



STANFORD UNIVERSITY

Institute for Research on Education Policy & Practice

WORKING PAPER #: 2008-07

**Differential Growth in the Black-White Achievement Gap  
During Elementary School Among Initially High- and  
Low-Scoring Students**

Sean F. Reardon  
Stanford University

March 2008

PRELIMINARY DRAFT:  
comments & suggestions welcome

Direct correspondence to [sean\\_reardon@stanford.edu](mailto: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.

This paper is part of a series of working papers focused on significant education policy and practice issues.  
Informing change & promoting innovation through rigorous & collaborative research

# **Differential Growth in the Black-White Achievement Gap During Elementary School Among Initially High- and Low-Scoring Students**

sean f. reardon  
*Stanford University*

## **Abstract**

The black-white cognitive test score gap is a stubborn feature of U.S. schooling and society. In this paper, I use data from a nationally representative sample of children enrolled in kindergarten in the fall of 1998 to examine the extent to which black-white test score gaps grow differently among initially high- and low-achieving students. Two methodological challenges complicate such analyses: the presence of measurement error in the test scores and ambiguity regarding the interval-scaled nature of test score metrics. I suggest approaches to overcoming these challenges. I find that reading and math test scores diverge more between kindergarten and fifth grade among students who enter kindergarten with high levels of reading and math skill than among students who enter with low levels of reading skill. In fact, the gaps grow roughly twice as fast for students who begin school with scores one standard deviation above the mean as for those who begin one standard deviation below the mean.

## 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 (see, for example, Fryer and Levitt 2004; Hedges and Nowell 1999; Jencks and Phillips 1998; Neal 2005; Reardon and Robinson 2007). 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 methodological complexities arising from differences among studies in the tests and metrics used to measure the gap and the need to account for measurement error in test scores.

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%.<sup>1</sup> 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 2003; Neal and Johnson 1996).<sup>2</sup>

---

<sup>1</sup> 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.

<sup>2</sup> 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).

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 Vytlačil 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 (Reardon and Robinson 2007). 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 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 primarily the first and second of these questions, focusing on the

patterns of development of test score gaps within a given cohort. The other questions noted above, which deal with the sorting of students among schools, 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.<sup>3</sup>

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 details the data I use. In the third section, I describe my analytic approach. In particular, I discuss the implications of measurement error in test scores for the analyses. In addition, because conclusions regarding changes in the magnitude of the test score gaps may depend on the metric in which test scores are reported (Murnane, Willett, Bub, and McCartney 2006; Selzer, Frank, and Bryk 1994), I describe an approach that is insensitive to monotonic transformations of the test scores.

Following this, I describe and discuss my findings. The results indicate that reading and math test scores diverge more between kindergarten and fifth grade among students who enter kindergarten with high levels of reading and math skill than among students who enter with low levels of reading skill. I conclude with a brief discussion of the possible causes and implications of these results.

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

Prior research on the development of the black-white achievement gap comes from two

---

<sup>3</sup> A number of recent papers discuss the extent to which black-white test score gaps grow between and within schools (Cook and Evans 2000; Fryer and Levitt 2004; Hanushek and Rivkin 2006; Page, Murnane, and Willett 2008; Reardon 2007). 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, Willett, Bub, and McCartney 2006; see, for example, Phillips, Brooks-Gunn, Duncan, Klebanov, and Crane 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; see, for example, 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).

types of studies—studies that use longitudinal panel data on one or more cohorts of students,<sup>4</sup> and studies that rely on repeated cross-sectional data to infer developmental patterns.<sup>5</sup> 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 three-quarters and one half of a standard deviation, respectively (Fryer and Levitt 2004; Reardon 2007; Reardon and Galindo 2006). 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 2007; Reardon and Galindo 2006).<sup>6</sup> Table 1, taken from Reardon (2007) reports the magnitude and development of the black-white

---

<sup>4</sup> 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](http://www.nces.ed.gov/ecls)), the National Education Longitudinal Study (NELS) (see [www.nces.ed.gov/surveys/nels88](http://www.nces.ed.gov/surveys/nels88)), Prospects: The Congressionally Mandated Study of Educational Growth and Opportunity, and High School and Beyond (HSB) (see [www.nces.ed.gov/surveys/hsb](http://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).

<sup>5</sup> 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/](http://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 4<sup>th</sup>-, 8<sup>th</sup>-, and 12<sup>th</sup>-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.

<sup>6</sup> Some studies using the ECLS-K data 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), and find that the black-white gap increases very dramatically from kindergarten through fifth grade (Hanushek and Rivkin 2006; Murnane, Willett, Bub, and McCartney 2006). I report the gaps in scale score units for comparison in Table 1 below, although in general, the ECLS-K scale scores are inappropriate for measuring the change in gaps over time (Reardon 2007).

achievement gap from kindergarten through fifth grade.

Table 1 here

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, Willett, Bub, and McCartney 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.

#### *Does the Black-White Achievement Gap Grow Differentially Among High- and Lower-Skill Students?*

Most studies examining achievement disparities between groups focus on differences in mean achievement. There are, however, important reasons to examine the disparities across the full distribution of test scores. For example, underlying the debate regarding affirmative action in admissions to highly competitive colleges is the fact that black and Hispanic students are dramatically underrepresented in the upper end of the achievement distribution. As noted above, Neal (2005, see Figures 2a-2d) shows that roughly 5 percent of black students aged 13-17 years old in the 1990s had math scores in the top quartile of the white math score distribution. Such patterns suggest the importance of investigating not only differences in the black and white test score distributions, but also of investigating when and how such differences emerge.

Given evidence that there are no substantial differences in cognitive skill in early childhood between black and white children (Fryer and Levitt 2006), the substantial underrepresentation of black school-age children in the high end of the test score distribution implies that the mean growth rate of cognitive skill is lower for black children than for white children between birth and

adolescence. What is not clear, however, is whether the pattern of mean growth rate differences is exacerbated by differential growth rate differences for students with different levels of initial cognitive skill. In other words, do black students with high level of initial cognitive skill learn fall behind their similarly skilled white peers even faster than do black students with lower levels of initial skill?

There are several potential theoretical reasons to expect that black-white gaps might grow faster at the high end of the cognitive skill distribution than at the lower end. First is the relationship between racial segregation and within-school skill distributions. Given the high levels of black-white school segregation in the U.S. and the substantial black-white achievement gap when children enter school, the average black kindergarten student with a given level of math or reading skill attends a school with lower mean cognitive skill than the average similarly-skilled white kindergarten student. Initially high-skilled black students will be in schools where they are farther above the median student than similarly-skilled white students. If curriculum and instruction in schools are tailored to the median student in the school, then high skill black students will, on average, receive less challenging curriculum and instruction than their similar white peers, leading potentially to differential rates of achievement growth between such students. For students at the low end of the distribution, the opposite pattern may occur. Low-skilled black students will typically be in schools where the curriculum and instruction are targeted near, or slightly above, their skill level, while similarly-skilled white students will be in schools where they are well below the median student's skill level. From a Vygotskian (1978) perspective, we might expect the black student to learn more in such a case, though the relationship between peer skill composition and learning rates remains unclear (Angrist and Lang 2004; Boozer and Cacciola 2001; Hoxby and Weingarth 2006; Vigdor and Nechyba 2004).

There are other reasons why we might expect differences in learning rates at different parts of the skill distribution. If there are differences in the expectations and behaviors of teachers with

regard to black and white students, and if these differences vary by students skill level (teachers treat low-achieving black and white students similarly, but high-achieving black and white students differently), this could produce differential rates of achievement gap growth. If high-achieving black and white students differ more in terms of their family background and home resources than low-achieving black and white students, this could also lead to differential growth rates. Prior to investigating such mechanisms, however, it is necessary to ascertain the extent to which there is solid empirical evidence indicating such differential growth rates of the achievement gap.

Answering the question of whether gaps grow at different rates across the skill distribution empirically turns out to be more complex than it would seem, however, because any comparison of the magnitude of gaps or differences in growth rates relies on the assumption that the test metric used is interval-scaled. Clotfelter, Ladd, and Vigdor (2006) investigate whether the gap in scores between the 90<sup>th</sup> percentiles of the black and white test score distributions grows or narrows faster than the gap between the 10<sup>th</sup> percentiles of the distributions. They find that in math, racial test score gaps measured in standard deviation units generally narrow from grades three to eight at the 10<sup>th</sup> percentiles of the score distributions, and widen at the same time at the 90<sup>th</sup> percentiles of the distributions. They find no such pattern for reading. They interpret the math pattern as potentially a result of accountability pressures, arguing that the compression of the gap at the low end of the test score distribution is a result of policies that push schools to reduce the percentage of students scoring below certain thresholds. Likewise, they view the expansion of the gap at the high end as a result of the diversion of resources away from high-achieving minority students (because such students are in schools with many low-achieving students). While this is a plausible explanation, it is also possible that the results are an artifact of the tests used to measure the gaps. If the third- and eighth-grade tests are not both scored in interval-scaled metrics, and if the eighth-grade test metric is more sensitive to variation at the high end of the distribution than is the third-grade test, then the pattern they find would be observed in the absence of any true difference in the rate of the

gap growth.

In addition, measurement error in test scores will also tend to bias estimates of differential growth rates, because conditioning growth rates on scores measured with error will systematically bias estimates of differences in growth rates (Hanushek and Rivkin 2006; Reardon 2007).

Hanushek & Rivkin (2006) attempt to remove measurement error bias by conditioning 8<sup>th</sup> grade math scores on 3<sup>rd</sup> grade reading scores, arguing that the measurement error in the math and reading tests are uncorrelated. They find that the gap in math skills from third to eighth grade grows more rapidly among initially high-achieving (in reading) students than among initially low-achieving students. Their approach does not satisfactorily eliminate measurement error bias, however. Because the reading test is measured with error, their estimates of the differential growth of the black-white gap conditional on initial skill will be biased downward by regression to the mean. See Appendix for more detail.

In sum, relatively little empirical research has attempted to systematically address the question of whether achievement gaps within a cohort grow or narrow differentially across the range of skill distribution.<sup>7</sup> What research there is has generally not adequately addressed the complexities of measurement error and scale ambiguity. The analyses in this paper attempt to address these confounding issues.

## **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

---

<sup>7</sup> The comparison across cohorts relies much less on the assumption of interval scaling, since it is possible to compare the full test score distributions across cohorts. See for example, Hedges and Nowell (1999) and Ferguson (1998).

points during the years 1998-2004 (fall 1998, spring 1999, fall 1999, spring 2000, spring 2002, and spring 2004).<sup>8</sup> In addition to these cognitive developmental measures, the ECLS-K data include information gathered from parents, teachers, and school administrators regarding family, 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 5,604 white and 1,044 black students who were assessed at each of waves 1, 2, 4, 5, and 6. These students were sampled from 810 kindergarten schools (623 public and 187 private schools). In all analyses, I use ECLS-K panel sampling weights (weight *c1\_6fc0* 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). In addition, I use clustered standard errors to account for the school-level clustering of the sampling design.

The ECLS-K data set includes math and reading scores for each student at each wave measured in two different metrics: the T-score metric and the scale score metric. The T-scores are linear transformations of the IRT theta ( $\theta$ ) scores, scaled within each wave to have a sample mean of 50 and sample standard deviation of 10. The scale scores are a nonlinear transformation of the IRT  $\theta$  scores (for detail on the differences between these metrics, see Reardon 2007). In this paper, I rely primarily on the  $\theta$  scores, though I include analyses using the scale scores for comparison.

### 3. Analytic Strategy

In principle, to investigate whether achievement gaps grow at different rates among

---

<sup>8</sup> 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.

students with high initial math or reading skills than among those with lower initial skills, we could fit models of the form

$$T_{i5} = f(T_{iK}) + \delta(B_i) + \gamma(T_{iK}B_i) + \epsilon_i \quad [1]$$

where  $T_{i5}$  is the test score of student  $i$  in grade five;  $T_{iK}$  is the true score of student  $i$  in the fall of kindergarten (centered around its mean),  $f$  is some continuous function,  $B_i$  is indicator variable for race (black=1; white=0), and  $\epsilon_i$  is a random error term. In this model,  $\delta$  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  $\gamma$ , which indicates the extent to which the black-white difference in fifth grade scores varies with initial scores. A negative value of  $\gamma$  indicates that the black-white gap grows faster between initially high-achieving black and white students than among initially low-achieving students.

Two issues complicate the estimation and interpretation of  $\delta$  and  $\gamma$ . First, we do not observe a student's true skill  $T_{iK}$ , but rather an error-prone measure of  $T_{iK}$  (the observed test score  $t_{iK}$ ). Conditioning on an error-prone measure yields biased estimates of the parameters of [1] above. Second, the sign of  $\hat{\gamma}$  will be sensitive to the test metric used. A nonlinear monotonic transformation of the metric in which  $T_{i5}$  is measured may substantially alter the estimate of  $\gamma$ , even reversing its sign. Unless we are confident that  $T_{i5}$  is measured in an interval-scaled metric, inferences regarding  $\gamma$  are suspect.

#### *Eliminating bias due to measurement error*

Appendix A describes the bias in  $\delta$  and  $\gamma$  that result from measurement error in  $t_{iK}$ . Two remedies for this bias are available. First, we can “shrink”  $t_{iK}$  toward its conditional mean, using the Bayesian shrinkage estimator and rescaling the test into units representing the within-group standard deviation units of the true fall kindergarten scores:

$$t_{iK}^* = \frac{[(1-r_K)((1-B_i)\hat{\mu}_{wK} + B_i\hat{\mu}_{bK}) + r_K t_{iK}]}{\sqrt{r_K \text{Var}(t_{iK})}} \quad [2]$$

where  $r_K$  is the within-group reliability of  $t_{iK}$  as a measure of  $T_{iK}$  and  $\hat{\mu}_{wK}$  and  $\hat{\mu}_{bK}$  are the estimated mean values of  $T_{iK}$  for white and black students, respectively. If we use  $t_{iK}^*$  in Equation [1] above in place of  $t_{iK}$ , we eliminate bias in the estimates of  $\delta$  and  $\gamma$  due to measurement error in  $t_{iK}$  (details in Appendix). This approach requires that we know the within-group reliability  $r_K$  of  $t_{iK}$ .

Second, if we have a second measure of  $T_{iK}$ —another test score  $z_{iK}$  that measures the same cognitive skill as is measured by  $t_{iK}$ , for example—we can use this second measure as an instrument to identify the portion of the within-group variation in  $t_{iK}$  that is due to  $T_{iK}$  rather than measurement error. In effect, we can use a second test to estimate the reliability of the first test. This approach requires that we have a second test that measures the same cognitive skill as measured by  $t_{iK}$  (details in the Appendix).

In this paper, I rely on the first approach and report results based on a range of assumed reliabilities of the test scores.<sup>9</sup> The reported reliabilities of the ECLS-K tests range from 0.89 to 0.96 across waves and test subjects (Pollack, Narajian, Rock, Atkins-Burnett, and Hausken 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  $t_1$  and  $t_2$  are standardized test scores at two time points, the test-retest reliability  $r$  of the test is given by

$$r = \text{Corr}(t_1, t_2) - \text{Cov}(T_1, \Delta T) \quad [3]$$

where  $T_1$  is the true skill at time 1 and  $\Delta T$  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 [3] may be positive (if students with initially higher skills learn fastest) or negative (if

---

<sup>9</sup> It may be possible to employ the second approach by using the ECLS-K ARS scores (teachers' subjective ratings of students' math and reading skills) as  $z_{iK}$ . I plan to pursue this in future analyses.

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 (certainly too high, but comparison of the estimates under this assumption indicates the extent to which ignoring measurement error may bias the results) in order to examine the sensitivity of the conclusions to assumptions regarding the reliability of the tests.

### *Eliminating ambiguity due to uncertainty about test metric*

Even in the absence of measurement error, the parameters of Equation [1] depend on the metric in which  $T_{i5}$  is measured. Suppose the true value of  $\delta$  and  $\gamma$  in Equation [1] are -1 and 0, respectively, given some metric  $T_{i5}$ , implying that black students score, on average, one point lower in fifth grade than whites who started kindergarten with the same score in fifth grade, regardless the initial kindergarten score. Now suppose we replace  $T_{i5}$  with  $T'_{i5} = \ln(T_{i5})$ . The coefficient  $\gamma$  will now be negative, because the logarithmic transformation of  $T_{i5}$  shrinks differences at the high end of the metric relative to the low end of the metric. Unless we have an *a priori* reason to believe that a given version of the test metric is interval-scaled, estimates of  $\gamma$  will be uninterpretable, since they are a function of the (arbitrary) choice of a test metric.

I use two approaches to avoid erroneous conclusions based on uncertainty about the interval-nature of the test metric. First, I locally standardize the fifth grade scores and report black-white fifth grade gaps in local standard deviation units. Specifically, I divide the reliability-adjusted Fall kindergarten scores into 25, 50, or 100 quantiles.<sup>10</sup> Within each quantile  $q$ , I compute the “local” mean ( $\mu_{w5q}$ ) and standard deviation ( $\sigma_{w5q}$ ) of the white fifth grade test score distribution (using the ECLS-K panel weights *c1\_6fc0*). I then compute the locally-standardized fifth grade score

---

<sup>10</sup> I also estimate the same models using 25 and 100 quantiles. Results are substantively unchanged.

for each student as

$$t'_{i5} = \frac{(t_{i5} - \mu_{w5qi})}{\sigma_{w5qi}} \quad [4]$$

In each quantile, the fifth grade scores of white students have a mean of 0 and a standard deviation of 1. The fifth grade scores of black  $\gamma$  students are measured in local standard deviations of white students. To estimate  $\gamma$  from Equation X above, I fit regression models (using only the black students in the sample, since the white students' regression line is simply a horizontal line through the origin) of the form

$$t'_{i5} = \delta + \gamma(t_{iK}^*) + \epsilon_i. \quad [5]$$

The parameter  $\hat{\delta}$  from this model is the estimated black-white difference in fifth grade scores for students who have mean scores in the fall of kindergarten, expressed in standard deviations of initially-similar white students' fifth grade test scores. The parameter  $\hat{\gamma}$  is the estimate of how  $\delta$  changes with a one-standard deviation change in true fall kindergarten scores:  $\hat{\gamma} < 0$  indicates that black-white gaps grow faster among initially high-achieving students.

The locally-standardized methods eliminate, in principle, much of the potential bias that is due to ambiguity about the interval-nature of the outcome test metric. A second approach relies on quantile regression methods, and so is, in principle, insensitive to any monotonic transformation of the test metric. Specifically, I divide the reliability-adjusted fall kindergarten test score distribution into deciles; within each decile, I use estimate the median fifth grade score among black students. I then estimate where in the distribution of initially-similar white students' scores the median black student score would fall. This approach provides a readily interpretable description of the relative changes in black-white gaps among students of different initial ability.

#### 4. Results

Table 2 reports estimates from models of the type described in Equation [1], using the reliability-adjusted kindergarten scores  $t_{iK}^*$  in place of  $T_i$ . I estimate math and reading models

separately using two different outcome metrics (the ECLS-K T-scores and the ECLS-K scale scores, for comparison) and using four different reliability assumptions ( $r=0.7, 0.8, 0.9,$  and  $1.0$ ). In each case, I include a quadratic and cubic terms of  $t_{iK}^*$  to capture nonlinearities in the relationship between fifth grade and kindergarten scores. I fit 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).

Table 2 here

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 ( $\delta$  is always negative). The evidence regarding the extent to which the fifth grade gaps differ by fall kindergarten skills, however, is unclear. The coefficient on the interaction term ( $\gamma$ ) is negative in every model, but varies in size and statistical significance across subjects, test metrics, and reliabilities. In math, the estimates are never significantly different from zero, in either the T-score or scale score metric. In reading, the coefficient estimates are roughly stable and marginally significant regardless of the assumed reliability (they range from  $-.096$  to  $-.125$ ,  $p < .10$  in each case) in the theta metric; they vary much more and are generally far from significant in the scale score metric.

Although marginally significant, the magnitude of the interaction term in the reading T-score models 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  $0.33$  standard deviations, while the corresponding gap between students one standard deviation above the mean in kindergarten is  $0.55$  standard deviations. Nonetheless, the sensitivity of the estimates to the choice of metric leaves these results ambiguous.

Given that the results in Table 2 suggest that conclusions regarding the relative growth of achievement gaps may depend on the test metric used, Table 3 reports the gaps computed using the locally standardized fifth grade scores under a variety of specifications. Unfortunately, the results are still somewhat inconsistent between the T-scores and scale scores. In both math and reading, the estimated locally standardized fifth grade gap is larger among students with higher fall kindergarten scores (i.e.,  $\hat{\gamma} < 0$ ), however the slope is not always significantly different than zero. Moreover, the slopes are significantly different from zero and large in magnitude when the scale scores are used, but not when the T-scores are used. The highly skewed distribution of the scale scores in the fall of kindergarten may be partly responsible for this difference.

Table 3 here

In order to assess the sensitivity of the estimates to assumptions about the reliability of the tests, the choice of the number of quantiles used in the standardization, and to the effects of high-leverage outliers, Tables 4-6 report alternate estimates under a range of different specifications. Table 4 excludes a small number of cases that are more than  $\pm 2$  standard deviations from the mean fall kindergarten score because inspection of scatterplots suggests that a few black students with very high or low fall kindergarten scores may be exerting considerable leverage on the slope estimates. When these students are dropped from the models (the number dropped ranges from 26 to 59 black students dropped from the T-score models, and 5 to 17 dropped from the scale score models, out of the 1,057 in the sample), the estimated slopes are more negative in every instance and are almost all significantly different from zero ( $p < .05$ ).

Table 4 here

Table 5 and 6 report the estimated associations between the fifth grade standardized T-score and scale score gaps, respectively, and the fall kindergarten scores over a range of specifications of reliabilities, number of quantiles, and the domain of fall kindergarten scores. In Table 5, the standardized T-score gap slope is never significant in math when the full sample of

students is used. When the sample is limited to students within  $\pm 2$  standard deviations of the mean, however, the slopes are considerably more negative and all are significant ( $p < .05$ ) or marginally significant ( $p < .10$ ). In reading, a similar pattern holds, though the reading gap slopes are generally somewhat steeper in most specifications than the math slopes. When the sample is restricted, the slope is significantly different from zero in all but one of 12 specifications.

Table 5 here

Table 6 reports similar estimates using the scale scores. In general, the gap slopes are larger, and almost always significant, regardless of specification (except in reading using 100 quantiles), when the scale scores are used. The estimates based on the scale scores, however, are much more sensitive to changes in the assumed reliability and the number of quantiles used. I suspect this is because of the highly skewed scale score distribution, which makes the estimated slopes sensitive to students with high scores. Because of this sensitivity, I prefer the models using the T-scores.

Table 6 here

Figure 1 illustrates the estimates from one version of the model (reliability=.80; number of quantiles=50; sample includes only students within  $\pm 2$  standard deviations of the mean score). The figure shows that, on average, black students who enter kindergarten with average math or reading skills have scores more than one-half a standard deviation below their white counterparts who entered kindergarten with the same skills. Among students who enter kindergarten with scores one standard deviation below the mean, the estimated fifth grade gap is slightly smaller—roughly .40 standard deviations—while among those who enter at one standard deviation above the mean, the estimated fifth grade gap is almost twice as large—roughly .75 standard deviations. Initially high achieving black students fall behind their white peers at a rate twice as fast as do initially low-achieving students.

Figure 1 here

Figures 2 and 3 illustrate the same patterns slightly differently. For each decile of the reliability-adjusted fall kindergarten test score distribution, Figure 2 reports the percentile of the white fifth grade test score distribution that corresponds to the median black student's fifth grade test score. It is evident here that the median black student in a given decile falls well behind his or her white peers by fifth grade, and that this gap is largest for students in the higher fall kindergarten deciles.

Figures 2 & 3 here

## **5. Discussion**

After all the foregoing discussion of eliminating bias and assessing the sensitivity of the estimates to different specifications, two key robust findings emerge. First, among students entering kindergarten with the same math and reading skills, black students fall well behind their white peers. Black students who enter with average math and reading skills have, on average fifth grade scores that are half a standard deviation below their white peers, and place at roughly the 20-25<sup>th</sup> percentile of the white distribution.

Second, and of most interest for this paper, the black-white gap in both math and reading appears to grow fastest between students who enter kindergarten with above average math and reading skills. In fact, the gaps grow roughly twice as fast for students who begin school with scores one standard deviation above the mean as for those who begin one standard deviation below the mean.

It is beyond the scope of this paper to determine the causes of these differences. It is possible that schools contribute to this pattern—high achieving black students may encounter less challenging curriculum and instruction, have fewer resources in their schools, and may be subject to different sets of teacher expectations and behaviors than similarly high achieving white students. It is also possible that differences in the home or neighborhood environments of black and white children may contribute to this pattern. It is reasonable to think that high achieving students

depend more on out-of-school experiences and enrichment for their continued learning than lower achieving students (because their skill levels surpass the instructional content of their classrooms). Given the substantial income, wealth, and neighborhood inequality between black and white students, high achieving black students may have fewer such resources to draw on in their homes and neighborhoods than do white students.

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

We assume the following: Black and white students have true cognitive skill in the fall of kindergarten that are described by

$$T_i = \mu_w + (\mu_b - \mu_w)B_i + u_i \quad [A1]$$

where  $B_i$  is an indicator variable taking the value 0 for white students and 1 for black students and  $u_i \sim N(0,1)$  (that is,  $T_i$  is standardized to have a within-group standard deviation of 1). In this formulation,  $\mu_w$  and  $\mu_b$  are the white and black mean true scores; let  $\delta = \mu_b - \mu_w$  indicate the true black-white difference, expressed in within-group standard deviation units.

In practice,  $T_i$  is measured with (an unknown amount of) error by a test score  $t_i$ . Furthermore,  $t_i$  may be scaled by an (unknown) scaling factor  $c$ :

$$t_i = c(T_i + u_i) \quad [A2]$$

where  $e_i \sim N(0, \sigma)$  is the measurement error (scaled in within-group standard deviation units) in the observed score for student  $i$ . The reliability  $r_t$  of  $t$  as a measure of  $T$  is therefore  $r_t = \frac{1}{1+\sigma}$ . Note that  $Var(t_i) = c^2(1 + \sigma) = \frac{c^2}{r_t}$ ; we cannot empirically determine  $c$  unless we know  $\sigma$  or  $r$ , and vice-versa.

Let  $Y_i$  indicate some outcome of interest (fifth grade test scores in our example), measured with error by  $y_i = Y_i + v_i$ ,  $v_i \sim N(0, \nu)$ . We want to estimate the parameters of the model

$$Y_i = \gamma_0 + \gamma_1(T_i) + \gamma_2(B_i) + \gamma_3(B_i T_i) + \epsilon_i \quad [A3]$$

That is, we want to estimate the race-specific relationships between some outcome  $Y_i$  and students' true cognitive skill (where true skill is scaled to have a within-group standard deviation of 1). If we fit the model

$$y_i = \gamma_0 + \gamma_1(t_i) + \gamma_2(B_i) + \gamma_3(B_i t_i) + \epsilon_i \quad [A4]$$

via OLS, however, we will obtain biased estimates:

$$E[\hat{\gamma}_0] = \gamma_0 + [(1 - r_t)\gamma_1\mu_w - \lambda\mu_w] \quad [A5]$$

$$E[\hat{\gamma}_1] = \gamma_1 + \left[ \left( \frac{r_t}{c} - 1 \right) \gamma_1 + \frac{\lambda}{c} \right] \quad [A6]$$

$$E[\hat{\gamma}_2] = \gamma_2 + [(1 - r_t)(\gamma_1 \delta - \gamma_3 \mu_b) - \lambda \delta] \quad [A7]$$

$$E[\hat{\gamma}_3] = \gamma_3 + \left[ \left( \frac{r_t}{c} - 1 \right) \gamma_3 \right] \quad [A8]$$

where  $\lambda = \frac{Cov(e_{it}, v_i)}{1 + \sigma}$ . The terms in brackets on the right-hand side of each of [A5-A8] expressions indicate the expected bias of the each of the estimated  $\gamma$ 's. Note that the absence of measurement error in  $t_i$  (i.e.,  $r_t = 1$ ) also implies  $\lambda = 0$ , so the bias in each term is zero in the absence of measurement error. When  $r < 1$ , however, the bias in each  $\hat{\gamma}$  is, in general, non-zero. The biases arise from three factors: 1)  $r < 1$  (measurement error in score  $t_i$ ); 2)  $\delta \neq 0$  (the two groups have different mean values of  $T$ ); and 3)  $\lambda \neq 0$  (the error in outcome  $y_i$  is correlated with error in  $t_i$  (which occurs, for example if  $y_i$  measures a gain score in  $T_i$ ).<sup>11</sup>

If we know, or assume, the reliability of  $x$ , and if  $\lambda = 0$ , we can obtain unbiased estimates of each of the  $\gamma$ 's by 1) shrinking  $t_i$  toward its group (race) mean; 2) centering it on zero; 3) dividing it by the scaling factor  $\sqrt{rVar(t_i)}$ :

$$t_i^* = \frac{(1-r_t)c(\mu_w + \delta B_i) + r_t t_i - c(\mu_w + \delta \bar{B})}{\sqrt{rVar(t_i)}} \quad [A9]$$

and then 4) regressing  $y_i$  on  $t_i^*$  via OLS. That is, if we fit the model

$$y_i = \gamma_0^* + \gamma_1^*(t_i^*) + \gamma_2^*(B_i) + \gamma_3^*(B_i t_i^*) + \epsilon_i^* \quad [A10]$$

via OLS, we have

$$E[\hat{\gamma}_0^*] = \gamma_0 + \lambda \mu_w \quad [A11]$$

$$E[\hat{\gamma}_1^*] = \gamma_1 + \frac{\lambda}{c} \quad [A12]$$

---

<sup>11</sup> Note that in the case where  $y_i$  measures the change in  $T$  from time 1 to time 2, and where  $T_i$  is the value of  $T$  at time 1, we have

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

which yields:

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

$$E[\hat{\gamma}_2^*] = \gamma_2 + \lambda\delta \quad [\text{A13}]$$

$$E[\hat{\gamma}_3^*] = \gamma_3 \quad [\text{A14}]$$

Under the assumption that  $\lambda = 0$ , then, regressing  $y_i$  on  $t_i^*$  via OLS yields unbiased estimates of the parameters of [A3]. (note that even in the case where  $\lambda = 0$ , we will still obtain an unbiased estimate of  $\gamma_3$  in this case).

## References

- Angrist, Joshua and Kevin Lang. 2004. "Does school integration generate peer effect? Evidence from Boston's Metco Program." *American Economic Review* 94:1613-1634.
- Bollinger, Christopher. 2003. "Measurement error in human capital and the black-white wage gap." *The Review of Economics and Statistics* 85:578-585.
- Boozer, Michael and Stephen Cacciola. 2001. "Inside the Black Box of Project STAR: Estimation of Peer Effects using Experimental Data." Yale University.
- 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., Helen F. Ladd, and Jacob L. Vigdor. 2006. "The academic achievement gap in grades three to eight." National Bureau of Economic Research, Cambridge, MA.
- Cook, Michael D. and William N. Evans. 2000. "Families or schools? Explaining the convergence in white and black academic performance." *Journal of Labor Economics* 18:729-54.
- 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.

- . 2006. "Testing for racial differences in the mental ability of young children." National Bureau of Economic Research.
- 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. 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.
- Hoxby, Caroline M. and Gretchen Weingarth. 2006. "Taking Race Out of the Equation: School Reassignment and the Structure of Peer Effects." *Unpublished manuscript*.
- Jencks, Christopher and Meredith Phillips. 1998. "The Black-White Test Score Gap." Washington D.C.: Brookings Institution Press.
- LoGerfo, Laura, Austin Nichols, and Sean F. Reardon. 2006. "Achievement Gains in Elementary and High School." Urban Institute, Washington, DC.
- 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.
- Page, Lindsay C., Richard J. Murnane, and John B. Willett. 2008. "Understanding Trends in the Black-

- White Achievement Gap: The Importance of Decomposition Methodology." in *Annual Meeting of the American Educational Research Association*. New York, NY.
- 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. 2007. "Thirteen ways of looking at the black-white test score gap."
- 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 Trends in Racial/Ethnic and Socioeconomic Academic Achievement Gaps." in *Handbook of Research in 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.
- Vigdor, Jacob and Thomas Nechyba. 2004. "Peer effects in Elementary School: Learning from 'Apparent' Random Assignment." *Unpublished manuscript*.

Vygotsky, Lev S. 1978. *Mind and society: The development of higher psychological processes*.  
Cambridge, MA: Harvard University Press.

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)
$P_{b>n}$										
( $r=.7$ )	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

Source: Reardon (2007). Standard errors in parentheses. See Reardon (2007) for detailed description of gap measures. N=6,710.

Table 2: 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 Subject:	Math				Reading				
		Test Metric:	T-Score		Scale Score		T-Score		Scale Score	
			Model:	M1(T)	M2(T)	M1(S)	M2(S)	R1(T)	R2(T)	R1(S)
$r=0.70$	Black		-0.277 *** (0.053)	-0.315 *** (0.060)	-0.238 *** (0.060)	-0.301 *** (0.059)	-0.357 *** (0.059)	-0.394 *** (0.059)	-0.348 *** (0.063)	-0.383 *** (0.059)
	Standardized Fall K Score*Black			-0.084 (0.068)		-0.150 + (0.077)		-0.125 + (0.068)		-0.130 + (0.071)
$r=0.80$	Black		-0.360 *** (0.052)	-0.395 *** (0.058)	-0.348 *** (0.058)	-0.385 *** (0.057)	-0.408 *** (0.058)	-0.440 *** (0.058)	-0.416 *** (0.062)	-0.440 *** (0.058)
	Standardized Fall K Score*Black			-0.081 (0.062)		-0.091 (0.070)		-0.113 + (0.062)		-0.094 (0.066)
$r=0.90$	Black		-0.424 *** (0.051)	-0.457 *** (0.056)	-0.434 *** (0.056)	-0.454 *** (0.055)	-0.448 *** (0.058)	-0.477 *** (0.057)	-0.468 *** (0.061)	-0.485 *** (0.058)
	Standardized Fall K Score*Black			-0.078 (0.057)		-0.049 (0.065)		-0.103 + (0.057)		-0.068 (0.061)
$r=1.00$	Black		-0.476 *** (0.051)	-0.507 *** (0.054)	-0.502 *** (0.056)	-0.509 *** (0.054)	-0.481 *** (0.057)	-0.506 *** (0.056)	-0.509 *** (0.061)	-0.521 *** (0.057)
	Standardized Fall K Score*Black			-0.075 (0.053)		-0.019 (0.060)		-0.096 + (0.054)		-0.050 (0.058)

$N=6,683$ . Robust standard errors are in parentheses. †  $p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$

Table 3: 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 Subject:	Math				Reading				
		Test Metric:	T-Score		Scale Score		T-Score		Scale Score	
			Model:	M1(T)	M2(T)	M1(S)	M2(S)	R1(T)	R2(T)	R1(S)
$r=0.70$	Black		-0.356 *** (0.056)	-0.411 *** (0.080)	-0.241 *** (0.060)	-0.486 *** (0.089)	-0.468 *** (0.063)	-0.535 *** (0.071)	-0.449 *** (0.064)	-0.562 *** (0.076)
	Standardized Fall K Score*Black			-0.086 (0.081)		-0.409 *** (0.110)		-0.150 + (0.079)		-0.298 ** (0.100)
$r=0.80$	Black		-0.476 *** (0.057)	-0.545 *** (0.073)	-0.440 *** (0.059)	-0.577 *** (0.082)	-0.503 *** (0.064)	-0.562 *** (0.070)	-0.530 *** (0.066)	-0.633 *** (0.074)
	Standardized Fall K Score*Black			-0.114 (0.076)		-0.244 ** (0.094)		-0.139 * (0.068)		-0.293 ** (0.090)
$r=0.90$	Black		-0.541 *** (0.058)	-0.596 *** (0.073)	-0.534 *** (0.060)	-0.648 *** (0.083)	-0.539 *** (0.063)	-0.592 *** (0.069)	-0.588 *** (0.065)	-0.673 *** (0.072)
	Standardized Fall K Score*Black			-0.096 (0.071)		-0.215 * (0.092)		-0.134 * (0.067)		-0.255 ** (0.087)
$r=1.00$	Black		-0.597 *** (0.059)	-0.657 *** (0.073)	-0.649 *** (0.063)	-0.739 *** (0.084)	-0.588 *** (0.063)	-0.636 *** (0.067)	-0.665 *** (0.067)	-0.737 *** (0.073)
	Standardized Fall K Score*Black			-0.111 (0.070)		-0.179 * (0.090)		-0.129 * (0.064)		-0.227 ** (0.082)

$N=1,057$  black students. Robust standard errors are in parentheses. †  $p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ ; Local standardization based on 50 quantiles of reliability-adjusted Fall kindergarten score.

Table 4: 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 (among sample with fall kindergarten scores within +/- 2 s.d. of mean)

Assumed Reliability	Test Subject:	Math				Reading				
		T-Score		Scale Score		T-Score		Scale Score		
		Model:	M1(T)	M2(T)	M1(S)	M2(S)	R1(T)	R2(T)	R1(S)	R2(S)
$r=0.70$	Black		-0.342 *** (0.057)	-0.438 *** (0.082)	-0.240 *** (0.060)	-0.495 *** (0.092)	-0.450 *** (0.063)	-0.539 *** (0.073)	-0.444 *** (0.064)	-0.563 *** (0.082)
	Standardized Fall K Score*Black			-0.162 + (0.089)		-0.422 *** (0.116)		-0.208 * (0.087)		-0.303 * (0.118)
$r=0.80$	Black		-0.471 *** (0.058)	-0.563 *** (0.075)	-0.440 *** (0.060)	-0.582 *** (0.085)	-0.491 *** (0.065)	-0.562 *** (0.071)	-0.525 *** (0.066)	-0.640 *** (0.078)
	Standardized Fall K Score*Black			-0.171 * (0.086)		-0.250 * (0.098)		-0.182 * (0.078)		-0.309 ** (0.107)
$r=0.90$	Black		-0.536 *** (0.059)	-0.611 *** (0.075)	-0.537 *** (0.060)	-0.673 *** (0.088)	-0.526 *** (0.064)	-0.591 *** (0.071)	-0.583 *** (0.065)	-0.683 *** (0.078)
	Standardized Fall K Score*Black			-0.150 + (0.078)		-0.250 * (0.101)		-0.176 * (0.077)		-0.277 * (0.108)
$r=1.00$	Black		-0.592 *** (0.060)	-0.672 *** (0.076)	-0.652 *** (0.063)	-0.766 *** (0.088)	-0.574 *** (0.064)	-0.635 *** (0.070)	-0.660 *** (0.068)	-0.744 *** (0.079)
	Standardized Fall K Score*Black			-0.175 * (0.080)		-0.219 * (0.098)		-0.171 * (0.075)		-0.244 * (0.103)

*N* ranges from 998 to 1,052 black students across models. Robust standard errors are in parentheses. †  $p < .10$ ; \*  $p < .05$ ; \*\*  $p < .01$ ; \*\*\*  $p < .001$ ; Local standardization based on 50 quantiles of reliability-adjusted Fall kindergarten score.

Table 5: Estimated Association Between Black-White Locally Standardized Fifth Grade T-Score Gap and Standardized Fall Kindergarten Score, by Test Subject, Assumed Reliability, Sample, and Number of Quantiles Used for Standardization

		Full Sample				Sample Within +/-2 s.d. of Mean Fall K Score			
		rel.=.70	rel.=.80	rel.=.90	rel.=1.0	rel.=.70	rel.=.80	rel.=.90	rel.=1.0
#Quantiles									
Math									
	25	-0.053 (0.078)	-0.087 (0.073)	-0.107 (0.070)	-0.111 <sup>+</sup> (0.067)	-0.157 <sup>+</sup> (0.083)	-0.181 <sup>*</sup> (0.079)	-0.180 <sup>*</sup> (0.079)	-0.192 <sup>*</sup> (0.078)
	50	-0.086 (0.081)	-0.114 (0.076)	-0.096 (0.071)	-0.111 (0.070)	-0.162 <sup>+</sup> (0.089)	-0.171 <sup>*</sup> (0.086)	-0.150 <sup>+</sup> (0.078)	-0.175 <sup>*</sup> (0.080)
	100	-0.111 (0.082)	-0.108 (0.076)	-0.109 (0.077)	-0.126 <sup>+</sup> (0.072)	-0.168 <sup>+</sup> (0.095)	-0.179 <sup>*</sup> (0.083)	-0.146 <sup>+</sup> (0.087)	-0.177 <sup>*</sup> (0.084)
	N	1057	1057	1057	1057	1021	1011	1006	998
Reading									
	25	-0.147 <sup>+</sup> (0.076)	-0.125 <sup>+</sup> (0.069)	-0.131 <sup>*</sup> (0.066)	-0.132 <sup>*</sup> (0.062)	-0.201 <sup>*</sup> (0.085)	-0.178 <sup>*</sup> (0.077)	-0.180 <sup>*</sup> (0.075)	-0.175 <sup>*</sup> (0.073)
	50	-0.150 <sup>+</sup> (0.079)	-0.139 <sup>*</sup> (0.068)	-0.134 <sup>*</sup> (0.067)	-0.129 <sup>*</sup> (0.064)	-0.208 <sup>*</sup> (0.087)	-0.182 <sup>*</sup> (0.078)	-0.176 <sup>*</sup> (0.077)	-0.171 <sup>*</sup> (0.075)
	100	-0.105 (0.090)	-0.159 <sup>*</sup> (0.071)	-0.108 (0.073)	-0.135 <sup>*</sup> (0.062)	-0.196 <sup>*</sup> (0.096)	-0.197 <sup>*</sup> (0.082)	-0.134 (0.086)	-0.164 <sup>*</sup> (0.074)
	N	1057	1057	1057	1057	1031	1022	1017	1012

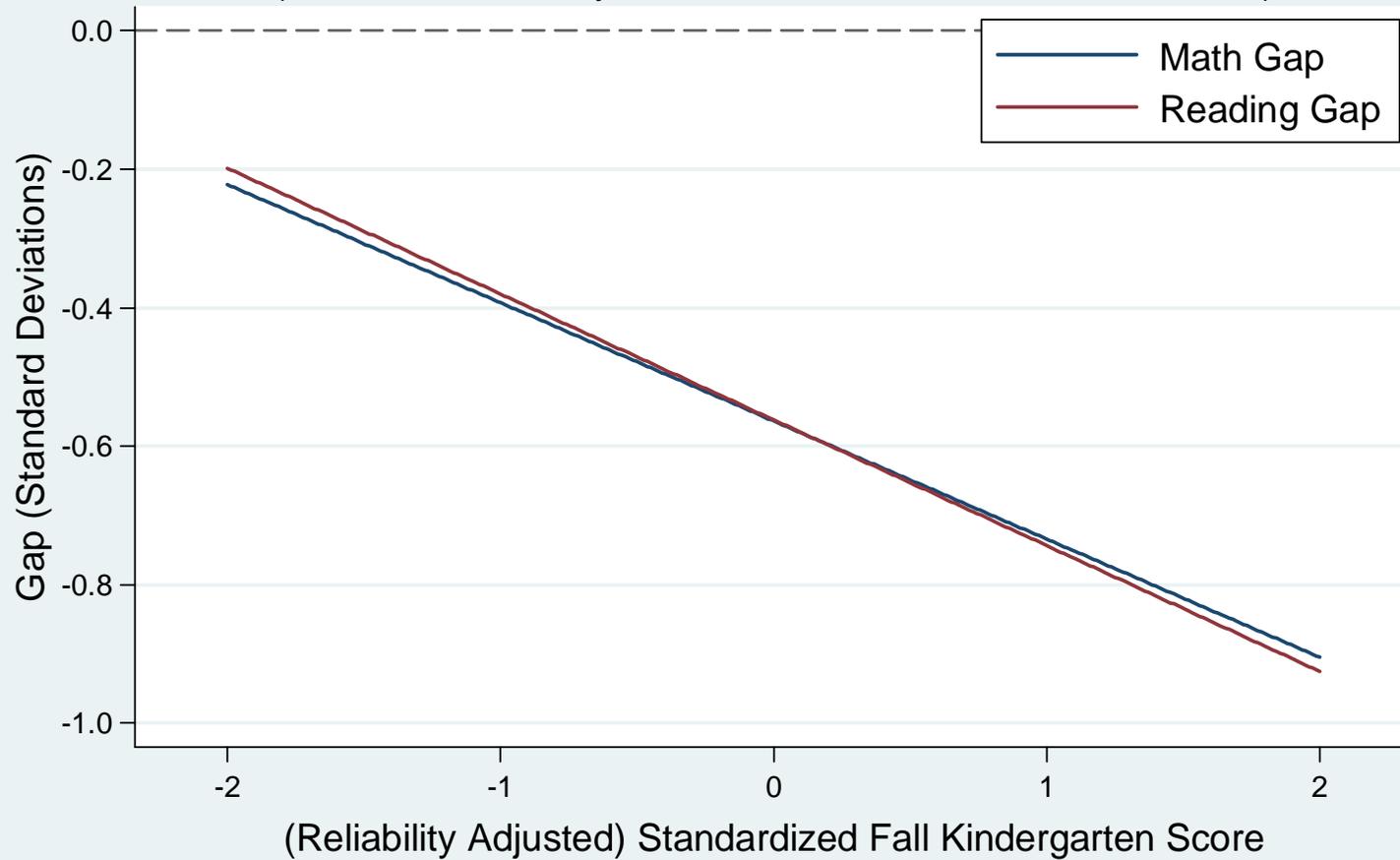
Note: Cells contain estimated gamma coefficient from model [5] (see text). Robust standard errors, corrected for school clustering, in parentheses. <sup>+</sup>  $p < .10$ ; <sup>\*</sup>  $p < .05$ .

Table 6: Estimated Association Between Black-White Locally Standardized Fifth Grade Scale Score Gap and Standardized Fall Kindergarten Score, by Test Subject, Assumed Reliability, Sample, and Number of Quantiles Used for Standardization

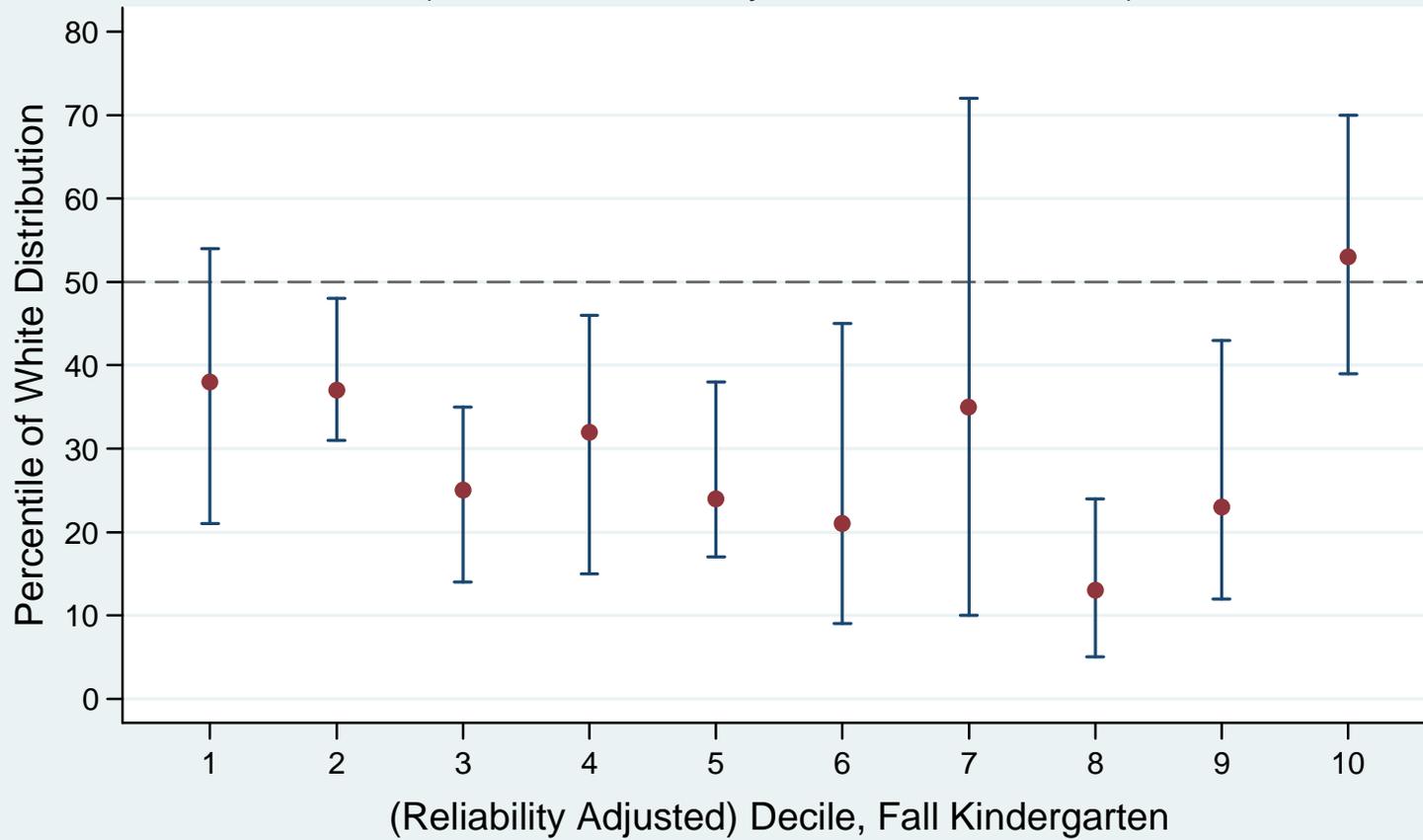
		Full Sample				Sample Within +/-2 s.d. of Mean Fall K Score			
		rel.=.70	rel.=.80	rel.=.90	rel.=1.0	rel.=.70	rel.=.80	rel.=.90	rel.=1.0
#Quantiles									
Math									
	25	-0.291 ** (0.101)	-0.204 * (0.090)	-0.201 * (0.090)	-0.189 * (0.086)	-0.304 ** (0.106)	-0.214 * (0.095)	-0.235 * (0.099)	-0.228 * (0.095)
	50	-0.409 *** (0.110)	-0.244 ** (0.094)	-0.215 * (0.092)	-0.179 * (0.090)	-0.422 *** (0.116)	-0.250 * (0.098)	-0.250 * (0.101)	-0.219 * (0.098)
	100	-0.494 *** (0.116)	-0.244 * (0.111)	-0.251 * (0.101)	-0.227 * (0.095)	-0.505 *** (0.123)	-0.246 * (0.116)	-0.292 ** (0.109)	-0.272 ** (0.102)
	N	1057	1057	1057	1057	1052	1052	1047	1045
Reading									
	25	-0.317 *** (0.094)	-0.274 ** (0.087)	-0.252 ** (0.083)	-0.227 ** (0.079)	-0.337 ** (0.110)	-0.297 ** (0.103)	-0.281 ** (0.102)	-0.249 (0.098) *
	50	-0.298 ** (0.100)	-0.293 ** (0.090)	-0.255 ** (0.087)	-0.227 ** (0.082)	-0.303 * (0.118)	-0.309 ** (0.107)	-0.277 * (0.108)	-0.244 (0.103)
	100	-0.084 (0.131)	-0.234 * (0.099)	-0.181 + (0.098)	-0.195 (0.079)	-0.047 (0.157)	-0.233 + (0.119)	-0.175 (0.124)	-0.185 (0.098)
	N	1057	1057	1057	1057	1047	1044	1042	1040

Note: Cells contain estimated gamma coefficient from model [5] (see text). Robust standard errors, corrected for school clustering, in parentheses. +  $p < .10$ ; \*  $p < .05$ .

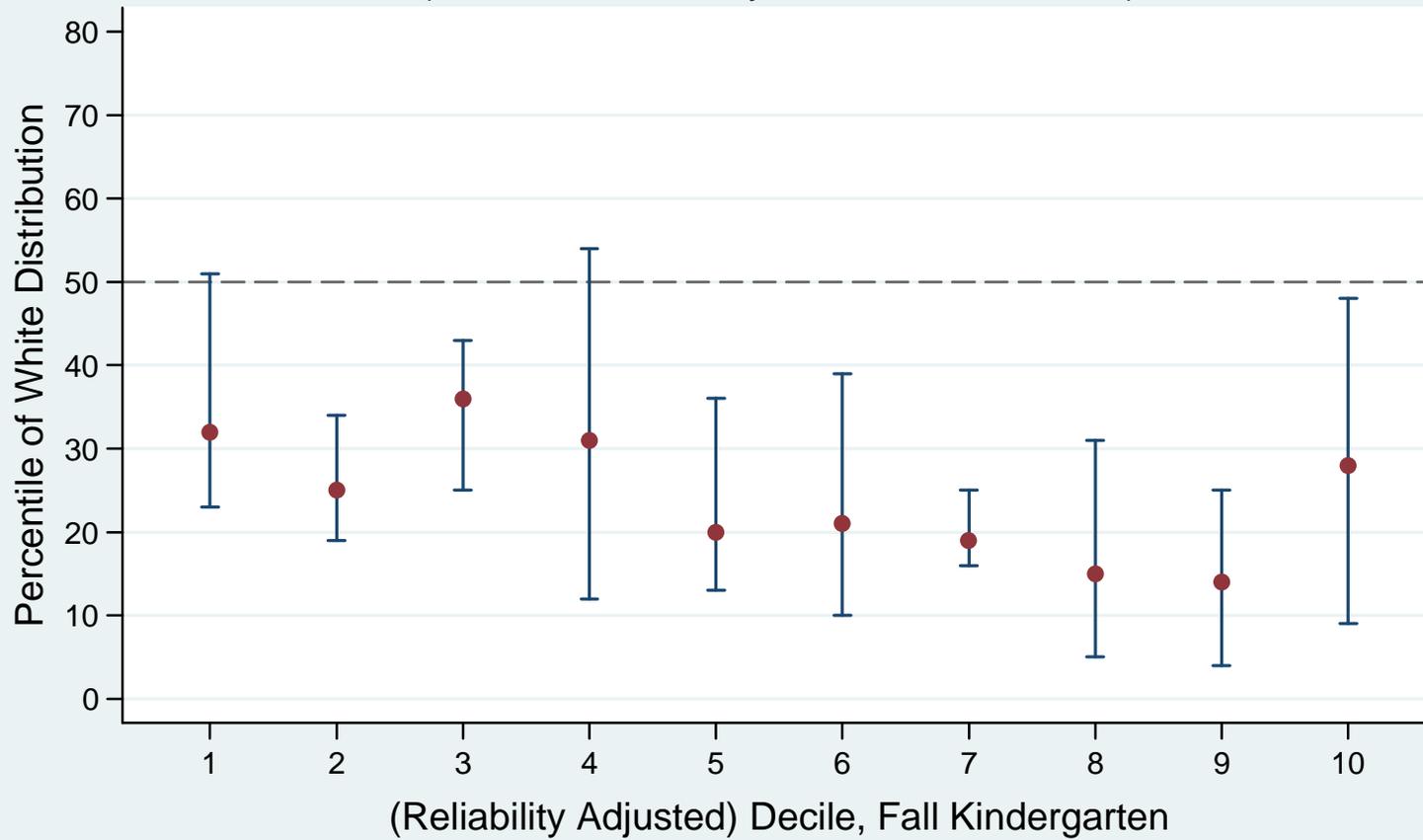
Estimated Standardized Fifth Grade Black-White Gaps,  
by Estimated Fall Kindergarten Standardized Score, Math and Reading  
(Assumed Reliability of Fall K Score = .80; # Quantiles = 50)



Estimated Location of Median Black Student  
in White Fifth Grade Math T-score Distribution  
by Quintile of Fall Kindergarten Test Score  
(Assumed Reliability of Fall K Score = .80)



Estimated Location of Median Black Student  
in White Fifth Grade Reading T-score Distribution  
by Quintile of Fall Kindergarten Test Score  
(Assumed Reliability of Fall K Score = .80)



Institute for Research on Education Policy & Practice

Stanford University  
520 Galvez Mall, 5th Floor  
Stanford, CA 94305-3084

T 650-736-1258  
F 650-723-9931  
[irepp@suse.stanford.edu](mailto:irepp@suse.stanford.edu)  
[www.irepp.net](http://www.irepp.net)