957 or later than 1964 are dropped from our sample, as these birth years do not fit into the reported age range of the survey.

Additionally, those who were born after 1961 were excluded from analyses. Those born in 1961 or earlier were at least 18 at the time of taking the AFQT and therefore were more likely to have already entered the labor force, which introduces the potential for bias in using AFQT to measure achievement. See Neal and Johnson (1996) for a full explanation.

Ever Incarcerated

In order to construct this variable, we use the fact that the residence of a respondent is recorded each time they are surveyed. One of the categories for type of residence is "jail." Therefore, the variable "ever incarcerated" is equal to one if for any year of the survey the individual's type of residence was "jail". We also include in our measure those who were not incarcerated at any point during the survey but who had been sentenced to a corrective facility before the initial 1979 survey.

Family Income

To construct family income, we use the total net family income variables from 1979, 1980, and 1981. We convert all incomes into 1979 dollars, and then use the most recent income available.

Numerous Reading Materials

We classify a person as having "numerous reading materials" if they had magazines, newspapers, and a library card present in their home environment at age 14.

Parent Occupation

To construct the dummies for having a mother (father) with a professional occupation, we use the variable which gives the occupational code of the adult female (male) present in the household at age 14. We classify mothers (fathers) as professionals if they have occupational codes between 1 and 245. This corresponds to the following two occupational categories: professional, technical, and kindred; and managers, officials, and proprietors.

Physical Health Component Score

This variable is constructed within the data set using the questions asked by the SF-12 portion of the 2006 administration of the surveys. For the analysis, the physical component score (PCS) is standardized across all individuals for whom a score is available. Those without a valid PCS are not included in the analysis.

Race

A person's race is coded using a set of mutually exclusive dummy variables from the racial/ethnic cohort of the individual from the screener. Individuals are given a value of one in one of the three dummy variables - white, black, or Hispanic. All respondents have a value for this race measure.

Sex

A person's sex was coded as a dummy variable equal to one if the person is male and zero if the person is female. Preference was given for the reported sex in 1982; if this was unavailable, the sex reported in 1979 was used.

Unemployed

The variable "unemployed" is a binary variable that is equal to one if the person's employment status states that they are unemployed. Those whose employment status states that they are not in the labor force are excluded from labor force participation analyses.

Wage

Job and wage information are given for up to five jobs per person in 2006, which was the latest year for which published survey results were available. The data contains the hourly compensation and the number of hours worked for each of these jobs, as well as an indicator variable to determine whether each particular job is a current job. The hourly wage from all current jobs is weighted by the number of hours worked at that job in order to determine an individual's overall hourly wage.

Neal and Johnson (1996) considered wage reports invalid if they were over $75. We do the same, but adjust this amount for inflation; therefore, wages over $115 (the 2006 equivalent of $75 in 1990) are considered to be invalid. Wage is also considered to be missing/invalid if the individual does not have a valid job class for any of the five possible jobs. Individuals with invalid or missing wages are not included in the wage regressions, which use the log of the wage measure as the dependent variable.

B. National Longitudinal Survey of Youth 1997 (NLSY97)

The National Longitudinal Survey of Youth, 1997 Cohort (NLSY97) is a panel data set with data from approximately 9,000 individuals born between 1980 and 1984 who were first surveyed in 1997 when they were between the ages of 13 and 17.

AFQT Score

The Armed Forces Qualification Test (AFQT) is a subset of four tests given as part of the Armed Services Vocational Aptitude Battery (ASVAB). In the NLSY97 data set, an ASVAB math-verbal percent score was constructed. The NLS staff states that the formula they used to construct this score is similar to the AFQT score created by the Department of Defense for the NLSY79, but that it is not the official AFQT score.

The AFQT percentile score created by the NLS was standardized by student age within three-month birth cohorts. We then standardized the scores across the entire sample of valid test scores.

The variable AFQT^2 is simply constructed by squaring the standardized AFQT score.

Age

Because wage information was collected in either 2006 or 2007 (discussed below), the age variable needed to be from the year in which the wage data was collected. The age variable was constructed first as two separate age variables - the person's age in 2006 and the person's age in 2007 - using the person's birth year as reported in the baseline (1997) survey. The two age variables are then combined, with the age assigned to be the one from the year in which the wage was collected.

All age cohorts were included in the labor force analyses. Because participants were younger during the baseline year of the survey when the AFQT data were collected - all were under the age of 18 - they were unlikely to have entered the labor force yet.

Ever Incarcerated

In the NLSY97, during each yearly administration of the survey, individuals are asked what their sentence was for any arrests (up to 9 arrests are asked about). Individuals who reported that they were sentenced to "jail", an "adult corrections institution", or a "juvenile corrections institution" for any arrest in any of the surveys were given a value of one for this variable; otherwise this variable was coded as zero.

Race

A person's race is coded using a set of mutually exclusive dummy variables from the racial/ethnic cohort of the individual from the screener. Individuals are given a value of one in one of the four dummy variables - white, black, Hispanic, or mixed race. All respondents have a value for this race measure.

Sex

A person's sex was coded as a dummy variable equal to one if the person is male and zero if the person is female.

Unemployed

The variable "unemployed" is a binary variable that is equal to one if the person's employment status states that they are unemployed. Those whose employment status states that they are not in the labor force are excluded from labor force participation analyses.

Wage

Jobs and wage information is given for up to 9 jobs in 2007 and up to 8 jobs in 2007. We are given the hourly compensation and the number of hours worked for each of these jobs, as well as a variable to determine whether each particular job is a current job. The hourly wage from all current jobs is weighted by the number of hours worked at that job in order to determine an individual's overall hourly wage.

Once again, wages over $115 in 2006 and $119 in 2007 (the equivalent of $75 in 1990) are considered to be invalid. Wage is also considered to be missing/invalid if the individual does not have a valid job class for any of the possible jobs. Individuals with invalid or missing wages are not included in the wage regressions, which use the log of the wage measure as the dependent variable.

Wage in 2007 is converted to 2006 dollars so that the two wage measures are comparable. We use the 2007 wage measure for any individuals for whom it is available; otherwise, we use the 2006 wage measure.

C. College & Beyond, 1976 Cohort (C&B)

The College and Beyond Database contains data on 93,660 full-time students who entered thirty-four colleges and universities in the fall of 1951, 1976, or 1989. For this analysis, we focus on the cohort from 1976. The C&B data contain information drawn from students' applications and transcripts, SAT and ACT scores, as well as information on family demographics and socioeconomic status. The C&B database also includes responses to a survey administered in 1996 to all three cohorts that provides detailed information on post-college labor market outcomes. The response rate to the 1996 survey was approximately 80 percent.

Income

Income information is reported as fitting into one of a series of income ranges, but these ranges were different in the 1995 and 1996 surveys. For all the possible ranges in each survey year, the individual's income was assigned to the midpoint of the range (i.e. $40,000 for the $30,000-50,000 range); for less than $10,000, income was assigned to be $5,000 (1995 survey). Income less than $1,000 income was assigned to be missing because an individual could not have made this sum of money working full-time (1996 survey). For more than $200,000, income was assigned to be $250,000. If available, income reported for 1995 (the 1996 survey) was used; otherwise 1994 annual income (collected in 1995) was used. Individuals with invalid or missing wages are not included in the income regressions, which use the log of the income measure as the dependent variable.

Race

A person's race is coded using a set of mutually exclusive dummy variables from the racial/ethnic cohort of the individual from the screener. Individuals are given a value of one in one of the five dummy variables - white, black, Hispanic, other race, or missing the race variable.

SAT Score

The SAT score of an individual is coded as the true value of the combined math and verbal scores, with possible scores ranging between 400 (200 per section) and 1600 (800 per section). Individuals with missing scores are assigned a score of zero and are accounted for using a missing score dummy variable. The square of SAT score was also included in regressions that controlled for educational achievement.

Sex

A person's sex was coded as a dummy variable equal to one if the person is male and zero if the person is female.

Unemployed

Determining who was unemployed in this data set required a few steps. First, we had to determine who was not working at the time of the survey. This is coded within two variables, one for each survey (1995 and 1996). If an individual reports that they are not working because they are retired or for another reason, we then consider a later question, where they are asked about any times at which they were out of work for 6 months or longer. For those people who stated that they were not currently working, we considered any period of time that included the year of the survey in which they stated they were not working. We then considered the reason they gave for being out of work during that period. If the person stated that they were retired, a student, had family responsibilities, had a chronic illness, or did not need/want to work, we considered them out of the labor force. If a person was not out of the labor force but was not currently working because they were laid off or suitable work was not available, we considered that individual unemployed. Because only 39 people from the entire sample could be considered unemployed, we did not perform analyses using this variable.

D. Early Childhood Longitudinal Study, Birth Cohort (ECLS-B)

The Early Childhood Longitudinal Study, Birth Cohort (ECLS-B) is a nationally representative sample of over 10,000 children born in 2001. The first wave of data collection was performed when most of the children were between eight and twelve months of age. The second wave interviewed the same set of children around their second birthday; the third wave was conducted when the children were of preschool age (approximately 4 years old). The data set includes an extensive array of information from parent surveys, interviewer observation or parent-child interactions, and mental and motor proficiency tests. Further details on the study design and data collection methods are available at the ECLS website (http://nces.ed.gov/ecls).

From the total sample, 556 children had no mental ability test score in the first wave. Test scores are missing for an additional 1,326 children in the second wave and 1,338 children in the third wave. All subjects with missing test scores are dropped from the analysis. This is the only exclusion we make from the sample.[55] Throughout the analysis, the results we report are weighted

[55] In cases where there are missing values for another of these covariates, we set these missing observations equal to zero and add an indicator variable to the specification equal to one if the observation is missing and equal to zero otherwise. We obtain similar results for the first wave when we include all children with an initial test score, including those who subsequently are not tested.

to be nationally representative using sampling weights included in the data set.[56]

Bayley Short Form - Research Edition (BSF-R)

The BSF-R is an abbreviated version of the Bayley Scale of Infant Development (BSID) that was designed for use in the ECLS to measure the development of children early in life in five broad areas: exploring objects (e.g., reaching for and holding objects), exploring objects with a purpose (e.g., trying to determine what makes the ringing sound in a bell), babbling expressively, early problem solving (e.g., when a toy is out of reach, using another object as a tool to retrieve the toy), and naming objects.[57] The test is administered by a trained interviewer and takes twenty-five to thirty-five minutes to complete. A child's score is reported as a proficiency level, ranging from zero to one on each of the five sections. These five proficiency scores have also been combined into an overall measure of cognitive ability using standard scale units. Because this particular test instrument is newly designed for ECLS-B, there is little direct evidence regarding the correlation between performance on this precise test and outcomes later in life. For a discussion of the validity of this instrument, see Fryer and Levitt (forthcoming). The BSF-R scores have been standardized across the population of children with available scores to have a mean of zero and a standard deviation of one.

Early Reading and Math Scores

As the BSF-R is not developmentally appropriate for preschool-aged children, in order to measure mental proficiency in the third wave (4 years old), a combination of items were used from several assessment instruments. The test battery was developed specifically for use in the ECLS-B and included items from a number of different assessments, including the Peabody Picture Vocabulary Test (PPVT), the Preschool Comprehensive Test of Phonological and Print Processing (Pre-CTOPPP), the PreLAS 2000, and the Test of Early Mathematics Ability-3 (TEMA-3), as well as questions from other studies, including the Family and Child Experiences Study (FACES), the Head Start Impact Study, and the ECLS-K. The assessment battery was designed to test language and literacy skills (including English language skills, emergent literacy, and early reading), mathematics ability, and color knowledge. The cognitive battery was available in both English and Spanish; children who spoke another language were not assessed using the cognitive battery.

The preschool cognitive scores are estimated using Item Response Theory (IRT) modeling based on the set of questions that was administered to each student. The study used IRT modeling to create skill-specific cluster scores that estimate what a student's performance within a given cluster would have been had the entire set of items been administered. Additionally, scores have been converted to a proficiency probability score that measures a child's proficiency within a given skill domain and standardized T-scores that measure a child's ability in comparison to his peers.

Age

Child's age is coded in three sets of variables, one for each wave of the survey. For the 9-month wave, dummy variables were created for each of the possible one-month age ranges between 8 months and 16 months (inclusive). Children who were younger than 8 months were included in the 8-month variable and children who were older than 16 months were included in the 16-month variable. For the 2-year wave, dummy variables were created for each of the possible one-month age ranges between 23 months and 26 months (inclusive). Children who were younger than 23

[56] A comparison of the ECLS-B sample characteristics with known national samples, such as the U.S. Census and the Center for Disease Control's Vital Statistics, confirms that the sample characteristics closely match the national average.

[57] See Nord et al. (2006) for further details.

months were included in the 23-month variable, while children who were older than 26 months were included in the 26-month variable. For the preschool wave, dummy variables were created for each of the possible one-month age ranges between 47 months and 60 months (inclusive). Children who were younger than 47 months were included in the 47-month variable and children who were older then 60 months were included in the 60-month variable.

Race

Race is defined in a mutually exclusive set of dummy variables, with a child being assigned a value of one for one of white, black, Hispanic, Asian, or other race.

Region

Dummy variables were created for each of four regions of the country: Northeast, Midwest, South, and West.

Sex

The variable for a child's sex is a binary variable that is equal to one if the child is female and zero if the child is male.

Family Structure

This is coded as a set of four dummy variables, each representing a different possible set of parents with whom the child lives: two biological parents, one biological parent, one biological parent and one non-biological parent, and other.

Mother's Age

A continuous variable was created for the age of the child's mother. Analyses including this variable also included squared, cubic, quartic, and quintic terms. The cubic, quartic, and quintic terms were divided by 100,000 before their inclusion in the regressions.

Number of Siblings

Number of siblings is coded as a set of dummy variables, each one representing a different number of siblings. All children with 6 or more siblings are coded in the same dummy variable.

Parent as Teacher Score

The "parent as teacher" score is coded based on interviewer observations of parent-child interactions in a structured problem-solving environment and is based on the Nursing Child Assessment Teaching Scale (NCATS). The NCATS consists of 73 binary items that are scored by trained observers. The parent component of the NCATS system has 50 items that focus on the parent's use of a "teaching loop," which consists of four components: (1) getting the child's attention and setting up expectations for what is about to be done; (2) giving instructions; (3) letting the child respond to the teaching; and (4) giving feedback on the child's attempts to complete the task. The parent score ranges from 0 to 50. Analyses including this variable also included squared, cubic, quartic, and quintic terms. The cubic, quartic, and quintic terms were divided by 100,000 before their inclusion in the regressions.

Socioeconomic Status

Socioeconomic status is constructed by ECLS and includes parental income, occupation, and education. It is coded as a set of five mutually exclusive and exhaustive dummy variables, each one representing a different socioeconomic status quintile.

Birthweight

The birthweight of the child was coded in a set of four dummy variables: under 1500 grams, 1500-2500 grams, 2500-3500 grams, and over 3500 grams.

Multiple Birth Indicator

A set of dummy variables were created to indicate how many children were born at the same time as the child: single birth, twin birth, or triplet or higher order birth.

Premature Births

Premature births are considered in two different ways. First, a dummy variable is created to classify the child as being born prematurely or not. Then a set of dummy variables were created to capture how early the child was born: less than 7 days, 8-14 days, 15-21 days, etc. in seven day increments up to 77 days premature. Any births more than 77 days premature are coded in the 71-77 days premature dummy variable.

E. Collaborative Perinatal Project (CPP)

The Collaborative Perinatal Project (CPP) consists of over 31,000 women who gave birth in twelve medical centers between 1959 and 1965. All medical centers were in urban areas; six in the Northeast, four in the South, one in the West, and one in the north-central region of the U.S. Some institutions selected all eligible women, while others took a random sample.[58] The socioeconomic and ethnic composition of the participants is representative of the population qualifying for medical care at the participating institutions. These women were re-surveyed when their children were eight months, four years, and seven years old. Follow-up rates were remarkably high: eighty-five percent at eight months, seventy-five percent at four years, and seventy-nine percent at seven years. We only include students in our analysis that had score results for all three tests.[59] Our analysis uses data on demographics, measures of home environment, and prenatal factors. In all cases, we use the values collected in the initial survey for these background characteristics.[60]

Bayley Scales of Infant Development (BSID)

The Bayley Scales of Infant Development (BSID) can be used to measure the motor, language, and cognitive development of infants and toddlers (under three years old). It is therefore used only in the first wave of the CPP. The assessment consists of 45-60 minutes of developmental play tasks administered by a trained interviewer. For use in this analysis, scores were standardized across the entire population. Individuals with scores lower than ten standard deviations below the mean are considered to have missing scores.

Stanford-Binet Intelligence Scales

The Stanford-Binet Intelligence Scales were used as the main measure of cognitive ability for the second wave of the CPP when the children were four years-old. The scores are standardized across the entire sample of available scores.

Wechsler Intelligence Scale for Children (WISC)

The Wechsler Intelligence Scale for Children (WISC) was used as the main measure of cognitive ability for the third wave of the CPP when the children were seven years-old. The scores are standardized across the entire sample of available scores.

[58] Detailed information on the selection methods and sampling frame from each institution can be found in Niswander and Gordon (1972). Over 400 publications haev emanated from the CPP; for a bibliography, see http://www.niehs.nih.gov/research/atniehs/labs/epi/studies/dde/biblio.cfm. The most relevant of these papers is Bayley (1965), which, like our reanalysis, finds no racial test score gaps among infants.

[59] Analyzing each wave of the data's test scores, not requiring that a student have all three scores, yields similar results.

[60] It must be noted, however, that there are a great deal of missing data on covariates in CPP; in some cases more than half of the sample has missing values. We include indicator variables for missing values for each covariate in the analysis.

Age

For the first wave of the study (8 months), age is coded as a set of dummy variables representing 5 age ranges: less than 7.5 months, 7.5-8.5 months, 8.5-9 months, 9-10 months, and over 10 months.

In the second (4 years) and third (7 years) waves of the study, age is coded as a continuous variable and given as age of the child in months at the time of the follow-up survey and testing.

Race

Race is defined in a mutually exclusive set of dummy variables, with a child being assigned a value of one for one of white, black, Hispanic, or other race. Preference is given for the race reported when the child is 8 months; if no race is reported then, race is used as reported at 7 years, then at 3 years, then at 4 years.

Sex

The variable for a child's sex is a binary variable that is equal to one if the child is female and zero if the child is male. Preference is given for the sex reported when the child is 8 months; if no sex is reported then, sex is used as reported at 7 years, then at 3 years, then at 4 years.

Family Structure

A dummy variable is created to indicate whether both the biological mother and biological father are present.

Income

The cumulative income of the family during the first three months of pregnancy is coded as a set of dummy variables representing a range of incomes. Each family is coded within one of the following income ranges: less than $500, $500-1000, $1000-1500, $1500-2000, $2000-2500, or more than $2500.

Mother's Age

A continuous variable was created for the age of the child's mother. Analyses including this variable also included squared, cubic, quartic, and quintic terms. The quartic and quintic terms were divided by 1000 before their inclusion in the regressions.

Mother's Reaction to Child

A set of dummy variables for the mother's reaction to the child are included, indicating if the mother is indifferent, accepting, attentive, or over-caring toward the child, or if she behaves in another manner. These dummy variables are constructed by considering the mother's reaction to and interactions with the child, which are assessed by the interviewer. These dummy variables are not mutually exclusive, as a mother is coded as fitting into each category (negative, indifferent, accepting, attentive, caring, or other) if she fits into that category for any of the measures. Therefore, any mother who falls into different categories for the different measures will be coded with a value of one for multiple dummy variables in this set.

Number of Siblings

Number of siblings is coded as a set of dummy variables, each one representing a different number of siblings from zero to six-plus siblings. All children with 6 or more siblings are coded in the same dummy variable.

Parents' Education

A separate set of dummy variables are coded to represent the educational attainment of the child's mother and father. Each parent's education is coded as one of: high school dropout (less than 12 years of schooling), high school graduate (12 years of schooling), some college (more than 12 years of schooling but less than 16 years of schooling), or at least college degree (16 or more years of schooling).

Parents' Occupation

A separate set of dummy variables are coded to represent the field of work done by the mother and father of the child. Each parent's occupational status is coded as one of: no occupation, professional occupation, or non-professional occupation.

Birthweight

The birthweight of the child was given as an amount in pounds and ounces. This measure was first converted to an amount in ounces and the weight in ounces was then converted to a weight in grams. The birthweight of the child was coded in a set of four dummy variables: under 1500 grams, 1500-2500 grams, 2500-3500 grams, and over 3500 grams.

Multiple Birth Indicator

A set of dummy variables were created to indicate how many children were born at the same time as the child: single birth, twin birth, or triplet or higher order birth.

Prematurity

Premature births are considered in two different ways. First, a dummy variable is created to classify the child as being born prematurely or not. Then a set of dummy variables were created to capture how early the child was born, in weekly increments up to 11 weeks. Any children born more than 11 weeks premature were included in the dummy variable for 11 weeks premature. The amount of time that a child was born prematurely was determined by subtracting the gestation length of the child from 37, which is the earliest gestation period at which a birth is considered full-term.

F. Early Childhood Longitudinal Study, Kindergarten Cohort (ECLS-K)

The Early Childhood Longitudinal Study kindergarten cohort (ECLS-K) is a nationally representative sample of 21,260 children entering kindergarten in 1998. Thus far, information on these children has been gathered at seven separate points in time. The full sample was interviewed in the fall and spring of first grade. All of our regressions and summary statistics are weighted, unless otherwise noted, and we include dummy variables for missing data. We describe below how we combined and recoded some of the ECLS variables used in our analysis.

Math and Reading Standardized Test Scores

The primary outcome variables in this data set were math and reading standardized test scores from tests developed especially for the ECLS, but based on existing instruments including Children's Cognitive Battery (CCB), Peabody Individual Achievement Test - Revised (PIAT-R), Peabody Picture Vocabulary Test-3 (PPVT-3), Primary Test of Cognitive Skills (PTCS), and Woodcock-Johnson Psycho-Educational Battery - Revised (WJ-R). The test questions were administered to students orally, as an ability to read is not assumed.[61] The values used in the analyses are IRT scores provided by ECLS that we have standardized to have a mean of zero and standard deviation

[61] A "general knowledge" exam was also administered. The general knowledge test is designed to capture "children's knowledge and understanding of the social, physical, and natural world and their ability to draw inferences and comprehend implications." We limit the analysis to the math and reading scores, primarily because of the comparability of these test scores to past research in the area. In addition, there appear to be some peculiarities in the results of the general knowledge exam. See Rock and Stenner (2005) for a more detailed comparison of ECLS to previous testing instruments.

of one for the overall sample on each of the tests and time periods.[62] In all instances sample weights provided in ECLS-K are used.[63]

Socioeconomic Composite Measure

The socioeconomic scale variable (SES) was computed by ECLS at the household level for the set of parents who completed the parent interview in fall kindergarten or spring kindergarten. The SES variable reflects the socioeconomic status of the household at the time of data collection for spring kindergarten. The components used for the creation of SES were: father or male guardian's education, mother or female guardian's education, father or male guardian's occupation, mother or female guardian's occupation, and household income.

Number of Children's Books

Parents or guardians were asked, "How many books does your child have in your home now, including library books?" Answers ranged from 0 to 200.

Child's Age

We used the composite variable child's age at assessment provided by ECLS. The child's age was calculated by determining the number of days between the child assessment date and the child's date of birth. The number was then divided by 30 to calculate the age in months.

Birth Weight

Parents were asked how much their child weighed when they were born. We multiplied the number of pounds by 16 and added it to the ounces to calculate birth weight in ounces.

Mother's Age at First Birth

Mothers were asked how old they were at the birth of their first child.

G. Children of the National Longitudinal Survey of Youth (CNLSY)

There are 11,469 children in the original sample. We drop 2,413 children who do not have valid scores for an assessment. We drop 4 more children whose mothers have invalid birth years (before 1957 or after 1964), 459 more children whose mothers have invalid AFQT scores (or whose mothers had recorded problems with the test administration), and 568 more children whose mothers are from the military or low-income white oversamples, for an overall sample of 8,025 children.

We define the age group with 5-year-olds as those children between 60 and 71 months old (3,375 children). We define the age group with 6-10-year-olds as those children who are between 72 and 119 months old (7,699 children). We define the age group with 10-14-year-olds as those children who are between 120 and 179 months old (7,107 children). Note that many children have observations in multiple age groups because they participated in multiple assessments.

Income

We construct income as follows: For each child, we look at all of the incomes that the child's mother had between 1979 and 2006 which are available in the dataset. We use the income that is closest to the assessment year and convert it to 1979 dollars. If two incomes are equally close to the assessment year, then we use the earlier one.

[62] For more detail on the process used to generate the IRT scores, see Chapter 3 of the ECLS-K Users Guide. Our results are not sensitive to normalizing the IRT scores to have a mean of zero and standard deviation of one.

[63] Because of the complex manner in which the ECLS-K sample is drawn, different weights are suggested by the providers of the data depending on the set of variables used (BYPW0). We utilize the weights recommended for making longitudinal comparisons. None of our findings are sensitive to other choices of weights, or not weighting at all.

Demographic Variables

Free lunch, special education, and private school are defined as follows: The variable is 1 if the child was in the program in either the 1994 or 1995 school survey. The variable is 0 if the child was never in the program and if the child was recorded as not being in the program in the 1994 or 1995 school survey. The variable is missing otherwise.

Test Scores

Test scores are standardized within the sample by age group. Mother's AFQT score is standardized within the sample.

H. National Assessment of Educational Progress (NAEP)

All data is derived from the 2007 NAEP data. Note that there is a different sample of students for each of the 4 tests. In the full NAEP sample, there are 191,040 children who took the 4th grade reading test, 197,703 who took the 4th grade math test, 160,674 who took the 8th grade reading test, and 153,027 who took the 8th grade math test. Within the Trial Urban District Assessment (TUDA) subsample, there are 20,352 students who took the 4th grade reading test, 17,110 who took the 8th grade reading test, 21,440 who took the 4th grade math test, and 16,473 who took the 8th grade math test.

Test Scores

To calculate the overall test score, we take the mean of the 5 plausible test score values. For analysis that includes the entire NAEP sample, test scores are standardized across the entire sample. For analysis that includes only the district sample, test scores are standardized across the district (TUDA) subsample.

I. Chicago Public Schools

We use Chicago Public Schools (CPS) ISAT test score administrative data from the 2008-09 school year. In our data file, there are 177,001 students with reading scores and 178,055 students with math scores (grades 3-8). We drop 273 students for whom we are missing race information. This leaves us with 176,767 students with non-missing reading scores and 177,787 students with non-missing math scores.

Demographic Variables

We use 4 different CPS administrative files to construct demographic data. These files are the 2009-10 enrollment file, and 2008-09 enrollment file, a file from 2008-09 with records of all students in the school district, and a file from 2008-09 containing records for students in bilingual education. For the demographic variables that should not change over time (race, sex, age), we give use the variables from the 2009-10 enrollment file to construct these and then fill in missing values using the other three files in the order of precedence listed above. For the demographic variables that may vary from year to year (free lunch and ELL status), we use the same process but exclude the 2009-10 enrollment file since it is from a year that is not the same as the year in which the ISAT test score was administered. Note that we include both "free" and "reduced" lunch statuses for our construction of the free lunch variable.

School ID

In order to construct school ID, we use the school ID from the 2008-09 enrollment file but fill in missing values with the 2008-09 with records of all students in the school district. For the purposes of analysis, we assign a common school ID to the 928 students (about 0.5 percent of the sample) for whom we are still missing school ID information.

Test Scores

Illinois Standards Achievement Test (ISAT) scores for math, reading, science, and writing were pulled from a file listing scores for all students in Chicago Public Schools. Eighth graders do not take the science portion of the test and we decided to use only math and reading scores to keep the analysis consistent across districts. ISAT test scores are standardized to have mean 0 and standard deviation 1 within each grade.

J. Dallas Independent School District

We pull our Dallas TAKS scores from files provided by the Dallas Independent School District (DISD). There are 33,881 students for whom we have non-missing TAKS score data. We use two files to construct grade and school ID information for these students: the 2008-09 DISD enrollment file and the 2008-09 DISD transfers file (containing students who were either not in the school district at the time the enrollment file information was compiled or who ever transferred schools during the school year). We drop 15 students (about 0.04 percent of the sample) whose grade at the time of the tests cannot be definitively determined either because they skipped a grade during the school year or because their grade levels in the enrollment and transfers files conflict. This leaves us with a sample of 33,866 students in grades 3-5 with non-missing TAKS score data. Within this sample, there are no students with missing race data. This leaves us with 28,126 students in grades 3-5 with non-missing TAKS reading scores and 33,561 students in grades 3-5 with non-missing TAKS math scores.

Age

To calculate age in months, we calculate the exact number of days old each student was as of August 25, 2008 (the first day of the 2008-09 school year) and then divide by 30 and round down to the nearest integer number of months.

Demographic Variables

In order to construct demographic data, we use the demographic information from the 2008-09 enrollment file. For the race, sex, and age variables, we fill in missing information using the enrollment files from 2002-03 through 2007-08, giving precedence to the most recent files first.

Income

In order to construct the income variable, we use ArcGIS software to map each student's address from the 2008-09 enrollment file to a 2000 census tract block group. Then we assign each student's income as the weighted average income of all those who were surveyed in that census tract block group in 2000.

School ID

We construct school ID as follows: For students who attended only school during the 2008-09 school year, we assign them to that school. For students who attended more than one school according to the transfers file, we assign the school that they attended for the greatest number of days. If a student attended more than one school for equally long numbers of days, we use the school among these with the lowest school identification number.

Test Scores

Students in grades three through five take the Texas Assessment of Knowledge and Skills (TAKS). TAKS has a variety of subjects. We use scores from the reading and math sections of this exam. Unlike the Iowa Test of Basic Skills (ITBS) scores, the TAKS data that we have are not grade-equivalent scores. In order to ease interpretation of these scores, we standardize them by, for every subject and year, subtracting the mean and dividing by the standard deviation.

K. New York City Department of Education

We pull our NYC math and ELA scores from NYC Public Schools (NYCPS) test score administrative files. There are 427,688 students (in grades 3-8) with non-missing ELA score data and 435,560 students (in grades 3-8) with non-missing math score data. We drop 1,230 students for whom we are missing race information (about 0.3 percent of the sample). This leaves us with a sample of 426,806 students with non-missing ELA score data and 434,593 students with non-missing math score data.

Age

To calculate age in months, we calculate the exact number of days old each student was as of September 2, 2008 (the first day of the 2008-09 school year) and then divide by 30 and round down to the nearest integer number of months.

Demographic Variables

In order to construct demographic data, we use the demographic information from the 2008-09 enrollment file. For the race, sex, and age variables, we fill in missing information using the enrollment files from 2003-04 through 2007-08, giving precedence to the most recent files first.

Income

In order to construct the income variable, we use ArcGIS software to map each student's address from the 2008-09 enrollment file to a 2000 census tract block group. Then we assign each student's income as the weighted average income of all those who were surveyed in that census tract block group in 2000.

School ID

We assign school ID for each subject as the school ID recorded in the 2008-09 test score file for that subject. We use Human Resources files provided by NYCPS to link students to their teachers for ELA and math.

Test Scores

The New York state math and ELA tests, developed by McGraw-Hill, are high-stakes exams conducted in the winters of thrid through eighth grades. Students in third, fifth, and seventh grades must score proficient or above on both tests to advance to the next grade. The math test includes questions on number sense and operations, algebra, geometry, measurement, and statistics. Tests in the earlier grades emphasize more basic content such as number sense and operations, while later tests focus on advanced topics such as algebra and geometry. The ELA test is designed to assess students on three learning standards – information and understandings, literary response and expresssion, and critical analysis and evaluation – and includes multiple-choice and short-response sections based on a reading and listening section, along with a brief editing task.

In our analysis ELA and math scores are standardized by subject and by grade level to have mean 0 and standard deviation 1.

L. District Data: Washington, D.C.

We pull our DCCAS test scores from DC Public Schools (DCPS) test score administrative files from 2008-09. There are 20,249 students with non-missing reading scores and 20,337 students with non-missing math scores. We drop 6 observations because the students have two observations with conflicting test scores. This leaves us with a sample of 20,243 students with non-missing reading scores and 20,331 students with non-missing math scores, all from grades 3-8 and 10 (the full set of grades for which the DCCAS tests are administered).

Age

To calculate age in months, we calculate the exact number of days old each student was as of August 25, 2008 (the first day of the 2008-09 school year) and then divide by 30 and round down to the nearest integer number of months.

Demographic Variables

In order to construct demographic data, we use the demographic information from the 2008-09 enrollment file and use the DCCAS test score file from 2008-09 to fill in missing demographic information. For the race, sex, and age variables, we fill in missing information using the enrollment files from 2005-06 through 2007-08, giving precedence to the most recent files first.

Income

In order to construct the income variable, we use ArcGIS software to map each student's address from the 2008-09 enrollment file to a 2000 census tract block group. Then we assign each student's income as the weighted average income of all those who were surveyed in that census tract block group in 2000.

School ID

We assign school ID as the school ID recorded in the 2008-09 DCCAS test score file.

Test Scores

The DC CAS is the DC Comprehensive Assessment System and is administered each April to students in grades three through eight as well as tenth graders. It measures knowledge and skills in reading and math. Students in grades four, seven, and ten also take a composition test; students in grades five and eight also take a science test; and students in grades nine through twelve who take biology also take a biology test

DCCAS scores are standardized by subject and by grade level to have mean 0 and standard deviation 1.

M. National Education Longitudinal Study of 1988 (NELS)

We use the first three waves (1988, 1990, and 1992) of the NELS panel dataset for our analysis, when respondents were in 8th, 10th, and 12th grade, respectively. There were 19,645 students in the 8th grade cohort, 18,176 students in the 10th grade cohort, and 17,161 students in the 12th grade cohort. We use IRT-estimated number right scores for the analysis. In the base year, there are 23,648 students with non-missing math scores, 23,643 students with non-missing English scores, 23,616 students with non-missing science scores, and 23,525 students with non-missing history scores. In the first follow-up year, there are 17,793 students with non-missing math scores, 17,832 students with non-missing English scores, 17,684 students with non-missing science scores, and 17,591 students with non-missing history scores. In the second follow-up year, there are 14,236 students with non-missing math scores, 14,230 students with non-missing English scores, 14,134 students with non-missing science scores, and 14,063 students with non-missing history scores. If first follow-up and second follow-up scores are missing, we impute them from one another.

Age

We use birth year and birth month to calculate each student's age as of September 1988.

Income

The income variable is constructed using the income reported in the base year parent questionnaire. The variable in the dataset categorizes income into different ranges, and our income variable is coded as the midpoint of each range, with the exception of the lowest income category (which corresponds to no income), which we code as \$0, and the highest income category (which corresponds to an income of \$200,000 or more), which we code as \$200,000. We divide income by \$10,000.

Parent's Education

Parents' education refers to the highest level of education obtained by either parent.

School ID

In order to construct the base year school ID, we use the base year school ID variable but supplement it using the student ID when it is missing. The base year school ID is embedded in the student ID as all but the last two digits of the student ID.

Socioeconomic Status

We take the SES quartile variable directly from the dataset.

Figure 1

Emergence of Gaps in ECLS−B

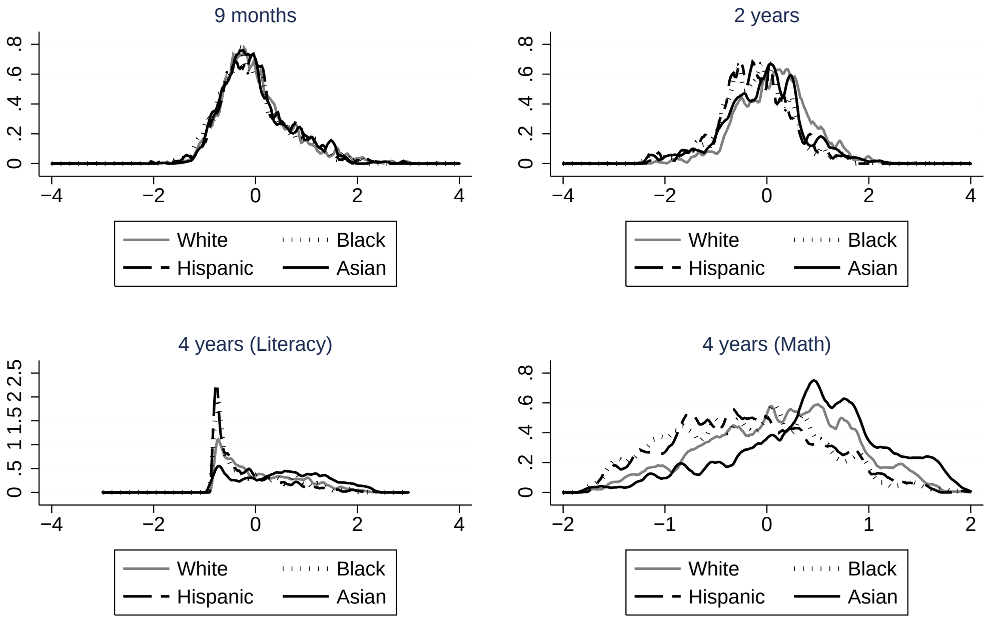


Figure 2
Emergence of Gaps in CPP

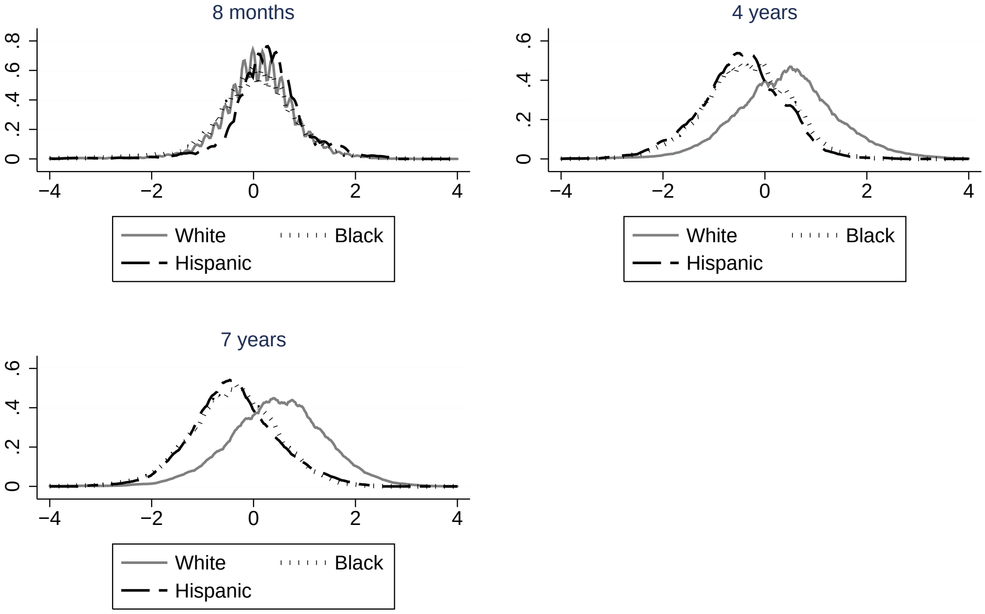
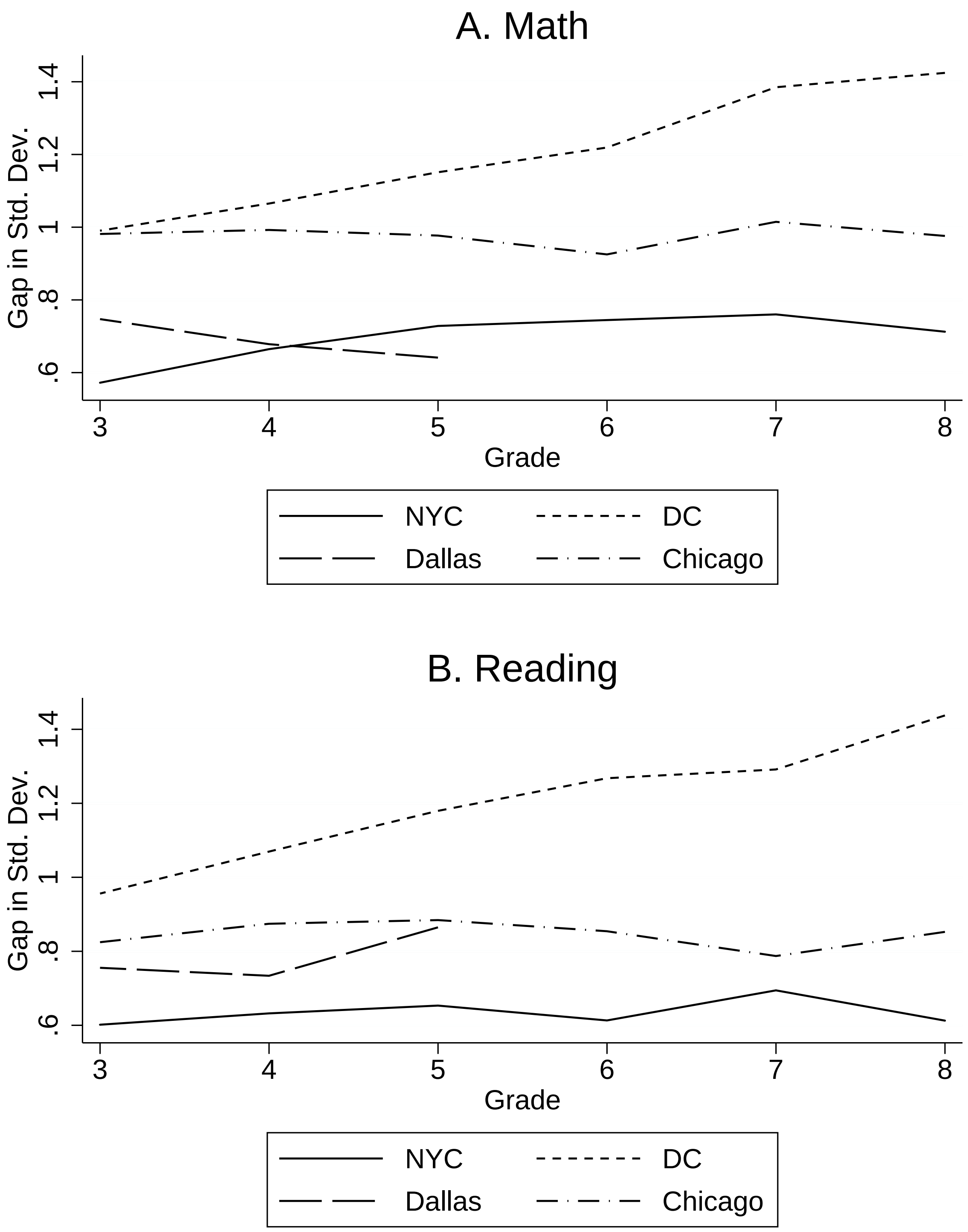


Figure 3

Black−White Achievement Gap (Raw) by Grade

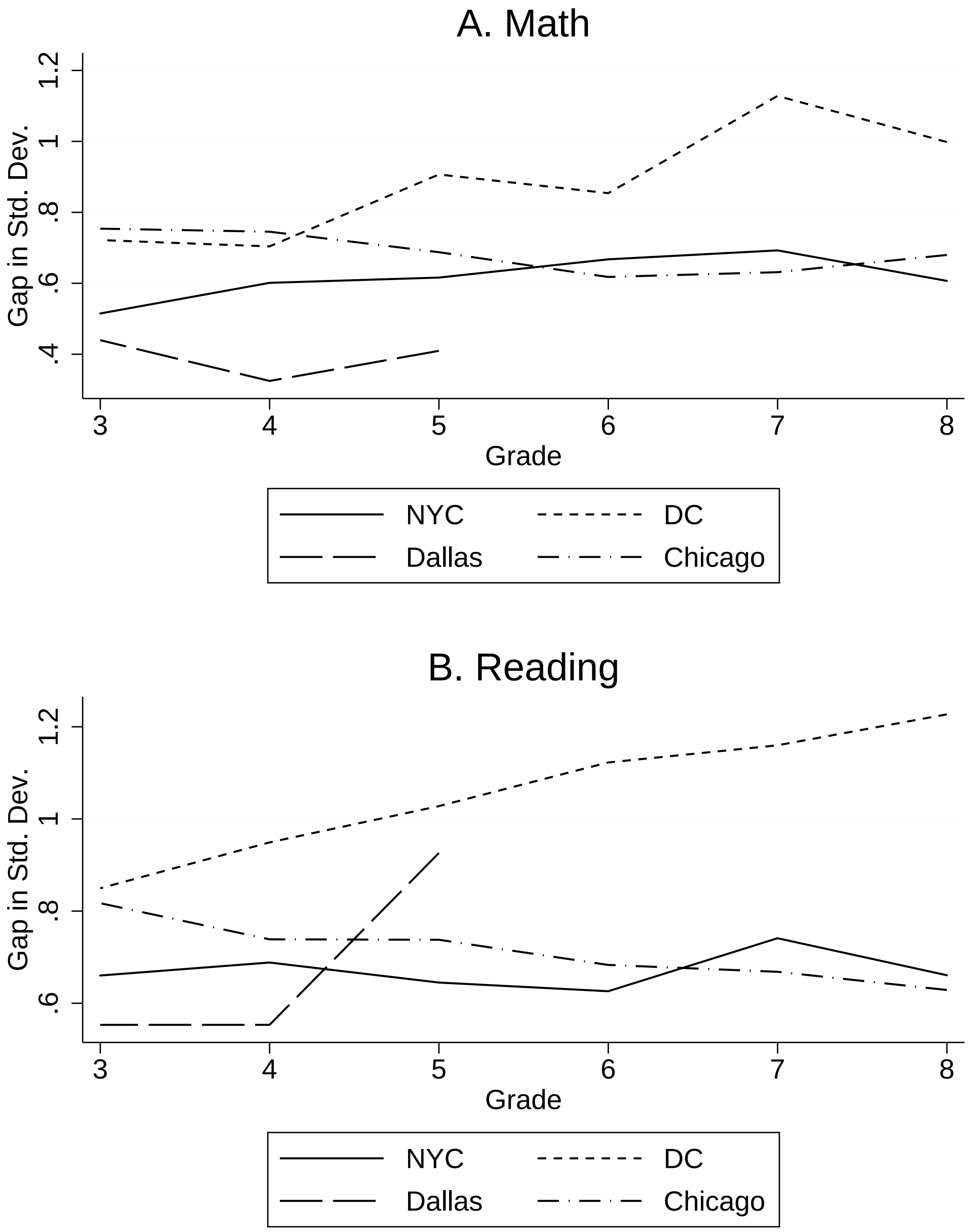


A. Math

B. Reading

Figure 4

Hispanic−White Achievement Gap (Raw) by Grade

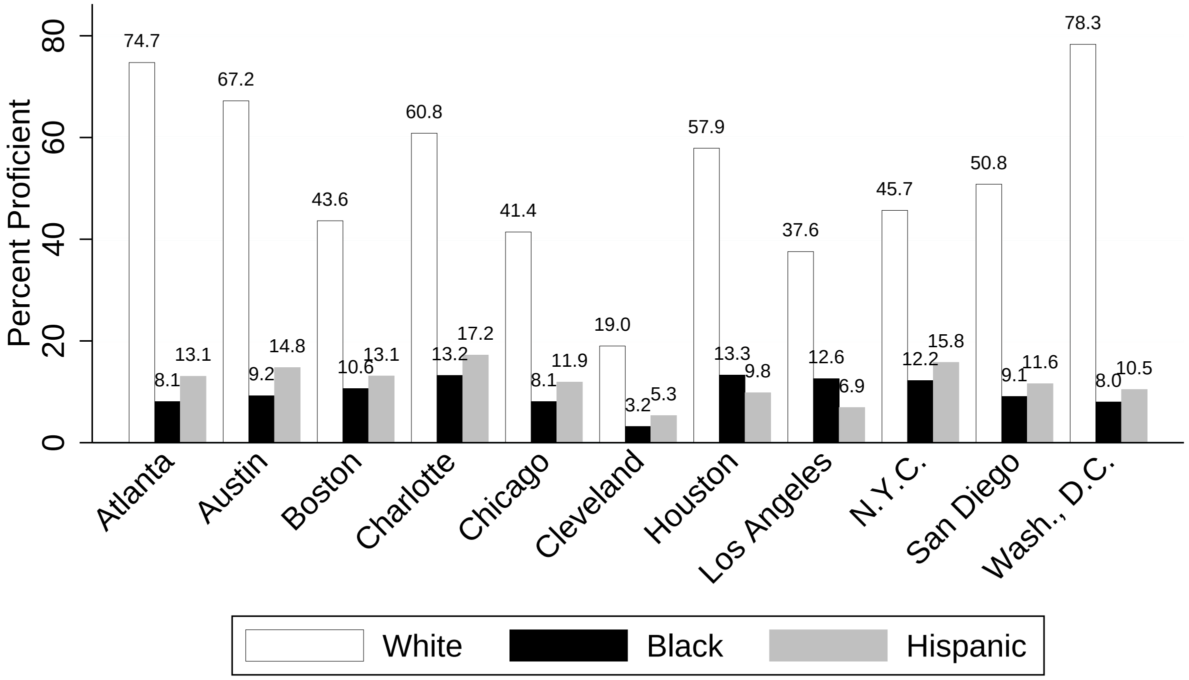


A. Math

B. Reading

Figure 5A

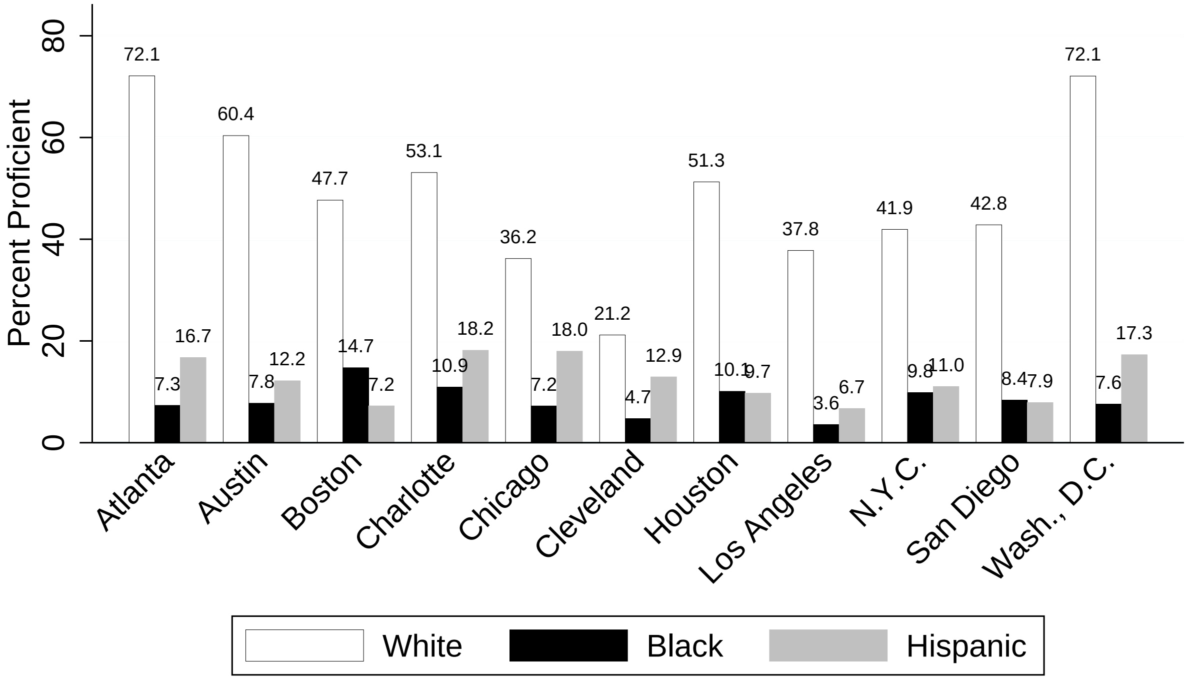
NAEP 2007 Proficiency Levels by City and Race: 4th Grade Reading



All means are calculated using sample weights.
N = 20352.

Figure 5B

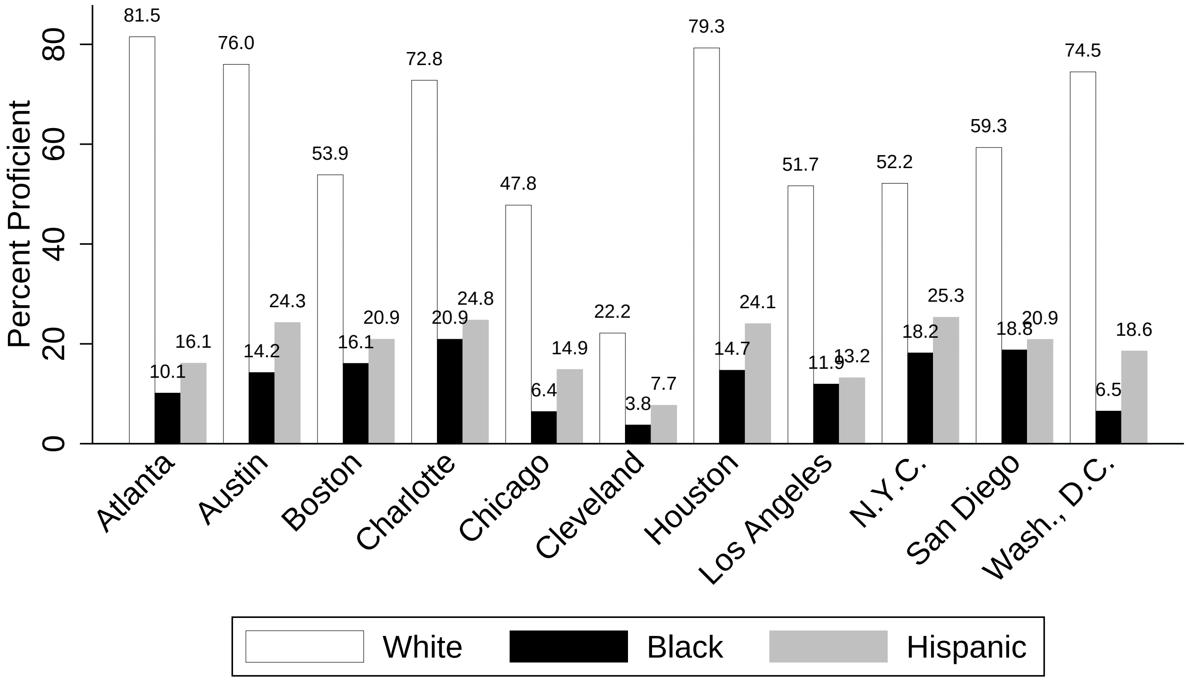
NAEP 2007 Proficiency Levels by City and Race: 8th Grade Reading



All means are calculated using sample weights.
N = 17110.

Figure 6A

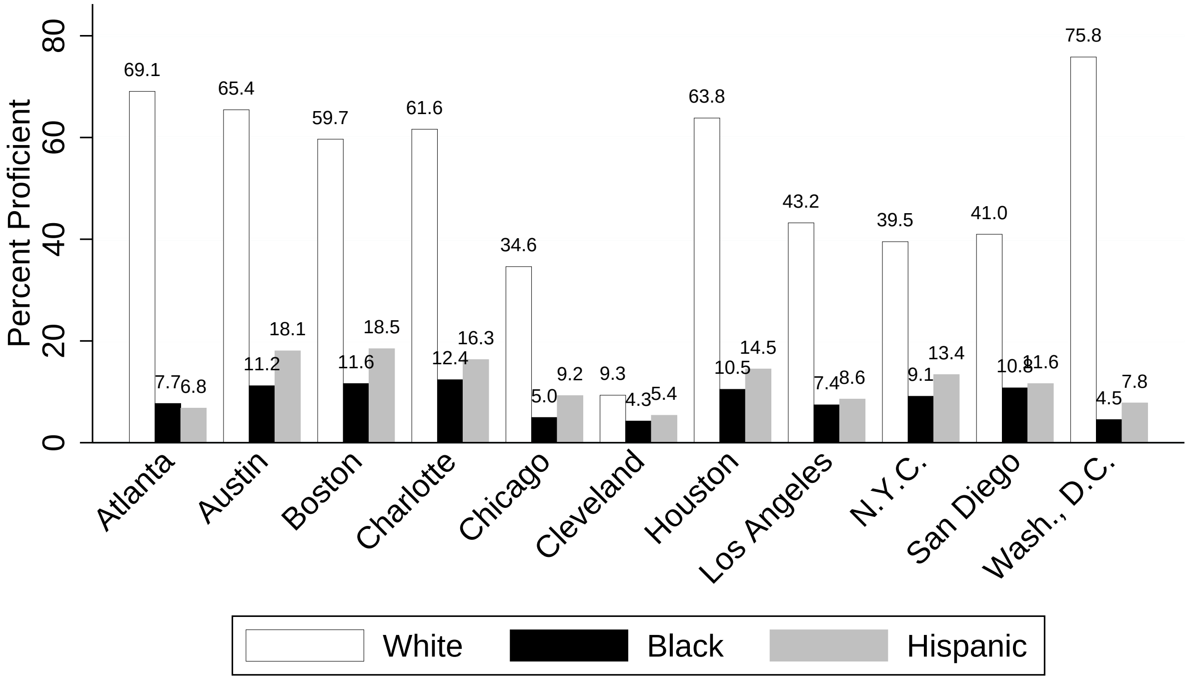
NAEP 2007 Proficiency Levels by City and Race: 4th Grade Math



All means are calculated using sample weights.
N = 21440.

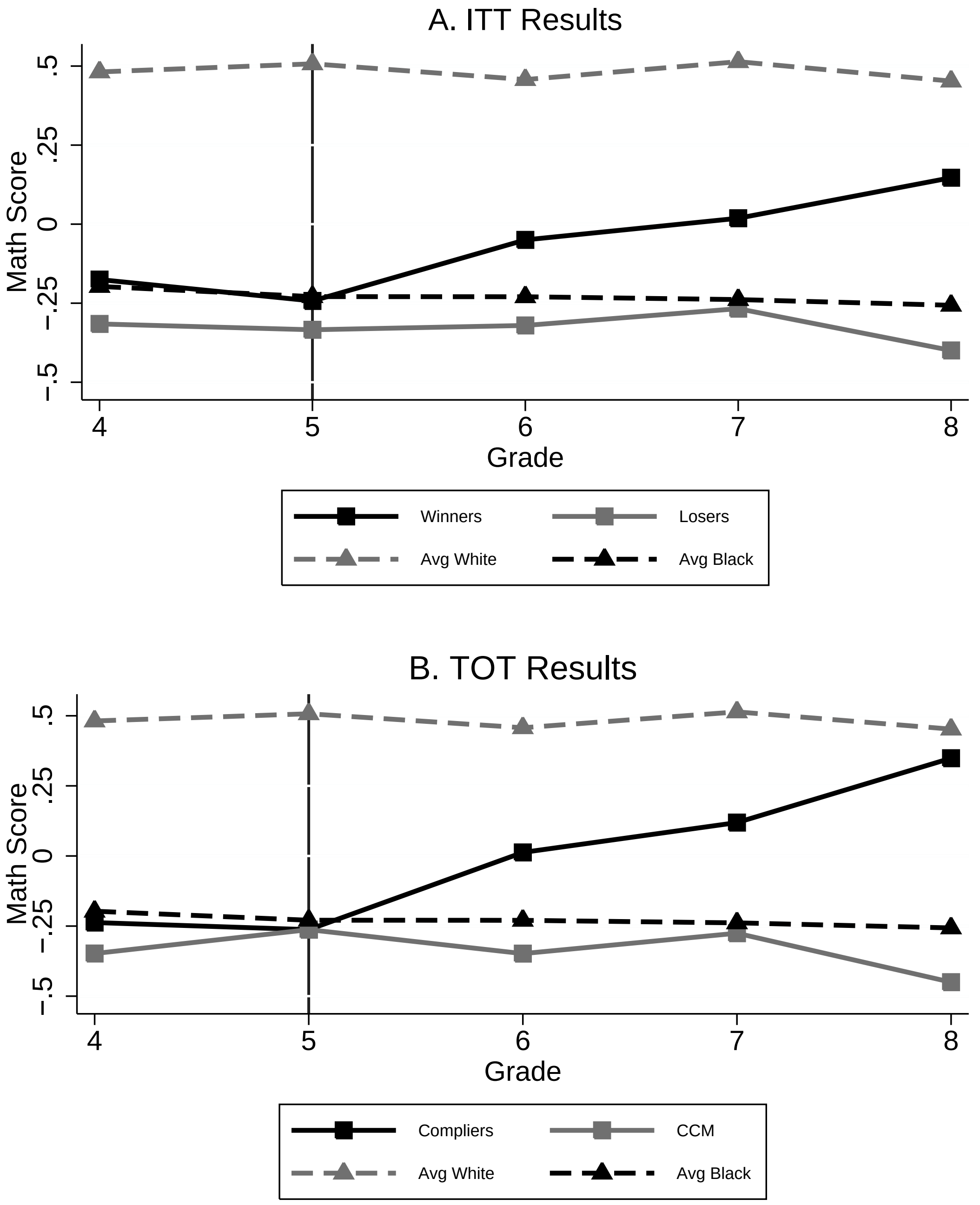
Figure 6B

NAEP 2007 Proficiency Levels by City and Race: 8th Grade Math



All means are calculated using sample weights.
N = 16473.

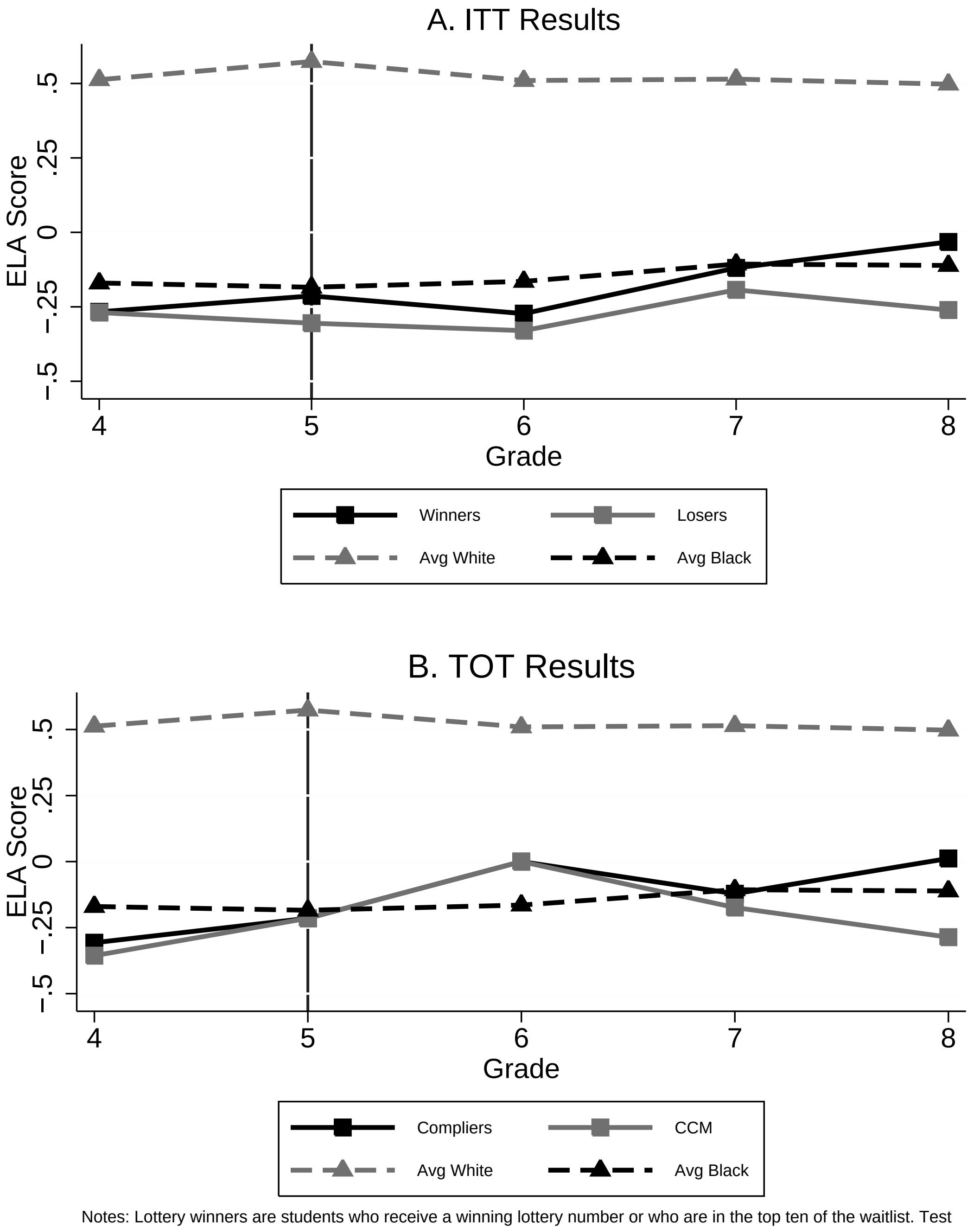
Figure 7: Student Achievement in HCZ − Math


A. ITT Results

B. TOT Results

Notes: Lottery winners are students who receive a winning lottery number or who are in the top ten of the waitlist. Test scores are standardized by grade to have mean zero and standard deviation one in the entire New York City sample. The CCM is the estimated test score for those in the control group who would have complied if they had received a winning lottery number.

Figure 8: Student Achievement in HCZ − ELA



A. ITT Results

B. TOT Results

Notes: Lottery winners are students who receive a winning lottery number or who are in the top ten of the waitlist. Test scores are standardized by grade to have mean zero and standard deviation one in the entire New York City sample. The CCM is the estimated test score for those in the control group who would have complied if they had received a winning lottery number.

Figure 9: Student Achievement in KIPP Lynn

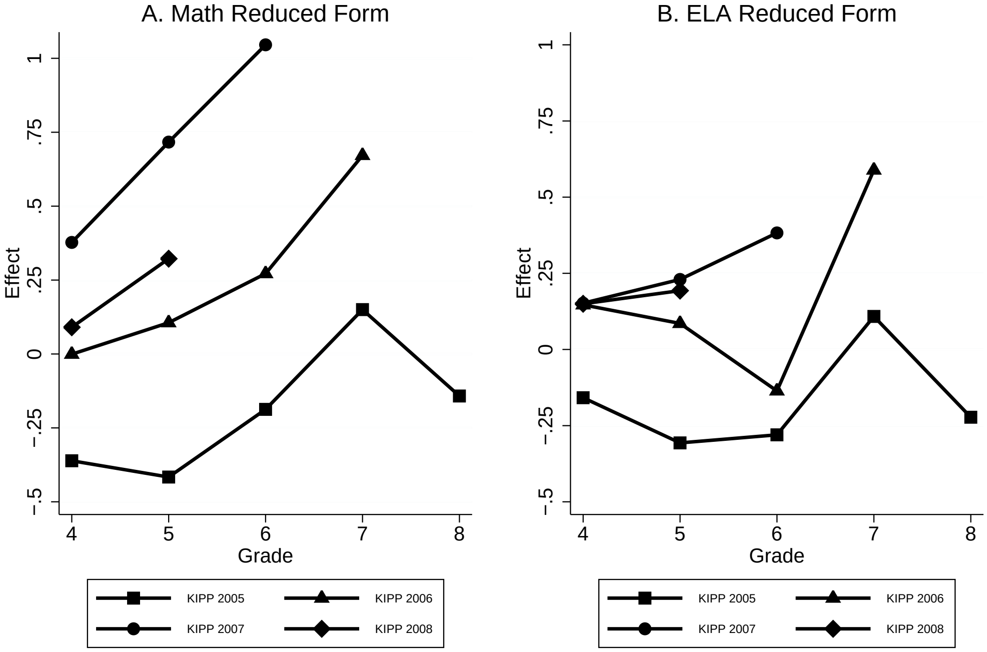


Table 1: The Importance of Educational Achievement
on Racial Differences in Labor Market Outcomes (NLSY79)

	Wage				Unemployment			
	Men		Women		Men		Women	
Black	–0.394	–0.109	–0.131	0.127	2.312	1.332	3.779	2.901
	(0.043)	(0.046)	(0.043)	(0.046)	(0.642)	(0.384)	(1.160)	(1.042)
Hispanic	–0.148	0.039	–0.060	0.161	2.170	1.529	2.759	2.181
	(0.049)	(0.047)	(0.051)	(0.051)	(0.691)	(0.485)	(0.973)	(0.871)
Age	0.027	0.012	–0.011	0.016	1.191	1.202	0.956	0.941
	(0.023)	(0.022)	(0.024)	(0.022)	(0.175)	(0.178)	(0.131)	(0.133)
AFQT		0.270		0.288		0.561		0.735
		(0.021)		(0.023)		(0.082)		(0.123)
AFQT^2		0.039		–0.009		1.005		1.276
		(0.019)		(0.020)		(0.151)		(0.161)
Obs.	1167	1167	1044	1044	1315	1315	1229	1229
R^2	0.068	0.206	0.009	0.135	0.022	0.050	0.040	0.058
% Reduction		72		197		75		32

NOTES: The dependent variable in columns 1 through 4 is the log of hourly wages of workers. The wage observations come from 2006. All wages are measured in 2006 dollars. The wage measure is created by multiplying the hourly wage at each job by the number of hours worked at each job that the person reported as a current job and then dividing that number by the total number of hours worked during a week at all current jobs. Wage observations below $1 per hour or above $115 per hour are eliminated from the data. The dependent variable in columns 5 through 8 is a binary variable indicating whether the individual is unemployed. The unemployment variable is taken from the individual's reported employment status in the raw data. In both sets of regressions, the sample consists of the NLSY79 cross-section sample plus the supplemental samples of blacks and Hispanics. Respondents who did not take the ASVAB test are included in the sample and a dummy variable is included in the regressions that include AFQT variables to indicate if a person did not have a valid AFQT score. This includes 134 respondents who had a problem with their test according to the records. All included individuals were born after 1961. The percent reduction reported in even-numbered columns represents the reduction in the coefficient on black when controls for AFQT are added. Standard errors are in parentheses.

Table 2: The Importance of Educational Achievement
on Racial Differences in Labor Market Outcomes (NLSY97)

	Wage				Unemployment			
	Men		Women		Men		Women	
Black	–0.179	–0.109	–0.153	–0.044	2.848	2.085	2.596	1.759
	(0.023)	(0.024)	(0.020)	(0.021)	(0.377)	(0.298)	(0.380)	(0.278)
Hispanic	–0.065	–0.014	–0.057	0.035	1.250	0.994	1.507	1.065
	(0.023)	(0.024)	(0.023)	(0.023)	(0.205)	(0.170)	(0.267)	(0.202)
Mixed race	0.007	0.009	–0.090	–0.057	3.268	3.216	1.317	1.278
	(0.143)	(0.145)	(0.072)	(0.065)	(1.661)	(1.618)	(0.975)	(0.911)
Age	0.064	0.062	0.039	0.039	0.934	0.937	1.084	1.081
	(0.006)	(0.006)	(0.006)	(0.006)	(0.038)	(0.038)	(0.048)	(0.048)
AFQT		0.089		0.148		0.664		0.595
		(0.011)		(0.012)		(0.049)		(0.052)
AFQT^2		–0.022		–0.035		1.248		1.140
		(0.012)		(0.012)		(0.095)		(0.107)
Obs.	3278	3278	3204	3204	3294	3294	3053	3053
R^2	0.047	0.065	0.029	0.081	0.032	0.051	0.026	0.049
% Reduction		39		71		41		52

NOTES: The dependent variable in columns 1 through 4 is the log of hourly wages of workers. The wage observations come from 2006 and 2007. All wages are measured in 2006 dollars. The wage measure for each year is created by multiplying the hourly wage at each job by the number of hours worked at each job that the person reported as a current job and then dividing that number by the total number hours worked during a week at all current jobs. If a person worked in both years, the wage is the average of the two wage observations. Otherwise the reported wage is from the year for which the individual has valid wage data. Wage observations below $1 per hour or above $115 per hour are eliminated from the data. The dependent variable in columns 5 through 8 is a binary variable indicating whether the individual is unemployed. The unemployment variable is taken from the individual's reported employment status in the raw data. The employment status from 2006 is used for determining unemployment. The coefficients in columns 5 through 8 are odds ratios from logistic regressions. Respondents who did not take the ASVAB test are included in the sample and a dummy variable is included to indicate if a person did not have a valid AFQT score in the regressions that include AFQT variables. The percent reduction reported in even- numbered columns represents the reduction in the coefficient on black when controls for AFQT are added. Standard errors are in parentheses.

Table 3: The Importance of Educational Achievement
on Racial Differences in Incarceration and Health Outcomes

	Incarceration								Physical Health			
	NLSY79				NLSY97				NLSY79			
	Men		Women		Men		Women		Men		Women	
Black	3.494	1.777	1.054	0.418	2.325	1.417	1.218	0.710	–0.151	0.011	–0.230	–0.111
	(0.549)	(0.304)	(0.484)	(0.226)	(0.245)	(0.159)	(0.244)	(0.148)	(0.053)	(0.061)	(0.068)	(0.076)
Hispanic	2.599	1.549	1.135	0.497	1.641	1.120	0.908	0.591	–0.140	–0.035	0.030	0.125
	(0.476)	(0.300)	(0.573)	(0.275)	(0.196)	(0.136)	(0.216)	(0.146)	(0.061)	(0.063)	(0.065)	(0.071)
Mixed Race					0.851	0.887	5.306	4.760				
					(0.511)	(0.557)	(2.428)	(2.207)				
Age	1.044	1.077	1.424	1.341	1.070	1.072	1.012	1.002	–0.035	–0.038	0.064	0.068
	(0.087)	(0.092)	(0.400)	(0.387)	(0.034)	(0.035)	(0.062)	(0.062)	(0.028)	(0.027)	(0.035)	(0.035)
AFQT		0.352		0.346		0.447		0.458		0.164		0.127
		(0.052)		(0.138)		(0.033)		(0.057)		(0.028)		(0.036)
AFQT^2		0.746		1.187		0.905		1.166		–0.023		–0.035
		(0.089)		(0.291)		(0.063)		(0.158)		(0.023)		(0.030)
Obs.	1989	1989	1894	1894	4599	4599	4385	4385	1588	1588	1576	1576
R^2	0.046	0.114	0.007	0.078	0.021	0.066	0.009	0.050	0.008	0.033	0.012	0.020
% Reduction		69		1178		69		233		107		52

NOTES: The dependent variable in columns 1 through 8 is a measure of whether the individual was ever incarcerated. In the NLSY79 data, this variable is equal to one if the individual reported their residence as jail during any of the yearly follow-up surveys or if they reported having been sentenced to a corrective institution before the baseline survey and is equal to zero otherwise. In the NLSY97 data, this variable is equal to one if the person reports having been sentenced to jail, an adult corrections institution, or a juvenile corrections institution in the past year during any of the yearly administrations of the survey and is equal to zero otherwise. The coefficients in columns 1 through 8 are odds ratios from logistic regressions. The dependent variable in columns 9 through 12 is the physical component score (PCS) reported in the NLSY79 derived from the 12-Item Short Form Health Survey of individuals over age 40. The PCS is standardized to have a mean of zero and a standard deviation of one. Individuals who do not have valid PCS data are not included in these regressions. In the NLSY79 regressions, included individuals were born after 1961. Respondents who did not take the ASVAB test are included in the sample and a dummy variable is included in the regressions that include AFQT variables to indicate if a person did not have a valid AFQT score. For NLSY79, this includes 134 respondent that had a problem with their test according to the records. The percent reduction reported in even-numbered columns represents the reduction in the coefficient on black when controls for AFQT are added. Standard errors are in parentheses.

Table 4: The Importance of Educational Achievement on Racial Differences in Labor Market Outcomes (C&B 76)

	Men		Women	
Black	–0.273	–0.152	0.186	0.286
	(0.042)	(0.047)	(0.035)	(0.031)
Hispanic	–0.038	–0.007	0.005	0.059
	(0.081)	(0.077)	(0.094)	(0.088)
Other race	0.153	0.147	0.271	0.270
	(0.066)	(0.062)	(0.048)	(0.049)
SAT		0.003		0.001
		(0.001)		(0.001)
SAT^2		–0.000		–0.000
		(0.000)		(0.000)
Obs.	11088	11088	8976	8976
R^2	0.007	0.015	0.004	0.012
% Reduction		44		53

NOTES: The dependent variable is the log of annual income. Annual income is reported as a series of ranges; each individual is assigned the midpoint of their reported income range as their annual income. Income data were collected for either 1994 or 1995. Individuals who report earning less than $1000 annually or who were students at the time of data collection are excluded from these regressions. Those individuals with missing SAT scores are included in the sample and a dummy variable is included in the regressions that include SAT variables to indicate that a person did not have a valid AFQT score. All regressions use institution weights and standard errors are clustered at the institution level. Standard errors are in parentheses.

Table 5: Racial Differences in the Mental Function Composite Score, ECLS-B and CPP

	Less than 1 year				2 years		4 years						7 years	
	CPP		ECLS-B		ECLS-B		CPP		ECLS-B, Math		ECLS-B, Literacy		CPP	
Black	–0.096	0.024	–0.077	0.006	–0.393	–0.219	–0.785	–0.296	–0.337	–0.130	–0.195	0.020	–0.854	–0.348
	(0.012)	(0.017)	(0.031)	(0.021)	(0.031)	(0.036)	(0.011)	(0.016)	(0.032)	(0.036)	(0.031)	(0.035)	(0.010	(0.016)
Hispanic	0.183	–0.039	–0.025	–0.021	–0.401	–0.262	–0.895	–0.542	–0.311	–0.174	–0.293	–0.103	–0.846	–0.545
	(0.034)	(0.040)	(0.029)	(0.018)	(0.028)	(0.032)	(0.032)	(0.039)	(0.029)	(0.034)	(0.028)	(0.033)	(0.031)	(0.038)
Asian	–	–	–0.027	–0.017	–0.237	–0.324	–	–	0.298	0.086	0.443	0.218	–	–
	–	–	(0.040)	(0.023)	(0.041)	(0.043)	–	–	(0.038)	(0.038)	(0.044)	(0.040)	–	–
Other Race	–0.171	–0.107	–0.023	0	–0.229	–0.135	–0.443	–0.271	–0.213	–0.066	–0.103	0.050	–0.345	–0.208
	(0.067)	(0.060)	(0.041)	(0.025)	(0.045)	(0.043)	(0.062)	(0.057)	(0.050)	(0.044)	(0.048)	(0.046)	(0.061)	(0.057)
Obs.	31,116	31,116	7468	7468	7468	7468	31,116	31,116	7468	7468	7468	7468	31,116	31,116
R^2	0.000	0.240	0.001	0.766	0.066	0.306	0.000	0.320	0.051	0.425	0.040	0.380	0.180	0.320
Controls	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y

NOTES: The dependent variable is the mental composite score, which is normalized to have a mean of zero and a standard deviation of one in each wave for the full, unweighted sample in CPP and the full sample with wave 3 weights in ECLS-B. Non-Hispanic whites are the omitted race category in each regression and all race coefficients are relative to that group. The unit of observation is a child. Estimation is done using weighted least squares for the ECLS-B sample (columns 3-6 and 9-12) using sample weights provided in the third wave of the data set. Estimation is done using ordinary least squares for the CPP sample (columns 1-2, 7-8, and 13-14). In addition to the variables included in the table, indicator variables for children with missing values on each covariate are also included in the regressions. Standard errors are in parentheses. Columns 1 through 4 present results for children under one year; Columns 5 and 6 present results for 2-year-olds; Columns 7 through 12 present results for 4-year-olds; Columns 13 and 14 present results for 7-year-olds.

Table 6: Early Childhood Interventions to Increase Achievement

Early Childhood Interventions	Ages Treated	Impact	Study
Abecedarian Project	Birth - 5 years	5 points on Wechsler Intelligence Scale at age 12; 5-7 points on various subscales of WJ-R	Campbell and Ramey (1994)
Baby College (HCZ)	Prenatal - 3 years		
Early Head Start	Prenatal - 3 years		
Early Training Project	4 - 6 years	2-5 points on Stanford-Binet IQ scores at the end of 4th grade	Gray and Klaus (1970)
Educare	Birth - 5 years		
Harlem Gems	4 - 5 years		
Harlem Study	2 - 3 years		
Head Start	3 - 5 years	0.09 standard deviations on PPVT receptive vocabulary after 1st grade; 0.08 standard deviations on WJ-III oral comprehension after 1st grade	Puma et al. (2010)
Houston Parent-Child Development Centers	1 - 2 years		
Infant Health and Development Program	Birth - 3 years	0.19 standard deviations on PPVT; 0.21 standard deviations on receptive language; 0.20 standard deviations on vocabulary, 0.16 standard deviations on reasoning, 0.22 standard deviations on visual-motor and spatial; 0.09 standard deviations on visual motor integration	Brooks-Gunn, Liaw, and Klebanov (1992)
Milwaukee Project	Birth - 6 years	23 points on Stanford-Binet IQ scores at age 6	Garber (1988)
Mother-Child Home Program	3 - 4 years		
Nurse Family Partnership	Prenatal - 2 years	4 points on Mental Development Index scores at age 2	Olds et al. (2002)
Parents as Teachers	Prenatal - 5 years		
Perry Preschool	3 - 4 years	Heckman et al. (2009) report 7-10 percent rate of return on program investment	Schweinhart, Barnes, and Weikart (1993)
Prenatal/Early Infancy Project	Prenatal - 2 years		
Syracuse University Family Development	Prenatal - 5 years		Lally, Mangione, and Honig (1987)
The Three Year Old Journey	3 years		
Tulsa Pre-K Program	4 years	Ranging from 0.38 to 0.79 standard deviations on WJ-R	Gormley et al. (2005)
Yale Experiment	Birth - 2 years		

NOTES: The set of interventions included in this table was generated in two ways. First, we used Heckman (1999) and Heckman et al. (2009) as the basis for a thorough literature review on early childhood intervention programs. We investigated all of the programs included in these papers, and then examined the papers written on this list of programs for additional programs. Second, we examined all of the relevant reports available through the IES What Works Clearinghouse. From this original list, we included twenty of the most credibly evaluated, largest scale programs in our final list.

Table 7: The Evolution of the Achievement Gap (ECLS), K-8

A. Math

	Fall K		Spring 1st		Spring 3rd		Spring 5th		Spring 8th		Spring 8th (Adjusted)	
Black	-0.393	-0.100	-0.440	-0.179	-0.498	-0.284	-0.539	-0.304	-0.522	-0.256	-0.961	-0.422
	(0.029)	(0.035)	(0.034)	(0.042)	(0.033)	(0.040)	(0.033)	(0.048)	(0.034)	(0.058)	(0.055)	(0.093)
Hispanic	-0.427	-0.104	-0.314	-0.086	-0.292	-0.074	-0.253	-0.062	-0.240	-0.014	-0.475	-0.030
	(0.024)	(0.030)	(0.025)	(0.027)	(0.025)	(0.029)	(0.025)	(0.032)	(0.025)	(0.042)	(0.045)	(0.078)
Asian	0.106	0.171	0.016	0.120	0.044	0.104	0.141	0.161	0.138	0.186	0.363	0.392
	(0.064)	(0.046)	(0.057)	(0.055)	(0.062)	(0.053)	(0.052)	(0.041)	(0.059)	(0.054)	(0.115)	(0.117)
Other Race	-0.232	-0.016	-0.215	0.015	-0.237	-0.000	-0.215	-0.048	-0.206	0.012	-0.358	0.084
	(0.052)	(0.049)	(0.047)	(0.042)	(0.044)	(0.051)	(0.047)	(0.068)	(0.050)	(0.076)	(0.093)	(0.150)
Controls	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
School FEs	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Obs.	7576	7576	7576	7576	7576	7576	7576	7576	7576	7576	7576	7576
R^2	0.116	0.533	0.106	0.564	0.127	0.627	0.141	0.682	0.136	0.667	0.135	0.665

B. Reading

	Fall K		Spring 1st		Spring 3rd		Spring 5th		Spring 8th		Spring 8th (Adjusted)	
Black	-0.246	0.009	-0.270	-0.022	-0.391	-0.160	-0.453	-0.246	-0.503	-0.168	-0.918	-0.284
	(0.031)	(0.037)	(0.034)	(0.037)	(0.035)	(0.044)	(0.034)	(0.045)	(0.036)	(0.051)	(0.060)	(0.090)
Hispanic	-0.267	-0.073	-0.160	0.003	-0.199	-0.028	-0.189	-0.007	-0.183	-0.000	-0.382	-0.004
	(0.028)	(0.033)	(0.033)	(0.029)	(0.033)	(0.035)	(0.031)	(0.032)	(0.030)	(0.035)	(0.055)	(0.065)
Asian	0.194	0.218	0.199	0.273	0.041	0.068	0.061	0.096	0.082	0.071	0.197	0.182
	(0.059)	(0.050)	(0.042)	(0.043)	(0.042)	(0.043)	(0.036)	(0.043)	(0.040)	(0.046)	(0.088)	(0.100)
Other Race	-0.175	-0.002	-0.164	0.058	-0.217	0.003	-0.188	-0.046	-0.169	0.036	-0.345	0.065
	(0.063)	(0.056)	(0.050)	(0.043)	(0.048)	(0.043)	(0.049)	(0.044)	(0.043)	(0.053)	(0.082)	(0.097)
Controls	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
School FEs	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Obs.	7091	7091	7091	7091	7091	7091	7091	7091	7091	7091	7091	7091
R^2	0.050	0.501	0.047	0.589	0.085	0.637	0.108	0.680	0.129	0.687	0.121	0.679

NOTES: The dependent variable in each column is test score from the designated subject and grade. Odd-numbered columns estimate the raw racial test score gaps and do not include any other controls. Specifications in the even-numbered columns include controls for socioeconomic status, number of books in the home (linear and quadratic terms), gender, age, birth weight, dummies for mother's age at first birth (less than twenty years old and at least thirty years old), a dummy for being a Women, Infants, Children (WIC) participant, missing dummies for all variables with missing data, and school fixed effects. Test scores are IRT scores, normalized to have mean zero and standard deviation one in the full, weighted sample. Non-Hispanic whites are the omitted race category, so all of the race coefficients are gaps relative to that group. The sample is restricted to students from whom data were collected in every wave from fall kindergarten through spring eighth grade, as well as students who have non-missing race and non-missing gender. Panel weights are used. The unit of observation is a student. Robust standard errors are located in parentheses.

Table 8: The Evolution of the Achievement Gap (ECLS), K-8: Accounting for Teachers

A. Math

	Fall K		Spring 1st		Spring 3rd		Spring 5th		Spring 8th		Spring 8th (Adjusted)	
Black	-0.100	-0.085	-0.183	-0.111	-0.284	-0.309	-0.324	-0.261	-0.258	-0.239	-0.428	-0.449
	(0.035)	(0.043)	(0.042)	(0.059)	(0.040)	(0.059)	(0.046)	(0.055)	(0.058)	(0.088)	(0.093)	(0.153)
Hispanic	-0.104	-0.049	-0.087	-0.067	-0.074	-0.050	-0.067	-0.088	-0.015	-0.064	-0.030	-0.118
	(0.030)	(0.036)	(0.027)	(0.035)	(0.029)	(0.042)	(0.032)	(0.037)	(0.042)	(0.044)	(0.078)	(0.084)
Asian	0.171	0.198	0.092	0.076	0.104	0.120	0.151	0.100	0.184	0.108	0.385	0.240
	(0.046)	(0.052)	(0.050)	(0.061)	(0.053)	(0.057)	(0.041)	(0.047)	(0.054)	(0.060)	(0.118)	(0.125)
Other Race	-0.016	0.063	0.012	-0.014	0.000	0.037	-0.051	-0.041	0.009	-0.014	0.080	-0.008
	(0.049)	(0.055)	(0.042)	(0.049)	(0.051)	(0.057)	(0.068)	(0.052)	(0.076)	(0.100)	(0.150)	(0.177)
Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
School FEs	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Teacher FEs	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Obs.	7576	7576	7514	7514	7526	7526	7484	7484	7511	7511	7511	7511
R^2	0.533	0.688	0.546	0.763	0.619	0.812	0.671	0.842	0.663	0.873	0.662	0.858

B. Reading

	Fall K		Spring 1st		Spring 3rd		Spring 5th		Spring 8th		Spring 8th (Adjusted)	
Black	0.009	0.015	-0.025	-0.011	-0.160	-0.294	-0.245	-0.178	-0.169	-0.126	-0.285	-0.233
	(0.037)	(0.042)	(0.037)	(0.051)	(0.044)	(0.050)	(0.045)	(0.048)	(0.051)	(0.050)	(0.090)	(0.091)
Hispanic	-0.073	-0.052	-0.002	-0.026	-0.028	-0.050	-0.004	-0.019	-0.002	-0.046	-0.008	-0.081
	(0.033)	(0.042)	(0.029)	(0.036)	(0.035)	(0.050)	(0.032)	(0.038)	(0.035)	(0.037)	(0.065)	(0.075)
Asian	0.218	0.239	0.257	0.208	0.068	0.022	0.094	0.017	0.069	0.031	0.180	0.093
	(0.050)	(0.056)	(0.042)	(0.061)	(0.043)	(0.050)	(0.043)	(0.057)	(0.046)	(0.057)	(0.100)	(0.121)
Other Race	-0.002	-0.010	0.055	0.077	0.003	0.010	-0.045	-0.032	0.036	0.021	0.065	0.041
	(0.056)	(0.050)	(0.043)	(0.056)	(0.043)	(0.049)	(0.044)	(0.046)	(0.052)	(0.061)	(0.097)	(0.119)
Controls	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y	Y
School FEs	Y	N	Y	N	Y	N	Y	N	Y	N	Y	N
Teacher FEs	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Obs.	7091	7091	7032	7032	7044	7044	7009	7009	7035	7035	7035	7035
R^2	0.501	0.671	0.568	0.767	0.629	0.814	0.665	0.832	0.683	0.832	0.675	0.809

NOTES: The dependent variable in each column is test score from the designated subject and grade. All specifications include controls for race, socioeconomic status, number of books in the home (linear and quadratic terms), gender, age, birth weight, dummies for mother's age at first birth (less than twenty years old and at least thirty years old), a dummy for being a Women, Infants, Children (WIC) participant, and missing dummies for all variables with missing data. Odd-numbered columns include school fixed effects, whereas even-numbered columns include teacher fixed effects. Test scores are IRT scores, normalized to have mean zero and standard deviation one in the full, weighted sample. Non-Hispanic whites are the omitted race category, so all of the race coefficients are gaps relative to that group. The sample is restricted to students from whom data were collected in every wave from fall kindergarten through spring eighth grade and students for whom teacher data was available in the relevant grade, as well as students who have non-missing race and non-missing gender. Panel weights are used. The unit of observation is a student. Robust standard errors are located in parentheses.

Table 9: Sensitivity Analysis for Losing Ground, ECLS (Fall K vs. Spring 8th)

	Math			Reading		
	Fall K	Spring 8th (Adjusted)	Lost Ground	Fall K	Spring 8th (Adjusted)	Lost Ground
Baseline	-0.063 (0.030)	-0.419 (0.057)	-0.356 (0.047)	0.076 (0.028)	-0.407 (0.064)	-0.483 (0.060)
Unweighted	-0.056 (0.019)	-0.407 (0.037)	-0.351 (0.032)	0.070 (0.020)	-0.457 (0.039)	-0.527 (0.037)
By gender:						
Males	-0.047 (0.046)	-0.446 (0.082)	-0.399 (0.066)	0.092 (0.040)	-0.374 (0.087)	-0.466 (0.085)
Females	-0.077 (0.038)	-0.385 (0.079)	-0.307 (0.068)	0.058 (0.039)	-0.430 (0.093)	-0.488 (0.083)
By SES quintile:						
Bottom	0.037 (0.049)	-0.209 (0.112)	-0.246 (0.094)	0.018 (0.051)	-0.346 (0.151)	-0.364 (0.138)
Second	-0.085 (0.059)	-0.320 (0.115)	-0.236 (0.103)	-0.006 (0.042)	-0.227 (0.124)	-0.221 (0.116)
Third	-0.113 (0.057)	-0.547 (0.110)	-0.433 (0.099)	0.079 (0.057)	-0.511 (0.132)	-0.590 (0.128)
Fourth	-0.075 (0.079)	-0.465 (0.137)	-0.390 (0.106)	0.237 (0.080)	-0.392 (0.154)	-0.629 (0.141)
Top	0.035 (0.093)	-0.348 (0.179)	-0.383 (0.135)	0.125 (0.094)	-0.517 (0.192)	-0.643 (0.186)
By family structure:						
Two biological parents	-0.091 (0.054)	-0.504 (0.094)	-0.413 (0.069)	0.079 (0.049)	-0.471 (0.092)	-0.551 (0.096)
Single mother	0.035 (0.046)	-0.264 (0.113)	-0.299 (0.102)	0.126 (0.047)	-0.154 (0.132)	-0.280 (0.122)
Teen mother at birth	-0.061 (0.050)	-0.361 (0.094)	-0.300 (0.079)	-0.021 (0.040)	-0.364 (0.121)	-0.343 (0.111)
Mother in her 20s at birth	-0.066 (0.044)	-0.463 (0.087)	-0.397 (0.072)	0.127 (0.042)	-0.440 (0.090)	-0.567 (0.086)
Mother over 30 at birth	-0.038 (0.063)	-0.335 (0.199)	-0.297 (0.165)	0.235 (0.076)	-0.201 (0.218)	-0.436 (0.234)
By region:						
Northeast	0.056 (0.057)	-0.058 (0.139)	-0.113 (0.122)	0.124 (0.057)	-0.320 (0.132)	-0.444 (0.129)
Midwest	-0.148 (0.072)	-0.604 (0.106)	-0.457 (0.097)	0.003 (0.068)	-0.422 (0.156)	-0.425 (0.156)
South	-0.065 (0.043)	-0.403 (0.081)	-0.338 (0.066)	0.044 (0.039)	-0.410 (0.086)	-0.454 (0.081)
West	0.007 (0.072)	-0.513 (0.200)	-0.520 (0.164)	0.227 (0.095)	-0.122 (0.268)	-0.349 (0.236)
By location type:						
Central city	-0.070 (0.049)	-0.466 (0.089)	-0.396 (0.072)	0.063 (0.045)	-0.439 (0.105)	-0.502 (0.097)
Suburban	-0.070 (0.054)	-0.369 (0.099)	-0.299 (0.081)	0.115 (0.053)	-0.338 (0.113)	-0.454 (0.109)
Rural	-0.101 (0.050)	-0.526 (0.163)	-0.425 (0.161)	-0.052 (0.046)	-0.566 (0.149)	-0.514 (0.155)
By school type:						
Public school	-0.073 (0.031)	-0.418 (0.061)	-0.345 (0.051)	0.071 (0.029)	-0.397 (0.067)	-0.468 (0.062)
Private school	0.006 (0.114)	-0.369 (0.172)	-0.376 (0.118)	0.075 (0.112)	-0.420 (0.228)	-0.495 (0.216)
School > 50% black	-0.261 (0.154)	-0.887 (0.318)	-0.626 (0.235)	-0.084 (0.119)	-0.550 (0.267)	-0.467 (0.287)
School > 50% white	-0.123 (0.060)	-0.409 (0.145)	-0.286 (0.135)	0.027 (0.082)	-0.423 (0.117)	-0.449 (0.140)

NOTES: Specifications in this table include controls for race, socioeconomic status, number of books in the home (linear and quadratic terms), gender, age, birth weight, dummies for mother's age at first birth (less than twenty years old and at least thirty years old), a dummy for being a Women, Infants, Children (WIC) participant, and missing dummies for all variables with missing data. Only the coefficients on black are reported. The sample is restricted to students from whom data were collected in every wave from fall kindergarten through spring eighth grade, as well as students who have non-missing race and non-missing gender. Panel weights are used (except in the specification). The top row shows results from the baseline specification across the entire sample, the second row shows the results when panel weights are omitted, and the remaining rows correspond to the baseline specification restricted to particular subsets of the data.

Table 10: Unadjusted Means on Questions Assessing Specific Sets of Skills, ECLS

	Fall K				Spring 8th			
	White	Black	Hispanic	Asian	White	Black	Hispanic	Asian
Math								
Count, number, shapes	0.964 (0.102)	0.896 (0.184)	0.851 (0.242)	0.965 (0.103)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Relative size	0.660 (0.314)	0.400 (0.313)	0.398 (0.339)	0.668 (0.325)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Ordinality, sequence	0.271 (0.334)	0.088 (0.201)	0.102 (0.218)	0.333 (0.385)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Add/subtract	0.051 (0.139)	0.009 (0.047)	0.011 (0.050)	0.088 (0.191)	1.000 (0.003)	0.998 (0.006)	0.999 (0.004)	1.000 (0.002)
Multiply/divide	0.003 (0.028)	0.000 (0.006)	0.000 (0.012)	0.006 (0.049)	0.990 (0.055)	0.955 (0.121)	0.977 (0.087)	0.989 (0.050)
Place value	0.000 (0.002)	0.000 (0.000)	0.000 (0.000)	0.000 (0.003)	0.940 (0.187)	0.769 (0.324)	0.877 (0.259)	0.947 (0.189)
Rate and measurement	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.000 (0.000)	0.762 (0.324)	0.405 (0.364)	0.606 (0.372)	0.822 (0.307)
Fractions	–	–	–	–	0.460 (0.415)	0.124 (0.268)	0.279 (0.371)	0.609 (0.426)
Area and volume	–	–	–	–	0.204 (0.323)	0.040 (0.163)	0.094 (0.223)	0.376 (0.404)
Reading								
Letter recognition	0.758 (0.279)	0.591 (0.330)	0.570 (0.346)	0.782 (0.298)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Beginning sounds	0.366 (0.340)	0.217 (0.293)	0.214 (0.287)	0.450 (0.377)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Ending sounds	0.210 (0.279)	0.113 (0.209)	0.108 (0.202)	0.298 (0.342)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)	1.000 (0.000)
Sight words	0.039 (0.139)	0.012 (0.063)	0.015 (0.089)	0.094 (0.242)	1.000 (0.001)	0.999 (0.003)	1.000 (0.001)	1.000 (0.001)
Words in context	0.018 (0.090)	0.004 (0.029)	0.007 (0.055)	0.051 (0.167)	0.992 (0.023)	0.970 (0.046)	0.987 (0.029)	0.995 (0.016)
Literal inference	0.004 (0.043)	0.000 (0.006)	0.001 (0.022)	0.013 (0.062)	0.955 (0.104)	0.851 (0.187)	0.926 (0.136)	0.969 (0.074)
Extrapolation	0.001 (0.014)	0.000 (0.000)	0.000 (0.007)	0.001 (0.009)	0.887 (0.202)	0.671 (0.303)	0.824 (0.249)	0.914 (0.161)
Evaluation	0.001 (0.009)	0.000 (0.001)	0.000 (0.005)	0.002 (0.010)	0.737 (0.261)	0.462 (0.298)	0.639 (0.282)	0.776 (0.243)
Evaluating nonfiction	–	–	–	–	0.363 (0.367)	0.113 (0.244)	0.227 (0.316)	0.441 (0.398)
Evaluating complex syntax	–	–	–	–	0.079 (0.141)	0.020 (0.063)	0.043 (0.097)	0.107 (0.155)

NOTES: Entries are unadjusted mean scores on specific areas of questions in kindergarten fall and eighth grade spring. They are proficient probability scores, which are constructed using IRT scores and provide the probability of mastery of a specific set of skills. Dashes indicate areas that were not included in kindergarten fall exams. Standard deviations are located in parentheses.

Table 11: Determinants of PIAT Math and Reading Recognition Scores, Elementary School (CNLSY79)

A. Math

	Age 5		Age 6		Age 7		Age 8		Age 9		Age 10	
Black	-0.579	-0.147	-0.622	-0.137	-0.651	-0.129	-0.661	-0.197	-0.639	-0.132	-0.649	-0.146
	(0.040)	(0.046)	(0.039)	(0.044)	(0.039)	(0.044)	(0.038)	(0.044)	(0.038)	(0.042)	(0.039)	(0.044)
Hispanic	-0.466	-0.147	-0.598	-0.193	-0.503	-0.074	-0.527	-0.127	-0.417	-0.026	-0.555	-0.135
	(0.045)	(0.047)	(0.044)	(0.046)	(0.045)	(0.046)	(0.044)	(0.046)	(0.045)	(0.045)	(0.044)	(0.046)
Mother's AFQT score		0.234		0.269		0.354		0.289		0.332		0.312
		(0.022)		(0.021)		(0.021)		(0.021)		(0.020)		(0.021)
Controls	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Obs.	3118	3118	3208	3208	3228	3228	3217	3217	3199	3199	3107	3107
R^2	0.101	0.193	0.125	0.248	0.124	0.265	0.155	0.254	0.146	0.286	0.157	0.284

B. Reading Recognition

	Age 5		Age 6		Age 7		Age 8		Age 9		Age 10	
Black	-0.174	0.395	-0.207	0.246	-0.331	0.193	-0.557	-0.083	-0.525	0.012	-0.590	-0.093
	(0.042)	(0.045)	(0.039)	(0.044)	(0.040)	(0.043)	(0.040)	(0.044)	(0.040)	(0.043)	(0.040)	(0.046)
Hispanic	-0.402	0.017	-0.349	0.037	-0.273	0.158	-0.442	-0.025	-0.290	0.124	-0.435	-0.001
	(0.047)	(0.047)	(0.045)	(0.046)	(0.046)	(0.047)	(0.045)	(0.048)	(0.047)	(0.046)	(0.046)	(0.049)
Mother's AFQT score		0.314		0.276		0.329		0.308		0.337		0.337
		(0.021)		(0.022)		(0.021)		(0.021)		(0.021)		(0.021)
Controls	N	Y	N	Y	N	Y	N	Y	N	Y	N	Y
Obs.	3052	3052	3174	3174	3224	3224	3216	3216	3195	3195	3106	3106
R^2	0.069	0.234	0.110	0.229	0.078	0.246	0.092	0.224	0.083	0.266	0.093	0.235

NOTES: The dependent variable in each column is the Peabody Individual Achievement Test (PIAT) score for the designated subject and age. All specifications include dummies for the child's age in months and dummies for the year in which the assessment was administered. Odd-numbered columns estimate the raw racial test score gaps and also include a dummy for missing race. Non-black, non-Hispanic respondents are the omitted race category, so all of the race coefficients are gaps relative to that group. Specifications in the even-numbered columns include controls for gender, free lunch status, special education status, a dummy for attending a private school, parents' income, the Home Observation for Measurement of Environment (HOME) inventory, which is an inventory of measures related to the quality of the home environment, mother's AFQT score (standardized across the entire sample of mothers in our dataset), and dummies for the mother's birth year. Also included are missing dummies for all variables with missing data. Robust standard errors are located in parentheses. See data appendix for details of the sample construction.

Table 12: Determinants of PIAT Math and Reading Recognition Scores, Middle School (CNLSY79)

A. Math

	Age 11		Age 12		Age 13		Age 14	
Black	-0.681	-0.193	-0.729	-0.253	-0.685	-0.192	-0.781	-0.250
	(0.039)	(0.044)	(0.040)	(0.046)	(0.040)	(0.043)	(0.056)	(0.060)
Hispanic	-0.520	-0.112	-0.558	-0.148	-0.489	-0.084	-0.577	-0.111
	(0.047)	(0.049)	(0.046)	(0.049)	(0.049)	(0.049)	(0.066)	(0.068)
Mother's AFQT score		0.318		0.325		0.350		0.351
		(0.022)		(0.022)		(0.021)		(0.031)
Controls	N	Y	N	Y	N	Y	N	Y
Obs.	3022	3022	2824	2824	2738	2738	1443	1443
R^2	0.160	0.292	0.163	0.288	0.151	0.302	0.173	0.328

B. Reading Recognition

	Age 11		Age 12		Age 13		Age 14	
Black	-0.583	-0.069	-0.600	-0.119	-0.579	-0.067	-0.697	-0.251
	(0.040)	(0.045)	(0.043)	(0.046)	(0.042)	(0.045)	(0.058)	(0.063)
Hispanic	-0.332	0.106	-0.350	0.064	-0.275	0.153	-0.408	-0.013
	(0.048)	(0.049)	(0.048)	(0.050)	(0.051)	(0.051)	(0.066)	(0.069)
Mother's AFQT score		0.343		0.329		0.362		0.324
		(0.022)		(0.022)		(0.022)		(0.031)
Controls	N	Y	N	Y	N	Y	N	Y
Obs.	3012	3012	2830	2830	2740	2740	1452	1452
R^2	0.105	0.266	0.093	0.236	0.093	0.270	0.135	0.271

NOTES: The dependent variable in each column is the Peabody Individual Achievement Test (PIAT) score for the designated subject and age. All specifications include dummies for the child's age in months and dummies for the year in which the assessment was administered. Odd-numbered columns estimate the raw racial test score gaps and also include a dummy for missing race. Non-black, non-Hispanic respondents are the omitted race category, so all of the race coefficients are gaps relative to that group. Specifications in the even-numbered columns include controls for gender, free lunch status, special education status, a dummy for attending a private school, parents' income, the Home Observation for Measurement of Environment (HOME) inventory, which is an inventory of measures related to the quality of the home environment, mother's AFQT score (standardized across the entire sample of mothers in our dataset), and dummies for the mother's birth year. Also included are missing dummies for all variables with missing data. Robust standard errors are located in parentheses. See data appendix for details regarding sample construction.

Table 13: Racial Achievement Gap in Urban Districts

A. Math

	New York City			Washington, DC			Dallas			Chicago		
Black	-0.696	-0.536	-0.346	-1.162	-0.747	-0.657	-0.690	-0.678	-0.528	-0.978	-0.740	-0.522
	(0.024)	(0.020)	(0.005)	(0.089)	(0.049)	(0.029)	(0.124)	(0.108)	(0.031)	(0.049)	(0.032)	(0.011)
Hispanic	-0.615	-0.335	-0.197	-0.830	-0.401	-0.461	-0.392	-0.230	-0.079	-0.687	-0.435	-0.254
	(0.023)	(0.018)	(0.005)	(0.114)	(0.053)	(0.034)	(0.121)	(0.104)	(0.030)	(0.046)	(0.028)	(0.010)
Asian	0.266	0.335	0.345	-0.056	0.105	0.058	0.216	0.270	0.348	0.270	0.423	0.337
	(0.022)	(0.021)	(0.005)	(0.100)	(0.053)	(0.046)	(0.131)	(0.118)	(0.063)	(0.053)	(0.050)	(0.015)
Other race	-0.566	-0.420	-0.247	-0.155	-0.015	0.021	-0.407	-0.405	-0.226	-0.256	-0.194	-0.251
	(0.032)	(0.028)	(0.018)	(0.188)	(0.153)	(0.164)	(0.180)	(0.177)	(0.122)	(0.084)	(0.072)	(0.051)
Controls	N	Y	Y	N	Y	Y	N	Y	Y	N	Y	Y
School FEs	N	N	Y	N	N	Y	N	N	Y	N	N	Y
Obs.	434593	434593	434593	20331	20331	20331	33561	33561	33561	177787	177787	177787
R^2	0.131	0.283	0.362	0.111	0.285	0.405	0.030	0.084	0.149	0.108	0.145	0.240
% Reduction		22.9	35.6		35.7	12.1		1.8	22.2		24.3	29.5

B. Reading

	New York City			Washington, DC			Dallas			Chicago		
Black	-0.634	-0.455	-0.285	-1.163	-0.708	-0.599	-0.782	-0.761	-0.561	-0.846	-0.587	-0.381
	(0.025)	(0.020)	(0.005)	(0.073)	(0.044)	(0.030)	(0.137)	(0.119)	(0.031)	(0.046)	(0.029)	(0.012)
Hispanic	-0.670	-0.328	-0.194	-1.004	-0.410	-0.444	-0.680	-0.473	-0.278	-0.714	-0.433	-0.253
	(0.024)	(0.019)	(0.005)	(0.097)	(0.049)	(0.035)	(0.133)	(0.114)	(0.029)	(0.045)	(0.027)	(0.010)
Asian	0.007	0.103	0.121	-0.410	-0.172	-0.204	-0.195	-0.104	0.002	0.029	0.194	0.151
	(0.023)	(0.019)	(0.005)	(0.101)	(0.052)	(0.048)	(0.133)	(0.114)	(0.062)	(0.051)	(0.035)	(0.015)
Other race	-0.559	-0.395	-0.249	-0.251	-0.063	-0.052	-0.497	-0.496	-0.290	-0.105	-0.034	-0.091
	(0.031)	(0.027)	(0.019)	(0.161)	(0.102)	(0.167)	(0.187)	(0.188)	(0.121)	(0.081)	(0.067)	(0.053)
Controls	N	Y	Y	N	Y	Y	N	Y	Y	N	Y	Y
School FEs	N	N	Y	N	N	Y	N	N	Y	N	N	Y
Obs.	426806	426806	426806	20243	20243	20243	28126	28126	28126	176767	176767	176767
R^2	0.087	0.273	0.335	0.095	0.282	0.380	0.030	0.115	0.180	0.069	0.126	0.205
% Reduction		28.4	37.3		39.1	15.4		2.6	26.2		30.6	35.1

NOTES: The dependent variable in each column is the state assessment in that subject taken during the 2008-09 school year. For New York City, these are the New York State mathematics and English Language Arts (ELA) exams. For Washington, DC, these are the District of Columbia Comprehensive Assessment System (DC-CAS) mathematics and reading exams. For Dallas, these are the Texas Assessment of Knowledge and Skills (TAKS) mathematics and reading exams (English versions). For Chicago, these are the Illinois Standards Achievement Test (ISAT) mathematics and reading exams. All test scores are standardized to have mean zero and standard deviation one within each grade. Non-Hispanic whites are the omitted race category, so all of the race coefficients are gaps relative to that group. The New York City and Chicago specifications include students in grades three through eight. Washington, DC, includes students in grades three through eight and ten. Dallas includes students in grades three through five. The first specification for each city estimates the raw racial test score gap in each city and does not include any other controls. The second specification for each city includes controls for gender, free lunch status, English language learner (ELL) status, special education status, age in years (linear, quadratic, and cubic terms), census block group income quintile dummies, and missing dummies for all variables with missing data. The third specification includes the same set of controls as well as school fixed effects. Age, special education status, and income data are not available in the Chicago data. Standard errors, located in parentheses, are clustered at the school level. Percent reduction refers to the percent by which the magnitude of the coefficient on black is reduced relative to the coefficient on black in the preceding column. See data appendix for details regarding sample and variable construction.

Table 14: Racial Achievement Gap in Urban Districts:
Accounting for Teachers

A. NYC

	Math		Reading	
Black	-0.350	-0.280	-0.286	-0.214
	(0.005)	(0.005)	(0.006)	(0.005)
Hispanic	-0.198	-0.149	-0.193	-0.139
	(0.005)	(0.005)	(0.005)	(0.005)
Asian	0.350	0.331	0.124	0.110
	(0.005)	(0.005)	(0.006)	(0.005)
Other race	-0.246	-0.195	-0.251	-0.204
	(0.019)	(0.018)	(0.020)	(0.019)
Controls	Y	Y	Y	Y
School FEs	Y	N	Y	N
Teacher FEs	N	Y	N	Y
Obs.	398062	398062	391854	391854
R^2	0.359	0.477	0.332	0.445
% Reduction		20.0		25.0

B. Dallas

	Math		Reading	
Black	-0.530	-0.525	-0.563	-0.546
	(0.031)	(0.032)	(0.031)	(0.031)
Hispanic	-0.079	-0.099	-0.278	-0.270
	(0.030)	(0.030)	(0.029)	(0.030)
Asian	0.347	0.313	-0.004	-0.025
	(0.063)	(0.063)	(0.063)	(0.063)
Other race	-0.227	-0.155	-0.289	-0.244
	(0.122)	(0.121)	(0.121)	(0.121)
Controls	Y	Y	Y	Y
School FEs	Y	N	Y	N
Teacher FEs	N	Y	N	Y
Obs.	33507	33507	27949	27949
R^2	0.149	0.255	0.181	0.274
% Reduction		0.9		3.0

NOTES: The dependent variable in each column is the state assessment in that subject taken during the 2008-09 school year. For New York City, these are the New York State mathematics and English Language Arts (ELA) exams. For Dallas, these are the Texas Assessment of Knowledge and Skills (TAKS) mathematics and reading exams (English versions). All test scores are standardized to have mean zero and standard deviation one within each grade. Non-Hispanic whites are the omitted race category, so all of the race coefficients are gaps relative to that group. The New York City specifications include students in grades three through eight. The Dallas specifications include students in grades three through five. All specifications include controls for gender, free lunch status, English language learner (ELL) status, special education status, age in years (linear, quadratic, and cubic terms), census block group income quintile dummies, and missing dummies for all variables with missing data. Odd-numbered columns include school fixed effects, whereas even-numbered columns include teacher fixed effects. The samples are restricted to students for whom teacher data in the relevant subject are available. Standard errors are located in parentheses. Percent reduction refers to the percent by which the magnitude of the coefficient on black is reduced relative to the coefficient on black in the preceding column. See data appendix for details regarding sample and variable construction.

Table 15: Unadjusted Means on Questions Assessing Specific Sets of Skills, NYC

A. Elementary School

	3rd Grade		4th Grade		5th Grade	
	Black	White	Black	White	Black	White
Math						
Math st. 1: number sense/operations	-0.192 (1.053)	0.338 (0.812)	-0.233 (1.032)	0.393 (0.801)	-0.229 (1.006)	0.378 (0.836)
Math st. 2: algebra	-0.196 (1.088)	0.274 (0.777)	-0.172 (1.079)	0.294 (0.769)	-0.221 (1.081)	0.306 (0.790)
Math st. 3: geometry	-0.130 (1.048)	0.220 (0.824)	-0.178 (1.046)	0.311 (0.853)	-0.231 (1.028)	0.380 (0.849)
Math st. 4: measurement	-0.210 (1.102)	0.258 (0.796)	-0.242 (1.009)	0.418 (0.838)	-0.265 (1.044)	0.363 (0.807)
Math st. 5: statistics/probability	-0.200 (1.066)	0.283 (0.815)	-0.213 (1.030)	0.316 (0.871)	-0.227 (1.006)	0.368 (0.894)
ELA						
ELA st. 1: information and understanding	-0.105 (1.024)	0.322 (0.843)	-0.093 (1.010)	0.383 (0.835)	-0.134 (0.996)	0.383 (0.844)
ELA st. 2: literary response and expression	-0.138 (1.015)	0.374 (0.807)	-0.167 (0.976)	0.462 (0.917)	-0.095 (1.019)	0.304 (0.840)
ELA st. 3: critical analysis and evaluation	-0.102 (0.996)	0.349 (0.931)	-0.102 (1.025)	0.350 (0.825)	-0.171 (1.021)	0.369 (0.859)

B. Middle School

	6th Grade		7th Grade		8th Grade	
	Black	White	Black	White	Black	White
Math						
Math st. 1: number sense/operations	-0.261 (0.953)	0.452 (0.890)	-0.233 (0.962)	0.433 (0.877)	-0.225 (0.975)	0.366 (0.915)
Math st. 2: algebra	-0.209 (1.037)	0.393 (0.794)	-0.241 (0.981)	0.402 (0.894)	-0.274 (0.945)	0.431 (0.907)
Math st. 3: geometry	-0.249 (0.970)	0.397 (0.910)	-0.218 (1.003)	0.360 (0.873)	-0.262 (1.000)	0.390 (0.864)
Math st. 4: measurement	-0.215 (1.021)	0.349 (0.849)	-0.286 (0.906)	0.497 (0.939)	-0.198 (1.029)	0.313 (0.840)
Math st. 5: statistics/probability	-0.222 (0.994)	0.425 (0.846)	-0.235 (0.984)	0.465 (0.826)	–	–
ELA						
ELA st. 1: information and understanding	-0.111 (1.008)	0.322 (0.866)	-0.096 (0.967)	0.406 (0.806)	-0.163 (0.945)	0.456 (0.867)
ELA st. 2: literary response and expression	-0.176 (0.957)	0.448 (0.870)	-0.126 (0.973)	0.438 (0.834)	-0.060 (0.972)	0.360 (0.887)
ELA st. 3: critical analysis and evaluation	-0.099 (1.016)	0.286 (0.815)	-0.130 (0.973)	0.419 (0.852)	-0.036 (1.006)	0.224 (0.935)

NOTES: Entries are unadjusted mean percentage of items correct on specific areas of questions on the New York State assessments in mathematics and English Language Arts (ELA) in third through eighth grades in New York City, which are then standardized across the entire sample of test takers for each grade, so that units are standard deviations relative to the mean. Dashes indicate that Statistics/Probability was not included in the eighth grade mathematics exam. Standard deviations are located in parentheses.

Table 16: Evolution of the Achievement Gap over Time, NELS

A. Math

	8th Grade				10th Grade				12th Grade			
Black	-0.754	-0.526	-0.400	-0.343	-0.734	-0.500	-0.410	-0.288	-0.778	-0.543	-0.445	-0.581
	(0.025)	(0.021)	(0.021)	(0.031)	(0.038)	(0.034)	(0.032)	(0.060)	(0.045)	(0.042)	(0.045)	(0.089)
Hispanic	-0.581	-0.349	-0.236	-0.200	-0.573	-0.301	-0.220	-0.166	-0.544	-0.267	-0.212	-0.259
	(0.025)	(0.022)	(0.023)	(0.034)	(0.035)	(0.032)	(0.032)	(0.064)	(0.039)	(0.036)	(0.037)	(0.105)
Asian	0.186	0.134	0.170	0.127	0.251	0.168	0.132	0.018	0.235	0.145	0.119	-0.118
	(0.054)	(0.045)	(0.032)	(0.048)	(0.056)	(0.051)	(0.043)	(0.082)	(0.065)	(0.057)	(0.052)	(0.087)
Controls	N	Y	Y	Y	N	Y	Y	Y	N	Y	Y	Y
School FEs	N	N	Y	N	N	N	Y	N	N	N	Y	N
Teacher FEs	N	N	N	Y	N	N	N	Y	N	N	N	Y
Obs.	23648	23648	23648	10981	17793	17793	17793	7316	14236	14236	14236	5668
R^2	0.099	0.253	0.354	0.509	0.102	0.277	0.464	0.761	0.103	0.281	0.471	0.829

B. English

	8th Grade				10th Grade				12th Grade		
Black	-0.686	-0.495	-0.399	-0.368	-0.641	-0.435	-0.377	-0.336	-0.661	-0.479	-0.430
	(0.025)	(0.022)	(0.023)	(0.034)	(0.048)	(0.042)	(0.038)	(0.056)	(0.044)	(0.042)	(0.071)
Hispanic	-0.572	-0.358	-0.242	-0.206	-0.504	-0.264	-0.211	-0.113	-0.479	-0.270	-0.250
	(0.029)	(0.026)	(0.025)	(0.037)	(0.037)	(0.033)	(0.036)	(0.061)	(0.036)	(0.040)	(0.042)
Asian	-0.082	-0.123	-0.072	-0.103	0.024	-0.048	-0.050	-0.134	0.081	0.003	0.020
	(0.048)	(0.040)	(0.032)	(0.050)	(0.057)	(0.049)	(0.045)	(0.083)	(0.062)	(0.051)	(0.049)
Controls	N	Y	Y	Y	N	Y	Y	Y	N	Y	Y
School FEs	N	N	Y	N	N	N	Y	N	N	N	Y
Teacher FEs	N	N	N	Y	N	N	N	Y	N	N	N
Obs.	23643	23643	23643	11158	17832	17832	17832	8962	14230	14230	14230
R^2	0.080	0.211	0.293	0.409	0.079	0.226	0.417	0.638	0.084	0.219	0.414

C. History

	8th Grade				10th Grade				12th Grade		
Black	-0.660	-0.453	-0.340	-0.311	-0.599	-0.390	-0.332	-0.303	-0.621	-0.429	-0.302
	(0.028)	(0.024)	(0.023)	(0.038)	(0.041)	(0.037)	(0.035)	(0.078)	(0.047)	(0.042)	(0.047)
Hispanic	-0.590	-0.369	-0.233	-0.248	-0.518	-0.275	-0.156	-0.238	-0.501	-0.266	-0.210
	(0.028)	(0.026)	(0.026)	(0.043)	(0.037)	(0.033)	(0.035)	(0.080)	(0.041)	(0.040)	(0.042)
Asian	-0.020	-0.066	0.003	-0.022	0.030	-0.049	0.018	-0.008	0.093	0.008	0.041
	(0.052)	(0.045)	(0.033)	(0.053)	(0.058)	(0.050)	(0.049)	(0.112)	(0.068)	(0.057)	(0.062)
Controls	N	Y	Y	Y	N	Y	Y	Y	N	Y	Y
School FEs	N	N	Y	N	N	N	Y	N	N	N	Y
Teacher FEs	N	N	N	Y	N	N	N	Y	N	N	N
Obs.	23525	23525	23525	10297	17591	17591	17591	4567	14063	14063	14063
R^2	0.079	0.200	0.316	0.407	0.072	0.208	0.423	0.625	0.082	0.217	0.432

D. Science

	8th Grade				10th Grade				12th Grade			
Black	-0.792	-0.589	-0.437	-0.434	-0.848	-0.640	-0.465	-0.505	-0.911	-0.731	-0.574	-0.560
	(0.024)	(0.022)	(0.023)	(0.033)	(0.038)	(0.036)	(0.034)	(0.064)	(0.041)	(0.037)	(0.058)	(0.141)
Hispanic	-0.588	-0.377	-0.246	-0.203	-0.627	-0.382	-0.264	-0.106	-0.617	-0.389	-0.294	-0.245
	(0.026)	(0.025)	(0.025)	(0.039)	(0.034)	(0.032)	(0.033)	(0.069)	(0.038)	(0.037)	(0.040)	(0.148)
Asian	-0.045	-0.079	0.015	0.024	0.042	-0.040	0.023	-0.098	0.019	-0.056	-0.046	-0.054
	(0.053)	(0.046)	(0.033)	(0.052)	(0.061)	(0.054)	(0.044)	(0.110)	(0.056)	(0.046)	(0.050)	(0.106)
Controls	N	Y	Y	Y	N	Y	Y	Y	N	Y	Y	Y
School FEs	N	N	Y	N	N	N	Y	N	N	N	Y	N
Teacher FEs	N	N	N	Y	N	N	N	Y	N	N	N	Y
Obs.	23616	23616	23616	10575	17684	17684	17684	6148	14134	14134	14134	3715
R^2	0.099	0.210	0.310	0.375	0.113	0.253	0.444	0.648	0.127	0.256	0.448	0.772

NOTES: The dependent variable in each column is the NELS test score in the designated subject and grade. Test scores are IRT scores, normalized to have mean zero and standard deviation one in each grade. Non-Hispanic whites are the omitted race category, so all of the race coefficients are gaps relative to that group. The first specification for each grade and subject estimates the raw racial test score gap in that grade and only include race dummies and a dummy for missing race. The second specification for each grade and subject includes controls for gender, age (linear, quadratic, and cubic terms), family income, and dummies that indicate parents' level of education, as well as missing dummies for all variables with missing data. The third specification includes the same set of controls as well as school fixed effects. For grades eight through twelve of math and science, and for grades eight and ten of English and history, the fourth specification includes the same set of controls as well as teacher fixed effects. For grade twelve of English and history, teacher data were not collected in the second follow-up year of the NELS, so teacher fixed effects cannot be included. Standard errors, located in parentheses, are clustered at the school level.

Table 17: School-Age Interventions to Increase Achievement

Program	**Grades Treated**	**Treatment**	**Impact**	**Study**
Career Academies	9th - 12th	Small school model that combines academic and technical curricula and provides students with work-based learning opportunities	Eleven percent higher earnings per year (ages 18-27)	Kemple (2008)
Comer School Development Program	K - 12th	Whole-school reform model that aims to improve intra-school relations and climate in order to improve academic achievement.	No achievement effects (7th-8th grades)	Cook et al. (1999)
Experience Corps	1st - 3rd	This program trains older adults (55+) to tutor and mentor elementary school children who are at risk of academic failure.	0.13 standard deviation on reading comprehension; 0.16 standard deviation on general reading skills	Morrow-Howell et al. (2009)
Language Essentials for Teachers of Reading and Spelling (LETRS)	K - 12th	Teachers received professional development during the summer and following school year focused around the LETRS model of language instruction	No significant impact (2nd grade)	Garet et al. (2008)
Learnfare	7th - 12th	This conditional cash transfer program sanctions a family's welfare grant if teenagers in the family do not meet required school attendance goals.	Increased school enrollment and attendance (ages13-19)	Dee (2009)
Mastery Learning	K - 12th	This group-based, teacher-paced instructional model requires that students master a particular objective before moving to a new objective. Students are evaluated on absolute scales as opposed to norm-referenced scales.	0.78 standard deviations on achievement tests (on average)	Guskey and Gates (1985)
National Guard Youth ChalleNGe Program	10th-12th	This 17-month program for high school dropouts has residential and post-residential phases. The residential phase provides students with a highly structured "quasi-military" experience and the post-residential phase provides students with mentoring.	Increased percentage earned a high school diploma or GED within 9 months	Bloom, Gardenhire-Crooks, and Mandsager (2009)
NYC voucher program	K - 4th	This program provided low-income students in NYC with vouchers worth up to $1,400 per year for three years to attend private schools.	No significant impact	Krueger and Zhu (2002)
Project CRISS	4th - 12th	This teacher professional development model aims to give teachers more effective strategies for teaching reading and writing that focus on student-owned reading strategies.	No significant impact (5th grade)	James-Burdumy et al. (2009)
Quantum Opportunity Program	9th - 12th	This program had high school students participate in 250 hours of educational services, 250 hours of development activities, and 250 hours of community service and provided students with financial incentives.	Thirty-three percent more graduated from high school	Taggart (1995)

Program	Grades Treated	Treatment	Impact	Study
Seattle Social Development Project	1st - 6th	Teachers received training to allow them to teach elementary school students social skills focused around problem-solving in conflict resolution.	No reported achievement outcomes	Hawkins et al. (2008)
Self-affirmation essay writing	7th - 8th	Students were given structured writing assignments that required them to write about their personal values and the importance of those values.	0.24 standard deviations on GPA for black students; 0.41 standard deviations on GPA for low-achieving black students	Cohen et al. (2009)
Success for All	K - 5th	This program is a school-wide program that focuses on early detection of and intervention around reading problems using a ability-level reading group instruction.	0.36 standard deviations on phonemic awareness; 0.24 standard deviations on word identification; 0.21 standard deviations on passage comprehension (2nd grade)	Borman et al. (2007)
Summer Training and Education Program (STEP)	9th - 10th	This program provided summer reading and math remediation along with life skills instruction to academically struggling low-income students.	No long-term impact (ages 14-15)	Walker and Vilella-Velez (1992)
Supplemental reading instruction	9th	Students who were two to five years below grade level in reading were provided with full-year supplemental literacy courses that provided an average of eleven hours per month of supplemental instruction.	0.08 standard deviations on reading comprehension	Corrin et al. (2009)
Talent Development High School	9th - 12th	This comprehensive school reform model aims to establish a positive school climate and prepare all students academically for college. Two key features are the ninth-grade academy and upper grade career academies.	No significant impact on standardized tests	Kemple, Herlihy, and Smith (2005)
U.S. Department of Education Student Mentoring Program	4th - 8th	Students were matched with adult or peer mentors with whom they met weekly for six months to discuss academics, relationships, and future plans.	No significant impact	Bernstein et al. (2009)

NOTES: The set of interventions included in this table were generated using a two-step search process. First, a keyword search for for "school-aged interventions" was performed in Google Scholar, JSTOR, and the National Bureau of Economic Research database. Second, we examined all of the available reports for the appropriate age groups from the What Works Clearinghouse of IES. From the original list, we narrowed our focus to those programs that contained credible identification and were large enough in scale to possibly impact achievement gaps overall.