

Thirteen Ways of Looking at the Black-White Test Score Gap

sean f. reardon
Stanford University

March, 2007

ROUGH DRAFT: for discussion only

*I was of three minds,
Like a tree
In which there are three blackbirds*

--Wallace Stevens, "*Thirteen Ways of Looking at a Blackbird*"

Direct correspondence to sean.reardon@stanford.edu. I appreciate the thoughtful comments of Steve Raudenbush, Derek Neal, participants in the University of Chicago Education Workshop, the University of Chicago Workshop on Black-White Inequality, and the Stanford Institute for Research on Education Policy and Practice Research Seminar. The errors are mine.

Thirteen Ways of Looking at the Black-White Test Score Gap

sean f. reardon
Stanford University

Abstract

The black-white cognitive test score gap is a stubborn feature of U.S. schooling and society. The patterns and causes of the development of black-white test score gaps as children age and progress through school, however, are not well understood, despite considerable recent study. In part, the absence of a detailed descriptive picture of the development of racial test score disparities is due to differences among studies in the tests and metrics used to measure the gap, the need to account for measurement error in test scores, and the complexity of interpreting between- and within-school test score disparities. In this paper, I use data from a nationally representative sample of children enrolled in kindergarten in the fall of 1998 to describe the patterns and development of black-white test score disparities through the first six years of children's schooling, to examine the extent to which these gaps grow differently among initially high- and low-achieving students, and to describe the extent to which these gaps grow among students attending the same or different schools.

Introduction

The black-white cognitive test score gap remains a stubborn feature of U.S. schooling and society. National studies consistently show that the average non-Hispanic black student scores well below the average non-Hispanic white student on standardized tests of math and reading skills, as does the average Hispanic student (see, for example, Fryer and Levitt 2004; Hedges and Nowell 1999; Jencks and Phillips 1998; Neal 2005). The patterns and causes of the development of black-white test score gaps as children age and progress through school, however, are not well understood, despite considerable recent study. In part, the absence of a detailed descriptive picture of the development of racial test score disparities is due to differences among studies in the tests and metrics used to measure the gap, the need to account for measurement error in test scores, and the complexity of interpreting between- and within-school test score disparities.

From a societal perspective, the black-white test score gap remains salient because of the long history of racial inequality in the United States and the importance of cognitive skills in processes of social stratification and social mobility. From a labor market perspective, achievement disparities are important primarily because test scores disparities in elementary and secondary school are highly predictive of corresponding disparities in subsequent labor market outcomes. Data from the most recent Annual Demographic Survey (March Supplement) of the Current Population Survey (CPS) show that the median black worker earns 28% less than the median white full-time male worker. For female full-time workers, the corresponding gap is 15%.¹ Recent estimates suggest that at least one half (and maybe all) of these wage disparities are attributable to differences in cognitive skills obtained prior to entering the labor force (Bollinger 2003; Carneiro, Heckman, and Masterov

¹ Source: Annual Demographic Survey (March Supplement) of the 2006 Current Population Survey (CPS), Table PINC-10. Wage and Salary Workers—People 15 Years Old and Over, By Total Wage and Salary Income in 2005, Work Experience in 2005, Race, Hispanic Origin, and Sex.

2003; Neal and Johnson 1996).²

In addition to concerns regarding the magnitude of the differences in mean test scores among individuals of different racial groups, a number of researchers have called attention to the effects of racial disparities at the upper end of the achievement distribution. Neal (2005, see Figures 2a-2d), for example, shows that roughly 5% of Black students aged 13-17 years old in the 1990s had math scores in the top quartile of the White math score distribution. This means that Black students are underrepresented by 80% in the top quartile of the distribution, a finding that has enormous implications for Black students' access to elite colleges and employment in jobs with the highest skill demands (and the highest pay). In addition, recent evidence indicates that the increase in the returns to education in the 1980s was largest for those in the top quartile of the achievement distribution (Heckman and Vytlacil 2001). Because Whites are substantially overrepresented in the highest quartile of the achievement distribution, this pattern suggests that racial disparities at the top of the achievement distribution have become increasingly salient in shaping labor market and social inequality.

Key Questions About Black-White Test Score Gaps

Recent research on the black-white achievement gap has called attention to five key questions regarding the gaps. First, how does the size of achievement gaps change as students progress through school (within cohorts)? Second, do achievement gaps grow faster or slower among students with initially higher achievement? Third, to what extent is the growth in achievement gaps attributable to differences in the growth rates of students attending the same or different schools? Fourth, how much of the achievement gaps and their growth over time can be

² With regard to wage gaps for women, the evidence is less clear because of differential selection into the labor force among women. Among women in the labor force, however, Black and Hispanic women earn, on average, the same or more than White women after controlling for AFQT scores (Bollinger 2003; Carneiro, Heckman, and Masterov 2003).

explained by racial differences in socioeconomic status? Fifth, how has the magnitude of racial and socioeconomic achievement gaps changed over time (across cohorts)?

In this paper, I address the first three of these questions, since they each help us to understand the patterns of development of test score gaps within a given cohort. Moreover, in addressing the three questions regarding the development of black-white achievement gaps, this paper responds to several recent papers which have provided conflicting evidence regarding the development of the black-white test score gap during elementary schooling. The fourth and fifth questions noted above, which deal with the relationship between family environment and test scores and the trends across cohorts in the patterns of achievement gaps, are certainly equally important, but beyond the scope of this paper.³

The first section of the paper briefly summarizes prior research on the development of black-white test score gaps during the course of elementary school. The second section of the paper briefly details the data I use. Next, because conclusions regarding changes in the magnitude of the test score gaps may depend on the metric in which test scores are reported (Murnane et al. 2006; Selzer, Frank, and Bryk 1994), I provide a detailed discussion of the metrics in which achievement gaps are measured. Following this are results from three sets of analyses.

First, the paper examines the trend in black-white test score gaps during the course of elementary schooling. Two key results are important here. First, conclusions about the pattern of development of test score gaps are indeed sensitive to the choice of a test metric. Second, metric-free comparisons of the difference in the black and white test score distributions shows that these distributions become slightly more unequal from kindergarten through fifth grade, with the most

³ The extent to which black-white differences in socioeconomic family characteristics can account for achievement gaps has been the subject of considerable research, though there remains significant disagreement (see, for example, Brooks-Gunn, Klebanov, and Duncan 1996; Fryer and Levitt 2002, 2004; Murnane et al. 2006; Phillips et al. 1998). Likewise, there has been considerable detailed analysis of the trends in the black-white gap over the last three decades (see, for example, Grissmer, Flanagan, and Williamson 1998; Hedges and Nowell 1999; Neal 2005); these studies find that the black-white gap narrowed until the late 1980s, when progress stalled or reversed before beginning to narrow again in the early 2000s (Reardon and Robinson 2007).

rapid divergence occurring in kindergarten (in math) and in second and third grades (in reading).

The second section of the analysis takes up the question of whether black-white gaps grow faster or slower for students with initially higher test scores. In this section, I present reliability-corrected estimates of black-white differences in growth rates under a range of plausible assumptions about the magnitude of the reliability of the test scores. The results of these analyses indicate that reading test scores diverge more between kindergarten and fifth grade among students who enter kindergarten with high levels of reading skill than among students who enter with low levels of reading skill. The pattern in math is similar in direction, though not statistically significant.

The third section of the analysis addresses a point of disagreement between two recent papers. Fryer and Levitt (2004; 2005) and Hanushek and Rivkin (2006), using the same data, come to very different conclusions regarding the extent to which black-white test score gaps are attributable to within- and between-school differences in the average performance of black and white students. I show that the resolution of this disagreement hinges on the interpretation of an ambiguous term in the decomposition of the black-white gap into three components. Because this ambiguous term is empirically large relative to the size of the gap, disagreements about the attribution of the source of this component lead to substantively important disagreements about the location of test score gaps. I then conduct new decomposition analyses of the gaps and discuss their implications.

1. Evidence on the Development of the Black-White Gap

Prior research on the development of the black-white achievement gap comes from two types of studies—studies that use longitudinal panel data on one or more cohorts of students,⁴ and

⁴ Examples of such studies include those using panel data from nationally representative samples—such as the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K) (see www.nces.ed.gov/ecls), the National Education Longitudinal Study (NELS) (see www.nces.ed.gov/surveys/nels88), Prospects: The Congressionally Mandated Study of

studies that rely on repeated cross-sectional data to infer developmental patterns.⁵ Almost all research on the topic concludes that the black-white achievement gap in math grows significantly during the school years, particularly in elementary school. Most research shows that the same is true for the black-white reading gap. The most commonly-cited (and probably the best) contemporary evidence on the development of the black-white gap in elementary school comes from the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K), which includes kindergarten through fifth grade assessment data on a nationally-representative sample of students who were enrolled in kindergarten in the fall of 1998. ECLS-K data show that the black-white gaps in both math and reading are sizeable at the start of kindergarten—about two-thirds and two-fifths of a standard deviation, respectively (Fryer and Levitt 2004; Reardon and Galindo 2006).⁶ Measured in standard deviation units, these gaps widen between kindergarten and fifth grade, by which time the math gap is about one full standard deviation and the reading gap is about three-quarters of a standard deviation (Reardon and Galindo 2006). Other studies using the ECLS-K data, however, report black-white gaps in the ECLS-K scale score metric (an unstandardized metric measuring the number of items a student answers correctly on the test—see below for more detail), and find that the black-white gap increases very dramatically from kindergarten through fifth grade (Hanushek

Educational Growth and Opportunity, and High School and Beyond (HSB) (see www.nces.ed.gov/surveys/hsb)—and those drawn from state administrative data sources in states like North Carolina, Texas, or Florida, each of which has administrative data systems allowing tracking of individual student test scores over multiple years (Clotfelter, Ladd, and Vigdor 2006; Hanushek and Rivkin 2006).

⁵ Most repeated cross-sectional studies of the development of the black-white gap rely on data from the National Assessment of Educational Progress (NAEP), also known as “the Nation’s Report Card” (see www.nces.ed.gov/nationsreportcard/about/). NAEP includes two different assessments of the math and reading skills of nationally-representative samples of students. The first of these—NAEP long-term trend (NAEP-LTT)—is given every four years to a nationally-representative sample of children aged 9, 13, and 17, which allows comparison of the scores of a sample of the 9-year-old cohort in one assessment year with the scores of a (different) sample of the same cohort 4 and 8 years later, at ages 13 and 17 (Ferguson 1998; Neal 2005; Phillips, Crouse, and Ralph 1998). The second of the NAEP assessments—referred to as “Main NAEP”—has been administered roughly every two years since 1990 to representative samples of 4th-, 8th-, and 12th-grade students, which allows a similar type of developmental comparison. Of course, differential immigration and dropout rates may complicate developmental inferences based on such repeated cross-sectional data.

⁶ The standardized gap measures I report in this paper differ slightly from those reported elsewhere for a variety of reasons, including slight differences in the samples, the measurement of gaps in pooled standard deviations rather than sample standard deviations, and my use of standardized *T*-scores rather than standardized scale scores (see below).

and Rivkin 2006; Murnane et al. 2006). These metric-related differences in inferences regarding the magnitude and rate of growth of the gap in different periods suggest the importance of understanding what the tests and metrics used measure.

Analyses of several other large studies have produced somewhat different results than those evident in ECLS-K. Data from the Prospects study (which includes longitudinal data collected 1991 to 1993 from three age cohorts of students) suggest that the black-white math gap grows in first and second grade and from seventh to ninth grade (though not from third to fifth grade), while the black-white reading gap grows in first to second and third to fifth grades, but not in seventh to ninth grade (Phillips, Crouse, and Ralph 1998). The Prospects data were collected almost a decade before ECLS-K, however, (and on cohorts of children born 9-16 years prior to the ECLS-K cohort), so may be of less current relevance than the ECLS-K sample.

A recent analysis of data from the National Institute of Child Health and Human Development Study of Early Child Care and Youth Development (SECCYD) finds that the black-white math gap—measured in standard deviation units—narrows slightly from kindergarten through third grade (from 1.1 to 1.0 standard deviations), while the black-white reading gap widens during the same period (from 1.0 to 1.2 standard deviations) (Murnane et al. 2006). Murnane and his colleagues argue that at least part of the difference in the patterns observed in SECCYD and ECLS-K may be due to differences in the tests used in the two studies, since the Woodcock-Johnson tests used in the SECCYD assess a broad range of skills while the ECLS-K tests are designed to measure skills taught in school.

Finally, analysis of data sets collected by state departments of education in several states provides yet another set of conflicting findings regarding the development of the black-white gaps during the schooling years. Data from four cohorts of students in Texas (cohorts in third grade from 1994-1997) indicate that the black-white gap in math grew modestly, in standard deviation

units, from third through eighth grade (from .59 to .70 standard deviations) (Hanushek and Rivkin 2006). Similar data from North Carolina (five cohorts of students in third grade from 1994-1999), however, indicate that the black-white math gap was relatively stable from third to eighth grade (changing from 0.77 to 0.81 standard deviations); the black-white reading gap likewise increased only very modestly (from 0.69 to 0.77 standard deviations) (Clotfelter, Ladd, and Vigdor 2006). It is unclear whether the relatively small differences in the rate of growth of the math gap between Texas and North Carolina are due to differences in the tests used in each state, differences in their black and white student populations, or to differences in the features of the two states' educational systems, curricula, and/or instructional practices.

Much of the analysis of the development of the black-white achievement gap is focused on the elementary school period. This is largely because the gap appears to change relatively little during high school. Evidence from NELS, which contains longitudinal data on a nationally representative sample of eighth graders in 1988, shows that the black-white math gap—measured in standard deviation units—is stable from eighth through twelfth grades, while the black-white reading gap appears to narrow very slightly during this period (LoGerfo, Nichols, and Reardon 2006).

Studies that rely on NAEP-LTT data conclude that the black-white math gap (though not the reading gap) widens from age 9 to 13 (Ferguson 1998; Neal 2005; Phillips, Crouse, and Ralph 1998). Evidence from these studies of the development of the gap from age 13 to 17 is less clear—the gaps generally do not appear to widen much in this period, but these results are less certain because differential dropout patterns may bias the estimates of the gaps at age 17. In addition, studies using NAEP do not all use the same measure of the gaps—some use the NAEP scale score metric (which is constant over time), while others report gaps in standard deviation units (a metric which rescales the scores at each wave relative to the standard deviation of the test). Phillips, Crouse, & Ralph (1998) conduct a meta-analysis of a number of cross-sectional estimates of the

black-white gaps, and find that the black-white gap in math widens, on average, during high school, but is unchanged in reading and vocabulary.

In sum, evidence on how the black-white achievement gap changes during schooling is somewhat unclear. Data from ECLS-K and SECCYD suggest the gap is large at the start of kindergarten, and grows in the early elementary grades (particularly from first to third grade in ECLS-K), though the patterns differ somewhat depending on the gap metric used. Data from NAEP suggests that the gap continues to grow from age 9 to 13 (fourth to eighth grades, roughly), but state-level data from Texas and North Carolina seem to contradict this finding, at least during the late 1990s and early 2000s, suggesting that the gap grows relatively little in standard deviation units over the latter half of elementary school. Finally, data from NAEP and NELS suggest the gaps change relatively little following eighth grade, though there is some uncertainty in these estimates, since most are based on analysis of repeated cross-sectional data.

2. Data

The analyses presented here rely on data from the Early Childhood Longitudinal Study-Kindergarten Cohort (ECLS-K), conducted by the National Center for Educational Statistics (NCES). ECLS-K is a longitudinal study of a nationally representative sample of roughly 21,400 students in kindergarten in the Fall of 1998 (thus, representing a cohort born in roughly 1992-93). Students in the sample were assessed in reading, mathematics, and general knowledge/science skills at six time points during the years 1998-2004 (fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004).⁷ In addition to these cognitive developmental measures, the ECLS-K data include information gathered from parents, teachers, and school administrators regarding family,

⁷ Throughout this paper, I refer to these six assessments by the modal grade of the students at each wave (fall kindergarten, spring kindergarten, fall first grade, spring first grade, spring third grade, and spring fifth grade) to facilitate interpretation. Moreover, because only a 25-30% subsample of the students were assessed in the third wave (fall first grade), I rely in this paper on the five waves when the full sample was assessed.

school, community, and student characteristics. In this paper, I focus on the reading and mathematics cognitive assessments.

The ECLS-K sample includes 11,805 non-Hispanic white and 3,240 non-Hispanic Black students. The main analytic sample used in this paper consists of 5,644 white and 1,066 black students who were assessed at each of waves 1, 2, 4, 5, and 6.⁸ These students were sampled from 812 kindergarten schools (625 public and 187 private schools). In some analyses I restrict the sample to students who attended the same school at each of the six waves of the ECLS-K study. This ‘stable school sample’ contains 4,098 white (73% of the main sample) and 711 black (67% of the main sample) students in 674 schools (523 public, 151 private). In all analyses, I use ECLS-K panel sampling weights (weight *w1_6f0* in the ECLS-K data) to account for non-random attrition from the sample. Nonetheless, there is some evidence that the sample weights do not fully account for non-random sample attrition, so that the results reported here may underestimate the extent to which the black-white gaps grow over time (Hanushek and Rivkin 2006).

3. Measures of Achievement Gaps

A description of the development of achievement disparities during the course of schooling requires we choose a metric in which to describe the magnitude of the disparities. Using the ECLS-K data, several gap metrics are possible: 1) the ‘scale score’ metric reported by ECLS-K, and used by many scholars (Hanushek and Rivkin 2006; Murnane et al. 2006); 2) wave-standardized standard deviation units (Fryer and Levitt 2004, 2005; Reardon and Galindo 2006); 3) the ECLS-K ‘theta’ metric, which is described in ECLS-K reports, but which has not been used previously because NCES has not made it available to researchers; and 4) several ‘metric-free’ measures (Ho and Haertel 2006; Neal 2005). Each of these metrics has a different interpretation, which I describe

⁸ Among these, 31 white and 19 black students are missing at least one wave of the reading assessment and 9 white and 3 black students are missing at least one wave of the math assessment.

below.

Students in the ECLS-K were given orally-administered, untimed, adaptive tests at each wave. At any given wave, a student was administered a subset of items—chosen to be at an appropriate level of difficulty for the given student based on the student’s performance on an initial set of routing items—from the full set of items on the math and reading tests. For each test, a three-parameter IRT model was used to estimate each student’s latent ability θ_{it} at each wave t . The IRT model assumes that each student’s probability of answering a given test item correctly is a function of the student’s ability and the characteristics of the item. Under the three-parameter IRT model used to scale the ECLS-K tests, the probability that student i answers question k correctly at wave t is given by

$$p_{itk} = \Pr(Y_{itk} = 1 | \theta_{it}, a_k, b_k, c_k) = c_k + \frac{(1 - c_k)}{1 + e^{-a_k(\theta_{it} - b_k)}}, \quad [1]$$

where a_k , b_k , and c_k are the *discrimination*, *difficulty*, and *guessability* parameters of item k , respectively (Lord and Novick 1968; Pollack et al. 2005). Given the pattern of students’ responses to the items on the test that they are given, the IRT model provides estimates of both the person-specific latent abilities at each wave (the θ_{it} ’s) and the item parameters (the a_k , b_k , and c_k for each item).

Given the estimated $\hat{\theta}_{it}$ scores, two types of scores are constructed for each student at each wave. First, ECLS-K provides a wave-standardized version of $\hat{\theta}_{it}$, called a T -score and denoted \hat{T}_{it} , which is $\hat{\theta}_{it}$ standardized to a mean of 50 and standard deviation of 10 at each wave. Because they are standardized at each wave, the T -scores are not ideal for studying longitudinal growth over time, though they can be used in repeated cross-sectional analyses to examine changes in the magnitude of the gaps in cross-sectional standard deviation units (see, e.g., Fryer and Levitt 2004, 2005; Reardon and Galindo 2006). In this paper, I restandardize the T -scores at each wave based on the

unweighted pooled black-white standard deviation each wave:

$$\hat{T}'_{it} = \left(\frac{\hat{T}_{it} - \bar{\hat{T}}_t^{bw}}{\hat{\sigma}_{bw}(T_t)} \right) = \left(\frac{\hat{\theta}_{it} - \bar{\hat{\theta}}_t^{bw}}{\hat{\sigma}_{bw}(\theta_t)} \right), \quad [2]$$

where $\bar{\hat{T}}_t^{bw}$ and $\bar{\hat{\theta}}_t^{bw}$ are the mean values of \hat{T}_{it} and $\hat{\theta}_{it}$ among black and white students in the sample and $\hat{\sigma}_{bw}(T_t)$ and $\hat{\sigma}_{bw}(\theta_t)$ are the unweighted pooled black-white standard deviations in the main analytic sample, respectively, at wave t .⁹ Differences between white and black students in the T' metric are thus measured in pooled black-white standard deviation units.

Second, ECLS-K provides an estimated ‘scale score,’ \hat{S}_{it} , for each student at each wave, which is the estimated number of questions the student would have gotten correct if he or she had been asked all of the items on the test. The estimated scale score is obtained by summing the predicted probabilities of a correct response over all items, given the student’s estimated $\hat{\theta}_{it}$ and the estimated item parameters:

$$\begin{aligned} \hat{S}_{it} &= f(\hat{\theta}_{it}) \\ &= \sum_k \Pr(Y_{itk} = 1 | \hat{\theta}_{it}, \hat{a}_k, \hat{b}_k, \hat{c}_k) \\ &= \sum_k \left[\hat{c}_k + \frac{(1 - \hat{c}_k)}{1 + e^{-\hat{a}_k(\hat{\theta}_{it} - \hat{b}_k)}} \right] \end{aligned} \quad [3]$$

Because the ECLS-K tests contain many more ‘difficult’ items than ‘easy’ items (as measured by the distribution of the estimated b_k parameters; see Figure 1), the relationship between θ and S is not linear (see Figure 2). Thus a unit difference in θ corresponds to a larger difference in S at $\theta=1$ than at $\theta=-1$, for example.

Although they have been used by a number of researchers to investigate patterns of math

⁹ The standard deviations are computed from the main analytic sample separately for black and white students from race- and wave-specific regressions, adjusting for the date at which students took the test, and weighted by the ECLS-K longitudinal panel weight $v1_6fi0$. The pooled standard deviation at each wave is the square root of the average of the squares of the black and white wave-specific standard deviations.

and reading achievement growth in the ECLS-K study (e.g., Downey, Hippel, and Broh 2004; Hanushek and Rivkin 2006; Lee and Burkham 2002; Murnane et al. 2006), the scale scores are not particularly useful for comparing the learning rates of different students (or of the same student over time), since the scale score test metric is an interval-scaled metric only with respect to the specific set of items on the ECLS-K tests. To see this, suppose we test three students, A, B, and C, who get 10, 20, and 30 items correct on the test, respectively. We might conclude that the differences in ability between students A and B and B and C are the same. If we altered the test, however, by adding 10 items that were too hard for student A but easy for students B and C, we would observe scores on the new test of 10, 30, and 40, respectively, and would conclude that students B and C were closer in ability than students A and B. Our changing conclusions would be entirely an artifact of the set of items we included on the test. In fact, if we added items instead that were too hard for A and B, but accessible to C, we would conclude just the opposite, that A and B were closer in ability than B and C. Unless we believe that the distribution of item difficulties on the test corresponds to some meaningful scale,¹⁰ it is hard to justify the use of the scale score metric for making anything other than ordinal comparisons.

Although the ECLS-K data released by NCES do not contain the estimated $\hat{\theta}_{it}$ scores, it is possible to recover them from the reported estimated scale scores and the estimated item parameters $\hat{a}_k, \hat{b}_k, \hat{c}_k$, by inverting the functions f_r and f_m shown in Figure 2:

$$\hat{\theta}_{it} = f^{-1}(\hat{S}_{it}) \quad [4]$$

¹⁰ Formally, to say that the scale score metric is an interval scale is to say that S is linearly related to some meaningful metric. What constitutes a meaningful metric is unclear, and open to debate, in the realm of cognitive skill development. One possible definition of a metric is to link cognitive skill to observable behavioral outcomes, such as the probability of answering a given test item correctly; this is the implicit definition used in the Rasch (one-item) IRT model. Under the Rasch model, a unit difference in latent ability, as measured by θ , corresponds to a constant difference in the log-odds of responding to any item correctly, regardless of the difficulty of the item or the level of ability. The Rasch scale is therefore interval with respect to the log-odds of answering test items correctly. The corresponding scale score S will be interval-scaled with respect to the Rasch model, however, only if the density of the item difficulty parameters is flat over the entire range of abilities of interest.

The $\hat{\theta}_{it}$ scores I use in this paper are constructed by applying Equation [4] to the estimated scale scores included in the ECLS-K fifth grade data files. The reading and math test functions f_r and f_m are constructed from the estimated item parameters $\hat{a}_k, \hat{b}_k, \hat{c}_k$ reported in Appendix B of Pollack et al (2005); the functions f_r^{-1} and f_m^{-1} are then constructed from these by numerical interpolation.¹¹

Given that the scale scores are difficult to interpret as an interval-scaled metric, it is useful to consider whether the θ metric can be considered interval-scaled. The θ metric can be considered interval-scaled in a behaviorally-meaningful sense if the items on the test fit the assumptions of a Rasch model. If a_k equals a constant a and $c_k=0$ for all items, then the model fits the assumptions of the Rasch model (also called the one-parameter IRT model), and we have

$$\ln\left(\frac{P_{itk}}{1-P_{itk}}\right) = a(\theta_{it} - b_k). \quad [5]$$

Under the Rasch model, θ is measured in an interval-scaled metric with respect to the logit metric of the probability of answering an item correctly. In other words, regardless of the difficulty of a given item or the ability of a student, a one-unit increase in θ will increase the log-odds of a correct response by the same amount. Thus, if the Rasch assumptions are met, θ can be interpreted as measured in an interval metric in a behaviorally-meaningful sense.

Although the ECLS-K IRT model is not based on the Rasch assumptions, we can assess how well the items fit the Rasch assumptions by examining the item parameters. Figures 3 and 4 illustrate the fitted item characteristic curves (the probability of a correct response plotted against θ) for the math and reading test items, respectively. Under the Rasch assumptions, all item characteristic curves would be parallel ogive functions with asymptotes at 0 and 1. In the math and reading tests, there are 153 and 187 items, respectively. Of these, 35 and 77 have non-zero

¹¹ Details on the Stata code used to compute these estimates are available on request.

guessability parameters, respectively. Among the 118 math items with $c_k=0$, the average slope of the item characteristic curves at b_k (where the curves are steepest) is .95, with a standard deviation of 0.29. For the 110 reading items with $c_k=0$, the average slope is 1.0, with a standard deviation of 0.42. This variation does not suggest that the items fit the Rasch assumptions well: even among the items with $c_k=0$, the average item has a discrimination parameter that is 30% or 42% different from the average item discrimination in math and reading, respectively. At best, then, we might describe the ECLS-K θ metrics as “approximately Rasch,” meaning that the model only approximately fits the Rasch assumptions.

Metric-Free Test Gap Measures

In addition to describing the black-white gaps in the three ECLS-K test metrics described above, I describe the gaps with several ‘metric-free’ measures. These measures depend only on the ordinal nature of the test scores, so do not depend on assumptions about the interval-scaling of the test metric or the distribution of the test scores. One approach to describing gaps in metric-free way is to construct percentile-percentile (PP) plots, which plot the percentiles of the black distribution against the percentiles of the white distribution. If the black and white distributions are identical, the PP curve will lie on the 45-degree line; the deviation of the PP curve from the 45-degree line can be used as a measure of the extent to which the black and white distributions do not overlap.

I present results here based on two metric-free gap measures derived from the PP curves. First, I estimate the probability that a randomly chosen black student will have a test score higher than a randomly chosen white student. Formally, this is computed as

$$P_{b>w} = \Pr(Y_b > Y_w) = \int_0^1 F_w(F_b^{-1}(p_b)) dp_b, \quad [6]$$

where F_w and F_b are the white and black cumulative density functions (given a score x , they return

the proportion of white or black students with scores less than or equal to x), and p_b indicates a proportion of black students. This measure, when multiplied by 100, can also be interpreted as the average percentile in the white distribution of a black student (Ho and Haertel 2006; Neal 2005). A symmetric formulation produces $P_{w>b} = 1 - P_{b>w}$.

The measure $P_{b>w}$ can be converted to what Ho and Haertel (2006) term a standardized metric-free gap measure by computing $\sqrt{2}\Phi'(P_{b>w})$, where Φ is the probit function. This measure corresponds to gap we would estimate if we first transformed test scores via a monotone transformation so that both the white and black scores were normally distributed and we then computed the black-white gap in pooled standard deviation units of this new metric. This measure can be thought of as a metric-free pseudo effect size.

Correcting ECLS-K test scores for measurement error

Like any test, the ECLS-K tests do not measure students' math and reading skills without error. In describing black-white test score gaps and the association between growth rates and initial scores, we would like to know students' true scores, rather than the observed scores. Given the reliability r of a test, the observed score x_i , the Bayesian conditional shrinkage estimator of the true score is given by

$$x_i^* = (1-r)(\mu_w(1-B_i) + \mu_b B_i) + r x_i, \quad [7]$$

where μ_w and μ_b are the mean black and white scores and B_i is a dummy variable indicating if a student is black (see Appendix for details and discussion). Under the assumption of normally-distributed, independent errors, the conditional shrinkage estimator is an unbiased estimator of the true score. Note that shrinking x_i toward its conditional mean does not affect the estimate of the black and white means, though it does reduce the pooled standard deviation (multiplying it by a factor r), so that the standardized difference in the x^* metric is larger by a factor of $1/r$ than the

difference in the original x metric.

The reported reliabilities of the ECLS-K tests range from 0.89 to 0.96 across waves and test subjects (Pollack et al. 2005, Tables 4-5, 4-9). These reliabilities, however, are the internal item-consistency reliabilities, rather than the test-retest reliabilities, which are likely considerably lower. One way of estimating the test-retest reliabilities, in principle, is to examine the correlation between repeated test scores of the same students. Under the assumption that the errors are independent, if X_1 and X_2 are standardized test scores at two time points, the test-retest reliability r of the test is given by

$$r = \text{Corr}(X_1, X_2) - \text{Cov}(T_1, \Delta T), \quad [8]$$

where T_1 is the true skill at time 1 and ΔT is the change in true skill between times 1 and 2. In the ECLS-K sample, the correlations between the observed fall and spring kindergarten test scores (in the theta or T -score metric) in this sample are 0.82 in math and 0.80 in reading. The second term in Equation [8] may be positive (if students with initially higher skills learn fastest) or negative (if students with initially lower skills learn fastest), implying that the correlation between repeated test scores may over or underestimate the reliability of the tests.

In this paper, I report results under a range of assumptions about the reliability of the tests: I assume reliabilities of 0.70, 0.80, 0.90, and 1.00 in order to examine the sensitivity of the conclusions to assumptions regarding the reliability of the tests. Failure to account for measurement error in the tests leads to two types of bias of interest. First, it leads to underestimates of the standardized and metric-free gap measures, since measurement error in the test will lead to an overestimate of the extent to which the black and white distributions overlap. Second, it leads to bias in the estimation of the association between initial ability and subsequent achievement growth (see Appendix).

4. The development of black-white gaps in kindergarten through fifth grade

Table 1 reports the estimated black-white test score gaps in math and reading at each of five time points: Fall and Spring of kindergarten, and Spring first-, third-, and fifth-grades. In math, the black-white gap is evident and large in the Fall of kindergarten: the gap is 0.32 units in the theta metric (the units of the theta metric are approximately logits),¹² 0.90 standard deviations (assuming a reliability of 0.80) in the standardized *T*-score metric, or 1.02 standard deviations in the metric-free effect size metric. The reading gap is about two-thirds as large as the math gap at the start of kindergarten (0.23 units in the theta metric, 0.62 standard deviations in *T*-score, or 0.73 metric-free standard deviations).

Regardless of the metric used, the black-white gaps grow from kindergarten to fifth grade, though the pattern and magnitude of growth differs by metric. In math, the gap generally grows rapidly during kindergarten (recall that there are only six months between kindergarten assessments, so even small differences between Fall and Spring kindergarten gaps indicate large differences in growth rates). In first grade, the gap narrows in the theta metric, widens in the scale score metric, and is unchanged in the other metrics. From first to fifth grade, the gap generally grows as well, though more slowly than in kindergarten, and with some variation across metrics (the gap grows rapidly from third to fifth grade as measured in the theta metric, but slowly or not at all in the metric-free measures. If we take the theta metric as approximately interval-scaled then, the black-white gap follows a somewhat perplexing pattern—it grows rapidly in kindergarten, and then not at all in first grade, but then grows at an increasing rate from first through fifth grade. In any of the standardized (relative) measures—the standardized *T*-scores and the metric-free measures—however, the gap grows most rapidly in kindergarten and then at a slowing rate through fifth grade.

¹² In a Rasch scale, the units of the theta metric are logits; the theta metric in the three-parameter IRT model is strictly interpretable as logits only if the model fits the Rasch assumptions (guessability=0 and discrimination is constant for all items).

One explanation for this is that the variation in scores grows from first to fifth grade, leading to the patterns we observe in Table 1.

In reading, the trends are somewhat different. In the theta metric, the black-white gap is stable through third grade, and increases only slightly from third to fifth grade (and this increase is not statistically significant). In the standardized T -score metric and the metric-free measures, however, the reading gap is relatively stable through first grade, but then widens sharply from first to third grade. Note that we would describe the trends very differently if we relied on the scale score metric here—in the scale score metric, the reading gap grows most rapidly in first grade.

5. The development of the black white gap among students of different initial ability

To investigate whether achievement gaps grow faster or slower conditional on students' true fall kindergarten math and reading skills, I fit a series of models of the form

$$T_{i5} = \gamma_0 + \gamma_1(T_{iK}^r) + \gamma_2(T_{iK}^r)^2 + \gamma_3(B_i) + \gamma_4(T_{iK}^r B_i) + \varepsilon_i, \quad [9]$$

where T_{i5} is the test score of student i in grade five; T_{iK}^r is the (standardized) reliability-adjusted estimate of student i 's true score in the fall of kindergarten (estimated by shrinking the observed score toward its race-specific mean; see Appendix), and B_i is indicator variable for race (black=1). In this model, γ_3 is the average difference in fifth grade scores between black and white students who have identical test scores at the sample mean in the Fall of kindergarten. The parameter of interest is γ_4 , which indicates the extent to which the black-white difference in fifth grade scores varies with initial scores. A negative value of γ_4 indicates that the black-white gap grows faster between initially high-achieving black and white students than among initially low-achieving students.

Table 2a reports estimates from models of the type described in Equation [9]. I estimate math and reading models separately for each of the three outcome metrics and using four different

reliability assumptions ($r=0.7, 0.8, 0.9, \text{ and } 1.0$). In each case, I fit a models with and without the interaction term (for all models, an additional interaction term between the square of the test score and the indicator variable for black was dropped because it was significant in none of the models).

As we expect given the results from Table 1 above, white students have higher test scores in math and reading in fifth grade than do black students with the same true skills in the fall of kindergarten, a conclusion that holds regardless of the test metric used or the level of reliability we assume. In math, the models provide no evidence that the difference in fifth grade scores conditional on fall kindergarten scores varies by kindergarten score. The coefficient on the interaction term is always negative, but its confidence interval includes zero for all test metrics and reliabilities. In reading, however, the coefficient on the interaction term in both the T -score and theta models differs reliably from zero when we assume the reliability of the test is 0.8 or 0.7 (the coefficients are similar in magnitude but fall just below conventional significance levels when we assume higher test reliability). The magnitude of the interaction term is relatively large. Assuming reliability of 0.8, for example, model R2(T) indicates that the fifth grade gap between black and white students whose reading skills in Fall kindergarten were one standard deviation below the mean is 0.359 standard deviations, while the corresponding gap between students one standard deviation above the mean in kindergarten is 0.583 standard deviations.

Although the results shown in Table 2 correct for bias due to the unreliability of the tests, they still depend on the assumption of interval scaling of the tests. If the test is not interval scaled, then differences in gains among initially high-achieving students are not necessarily comparable to gains among initially low-achieving students, and the interpretation of γ_3 and γ_4 in Equation [22] is unclear. As a specification, then, Table 2b reports results analogous to those in 2a, but using a locally-standardized version of the outcome score in each case. In these models, the fifth grade test score is standardized conditional on the estimated fall kindergarten true scores, so that the black-

white differences are now interpreted in terms of local standard deviations.¹³ Table 2b shows results largely consistent with Table 2a: the average black-white difference in fifth grade reading scores is larger among students within initially high reading skills than among those with initially low reading scores. This pattern is true across all three test metrics and the range of assumed reliabilities. In math, the estimated γ_4 coefficients are similar in sign and slightly smaller in magnitude than in reading, falling just below conventional significance levels.

6. Decomposing the achievement gaps

A central question in understanding black-white test score gaps is the extent to which such gaps can be attributed to differences in average school quality between schools attended by white and black students. If black students attend, on average, lower quality schools than white students, we would expect the between-school component of the black-white achievement gap to grow over time. If black and white students receive unequal instructional opportunities when attending the same schools, we would expect the within-school component of the black-white gap to grow over time.

It is difficult to disentangle the effects of school quality from the sorting processes that produce racially segregated schools and that may result in lower-ability students, regardless of race, into schools that have higher proportions of black students. Likewise, it is not clear that differences in black and white achievement gains can be attributed solely to schooling processes, given unequal family resources, neighborhood context, and opportunity structures (which may lead to unequal

¹³ The local standardization is done as follows: the reliability-adjusted Fall kindergarten scores are divided into 50 quantiles (results are unchanged using 25 or 100 quantiles). Within each quantile, I separately regress black and white fifth grade scores on fifth-grade test assessment dates and obtain estimates of the residual error variance of the test scores from these models. I then compute the unweighted pooled standard deviation within each quintile as the square root of the average of the black and white error variances. I then standardize the fifth grade scores within each Fall score quintile using the quintile-specific pooled standard deviation. This ‘locally standardized’ fifth grade score is used as the outcome variable in model [9], and γ_3 and γ_4 are now interpreted as describing the average locally standardized difference in black and white fifth grade scores. Because of the local standardization, the estimate of γ_4 is much less sensitive to violations of the interval scale assumption than if we simply used the original metric.

motivation even in the presence of equal home and school resources). So any attempt to decompose the black-white gap into between- and within-school components should be understood primarily as a descriptive exercise rather than a method for inferring relative causality.

Although it is appealing to “decompose the gap into between- and within-school components,” the mathematics of such a decomposition turn out to be less straightforward and unambiguous than it may sound. Both Fryer and Levitt (2004; 2005) and Hanushek and Rivkin (2006) provide decompositions of the black-white test score gap, but obtain very different descriptive conclusions, despite using the same ECLS-K data. Fryer and Levitt (2004; 2005) conclude that the black-white gap is primarily a within-school phenomenon, while Hanushek and Rivkin (2006) conclude that it is primarily a between-school phenomenon. In order to understand the source of their disagreement, I begin with a detailed discussion of the decomposition methods each employs.

Mathematical decompositions of the black-white gap

For simplicity, I consider a population made up of only black and white students. Let i index persons and s index schools, let Y_i be the test score and B_i be a dummy indicator for race (black=1) for person i , and let π_s and π indicate the proportion black in school s and the population, respectively. First, define the black-white gap in a population of students as the parameter δ in the equation

$$Y_i = \alpha + \delta B_i + \varepsilon_i. \quad [10]$$

Fryer and Levitt (2004; 2005) (hereafter FL) define the within-school component of the black-white gap as the coefficient β^e from the school fixed-effects model

$$Y_i = \beta_s + \beta^e B_i + \varepsilon_i, \quad [11]$$

where β_s is a school-specific intercept.¹⁴ This implies a decomposition of the form

$$\delta = \beta^{fe} + \Delta_b \quad [12]$$

This decomposition explicitly estimates the within-school component of the gap and infers the between-school component Δ_b by subtraction (Δ_b may be positive or negative, as HR point out).

Hanushek and Rivkin (2006) (hereafter HR), however, argue that the FL decomposition is inappropriate because it fails to adequately capture the component of the gap that is due to segregation among schools. They derive a decomposition of the overall observed gap that can be written as:¹⁵

$$\delta = \frac{Cov(Y_i, B_i - \pi_s)}{Var(B_i)} + \frac{Cov(Y_i, \pi_s)}{Var(B_i)}. \quad [13]$$

To see how the FL [12] and HR [13] decompositions differ from one another, I illustrate a more general decomposition. Consider the model

$$Y_i = \beta_0 + \beta_1 B_i + \beta_2 \pi_s + \varepsilon_i. \quad [14]$$

In this model, β_1 is the average difference in Y between black and white students in schools with the same racial composition (and hence, it is also equal to β^e , the average within-school difference in Y between black and white students), and β_2 is the association between racial composition and average test scores, conditional on individual race.

It is simple to show (see Appendix B) that we can write

$$\delta = \beta_1(1 - V) + \beta_1 V + \beta_2 V, \quad [15]$$

where V is a measure of segregation between black and white students defined as

$$V = \frac{Var(\pi_s)}{Var(B_i)}. \quad [16]$$

¹⁴ Actually, FL include a set of individual-level covariates in their models, but that is not relevant to the exposition here.

¹⁵ I have expressed the HR formula slightly differently than they do, but the decomposition remains the same.

Note that V is most easily interpreted as the difference in the average percentage black in schools of black and white students, though V is also a measure of segregation known variously as the variance ratio index of segregation, the normalized exposure index, η^2 , and the gap-based measure of segregation (Clotfelter 1999; James and Taeuber 1985; Reardon and Firebaugh 2002), and is often interpreted as the proportion of the variance in individual race that is accounted for by the racial composition of schools. It can take on values ranging from a minimum of 0, obtained only when all schools have the same racial composition (no segregation; π_i is constant), to a maximum of 1, obtained only when all schools are monoracial (complete segregation; $\pi_i=B_i$).

The three-part decomposition of the black-white gap given in [15] is a more general form of both the FL and HR decompositions. First, note that combining the first and second terms in [15] yields the FL decomposition:

$$\begin{aligned}\delta &= \beta_1(1-V) + \beta_1V + \beta_2V \\ &= \beta_1 + \beta_2V \\ &= \beta^{fe} + \beta_2V\end{aligned}\tag{17}$$

The FL within-school term is the sum of the first two terms in the decomposition [15], and the FL between-school term is the third term in [15]. Grouping the three terms in [15] differently yields the HR decomposition (see Appendix B):

$$\begin{aligned}\delta &= \beta_1(1-V) + (\beta_1 + \beta_2)V \\ &= \frac{Cov(Y_i, B_i - \pi_s)}{Var(B_i)} + \frac{Cov(Y_i, \pi_s)}{Var(B_i)}\end{aligned}\tag{18}$$

The HR between-school term is equal to the sum of the second and third terms of the decomposition [15], and the HR within-school term is equal to the first term of [15].

Thus, estimating δ from [10] and the parameters β_1 and β_2 of [14] is sufficient to obtain the three components of the decomposition in [15], and from this to construct the FL and HR

decompositions.¹⁶ The difference between HR and FL is in whether they attribute the $\beta_1 V$ term to the within-school or between-school portion of the gap. If β_1 is small relative to β_2 and V is small, this term will contribute relatively little to the overall gap – and the FL and HR decompositions will yield similar conclusions. In practice, however, neither β_1 nor V are small in practice, so this term substantially affects the FL and HR attributions of achievement gaps.

The choice between the FL and HR decompositions hinges on the interpretation of the $\beta_1 V$ term. In order to interpret the three terms in the decomposition [15], it is useful to examine a stylized picture. Figure 5 illustrates a stylized pattern of black and white student outcomes (test scores or gains in test scores over some time period) conditional on school racial composition and student race. In this figure, the average black student attends a school of racial composition indicated by the point B , while the average white student attends a school of racial composition indicated by W . As noted above, the segregation index V equals $A-W$. The average outcome among black and white students are denoted $Y0(B)$ and $Y0(W)$, respectively, and the test score gap $\delta = Y0(W) - Y0(B)$.

The parameters of model [14] describe the lines in Figure 5. The parameter β_2 is the slope of the solid lines (in Figure 5, I have drawn these lines parallel for simplicity; in practice, they need not be parallel, in which case β_2 in [14] corresponds to their weighted average slope) describing the association between the outcome and school racial composition, conditional on a student's race.

The parameter β_1 in [14] describes the distance between the black and white lines (or, again, the

¹⁶ If the proportion black in each school is estimated from the sample of students, then π_i will be measured with error (particularly if within-school samples are relatively small, as they are in ECLS-K). The estimate of β_2 from [14] will therefore be biased downward, and the estimate of V will be correspondingly biased upward. These biases will exactly cancel one another out in the $\beta_2 V$ term, and the estimate of β_1 will not be affected, so measurement error in π_i will not affect the FL decomposition. It will, however, affect the HR decomposition, because the middle term $\beta_1 V$ will be biased upward by the bias in V . In simulations using the number of schools and students in the ECLS sample I use here, I find that sampling variation biases V upward by roughly 5-10% (and therefore biases β_2 downward by 5-10%). I adjust the decomposition estimates presented later to account for this bias.

weighted average of this distance in the case where the lines are not parallel), describing the association between the outcome and student race, conditional on school racial composition. Finally, the dashed line indicates the mean outcome for students in schools of a given racial composition; the slope of this line is $\beta_1 + \beta_2$.

To interpret the three components of the decomposition in [14], it is useful to understand the change in Figure 5 that each implies. The first component of the decomposition, $\beta_1(1-V)$, is the amount by which the gap would be reduced if we eliminated the black-white gap within each school, but left the mean achievement within each school constant, left school segregation constant, and left all students in their same schools. This reduction would be obtained by raising the black mean in each school to the existing school mean and lowering the white mean in each school to the existing school mean. Such a change is shown in Figure 6, in which the mean achievement of both black and white students has been set to the prior school mean. The black-white gap is now equal to $(\beta_1 + \beta_2)V$, a reduction of $\beta_1(1-V)$ from the original gap. The HR within-school component of the gap, then, can be understood as the amount by which the gap would be reduced if we eliminated within-school gaps but left school mean achievement unchanged. Note, however, that such a procedure would necessarily change the association between outcomes and racial composition, conditional on a student's race.

Alternatively, we might imagine eliminating within-school gaps while leaving unchanged the association between racial composition and student outcomes, conditional on a student's race. This corresponds to leaving β_2 unchanged. Figure 7 illustrates such a change. In this example, the black mean achievement in each school has been raised to equal the prior white mean achievement. The black-white gap is now equal to β_2V , a reduction of β_1 from the original gap. This is the FL within-school component, and it can be understood as the portion of the gap that would be eliminated if

we equalized black and white mean achievement within schools while leaving unchanged the association between racial composition and student outcomes, conditional on a student's race. Note, however, that such a procedure would necessarily change the overall association between racial composition and student outcomes.

To accomplish the FL scenario, then, we would have to raise mean student achievement in predominantly black schools much more than in predominantly white schools (or lower it less). If we consider the outcome measure (test scores or test score gains) as something produced by schools, then FL within-school component of the gap can only be eliminated by increasing the 'productivity' of predominantly black schools relative to predominantly white schools—a change that is hard to characterize as a within-school change. Under this logic, the portion of the FL within-school component equal to $\beta_1 V$ cannot be affected except by between-school changes.

Instead of considering the implications of eliminating within-school differences in outcomes, however, we can consider the implications of alternative ways of eliminating such between-school differences. One way of eliminating the contribution of between-school differences to gaps is to eliminate segregation. If the slope β_2 can be interpreted causally—that is, if changing the racial composition of a school by an amount x would change a student's outcome by an amount $\beta_2 x$ —then eliminating segregation among schools would reduce the black-white gap by an amount $\beta_2 V$, and the remaining gap would be β_1 , the average within-school gap. This implies the FL decomposition: the between-school portion of the gap is that which would be eliminated by eliminating segregation but leaving within-school gaps unchanged, under the assumption that β_2 is a valid prediction of the effect of changing school racial composition.

The FL decomposition also implies, however, that if $\beta_2=0$, then there is no between-school contribution to the test score gap, even if $V>0$. Figure 8, however, illustrates a scenario where $\beta_2=0$

and segregation is non-zero. In this case, the overall gap is equal to β_1 , but to eliminate this gap would require a between-school remedy, as noted above.

A second way of eliminating the contribution of between-school differences would be to leave segregation unchanged (and students in their initial schools) and to eliminate the association between school mean achievement and racial composition, while leaving within-school gaps unchanged. To accomplish this, we would have to produce a pattern of results similar to Figure 9. Note that in Figure 9, $\beta_1 + \beta_2 = 0$, and the remaining gap is equal to $\beta_1(1-V)$.

To summarize, if we eliminate within-school gaps and leave segregation and school mean achievement unchanged, the gap is reduced by $\beta_1(1-V)$; this portion of the gap is unambiguously due to within-school differences. The remaining gap, $(\beta_1 + \beta_2)V$, can be eliminated only by changing either segregation or the association between school mean outcomes and racial composition, both of which are between-school processes, which implies the HR decomposition is correct. If, however, we eliminate segregation, the gap is reduced by β_2V , implying that we can call β_2V the portion of the gap that is due to segregation. In the absence of segregation, the remaining gap, β_1 , can be reduced only through changing within-school gaps, implying that the FL decomposition is correct. Finally, if we eliminate the association between school mean achievement and racial composition and leave segregation and within-school gaps unchanged, the gap is reduced by $(\beta_1 + \beta_2)V$, again implying the HR decomposition.

In each of these interpretations the $\beta_1(1-V)$ component is unambiguously due to within-school differences and the β_2V component is unambiguously due to segregation and between-school differences. The remaining term, β_1V , is due to an interaction of between-school segregation and within-school gaps, and so remains ambiguous. In general, we cannot unambiguously attribute this term to within- or between-school processes without a clear theory of the extent to which the

parameters of [14] indicated causal quantities.

Empirical decomposition results

Despite this ambiguity, an examination of the magnitudes of the three components of the decomposition may be informative. Tables 3a-3c report the estimated decompositions of the math and reading gaps in the ECLS-K sample. These estimates are based on the stable school sample—the roughly 70% of students in the main sample who remained in the same school through the 6 years of the study. As such, they are certainly not representative of the full population of black and white students in elementary school during the study period, but the patterns evident are nonetheless informative. The tables include decompositions of both cross-sectional gaps and changes in the gaps between waves of the assessment. Moreover, each of the three tables reports decompositions based on a different test metric—standard deviation gaps (Table 3a), theta scores (Table 3b), and scale scores (Table 3c).

Most striking about the results in Tables 3a-3c is the consistency of the decompositions across waves, test metrics, and test subjects. In math, roughly 20% of the gap at each wave and 20-25% of the change in the gap from kindergarten to fifth grade is unambiguously due to within-school black-white differences in test scores, a pattern that is nearly identical across the three test metrics. Likewise, roughly 40% of the gap at each wave and 25-40% of the change in the gap from kindergarten to fifth grade is unambiguously due to between-school black-white differences in test scores. With the exception of the fact that 25% of the change in the standard deviation and theta score gaps and 40% of the change in scale score gaps is attributable to between-school differences, there is no substantial difference in these results across test metrics.

The third, ambiguous, component of the math gap accounts for 40% of the cross-sectional gaps and 50% of the change in math gaps from kindergarten to fifth grade. Given that this

component is quite large, it is easy to see how the FL and HR conclusions differ so: based on these results, the FL decomposition implies that three-quarters of the growth of the standardized gap from kindergarten to fifth grade is accounted for by within-school growth of the gap. Conversely, the HR decomposition implies that more than three-quarters is due to between-school patterns.

The decomposition results for reading gaps are quite similar, though the unambiguously between-school component is generally larger in reading than in math (roughly 45% of the reading gap and growth in the gap is unambiguously between schools), while the unambiguously within-school component and the ambiguous component are both slightly smaller than in math.

Nonetheless, the ambiguous term in the decomposition still accounts for 40% of the total gap and its growth in reading, indicating that the FL and HR decompositions would yield very different conclusions.

7. Discussion

The black-white gap in math and reading appears to grow between kindergarten and fifth grade, regardless of the test metric we use to describe it. Nonetheless, the timing and magnitude of that growth varies considerably across gap metrics. Given these discrepancies, which metric(s) should we use? The gap metrics I have described are of two types—measures of absolute difference in mean test scores (the theta scores and the scale scores), and measures of relative difference in mean test scores (the standard deviation gaps, the $P_{b>w}$ measure, and the metric-free effect size measure). Measures of absolute difference rely heavily on the assumption of interval scaling, an assumption which the scale scores clearly do not meet, rendering results based on them suspect. The theta scores are preferable, since they can be considered approximately interval-scaled with respect to the log-odds of answering test items correctly.

In the theta score metric, the black-white math gap widens by more than 25% from

kindergarten to fifth grade, with most of that growth occurring between third and fifth grade. In reading, the gap grows by only half as much from kindergarten through fifth grade, again with most of the growth occurring toward the end of that period. These results imply that the black-white gaps grow more slowly after school entry than they do prior to school entry (since by fifth grade, students have been in school for more than half their life, but most of the gap was present at the start of school). However, Fryer and Levitt (2004) show that socioeconomic factors can explain virtually of the black-white gap present at the start of kindergarten, but cannot explain the growth of the gap during school, suggesting that the schooling may play a role in the growth of the gap following kindergarten entry.

Measures of relative difference in mean test scores describe the magnitude of the difference relative to the overall variation in test scores, and so can be understood as measures of the inequality of the black and white test score distributions. Among these measures, the metric-free measures are preferable, since they rely on no assumption of interval scaling. By these measures, the black-white math and reading gaps both grow rapidly between first and third grade (the math gap grows by 10%, the reading gap by 40%, during these two years), and relatively little during other periods. The difference between these metric-free gap patterns and the theta metric patterns is due, in part, to the changes in the variance of test scores over time. The variation in reading theta scores, for example, narrows from first to third grade, and then widens following third grade, a pattern that suggests either that schooling has an equalizing effect on reading skills during these years, or that the theta metric is not appropriately interval-scaled.

The second analysis in this paper addressed the question of whether the black-white gap grows or narrows faster among initially high scoring students than among initially low-scoring students. For both math and reading, the pattern of results suggests that gaps grow faster among initially high-scoring students, though this result reaches conventional levels of statistical significance

only in reading (and is significant at the 0.10 level in math). One possible explanation for this pattern may have to do with segregation patterns (Hanushek, Kain, and Rivkin 2002): black students with high skill levels in kindergarten may be more likely than equally high-scoring white students to be in schools where the median skill level is far below their skills (because they are more likely to be in schools with predominantly black student populations). If this is true, and if schools' curricula and instructional practices are aimed at the median student, then high-achieving black students may be disproportionately located in schools with less challenging curricula, leading to their lower achievement gains through elementary school.

In the final section of this paper, I investigated a discrepancy in findings regarding the extent to which the black-white gaps can be attributed to differences in achievement within and between schools. The disagreement in recent papers on this issue stems from a methodological ambiguity in the decomposition of gaps. While I show that roughly one-fifth of the math and reading gaps is due to within-school differences in test scores and that roughly two-fifths is due to between-school differences, I am not able to determine the extent to which these within- and between-school gap components are attributable to selection processes and/or to schooling effects. The within-school differences in growth rates may be due to differences in family background and unobserved student characteristics that predate entry among white and black students attending the same schools, or they may be due to differential instruction and treatment of white and black students attending the same schools, or to some combination of the two. Likewise, differences in achievement between students attending different schools may be a result of sorting processes (due to residential segregation, income differences, etc.) and/or to a correlation between schools' racial composition and their educational effectiveness. For example, schools with higher proportions of white (and middle class) students may also have better teachers, more resources, more parental involvement, and so on. The decompositions I have presented here are therefore descriptive, rather than causal.

Appendix A: Eliminating bias when conditioning on a test score measured with error

Assume white and black students have true scores on a pretest described by

$$T_i = \mu_w(1 - B_i) + \mu_b B_i + u_i \quad u_i \sim N(0, \tau), \quad [A1]$$

where μ_w and μ_b are the white and black mean true scores and B is an indicator for black (measured without error). Assume T is measured with error by a test, so that we observe score x :

$$x_i = T_i + e_i \quad e_i \sim N(0, \sigma) \quad [A2]$$

The *conditional* (within-group) *reliability* of x as a measure of T is

$$r_x = \frac{\tau}{\tau + \sigma}. \quad [A3]$$

(Note that the conditional reliability r_x is the proportion of within-group variance in x that is due to within-group variance in T . In general, the *unconditional reliability* of x will be smaller than the conditional reliability, because the unconditional variance of x will be larger than τ , since there will be an additional component due to the difference in group means).

Next, assume the true relationship between outcome Y and T and B is described by the structural model:

$$Y_i = \gamma_0 + \gamma_1 T_i + \gamma_2 B_i + \gamma_3 T_i B_i + \varepsilon_i \quad \varepsilon_i \sim N(0, \nu). \quad [A4]$$

Assume further that Y_i is measured with error by y_i :

$$y_i = Y_i + v_i \quad v_i \sim N(0, \nu). \quad [A5]$$

We wish to estimate the parameters γ_0 , γ_1 , γ_2 , and γ_3 , given the observed x_i and y_i . Consider three approaches to estimating these parameters: 1) regress y_i on B_i and x_i via OLS; 2) rRegress y_i on B_i and an estimate of T_i obtained from shrinking x_i toward the (conditional or unconditional) mean x_b based on the reliability r_x ; 3) regress y_i on B_i and an estimate of T_i obtained from instrumenting for x_i based on a second test score z_i and B . I will show that options 2 and 3 yield unbiased estimates of

the parameters γ_0 , γ_1 , γ_2 , and γ_3 , albeit under different assumptions.

If we regress y_i on x_i and B_i via OLS:

$$y_i = \gamma'_0 + \gamma'_1 x_{1i} + \gamma'_2 B_i + \gamma'_3 x_{1i} B_i + \varepsilon'_i, \quad [A6]$$

it is trivial to show that the expected values of the parameter estimates from [6] are given by:

$$\begin{aligned} E[\hat{\gamma}'_0] &= \gamma_0 + [(1-r_x)\gamma_1\hat{\mu}_w - \lambda\hat{\mu}_w] \\ E[\hat{\gamma}'_1] &= \gamma_1 + [(r_x-1)\gamma_1 + \lambda] \\ E[\hat{\gamma}'_2] &= \gamma_2 + [((1-r_x)\gamma_1 - \lambda)(\hat{\mu}_b - \hat{\mu}_w) - (1-r_x)\gamma_3\hat{\mu}_b] \\ E[\hat{\gamma}'_3] &= \gamma_3 + [(r_x-1)\gamma_3], \end{aligned} \quad [A7]$$

where

$$\lambda = \frac{Cov(e_i, v_i)}{\tau + \sigma}. \quad [A8]$$

The terms in brackets on the right-hand side of [A7] indicate the expected bias of the each of the estimated γ 's from [A6]. Note that the absence of measurement error in x (i.e., $r_x=1$) implies $\lambda=0$, so each of the estimates in [A7] have 0 bias. When $r_x < 1$, however, the bias in each γ' is, in general, non-zero. The biases arise from three factors: 1) $r_x < 1$ (measurement error in score x); 2) $\mu_w \neq \mu_b$ (the two groups have different mean values of T); and 3) $\lambda \neq 0$ (the error in outcome y is correlated with error in x (which occurs, for example if Y measures a gain score in T).¹⁷

If we know r_x , λ , μ_w , and μ_b , we can obtain unbiased estimates of the parameters of [A4] by substituting the estimated γ 's, and r_x , λ , μ_w , and μ_b into [A7] and solving for γ_0 , γ_1 , γ_2 , and γ_3 .

¹⁷ Note that in the case where y_i measures the change in x from time 1 to time 2, and where x_i is the value of x at time 1, we have

$$y_i = (T_{i2} - T_{i1}) + (e_{i2} - e_{i1}), \quad e_{i1} \perp e_{i2},$$

which yields:

$$\begin{aligned} \lambda &= \frac{Cov(e_{i1}, e_{i2} - e_{i1})}{\tau + \sigma} \\ &= r_x - 1 \end{aligned}$$

If we know the reliability of x , and if we know the reliability of x is constant over the range of T , (and if we assume T and ϵ are normally distributed), then the Bayesian shrinkage estimator of T_i is given by

$$T_i^* = (1 - r_x)(\mu_w(1 - B_i) + \mu_b B) + r_x x_i \quad [\text{A9}]$$

Given the r_x and the normality assumptions, T^* is an unbiased estimator of T . Now if we regress y_i on T_i^* via OLS:

$$y_i = \gamma_0^* + \gamma_1^* T_i^* + \gamma_2^* B_i + \gamma_3^* T_i^* B_i + \epsilon_i^* \quad [\text{A10}]$$

we get

$$\begin{aligned} E[\hat{\gamma}_0^*] &= \gamma_0 + \left[-\frac{\lambda}{r_x} \hat{\mu}_w \right] \\ E[\hat{\gamma}_1^*] &= \gamma_1 + \left[\frac{\lambda}{r_x} \right] \\ E[\hat{\gamma}_2^*] &= \gamma_2 \\ E[\hat{\gamma}_3^*] &= \gamma_3 . \end{aligned} \quad [\text{A11}]$$

Shrinking toward the conditional mean of x eliminates the bias in the estimated γ 's except for the bias in γ_0 and γ_1 that is due to the correlation of the errors in x and y .¹⁸ In many cases, we can reasonably assume the errors in x and y are uncorrelated, so there will be no bias in this case. If, however, we regress a gain score measured with error on initial status measured with error, then we

¹⁸ Note that if we shrink x toward its unconditional mean μ , rather than its conditional mean, we do not obtain unbiased estimates of the γ 's, even if we use the conditional reliability (using the unconditional reliability produces even more bias), except in the case where $\mu_w = \mu_b$, in which case shrinking to the conditional and unconditional means are identical:

$$\begin{aligned} E[\hat{\gamma}_0^{**}] &= \gamma_0 + \left[\gamma_1(1 - r_x)(\mu - \mu_w) - \frac{\lambda}{r_x}((1 - r_x)\mu + r_x \mu_w) \right] \\ E[\hat{\gamma}_1^{**}] &= \gamma_1 + \left[\frac{\lambda}{r_x} \right] \\ E[\hat{\gamma}_2^{**}] &= \gamma_2 + [\gamma_3(1 - r_x)(\mu - \mu_b) + (\gamma_1(1 - r_x) - \lambda)(\mu_b - \mu_w)] \\ E[\hat{\gamma}_3^{**}] &= \gamma_3 . \end{aligned}$$

will obtain biased estimates of the slopes and intercept, but unbiased estimates of the differences between groups in intercepts and slopes.

Finally, if we have a second test that measures T with error that is independent of the error in x , we can use the second test as an instrument for x . Suppose we have a second test z , such that

$$z_i = a + c(T_i + w_i), \quad w_i \sim N(0, \omega), \quad w_i \perp e_i. \quad [\text{A12}]$$

The constants a and c allow z to measure T in a metric that is a linear function of the metric of x .

The reliability of z as a measure of T is

$$r_z = \frac{\tau}{\tau + \omega}. \quad [\text{A13}]$$

Now use z as an instrument for T_i in the first-stage equation

$$\begin{aligned} x_i &= \beta_0 + \beta_1 z_i + \beta_2 B_i + \xi \\ &= \beta_0 + \beta_1 (a + c(T_i + w_i)) + \beta_2 B_i + \xi. \end{aligned} \quad [\text{A14}]$$

The OLS parameter estimates from this first-stage equation are

$$\begin{aligned} \hat{\beta}_0 &= (1 - r_z) \mu_w - \frac{a r_z}{c} \\ \hat{\beta}_1 &= \frac{r_z}{c} \\ \hat{\beta}_2 &= (1 - r_z) (\mu_b - \mu_w) \end{aligned} \quad [\text{A15}]$$

This yields:

$$\begin{aligned} \hat{x}_i &= (1 - r_z) (\mu_w (1 - B_i) + \mu_b B_i) + \frac{r_z}{c} (z_i - a) \\ &= (1 - r_z) (\mu_w (1 - B_i) + \mu_b B_i) + r_z (T_i + w_i) \end{aligned} \quad [\text{A16}]$$

Equation [A16] shows that instrumenting for x with z is equivalent to converting z to the metric of x and shrinking z to its conditional mean in the x metric. The advantage to using z as an instrument is that the error in z will be uncorrelated with the error in y , even when y is a gain score on x . If we then regress y on this predicted value of x :

$$y_i = \gamma_0^{iv} + \gamma_1^{iv} \hat{x}_i + \gamma_2^{iv} B_i + \gamma_3^{iv} \hat{x}_i B_i + \varepsilon_i^{iv} \quad [\text{A17}]$$

we get

$$\begin{aligned} E[\hat{\gamma}_0^{iv}] &= \gamma_0 \\ E[\hat{\gamma}_1^{iv}] &= \gamma_1 \\ E[\hat{\gamma}_2^{iv}] &= \gamma_2 \\ E[\hat{\gamma}_3^{iv}] &= \gamma_3. \end{aligned} \quad [\text{A18}]$$

Thus, under the assumption that z is a linear function of T plus some unknown measurement error, using z as an instrument for x yields unbiased estimates of the parameters of interest, regardless of the reliability of x or z .¹⁹ Note that we do not need to know or assume the values of a , c , or r_z in order that the IV strategy yield unbiased estimates of the γ 's. The key assumption of the instrumental variable strategy is that z must measure T . If, instead, z is a measure of some other score G that is correlated with T , conditional on B , then the parameters estimated in [A18] will be measures of the association between G and Y , rather than between T and Y , as desired.

Thus, both the conditional shrinkage estimator and the instrumental variable estimator will give us unbiased estimates of the parameters of interest in the presence of measurement error, but under different assumptions. The conditional shrinkage estimator depends on information (or assumptions) about the reliability of y as a measure of T . The instrumental variable estimator depends on the assumption that z is a linear measure of T .

¹⁹ Note that if we do not include B_i in the first-stage equation in [A14], we get

$$\hat{x}_i = (1 - r_z)\mu + r_z(T_i + w_i),$$

which is equivalent to shrinking z toward the unconditional mean. As noted above, shrinking toward the unconditional mean yields biased estimates.

Appendix B: Derivation of decomposition [15]

Note that the black-white gap defined in Equation [10] is

$$\delta = \frac{\text{Cov}(Y_i, B_i)}{\text{Var}(B_i)}. \quad [\text{B1}]$$

Likewise, note that the fixed-effects estimator of β^e in [11] is equal to

$$\hat{\beta}^{fe} = \frac{\text{Cov}(Y_i, B_i - \pi_s)}{\text{Var}(B_i - \pi_s)}. \quad [\text{B2}]$$

Now write model [14] as

$$Y_i = \beta_0 + \beta_1(B_i - \pi_s) + (\beta_1 + \beta_2)\pi_s + \varepsilon_i. \quad [\text{B3}]$$

Because the school mean value of B_i is π_s , the variables $B_i - \pi_s$ and π_s are orthogonal to one another,

the parameters β_1 and β_2 are given by

$$\beta_1 = \frac{\text{Cov}(Y_i, B_i - \pi_s)}{\text{Var}(B_i - \pi_s)} = \beta^{fe} \quad [\text{B4}]$$

and

$$\beta_1 + \beta_2 = \frac{\text{Cov}(Y_i, \pi_s)}{\text{Var}(\pi_s)}. \quad [\text{B5}]$$

Now we have

$$\begin{aligned} \delta &= \frac{\text{Cov}(Y_i, B_i)}{\text{Var}(B_i)} \\ &= \frac{\text{Cov}(Y_i, B_i - \pi_s)}{\text{Var}(B_i)} + \frac{\text{Cov}(Y_i, \pi_s)}{\text{Var}(B_i)} \\ &= \frac{\text{Cov}(Y_i, B_i - \pi_s)}{\text{Var}(B_i - \pi_s)} \cdot \frac{\text{Var}(B_i) - \text{Var}(\pi_s)}{\text{Var}(B_i)} + \frac{\text{Cov}(Y_i, \pi_s)}{\text{Var}(\pi_s)} \cdot \frac{\text{Var}(\pi_s)}{\text{Var}(B_i)} \\ &= \beta_1(1-V) + (\beta_1 + \beta_2)V \\ &= \beta_1(1-V) + \beta_1V + \beta_2V \end{aligned} \quad [\text{15}]$$

References

- Bollinger, Christopher. 2003. "Measurement error in human capital and the black-white wage gap." *The Review of Economics and Statistics* 85:578-585.
- Brooks-Gunn, Jeanne, Pamela K. Klebanov, and Greg J. Duncan. 1996. "Ethnic differences in children's intelligence test scores: Role of economic deprivation, home environment, and maternal characteristics." *Child Development* 67:396-408.
- Carneiro, Pedro, James J. Heckman, and Dimitry V. Masterov. 2003. "Labor market discrimination and racial differences in premarket factors " National Bureau of Economic Research, Cambridge, MA.
- Clotfelter, Charles T. 1999. "Public school segregation in metropolitan areas." *Land Economics* 75:487-504.
- Clotfelter, Charles T., Helen F. Ladd, and Jacob L. Vigdor. 2006. "The academic achievement gap in grades three to eight." National Bureau of Economic Research, Cambridge, MA.
- Downey, Douglas B., Paul T. von Hippel, and Beckett A. Broh. 2004. "Are Schools the Great Equalizer? School and Non-School Influences on Socioeconomic and Black/White Gaps in Reading Skills." *American Sociological Review* 69:613-635.
- Ferguson, Ronald F. 1998. "Test-Score Trends Along Racial Lines, 1971 to 1996: Popular Culture and Community Academic Standards." Pp. 348-390 in *America Becoming: Racial Trends and Their Consequences*, vol. 1, edited by N. J. Smelser, W. J. Wilson, and F. Mitchell. Washington, D.C.: National Academies Press.
- Fryer, Roland G. and Stephen D. Levitt. 2002. "Understanding the black-white test score gap in the first two years of school." National Bureau of Economic Research, Cambridge, MA.
- . 2004. "Understanding the black-white test score gap in the first two years of school." *The Review of Economics and Statistics* 86:447-464.

- . 2005. "The black-white test score gap through third grade." National Bureau of Economic Research, Cambridge, MA.
- Grissmer, David W., Ann Flanagan, and Stephanie Williamson. 1998. "Why did the Black-White score gap narrow in the 1970s and 1980s?" Pp. 182-228 in *The Black-White Test Score Gap*, edited by C. Jencks and M. Phillips. Washington, D.C.: Brookings Institution Press.
- Hanushek, Eric A., John F. Kain, and Steven G. Rivkin. 2002. "New evidence about *Brown v. Board of Education*: The complex effects of school racial composition on achievement." NBER.
- Hanushek, Eric A. and Steven G. Rivkin. 2006. "School quality and the black-white achievement gap." NBER.
- Heckman, James J. and Edward Vytlacil. 2001. "Identifying the role of cognitive ability in explaining the level of and change in the return to schooling." *Review of Economics and Statistics* 83:1-12.
- Hedges, Larry V. and Amy Nowell. 1999. "Changes in the black-white gap in achievement test scores." *Sociology of Education* 72:111-135.
- Ho, Andrew D. and Edward H. Haertel. 2006. "Metric-Free Measures of Test Score Trends and Gaps with Policy-Relevant Examples." Center for the Study of Evaluation, National Center for Research on Evaluation, Standards, and Student Testing, Graduate School of Education & Information Studies, Los Angeles, CA.
- James, David R. and Karl E. Taeuber. 1985. "Measures of segregation." *Sociological Methodology* 14:1-32.
- Jencks, Christopher and Meredith Phillips. 1998. "The Black-White Test Score Gap." Washington D.C.: Brookings Institution Press.
- Lee, Valerie E. and David T. Burkham. 2002. *Inequality at the Starting Gate: Social Background Differences in Achievement as Children Begin School*. Washington, DC: Economic Policy Institute.
- LoGerfo, Laura, Austin Nichols, and Sean F. Reardon. 2006. "Achievement Gains in Elementary

- and High School." Urban Institute, Washington, DC.
- Lord, F.M. and M.R. Novick. 1968. *Statistical Theory of Mental Test Scores*. Reading, MA: Addison Wesley.
- Murnane, Richard J., John B. Willett, Kristen L. Bub, and Kathleen McCartney. 2006. "Understanding trends in the black-white achievement gaps during the first years of school." *Brookings-Wharton Papers on Urban Affairs*.
- Neal, Derek A. 2005. "Why has Black-White skill convergence stopped?" University of Chicago.
- Neal, Derek A. and William R. Johnson. 1996. "The role of premarket factors in black-white wage differences." *The Journal of Political Economy* 104:869-895.
- 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." Pp. 103-148 in *The Black-White Test Score Gap*, edited by C. Jencks and M. Phillips. Washington, D.C.: Brookings Institution Press.
- Phillips, Meredith, James Crouse, and James Ralph. 1998. "Does the black-white test score gap widen after children enter school?" Pp. 229-272 in *The black-white test score gap*, edited by C. Jencks and M. Phillips. Washington, DC: Brookings Institution Press.
- Pollack, Judith M., Michelle Narajian, Donald A. Rock, Sally Atkins-Burnett, and Elvira Germino Hausken. 2005. "Early Childhood Longitudinal Study-Kindergarten Class of 1998-99 (ECLS-K), Psychometric Report for the Fifth Grade." U.S. Department of Education, National Center for Education Statistics, Washington, DC.
- Reardon, Sean F. and Glenn Firebaugh. 2002. "Measures of multi-group segregation." *Sociological Methodology* 32:33-67.
- Reardon, Sean F. and Claudia Galindo. 2006. "Patterns of Hispanic Students' Math and English Literacy Test Scores in the Early Elementary Grades." National Task Force on Early

Childhood Education for Hispanics.

Reardon, Sean F. and Joseph Robinson. 2007. "Patterns and Trend in Racial/Ethnic and Socioeconomic Academic Achievement Gaps." in *Handbook of Research on Education Finance and Policy*, edited by H. Ladd and E. Fiske.

Selzer, Michael H., Ken A. Frank, and Anthony S. Bryk. 1994. "The metric matters: the sensitivity of conclusions about growth in student achievement to choice of metric." *Educational Evaluation and Policy Analysis* 16:41-49.

Table 1: Black-White Math and Reading Test Score Gaps, Kindergarten through Fifth Grade, by Gap Measure and Wave

	Math					Reading				
	Fall K	Spring K	Spring 1	Spring 3	Spring 5	Fall K	Spring K	Spring 1	Spring 3	Spring 5
Theta Score	-0.32 (0.03)	-0.35 (0.03)	-0.32 (0.03)	-0.34 (0.02)	-0.41 (0.03)	-0.23 (0.03)	-0.24 (0.04)	-0.23 (0.03)	-0.24 (0.02)	-0.26 (0.02)
Standardized T-Score										
($r=.7$)	-1.02 (0.09)	-1.12 (0.09)	-1.11 (0.10)	-1.27 (0.09)	-1.37 (0.09)	-0.71 (0.10)	-0.72 (0.11)	-0.75 (0.10)	-1.03 (0.10)	-1.10 (0.09)
($r=.8$)	-0.90 (0.08)	-0.98 (0.08)	-0.97 (0.09)	-1.11 (0.08)	-1.20 (0.08)	-0.62 (0.08)	-0.63 (0.09)	-0.65 (0.09)	-0.90 (0.08)	-0.96 (0.08)
($r=.9$)	-0.80 (0.07)	-0.87 (0.07)	-0.87 (0.08)	-0.98 (0.07)	-1.07 (0.07)	-0.55 (0.07)	-0.56 (0.08)	-0.58 (0.08)	-0.80 (0.07)	-0.86 (0.07)
($r=1.0$)	-0.72 (0.06)	-0.79 (0.07)	-0.78 (0.07)	-0.89 (0.07)	-0.96 (0.06)	-0.50 (0.07)	-0.51 (0.08)	-0.52 (0.07)	-0.72 (0.07)	-0.77 (0.06)
Scale Score										
	-5.42 (0.46)	-7.99 (0.63)	-12.36 (0.94)	-17.98 (1.31)	-19.41 (1.30)	-4.06 (0.54)	-5.54 (0.84)	-11.45 (1.43)	-17.57 (1.61)	-17.58 (1.45)
$P_{b>w}$										
($r=.7$)	0.21	0.19	0.19	0.17	0.17	0.28	0.28	0.28	0.20	0.20
($r=.8$)	0.23	0.22	0.22	0.20	0.20	0.30	0.30	0.30	0.24	0.23
($r=.9$)	0.26	0.25	0.25	0.23	0.22	0.32	0.32	0.33	0.26	0.26
($r=1.0$)	0.28	0.27	0.27	0.25	0.25	0.34	0.34	0.34	0.29	0.28
Metric-Free Effect Size										
($r=.7$)	-1.17	-1.22	-1.23	-1.34	-1.37	-0.83	-0.83	-0.84	-1.17	-1.18
($r=.8$)	-1.02	-1.07	-1.08	-1.17	-1.21	-0.73	-0.74	-0.73	-1.02	-1.03
($r=.9$)	-0.91	-0.95	-0.95	-1.04	-1.08	-0.65	-0.66	-0.64	-0.89	-0.91
($r=1.0$)	-0.82	-0.86	-0.85	-0.94	-0.97	-0.58	-0.60	-0.57	-0.80	-0.81

Standard errors in parentheses. See text for detailed description of gap measures. N=6,710.

Table 2a: Estimated Black-White Difference in Spring Fifth Grade Test Scores, Conditional on Fall Kindergarten Test Scores, by Subject, Test Metric, and Assumed Reliability

Assumed Reliability	Test Metric: Model:	Math						Reading					
		Theta		T-Score		Scale Score		Theta		T-Score		Scale Score	
		M1(θ)	M2(θ)	M1(T)	M2(T)	M1(S)	M2(S)	R1(θ)	R2(θ)	R1(T)	R2(T)	R1(S)	R2(S)
$r=0.70$	Black	-0.121 ** (0.023)	-0.146 ** (0.028)	-0.287 ** (0.055)	-0.346 ** (0.068)	-8.09 ** (1.25)	-5.54 ** (1.25)	-0.129 ** (0.020)	-0.143 ** (0.021)	-0.387 ** (0.060)	-0.428 ** (0.062)	-11.34 ** (1.48)	-9.04 ** (1.33)
	Standardized Fall K Score*Black		-0.041 (0.026)		-0.096 (0.062)		-0.40 (1.37)		-0.040 * (0.019)		-0.119 * (0.057)		-0.59 (1.31)
$r=0.80$	Black	-0.156 ** (0.023)	-0.178 ** (0.027)	-0.371 ** (0.054)	-0.422 ** (0.064)	-9.45 ** (1.23)	-7.21 ** (1.20)	-0.145 ** (0.020)	-0.158 ** (0.020)	-0.435 ** (0.059)	-0.471 ** (0.061)	-12.09 ** (1.46)	-10.15 ** (1.31)
	Standardized Fall K Score*Black		-0.038 (0.025)		-0.090 (0.060)		0.16 (1.33)		-0.037 * (0.019)		-0.112 * (0.056)		-0.34 (1.30)
$r=0.90$	Black	-0.184 ** (0.023)	-0.203 ** (0.026)	-0.435 ** (0.053)	-0.481 ** (0.062)	-10.52 ** (1.21)	-8.53 ** (1.16)	-0.158 ** (0.020)	-0.169 ** (0.020)	-0.473 ** (0.059)	-0.505 ** (0.060)	-12.67 ** (1.44)	-11.03 ** (1.30)
	Standardized Fall K Score*Black		-0.036 (0.025)		-0.086 (0.059)		0.59 (1.31)		-0.036 † (0.018)		-0.106 † (0.055)		-0.15 (1.29)
$r=1.00$	Black	-0.206 ** (0.022)	-0.223 ** (0.025)	-0.487 ** (0.053)	-0.529 ** (0.060)	-11.37 ** (1.20)	-9.61 ** (1.13)	-0.168 ** (0.020) ^a	-0.178 ** (0.020)	-0.503 ** (0.059)	-0.533 ** (0.059)	-13.14 ** (1.43)	-11.74 ** (1.29)
	Standardized Fall K Score*Black		-0.035 (0.024)		-0.082 (0.057)		0.93 (1.29)		-0.034 † (0.018)		-0.102 † (0.054)		0.00 (1.28)

Notes: Assumed reliability refers to the assumed test-retest reliability of Fall kindergarten test scores. For each outcome variable, the first column reports the estimated average black-white difference in Spring fifth grade scores, conditional on Fall kindergarten scores. The second column reports coefficient estimates from models that include linear and quadratic Fall kindergarten test scores (coefficient estimates not shown) and an interaction term between the Fall kindergarten test score and black. Fall kindergarten scores are standardized to facilitate comparison across models. In additional models including the interaction of the quadratic test score term and black, the interaction term was significant in none of the models, and so only results from the simpler models are reported here.

Robust standard errors are in parentheses. † $p < .10$; * $p < .05$; ** $p < .01$

Table 2b: Estimated Locally Standardized Black-White Difference in Spring Fifth Grade Test Scores, Conditional on Fall Kindergarten Test Scores, by Subject, Test Metric, and Assumed Reliability

Assumed Reliability	Test Metric: Model:	Math						Reading					
		Theta		T-Score		Scale Score		Theta		T-Score		Scale Score	
		M1(θ)	M2(θ)	M1(T)	M2(T)	M1(S)	M2(S)	R1(θ)	R2(θ)	R1(T)	R2(T)	R1(S)	R2(S)
$r=0.70$	Black	-0.904 ** (0.070)	-0.997 ** (0.087)	-0.901 ** (0.070)	-0.992 ** (0.087)	-0.807 ** (0.072)	-1.090 ** (0.093)	-0.916 ** (0.073)	-1.017 ** (0.079)	-0.922 ** (0.072)	-1.025 ** (0.079)	-0.831 ** (0.074)	-0.995 ** (0.082)
	Standardized Fall K Score*Black		-0.130 (0.075)		-0.126 (0.075)		-0.405 ** (0.092)		-0.198 ** (0.072)		-0.201 ** (0.072)		-0.350 ** (0.080)
$r=0.80$	Black	-0.806 ** (0.070)	-0.891 ** (0.085)	-0.810 ** (0.070)	-0.895 ** (0.085)	-0.746 ** (0.070)	-0.957 ** (0.093)	-0.788 ** (0.073)	-0.861 ** (0.077)	-0.791 ** (0.073)	-0.864 ** (0.077)	-0.770 ** (0.071)	-0.884 ** (0.078)
	Standardized Fall K Score*Black		-0.131 (0.076)		-0.132 (0.076)		-0.336 ** (0.094)		-0.158 * (0.071)		-0.158 * (0.071)		-0.271 ** (0.079)
$r=0.90$	Black	-0.705 ** (0.070)	-0.769 ** (0.082)	-0.709 ** (0.070)	-0.775 ** (0.082)	-0.675 ** (0.070)	-0.791 ** (0.087)	-0.699 ** (0.069)	-0.757 ** (0.072)	-0.702 ** (0.069)	-0.759 ** (0.072)	-0.680 ** (0.070)	-0.762 ** (0.074)
	Standardized Fall K Score*Black		-0.109 (0.073)		-0.111 (0.073)		-0.201 * (0.088)		-0.139 * (0.067)		-0.137 ** (0.067)		-0.216 ** (0.074)
$r=1.00$	Black	-0.651 ** (0.069)	-0.715 ** (0.081)	-0.650 ** (0.069)	-0.714 ** (0.081)	-0.644 ** (0.069)	-0.733 ** (0.087)	-0.656 ** (0.072)	-0.710 ** (0.075)	-0.651 ** (0.072)	-0.705 ** (0.075)	-0.660 ** (0.073)	-0.722 ** (0.077)
	Standardized Fall K Score*Black		-0.117 (0.074)		-0.116 (0.074)		-0.167 (0.090)		-0.140 * (0.068)		-0.140 * (0.068)		-0.181 * (0.077)

Notes: Assumed reliability refers to the assumed test-retest reliability of Fall kindergarten test scores. For each outcome variable, the first column reports the estimated average black-white difference in locally standardized (standardized conditional on estimated true Fall kindergarten scores) Spring fifth grade scores, conditional on Fall kindergarten scores. The second column reports coefficient estimates from models that include linear and quadratic Fall kindergarten test scores (coefficient estimates not shown) and an interaction term between the Fall kindergarten test score and black. Fall kindergarten scores are standardized to facilitate comparison across models. In additional models including the interaction of the quadratic test score term and black, the interaction term was significant in none of the models, and so only results from the simpler models are reported here. Local standardization of Spring fifth grade scores is computed by dividing Fall kindergarten scores into 50 quantiles, and standardizing fifth grade scores within each quantile using the unweighted pooled standard deviation within each quantile (standard deviations are adjusted for variation among students in the time elapsed between wave 1 and wave 6 assessment). Results are unchanged if 25 or 100 quantiles are used. Robust standard errors are in parentheses. * $p < .05$; ** $p < .01$

Table 3a: Three-Part Decomposition of Black-White Test Score Gaps, Math and Reading Standardized T-Score Gaps, by Wave and Period

	Math Standardized T-score								Reading Standardized T-score							
	Total Gap	Within-School Component	Interaction Component	Between-School Component	Total Gap	Within-School Component	Interaction Component	Between-School Component	Total Gap	Within-School Component	Interaction Component	Between-School Component	Total Gap	Within-School Component	Interaction Component	Between-School Component
<u>Cross-Sectional Gaps</u>																
Fall K	0.785 (0.041) *	0.154 (0.022) *	20%	0.316 (0.045) *	40%	0.315 (0.063) *	40%	0.511 (0.040) *	0.091 (0.021) *	18%	0.187 (0.043) *	37%	0.232 (0.062) *	45%		
Spring K	0.837 (0.040) *	0.160 (0.021) *	19%	0.327 (0.044) *	39%	0.349 (0.061) *	42%	0.536 (0.039) *	0.098 (0.020) *	18%	0.201 (0.041) *	37%	0.237 (0.060) *	44%		
Spring 1	0.877 (0.039) *	0.172 (0.022) *	20%	0.351 (0.044) *	40%	0.354 (0.060) *	40%	0.609 (0.037) *	0.054 (0.020) *	9%	0.110 (0.041) *	18%	0.444 (0.058) *	73%		
Spring 3	0.985 (0.039) *	0.217 (0.021) *	22%	0.443 (0.044) *	45%	0.325 (0.061) *	33%	0.932 (0.039) *	0.164 (0.021) *	18%	0.336 (0.044) *	36%	0.431 (0.060) *	46%		
Spring 5	1.009 (0.039) *	0.208 (0.021) *	21%	0.426 (0.043) *	42%	0.375 (0.061) *	37%	0.924 (0.039) *	0.169 (0.022) *	18%	0.347 (0.044) *	38%	0.408 (0.061) *	44%		
<u>Changes in Gaps</u>																
Fall K - Spring K	0.052 (0.026) *	0.006 (0.014)	11%	0.012 (0.029)	23%	0.035 (0.040)	66%	0.026 (0.026)	0.007 (0.014)	26%	0.014 (0.029)	54%	0.005 (0.040)	20%		
Spring K - Spring 1	0.040 (0.027)	0.011 (0.015)	28%	0.024 (0.032)	59%	0.005 (0.043)	13%	0.072 (0.027) *	-0.044 (0.015) *	-61%	-0.090 (0.030) *	-125%	0.207 (0.041) *	286%		
Spring 1 - Spring 3	0.108 (0.027) *	0.045 (0.015) *	42%	0.092 (0.031) *	85%	-0.029 (0.042)	-27%	0.323 (0.028) *	0.111 (0.016) *	34%	0.225 (0.032) *	70%	-0.013 (0.044)	-4%		
Spring 3 - Spring 5	0.024 (0.021)	-0.008 (0.012)	-35%	-0.017 (0.024)	-72%	0.050 (0.032)	207%	-0.008 (0.023)	0.005 (0.013)	-61%	0.010 (0.027)	-124%	-0.023 (0.035)	285%		
<u>Total Gap Change</u>																
Fall K - Spring 5	0.225 (0.034) *	0.054 (0.019) *	24%	0.110 (0.038) *	49%	0.061 (0.052)	27%	0.413 (0.038) *	0.078 (0.022) *	19%	0.159 (0.044) *	39%	0.176 (0.059) *	43%		

Notes: Sample includes 4,766 students (696 black and 4,070 white) who remained in the same school and had valid math and reading scores at each wave. Math and reading scores are standardized by dividing by assessment date adjusted pooled standard deviation of test scores of full analytic sample (see text). All estimates weighted by ECLS-K panel weight *c1_6f0*. Standard errors in parentheses.

Table 3b: Three-Part Decomposition of Black-White Test Score Gaps, Math and Reading Theta Score Gaps, by Wave and Period

	Math Theta Score								Reading Theta Score							
	Total Gap	Within-School Component	Interaction Component	Between-School Component	Total Gap	Within-School Component	Interaction Component	Between-School Component	Total Gap	Within-School Component	Interaction Component	Between-School Component	Total Gap	Within-School Component	Interaction Component	Between-School Component
<u>Cross-Sectional Gaps</u>																
Fall K	0.346 (0.018) *	0.068 (0.010) *	20%	0.139 (0.020) *	40%	0.139 (0.028) *	40%	0.233 (0.018) *	0.042 (0.010) *	18%	0.085 (0.020) *	37%	0.106 (0.028) *	45%		
Spring K	0.357 (0.017) *	0.068 (0.009) *	19%	0.140 (0.019) *	39%	0.149 (0.026) *	42%	0.244 (0.018) *	0.045 (0.009) *	18%	0.091 (0.019) *	37%	0.108 (0.027) *	44%		
Spring 1	0.341 (0.015) *	0.067 (0.008) *	20%	0.136 (0.017) *	40%	0.138 (0.023) *	40%	0.254 (0.016) *	0.023 (0.008) *	9%	0.046 (0.017) *	18%	0.186 (0.024) *	73%		
Spring 3	0.363 (0.015) *	0.080 (0.008) *	22%	0.163 (0.016) *	45%	0.120 (0.022) *	33%	0.292 (0.012) *	0.052 (0.007) *	18%	0.106 (0.014) *	36%	0.135 (0.019) *	46%		
Spring 5	0.427 (0.017) *	0.088 (0.009) *	21%	0.180 (0.018) *	42%	0.159 (0.026) *	37%	0.299 (0.013) *	0.055 (0.007) *	18%	0.112 (0.014) *	38%	0.132 (0.020) *	44%		
<u>Changes in Gaps</u>																
Fall K - Spring K	0.011 (0.011)	0.000 (0.006)	3%	0.001 (0.013)	5%	0.011 (0.017)	92%	0.011 (0.012)	0.003 (0.007)	28%	0.006 (0.013)	54%	0.002 (0.018)	17%		
Spring K - Spring 1	-0.017 (0.011)	-0.002 (0.006)	10%	-0.003 (0.013)	21%	-0.011 (0.017)	69%	0.011 (0.012)	-0.022 (0.006) *	-204%	-0.045 (0.013) *	-417%	0.078 (0.018) *	721%		
Spring 1 - Spring 3	0.022 (0.010) *	0.013 (0.006) *	59%	0.027 (0.012) *	122%	-0.018 (0.016)	-81%	0.038 (0.011) *	0.029 (0.006) *	76%	0.060 (0.012) *	156%	-0.050 (0.017) *	-132%		
Spring 3 - Spring 5	0.064 (0.009) *	0.008 (0.005)	13%	0.017 (0.010)	26%	0.039 (0.013) *	61%	0.007 (0.007)	0.003 (0.004)	50%	0.007 (0.009)	98%	-0.003 (0.011)	-48%		
<u>Total Gap Change</u>																
Fall K - Spring 5	0.081 (0.015) *	0.020 (0.008) *	25%	0.041 (0.017) *	50%	0.020 (0.023)	25%	0.066 (0.016) *	0.013 (0.009)	20%	0.027 (0.018)	41%	0.026 (0.024)	39%		

Notes: Sample includes 4,766 students (696 black and 4,070 white) who remained in the same school and had valid math and reading scores at each wave. Math and reading scores are standardized by dividing by assessment date adjusted pooled standard deviation of test scores of full analytic sample (see text). All estimates weighted by ECLS-K panel weight *c1_6f0*. Standard errors in parentheses.

Table 3c: Three-Part Decomposition of Black-White Test Score Gaps, Math and Reading Scale Score Gaps, by Wave and Period

	Math Scale Score								Reading Scale Score							
	Total Gap	Within-School Component		Interaction Component		Between-School Component		Total Gap	Within-School Component		Interaction Component		Between-School Component			
<u>Cross-Sectional Gaps</u>																
Fall K	5.792 (0.337) *	1.121 (0.182) *	19%	2.292 (0.373) *	40%	2.379 (0.522) *	41%	3.677 (0.356) *	0.583 (0.194) *	16%	1.191 (0.397) *	32%	1.903 (0.552) *	52%		
Spring K	8.151 (0.432) *	1.583 (0.237) *	19%	3.237 (0.484) *	40%	3.331 (0.669) *	41%	5.475 (0.495) *	1.061 (0.267) *	19%	2.170 (0.546) *	40%	2.244 (0.768) *	41%		
Spring 1	13.187 (0.623) *	2.699 (0.347) *	20%	5.518 (0.709) *	42%	4.970 (0.965) *	38%	11.972 (0.811) *	1.128 (0.441) *	9%	2.307 (0.901) *	19%	8.537 (1.253) *	71%		
Spring 3	19.476 (0.767) *	4.330 (0.418) *	22%	8.851 (0.854) *	45%	6.295 (1.188) *	32%	21.384 (0.886) *	3.800 (0.485) *	18%	7.769 (0.992) *	36%	9.814 (1.369) *	46%		
Spring 5	20.535 (0.758) *	4.079 (0.410) *	20%	8.339 (0.838) *	41%	8.117 (1.171) *	40%	20.228 (0.820) *	3.559 (0.445) *	18%	7.276 (0.910) *	36%	9.394 (1.266) *	46%		
<u>Changes in Gaps</u>																
Fall K - Spring K	2.358 (0.260) *	0.462 (0.147) *	20%	0.944 (0.301) *	40%	0.952 (0.404) *	40%	1.798 (0.293) *	0.478 (0.161) *	27%	0.979 (0.329) *	54%	0.341 (0.455)	19%		
Spring K - Spring 1	5.036 (0.414) *	1.116 (0.237) *	22%	2.281 (0.485) *	45%	1.639 (0.642) *	33%	6.497 (0.552) *	0.067 (0.309)	1%	0.137 (0.632)	2%	6.293 (0.853) *	97%		
Spring 1 - Spring 3	6.289 (0.514) *	1.630 (0.285) *	26%	3.334 (0.582) *	53%	1.325 (0.799)	21%	9.412 (0.657) *	2.672 (0.372) *	28%	5.462 (0.760) *	58%	1.278 (1.019)	14%		
Spring 3 - Spring 5	1.060 (0.410) *	-0.250 (0.230)	-24%	-0.512 (0.470)	-48%	1.822 (0.637) *	172%	(1.156) (0.496) *	(0.241) (0.291)	21%	(0.493) (0.594)	43%	-0.421 (0.770)	36%		
<u>Total Gap Change</u>																
Fall K - Spring 5	14.743 (0.622) *	2.958 (0.344) *	20%	6.047 (0.703) *	41%	5.738 (0.962) *	39%	16.552 (0.723) *	2.976 (0.407) *	18%	6.084 (0.832) *	37%	7.491 (1.117) *	45%		

Notes: Sample includes 4,766 students (696 black and 4,070 white) who remained in the same school and had valid math and reading scores at each wave. Math and reading scores are standardized by dividing by assessment date adjusted pooled standard deviation of test scores of full analytic sample (see text). All estimates weighted by ECLS-K panel weight c1_6f0. Standard errors in parentheses.

Figure 1:

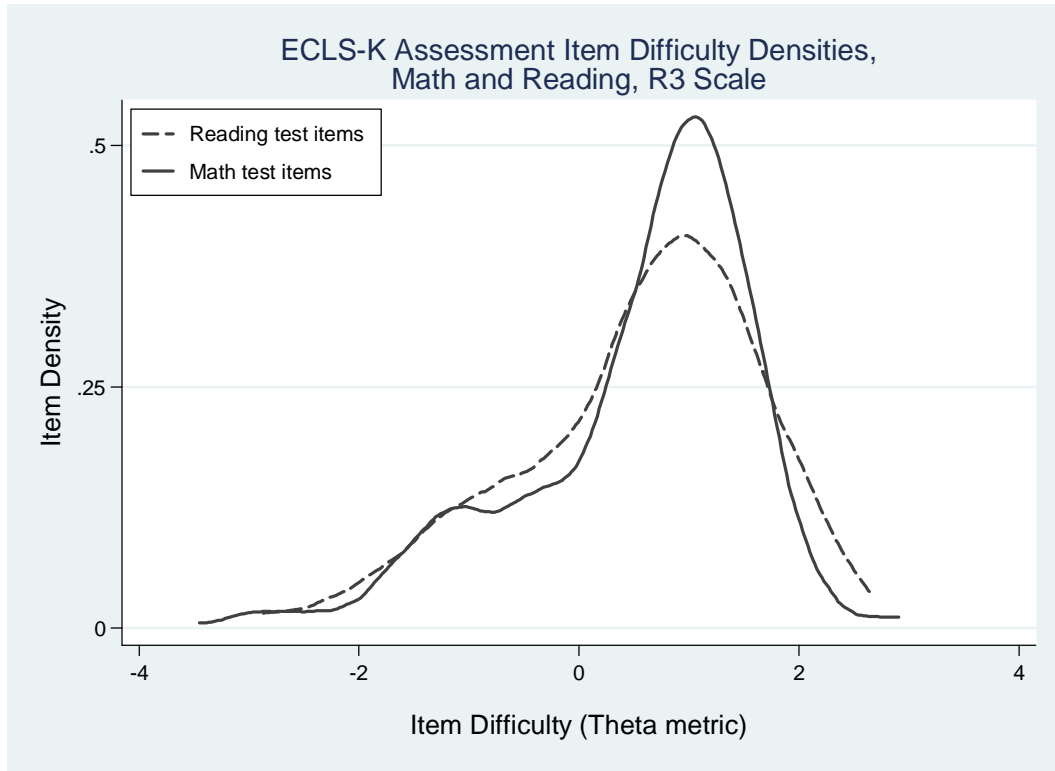


Figure 2:

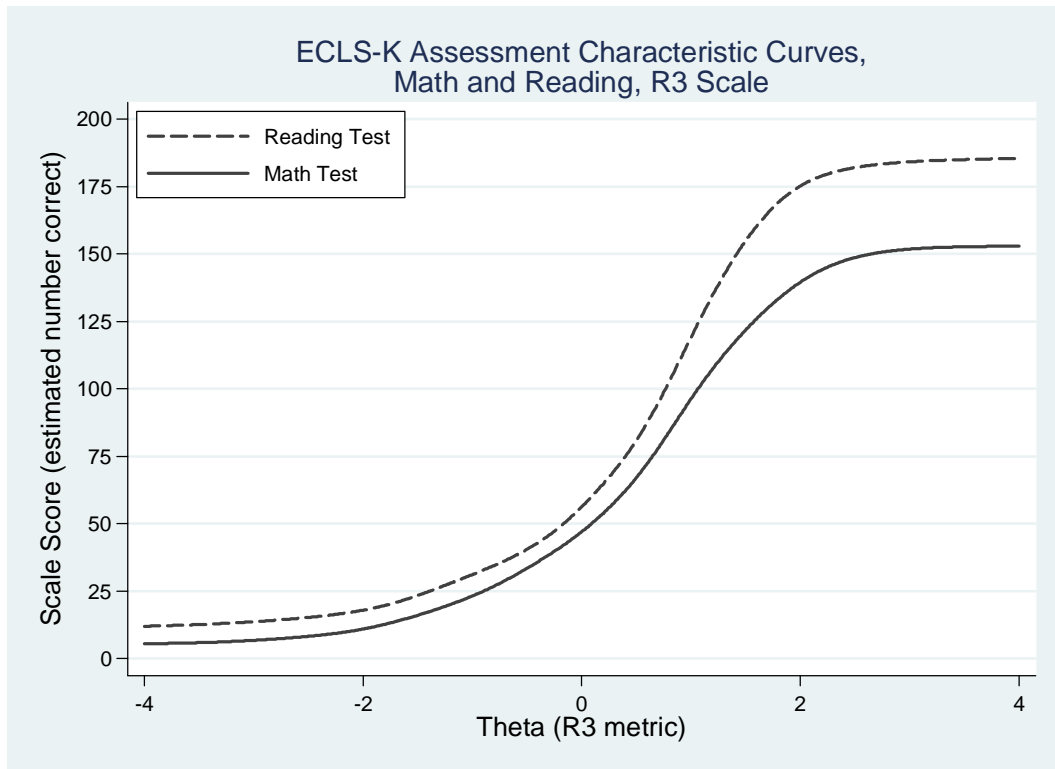


Figure 3:

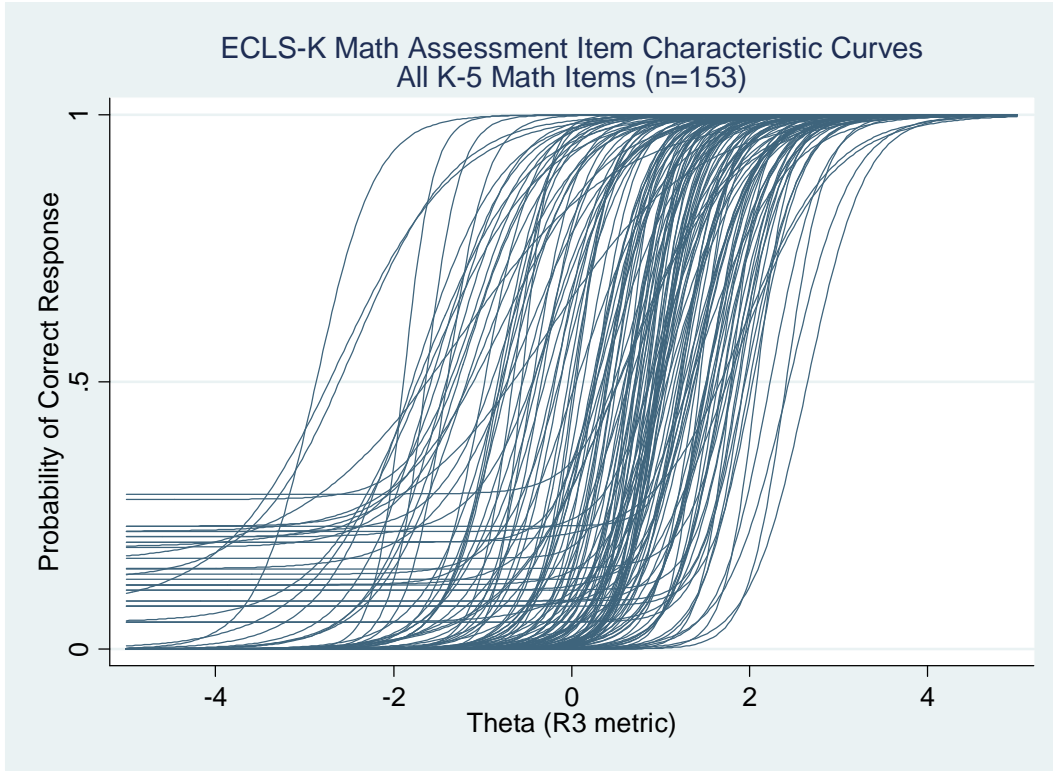


Figure 4:

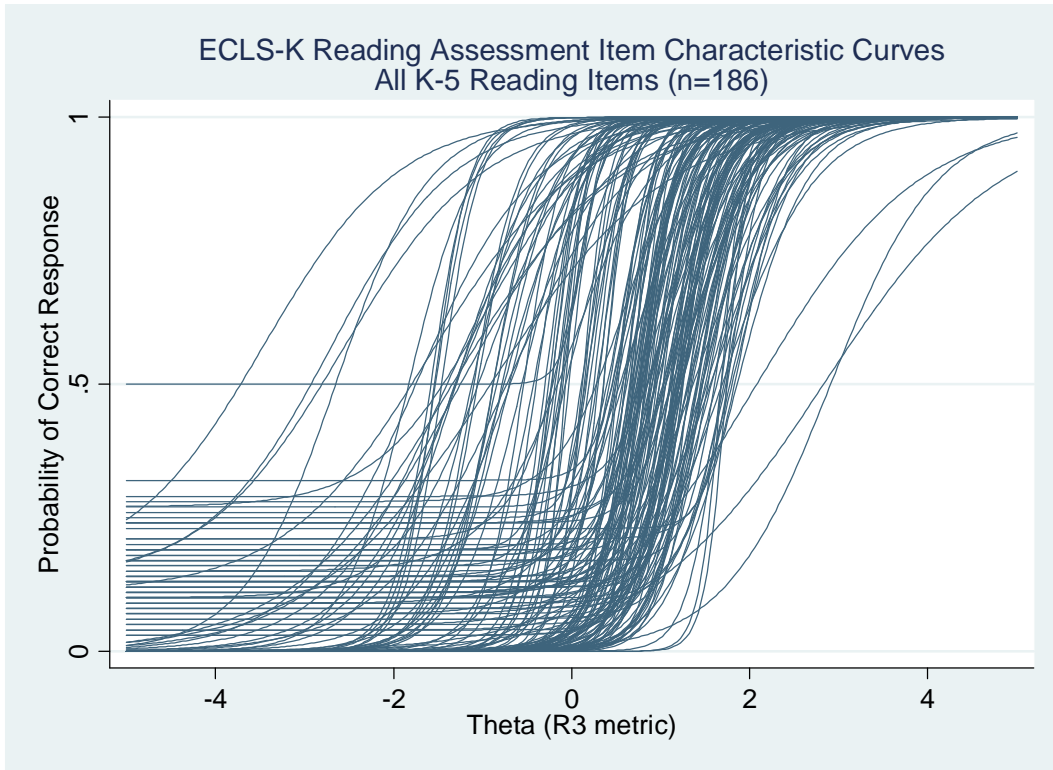


Figure 5:

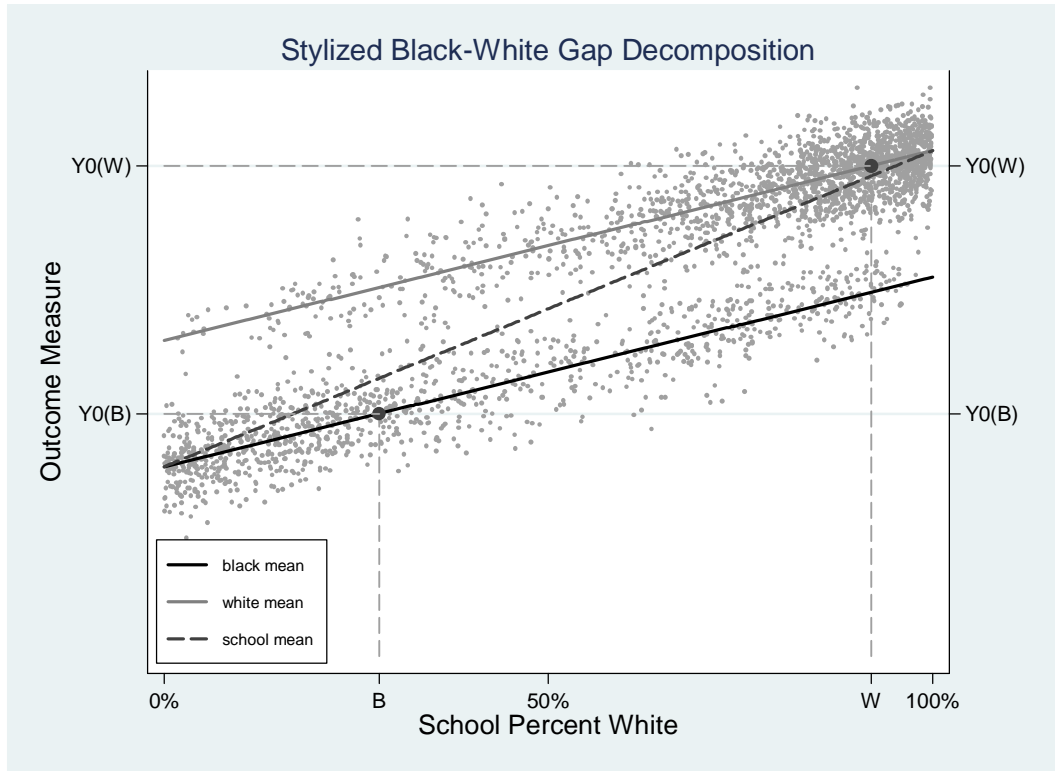


Figure 6:

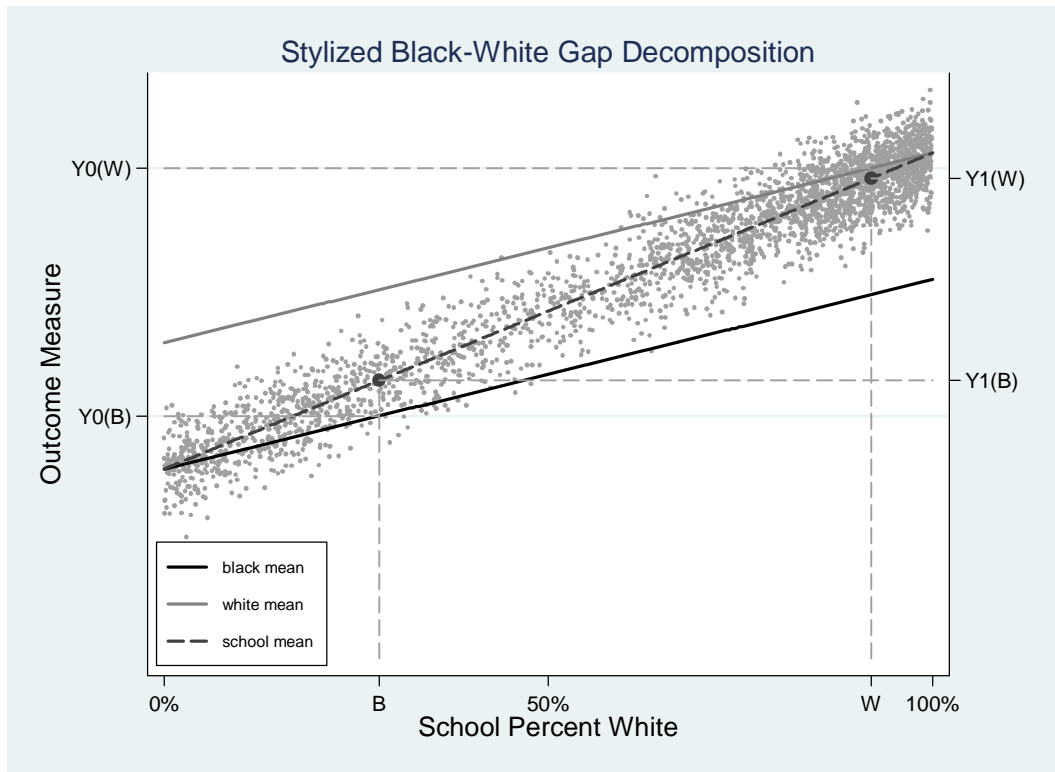


Figure 7:

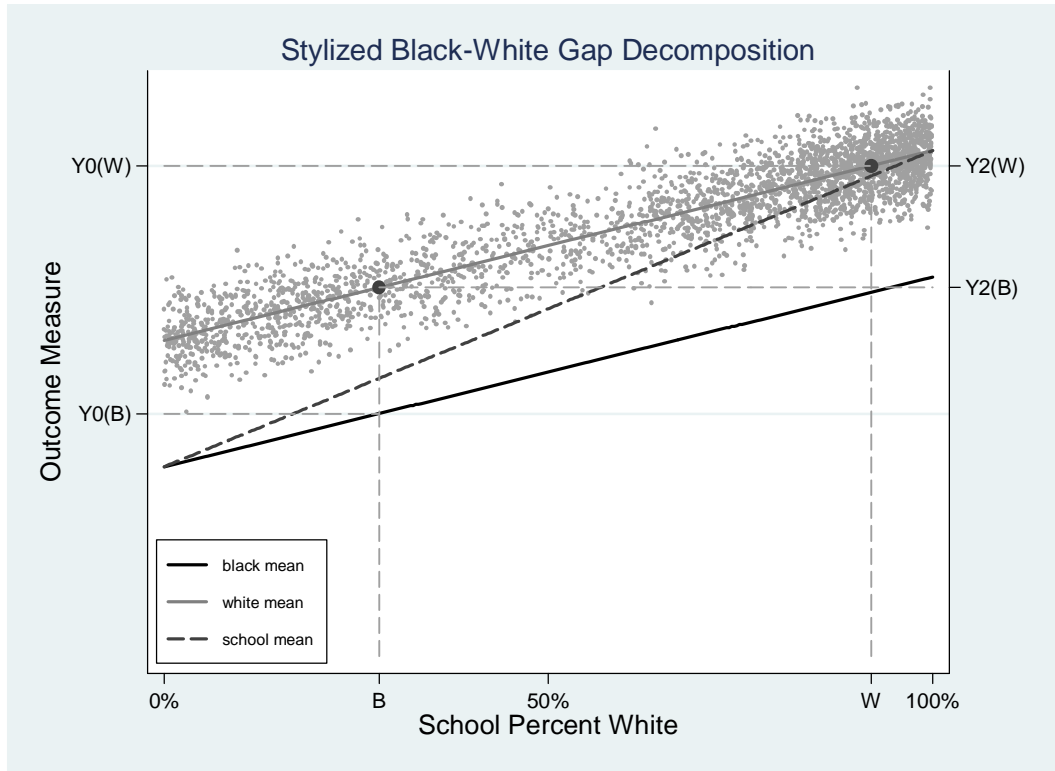


Figure 8:

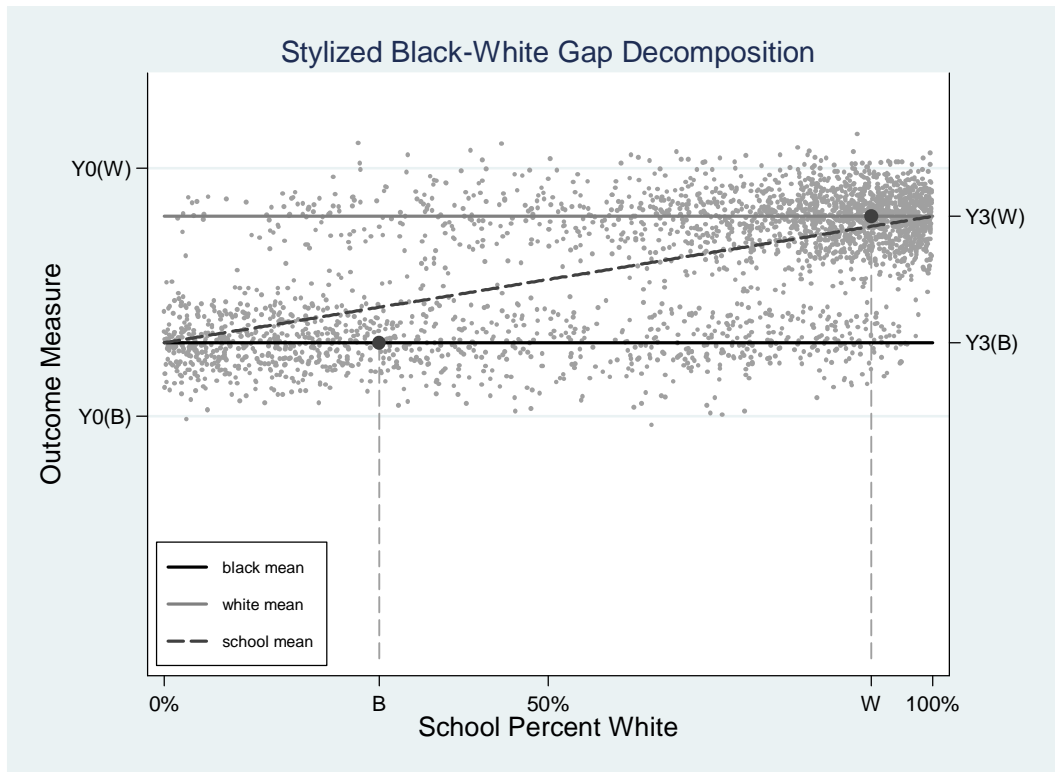


Figure 9:

