NBER WORKING PAPER SERIES

# THE EVOLUTION OF THE BLACK-WHITE TEST SCORE GAP IN GRADES K-3: THE FRAGILITY OF RESULTS

Timothy N. Bond
Kevin Lang

Working Paper 17960
http://www.nber.org/papers/w17960

NATIONAL BUREAU OF ECONOMIC RESEARCH
1050 Massachusetts Avenue
Cambridge, MA 02138
March 2012

We are grateful to Henry Braun, Sandy Jencks, Dan Koretz, Sunny Ladd, Ed Lazear, Michael Manove, Dick Murnane, Sean Reardon and participants in seminars and workshops at Baruch College, Boston University, Brown, Sciences Po and Stanford for helpful comments and suggestions. The usual caveat applies. The views expressed herein are those of the authors and do not necessarily reflect the views of the National Bureau of Economic Research.

NBER working papers are circulated for discussion and comment purposes. They have not been peer-reviewed or been subject to the review by the NBER Board of Directors that accompanies official NBER publications.

The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results
Timothy N. Bond and Kevin Lang
NBER Working Paper No. 17960
March 2012
JEL No. C18,I24,J15

## ABSTRACT

Although both economists and psychometricians typically treat them as interval scales, test scores are reported using ordinal scales. Using the Early Childhood Longitudinal Study and the Children of the National Longitudinal Survey, we examine the effect of order-preserving scale transformations on the evolution of the black-white reading test score gap from kindergarten entry through third grade. Plausible transformations reverse the growth of the gap in the CNLSY and greatly mitigate it in the ECLS-K during early school years. All growth from entry through first grade and a nontrivial proportion from first to third grade probably reflects scaling decisions.

Timothy N. Bond
Department of Economics
Boston University
270 Bay State Road
Boston, MA 02215
timbond@bu.edu

Kevin Lang
Department of Economics
Boston University
270 Bay State Road
Boston, MA  02215
and NBER
lang@bu.edu

# 1 Introduction

Economists who use test scores in their analyses have largely treated them as interval scales (like temperature). In reality, test scores are measured on ordinal scales (like utils). As with utility functions, any monotonic transformation of the test score scale is also potentially a valid scale. Surprisingly, there has been little attention to this issue among economists although there are some exceptions. Lang (2010) raises concerns about ordinality in the context of value-added measurement. Cascio and Staiger (2011) consider how changes in scaling affect estimates of the fade-out of teacher value-added. In this paper, we show that our conclusion about how the black-white test score gap evolves between kindergarten and third grade is sensitive to our choice of scale. We can find scale choices that show no increase in the gap over this period and choices that double the estimated increase compared with the published scale.

In utility theory, the solution to the absence of an interval scale is to monetize the scale. We calculate how much money the individual would need to be compensated to give up some good or the monetary equivalent of receiving some good such that the individual is indifferent between the money and the good. In contrast, economists have largely ignored the ordinality of test scores.

There are at least three potential responses to this ordinality:

1. Simply accept it and limit ourselves to conclusions that can be reached regardless of choice of scale. Unfortunately, this will often provide us with little insight into important questions. We do not learn much from concluding that the change in the black-white test score gap between kindergarten and third grade is somewhere between 0 and .6 standard deviations. In other cases, this approach may be adequate: the third grade test score gap is between .5 and .7 standard deviations.

2. Assume that we know a great deal about the distribution of the underlying latent variable that test scores measure. If we are confident that "ability" is normally distributed, then we can choose the scale that results in a test-score

distribution that best approximates the normal. We, at least, do not have strong priors about this distribution. Of course, the central limit theorem does explain why many phenomena in the real world have normal distributions. But many economists equate earnings with skill, and earnings are very skewed, and the wealth distribution is even more skewed. It is possible that the ability distribution is similarly skewed or skewed in the opposite direction.

3. Relate the test score to some desired or undesired outcome. If, for example, we care about the black-white test score gap because it translates into an earnings gap, then it makes sense, data permitting, to relate test scores to earnings as in Cunha and Heckman (2008). In general, the children in our sample are too young to permit us to base our scale on earnings, but we do choose scales which maximize the ability of earlier scores to predict performance on later tests. In most cases this approach suggests little growth in the gap between kindergarten and first grade but a significant widening of the gap by third grade. However, one such scale suggests no growth between kindergarten and third grade.

In this paper, we focus on whether scaling decisions affect how the measured black-white test score gap evolves as students progress through school.[1] Our findings should be placed in the context of the debate over when the black-white test score gap emerges and how it evolves during the school years. In their influential and controversial studies, Fryer and Levitt (2004, 2006) challenged the accepted view that a large black-white test score gap emerged in early childhood (Jencks and Phillips, 1998). Fryer and Levitt found that the gap in kindergarten is both modest and largely "explained" by a small number of socioeconomic characteristics. Murnane et al (2006) argue that the different findings reflect the use of different tests. Fryer and Levitt also find that the racial gap widens sharply in the early years of schooling.[2]

[1] Our concerns are in some ways similar to those raised in Koretz and Kim (2007) which focuses on whether there is a difference in the rate that blacks and whites with similar overall performance progress on different skills. They argue that blacks do not fall differentially behind on more advanced skills or catch-up on less advanced ones.

[2] Fryer (2010) finds that, depending on the measure used, the racial test gap in the ECLS-K either continues to expand through eighth grade or remains fairly constant from third through eighth grade.

Hanushek and Rivkin (2006) find a widening gap in Texas while Clotfelter, Ladd and Vigdor (2009) find in North Carolina that gaps widen among high-performing and narrow among low-performing students. Both studies, however, look at somewhat later grades than those used here and in Fryer/Levitt.

While Fryer and Levitt find that the gap at kindergarten entry is mostly or entirely explained by measures of family background, they also find that the increase in the gap is not. The extent to which family background, environmental measures and parental behaviors can explain the test score gap is controversial. This is in part because the influence of such factors varies among data sets[3] and in part because of conceptual issues. Jensen (1969) argues that controlling for such factors is subject to the "sociological fallacy:" family background may include heritable factors. Dickens and Flynn (2001) argue that the environment is endogenous to ability (for example, students who appear to have high cognitive ability may be placed in more challenging classes). Although they use their analysis to explain why environment may be more important than revealed in prior analyses, their argument also casts doubt on the interpretation of regression adjusted test score gaps.

Despite these caveats, we also examine the relation between family background and the test score gap. The inability of family background to explain the growth in the gap is suggestive evidence that schools play a large role in the widening of the gap. It is therefore important to determine whether this conclusion is robust to choice of scale. Most of the scales we derive show similar growth in the adjusted test score gap, but there is one notable exception which reduces the estimated growth in the gap between kindergarten entry and third grade. Perhaps most strikingly although our scales provide quite different estimates of the *unadjusted* gaps at entry and in third grade, there is almost no difference in the *adjusted* gaps at entry and only modest variation in the adjusted gaps in third grade. Thus the scales lead to very different conclusions about the importance of socioeconomic factors in "accounting for" the racial test score gap. In the next section, we show numerical examples of how scaling decisions can be important in interpreting the test gap. We then describe the data

[3] See the summary in Rouse, Brooks-Gunn and McLanahan (2005) and the analysis in Duncan and Magnuson (2005).

used for this study (section three) and present our approach (section four). Finally, we give our results in section five and then provide some concluding remarks.

## 2 Scaling Issues

Suppose that we have a very good test that is able to determine whether an individual has mastered each of three progressively difficult skills. We assume that the skills are cumulative either because skills are simply learned in this order or because skill 2 cannot be mastered before skill 1 (two-digit addition requires one-digit addition) and that there is no partial mastery. Such a test would produce scores of $a$ (no skills mastered), $b$ (only skill 1 mastered), $c$ (skills 1 and 2 mastered) and $d$ (all three skills mastered).

It might seem natural to assign the values 0, 1, 2 and 3 to these scores since these values correspond to the number of skills the individual has mastered. But there is no reason that the marginal value of all three skills should be equal. Skill 1 might be the ability to recite the alphabet, 2 the ability to recognize letters and 3 to read. Or skill 1 might be the ability to read and write English, skill 2 the ability to read and write Latin and skill 3 the ability to converse fluently in Latin. In the latter case, as economists, we are inclined to view the marginal value of 1 as much greater than that of 2 which is in turn much greater than 3, but there are surely other admissible scales.

Suppose we have a sample of twenty blacks and twenty whites. In each case, two people receive scores of $a$ and two scores of $d$. Overall, however, blacks do worse than whites. Of the remaining 16 blacks, 14 get a $b$ and two get a $c$, while among whites the figures are two $b$ and fourteen $c$.

If we use the naive scale, $0, 1, 2, 3$, then the difference in the means is .6 or about .73 standard deviations.

The theoretical lower limit is reached by making the gap between $b$ and $c$ arbitrarily small or, equivalently, sending $a$ and $d$ to minus and plus infinity. In this case, there is no gap since the total number of blacks and whites with a score of either $b$

or $c$ is identical.

At the other extreme, if we treat $a$ and $b$ and $c$ and $d$ as essentially equal, then we get a gap of 1.20 standard deviations. Without some external reference for determining the proper scale, all we can say is that the test-score gap is somewhere between 0 and 1.20 standard deviations. Of course, in some cases such an external standard may exist. Cunha, Heckman and Schennach (2010) tie a variety of objective and subjective measures of cognitive and noncognitive ability to later adult outcomes. More generally, test scores can be tied to other measures of performance.

The situation becomes, if anything, more difficult when we attempt to determine whether the gap increases or decreases as children progress through school. We suppose again that our testing situation is ideal. We have two vertically linked tests so that a score on one test is fully equivalent to that same score on the other test. Suppose further that on administering the second test, we observe that within each race/performance level, half of the individuals advanced exactly one level so that there is now one black and one white at each of levels $a$ and $e$.

If we follow the "natural" scale and assign consecutive integers to the levels so that an $e$ corresponds to a scaled score of 4, then the test-score gap remains at .6, but declines to .62 within-grade standard deviations because the within-grade variance has increased. In contrast, suppose that we believe that over the two tests, the score distribution should be approximately normal. Then we would assign scaled scores of $-1.89, -.75, .27, 1.27, 2.34$. The scaled-score gap would be virtually unchanged at roughly .61 in both periods, but because the variance of the scaled scores would have declined, the gap as a proportion of the standard deviation would grow from .61 to .72. This is similar to the situation documented in Murnane et al (2006) where the IRT gap remains constant but the gap as a proportion of the standard deviation grows.

The scale that minimizes the level of growth in this example is to set the scores for $b$, $c$, $d$ to be approximately equal. Any scale that sets $b$ and $c$ equal creates a gap of 0 on the first test, while any scale that sets $b$, $c$, and $d$ equal creates a gap of 0 on the second test. Alternatively, we can send $a$ and $e$ to minus and plus infinity, which would affect the variance of both tests while leaving the un-normalized levels of the gap unchanged. With both gaps equal to 0, there is no growth in the gap between the two tests. The scale that maximizes the level of growth sets $a, b$, and $c$ and $d$ and $e$ as approximately equal. Again, with $b$ and $c$ being approximately equal, we have no gap on the first test, however this arrangement gives the upper bound on the second test of 1.39. All we can say, therefore, is that the growth of the racial test gap between the two tests is somewhere between 0 and 1.39 standard deviations.

A frequently proposed solution is to use measures based on percentile ranks as opposed to test scores since percentiles are invariant to scale. Nevertheless, it should be recognized that they, too, are a monotonic transformation of the scale, one in which the value placed between ranks is constant across the distribution.

The most prominent percentile-based measure is the percentile-percentile (PP) curve.[4] This method plots the percentile associated with a given score for one group (typically the lower performer) against the percentile associated with that score for the other. If the PP curve does not cross the 45 degree line, the scores of one group are lower than the other in the sense of stochastic dominance. If one PP curve lies above another, this approach suggests that the gap is smaller for the comparison captured by the higher curve. Likewise, if the PP curves are identical, so, it would appear, is the gap.[5]

Unfortunately, as the following example shows, the conclusion from analysis of shifts (or lack thereof) of the PP curve can be misleading. Consider a test with 5 scores which, to emphasize the absence of an interval scale, we denote $a, b, c, d$ and $e$. When the test is first administered, the three white children initially score $a, c$, and $e$ while two black children score $a$ and one black child scores $b$. One year later the white children score $b, d$, and $e$, while two black children score $b$ and one black child scores $c$. It is easy to verify that the PP curve is unchanged. In our example, one white child improved from $a$ to $b$ and one from $c$ to $d$. In contrast, two black children improved from $a$ to $b$ and one from $b$ to $c$. The test gap is unchanged *only if* the value of improvement from $c$ to $d$ for one child equals the value of improvement

[4] The earliest reference appears to be Wilk and Gnanadesikan (1968). For examples of test gap measures based on the PP curve, see Braun (1988), Holland (2002) and Ho and Haertel (2006).

[5] See, for example, Ho (2009).

from $a$ to $b$ for one child *plus* the value of improvement from $b$ to $c$ for a second child. This is neither obviously false nor obviously true and cannot be resolved without some reference to the "correct" underlying scale.

It is clear from these examples that scaling issues can be of great importance in theory. We now proceed to explore to what extent they are in practice.

# 3 Data

We use two data sets: the Children of the National Longitudinal Survey of Youth (CNLSY) and the Early Childhood Longitudinal Study Kindergarten Class of 1998-1999 (ECLS-K). The principal advantage of the CNLSY is that it features two separate tests. The first, the Peabody Picture Vocabulary Test, was administered before school entry and is similar to those on which there are early test score gaps. The second, the Peabody Individual Achievement Test, served as part of the basis for the test administered in the ECLS-K, the test on which the early gap is much more modest. This helps us examine directly the importance of test differences for the conflicting findings in the literature. On the other hand, unlike the CNLSY, the ECLS-K sample is nationally representative, and all students take each of the tests administered and in the same grade. Moreover, it is the data set used by Fryer and Levitt.

## 3.1 Children of the National Longitudinal Survey of Youth

The CNLSY is a biennial survey of children of women in the National Longitudinal Survey of Youth 1979 cohort (NLSY79). The NLSY79 is a longitudinal survey that has followed a sample of 12,686 youths who were between the ages of 14 and 21 as of December 1978. The survey includes a nationally representative sample, as well as an oversample of blacks, Hispanics, military personnel, and poor whites, the latter two being dropped from the later surveys.

Beginning in 1986, the children of women surveyed in the NLSY79 were surveyed and assessed biennially. The assessments included a battery of tests of psychological, socioemotional, and cognitive ability, in addition to questions on the environment in which the child was raised. Children exit the sample at age 15, and enter a separate sample of young adults. As of 2008, a total of 11,495 children born to 4,929 unique female respondents had been surveyed.

Due to the way the sample was created, it is not nationally representative. Children born before 1982, when mothers in the NLSY79 were seventeen to twenty-five years old, will only be partially included in the sample as they were over four years old in the first survey year. Children born before 1972 will not be included at all although there should be very few such children. Children who were adopted into and out of the families of the mothers are not sampled. Children born after 1994 will only be partially observed in the sample and will thus be underrepresented.

Our sample consists of children from age three or four through third grade or roughly age nine and so underrepresents children of older mothers since children born after 1998, when the mothers would have been thirty-four through forty-one, will not have reached third grade. It also underrepresents children born before 1982, when the mothers were seventeen to twenty-five, since such children would be older than four in 1986.

Our focus is on the Peabody Individual Achievement Test (PIAT) Reading: Recognition and Comprehension subtests and on the Peabody Picture Vocabulary Test (PPVT). The Peabody Picture Vocabulary Test is a test of receptive vocabulary that is, according to the CNLSY User's Guide, designed to provide a quick estimate of scholastic aptitude. The User's Guide reports that the PPVT was administered when children were four or five and again when they were ten or eleven. It appears to us that, in fact, the earlier administration occurred between the ages of 36 and 60 months. In order to avoid measuring differences in human capital that could be caused by differences in kindergarten quality, when we examine young children we limit the analysis to children who took the exam at less than four years of age. Further limiting our sample to only black and white youths, we have a total of 1,655 observations of test scores taken between the ages of 3 and 4.[6] Each score is unique

[6] We dropped one observation that reported being in the third grade at age 3, who had a PPVT score well above all of the other 3-year-old scores.

to the individual; there are no repeat exam takers. In our sample, 1,072 children are white, and the remaining 583 are black.

The data show a racial test gap within this sample. On average, black children perform .97 standard deviations worse on the PPVT than do white children, based on the official scale. The highest score is 77 and was attained by one child, while two children scored 0, the lower bound.

Although not tied to a particular curriculum, the PIAT is designed to measure the types of skills typically taught in school. It covers a sufficiently wide range of material that the scores are not subject to boundary effects at the top although this is somewhat of a concern at the bottom. The PIAT was administered at each survey to all children age 5-14. Because the survey is conducted in alternate years, we typically observe a child in kindergarten and second grade or in first and third grade but not both.

Table 1 shows descriptive statistics for our sample of PIAT test scores. Although children typically take the test in only two grades, sample size is fairly consistent across the grades that we analyze, both in terms of total observations and the proportion of test-takers who are black. Since the testing material remains the same, scores rise steadily over time, from an average score of 17 in kindergarten to an average score of 39 by third grade. In each grade, the average score is higher among whites than among blacks, and this difference rises as children progress through school. The standard deviation of the test scores also rises. While there is at least one child who scores a 0 in each grade, in the later grades these children are severe outliers. In the third grade, for instance, the three lowest scores are 0, 2, and 3, while the fourth lowest score is 15. The highest score rises as children reach higher grades, the highest score in the sample is 81 achieved by a white child in third grade.

The gap between blacks' and whites' PIAT scores is quite modest in kindergarten, but expands over the first four years of school. Blacks initially perform .25 standard deviations worse than whites do on the PIAT but by third grade have fallen .61 standard deviations behind. These results are in line with those in Fryer and Levitt (2006) and our own from the ECLS-K which we describe in the next subsection. On the other hand, the *pre-kindergarten* gap on the PPVT is almost a full standard deviation, in line with the results in Jencks and Phillips (1998). These two findings are suggestive of the result in Murnane et al (2006) that the difference between the results in Fryer and Levitt and the prior literature reflects the differences in the tests.

## 3.2 Early Childhood Longitudinal Study

The ECLS-K is a nationally representative longitudinal survey that follows children who entered kindergarten in the 1998-1999 school year. Information was collected in the fall and spring of kindergarten, and the springs of first, third, fifth, and eighth grades.[7]

The children were surveyed and assessed on a variety of different dimensions, such as school experience, motor skill development, height, weight, and direct cognitive assessments of reading and mathematical skill. In each survey year, the student's parents and teacher were interviewed about the child's background, home, and school environment. Like Fryer and Levitt, we use the direct cognitive assessments as our measure of achievement. The tests were designed to measure the student's ability in reading, mathematics and general knowledge or science.[8] The material covered on the test remained the same through first grade, but was modified in later years to reflect the growing knowledge that should be gained in school. Children were first given a short "routing test" that directed them to a more comprehensive exam, the difficulty of which depended on their answers to the routing test. Overall scores are calculated using Item Response Theory (IRT), which the User's Manual states "uses the pattern of right, wrong, and omitted responses to the items actually administered in an assessment and the difficulty, discriminating ability, and 'guess-ability' of each item to place each child on a continuous ability scale." All scores are updated at each interview to expand the range to account for improved performance with age, but the revised scale is then applied to all tests. In principle, a 112 on the kindergarten entry test represents the same level of accomplishment as a 112 on the third grade

[7] An additional subsample includes a set of children who were initially interviewed in the fall of their first grade. These children are excluded from both our and Fryer and Levitt's analysis, since they do not have kindergarten test scores.

[8] Beginning in the third grade, the general knowledge test was replaced by a science test.

test.[9] For our analysis, we will focus only on the evolution of the test score gap through third grade but in some cases also draw on the fifth grade data to scale the earlier scores. Therefore, we use the scores that were released with the 5th grade data file.

Fryer and Levitt (2004, 2006) used the ECLS-K to study the evolution of the black-white test gap. We mimic their sample construction methods to make our results comparable. We focus on the reading scores because they show the most striking growth in the early years in the Fryer/Levitt study. We drop all students who are missing a valid reading score from kindergarten through third grade, and drop all students who do not have a valid entry for race. We also use the sampling weights associated with grades kindergarten through three for child assessment studies, and drop all children who do not have a valid set of these weights. For much of the analysis we use only the test score and race data, but in one table we control for sociodemographic characteristics.

Table 2 shows descriptive statistics for our ECLS-K sample. We have 11,414 observations of whom 62 percent are white and 17 percent are black. The IRT test score scales show a modest (.4 standard deviations) test-score gap at the beginning of kindergarten, rising steadily to a gap of three-quarters of a standard deviation towards the end of third grade. The second column of Table 2 shows the corresponding figures from Fryer and Levitt. Although our sample is somewhat larger with a higher proportion of whites and blacks than theirs, the test-score gap evolves in very similar ways in the two samples.

It is important to recognize that there is only a modest amount of overlap in the entry and third grade scores of the ECLS-K. About 95 percent of students received scores on the entry test that were below the lowest score on the third grade test. Still the remaining 5 percent scored better than at least some third graders and two students entering kindergarten scored above the third grade mean using the original test score scale.

[9] The scores are supposed to be an estimate of the number of questions the test taker would have answered correctly had she taken the entire test, rather than just the section to which she was routed.

# 4 Methods

We define the test score gap at a given grade or age as the difference between the mean test scores of whites and blacks divided by the standard deviation of test scores in that grade or at that age.

We begin by searching for the monotonic transformations of the original scale that maximize and minimize the growth of this gap. We impose the transformation

$$T(t+1) = T(t) + a_{t+1}^2 \tag{1}$$

where $t$ is the original scale, $T$ is the transformed scale and $\alpha_{t+1}$ is a real number. Since the gap is unchanged by a linear transformation, we must normalize two of the parameters. We set $T\left(0\right)$ equal to 0 and $T\left(t_{\max}\right)$ equal to $t_{\max}$ where $t_{\max}$ is the highest score observed in that grade.[10] Define $G_g$ to be the test gap in grade $g$.

$$G_g = \frac{N_w^{-1} \sum\limits_{i \in white} T\left(t_{ig}\right) - N_b^{-1} \sum\limits_{i \in black} T\left(t_{ig}\right)}{\sqrt{N^{-1} \sum \left(T\left(t_{ig}\right) - N^{-1} \sum T\left(t_{ig}\right)\right)^2}} \tag{2}$$

where $G_g$ is the gap in grade $g$, $N_w$, $N_b$ and $N$ are the sizes of the white, black and total sample. We choose the remaining values of $a$ using Newton-Raphson to minimize the objective function given by

$$D_{\min} = \min_a \left(G_3 - G_e\right) \tag{3}$$

where $D$ is the difference between the test gap in grade 3 and the test gap in kindergarten, and $a$ refers to the vector of coefficients. We define $D_{\max}$ similarly for the maximum. In practice, not all of the possible scores are observed each year in the data. We normalize $a_{t+1}$ to 0 if no member of the sample in that grade is observed to have an initial test score of $t+1$.

[10] In practice, our program sometimes converged faster (or only) when we normalized the two lowest scores, and then transformed the data afterwards to range from 0 to $t_{\max}$.

This nonparametric approach is useful for finding the bounds on the gap, but it produces scales that are typically step functions with one or two steps and likely implausible. Additionally it cannot be used when the test score is a continuous variable, as the ECLS-K assessment approximately is. Therefore, we focus on transformations that are both monotonic and smooth. We look at the path of the gap that can be formed by varying parameters in a sixth degree polynomial

$$T(t) = \beta_0 + \beta_1(t-c) + \beta_2(t-c)^2 + \beta_3(t-c)^3 + \beta_4(t-c)^4 + \beta_5(t-c)^5 + \beta_6(t-c)^6 \quad (4)$$

where $\beta_0 - \beta_6$ and $c$ are constants. This type of function is very flexible and can be used to approximate a wide array of continuous functions. This transformation, however, is not guaranteed to be monotonic. Our algorithm checks for monotonicity and rejects attempts to choose parameters that violate this condition. Needless to say, not all monotonic functions will be well approximated by even a monotonic six-degree polynomial. We therefore cannot rule out the possibility that some other transformation could generate results outside the range we present here.

Again, $G$ is unchanged by linear transformations. When showing the density of the test scores, we normalize the standard deviation of test scores to equal 1 and choose $\beta_0$ so that the mean of the test score distribution is 0. Note that the test score distribution is not required to be symmetric so that the median need not be 0. However, it is easiest to show the transformations on a scale similar to the one used for the original test scores. Therefore when showing the relation between the two scales, we fix the highest and lowest scores to be equal across scales.[11]

If the test score distributions on entry and in third grade were disjoint, then (subject to a minor caveat about the ability of a six-degree polynomial to simultaneously approximate two different distributions), we would find $D_{\max}$ by minimizing the test-score gap at entry and maximizing it in third grade. Conversely, to find $D_{\min}$ we would maximize $G_e$ and minimize $G_{3.}$

In practice, because the two test score distributions overlap, we cannot do the

[11] In practice, it was easier to do the estimation by setting the constant term to 0 and constraining the linear term and only subsequently transforming the estimated coefficients.

maximizations and minimizations separately.[12] Nevertheless, because there is not much overlap, the process of selecting the transformations comes close to mimicking this approach.

As we will see, in both data sets, the implications of $D_{\min}$ and $D_{\max}$ are very different. In the former case, the black-white gap is trivial when children first enter school but grows to be substantial by the end of third grade. In contrast, in the latter case, the black-white gap in the ECLS-K is modest but not trivial when children enter school and changes little over the next four years. In the CNLSY the gap under $D_{\min}$ actually shrinks.

These bounds are not very helpful. Therefore, to help us select among the possible transformations, including less extreme ones, we choose the transformations that have the most predictive power for future test scores. For the CNLSY, we maximize the correlation between the PPVT at age 3 and the PIAT reading test administered during school. For the ECLS-K, we maximize the correlation between the entry and third grade tests.

# 5 Results

## 5.1 Maximizing and Minimizing the Growth of the Gap

The first column of Table 3 shows the evolution of the black-white test gap in the PIAT, using the original scale provided with the exam. The gap shows a large increase over the first four years of education, beginning at a modest .25 standard deviations in kindergarten and rising to .61 standard deviations by third grade.

We can find the boundaries of the evolution of this gap by assigning a new set of monotonically increasing test scores chosen to either maximize or minimize the difference between the third grade and kindergarten test gap. Under the growth-minimizing scale, the black-white test gap *shrinks* by .18 standard deviations during

[12] It is not entirely obvious that we should treat the difference between getting exactly the first six and the first five questions right as identical regardless of when the student took the test, but we impose this assumption.

the first four years of education. Column (2) of Table 3 shows the evolution under this minimizing transformation. The test gap in kindergarten is similar to that of the baseline at .24 standard deviations. In contrast with the baseline, the gap immediately begins to decline to .17 standard deviations in first grade and .07 standard deviations in second grade. The gap remains roughly constant in third grade, ending at .06 standard deviations.

The evolution under the growth-maximizing scale is shown in the third column of Table 3. This transformation reduces the gap at kindergarten substantially to just .12 standard deviations. After kindergarten the evolution is similar to that in the baseline model, with blacks performing .64 standard deviations worse than whites in third grade, only slightly worse than they perform using the baseline scale. With this transformation, the gap grows by .52 standard deviations over the first four years of school.

The two extreme transformations produce test scales that differ noticeably from the baseline scale. The transformed scales are essentially step functions, with scores that are almost constant within tiers separated by large jumps. Though this may not be intuitively appealing, it is not unlike tests which have “proficiency” cutoffs. Suppose for instance that kindergartners only differed in their possession a few meaningful skills such as the ability to recognize letters, the ability to recognize words, and the ability to read for comprehension. Then this could be an appropriate scale to use at that grade. In fact, the PIAT reading test is designed somewhat like this. Students must pass a reading recognition test in order to advance to questions on a reading comprehension test. The modal score in both our kindergarten and first grade sample is 18, which is the highest score a student could achieve without advancing to the reading comprehension section.

Turning our attention to the ECLS-K, because the IRT scoring method produces an essentially continuous variable, we use a sixth-degree monotonic polynomial transformation on the entire IRT scale. This means that we apply the same transformation to each grade. Table 4 shows how the achievement gap on the ECLS-K reading assessment evolves from the beginning of kindergarten through the spring of third grade. The first column repeats the baseline pattern from Table 2. The second column shows the choice of transformation that minimizes the estimated growth in the gap. At kindergarten entry the gap is .46, only slightly higher than in the baseline. As discussed above, the scale that minimizes the growth in the test score gap should comes close to maximizing the entry gap. Thus it appears that the scale used in the ECLS-K comes close to maximizing that gap. The minimum possible growth in the gap is quite small. Using this scale, in third grade the gap is only .51 and thus noticeably less than the gap in the baseline. And the growth between entry and third grade is only .05. Note that, in principle, minimizing growth between entry and third grade could still generate large swings in the first grade gap. However, this does not occur. There is no noticeable change in the gap between any pair of tests when this scale is applied.

Column (3) of Table 4 shows the results of choosing the transformation that maximizes the growth of the gap between kindergarten and third grade. The transformed gap at the beginning of kindergarten is now only .11 standard deviations, which is .29 less than in the baseline. The transformed gap increases by .10 standard deviations to .21 between the fall and spring kindergarten tests and then rises a further .22 standard deviations by the spring of first grade so that the estimated gaps are similar to the baseline for the first and third grades. The end result is a growth of .62 standard deviations in the racial test gap in the first four years of education, almost twice that using the baseline scale. Note that the gap at the end of third grade is almost unchanged from the baseline, suggesting that the baseline scale comes close to maximizing the black-white gap at this stage.

Figure 1 shows the density function of test scores associated with each choice of scale for the ECLS-K at kindergarten entry. Note that in the baseline case, it is skewed with a long right tail. In contrast, visually, the resulting test score distribution from the minimizing transformation more closely approximates a normal distribution. The density associated with the maximizing transformation is somewhat aesthetically displeasing and possibly unattractive on other grounds. Most of the weight of this distribution is in a narrow band around its mode, and there are no scores substantially below this mode. Nevertheless, we do not find this representation of the scores altogether counterintuitive. It is plausible that most children do not have much in the way of reading, math and general knowledge skills and that the modest differences over much of the range are uninformative. On the other hand, there are a small number, best represented by the two who are already operating solidly at the third grade level, who are truly distinct from the rest of the pack. Moreover, in some respects the density of the growth-maximizing transformation is more aesthetically pleasing than the income or wealth distribution in the United States. It is less skewed than either. The 50-10 spread (measured in standard deviations) is plausibly larger than it is in the wealth distribution.[13]

How do these transformations affect the test score distributions in third grade? As previously noted, the transformation that minimizes the growth in the gap will be close to the one that minimizes the third grade gap while the choice of $T(t)$ that maximizes the growth of the gap produces a third-grade gap very close to the one in the baseline. Figure 2 shows the density of the test score distribution for the baseline scale and the two transformations. As in the case of the kindergarten scores, the key to minimizing the third grade gap, and thus growth, is compressing the middle of the distribution so that most students appear quite similar and spreading out the differences among very high and among very low scores. In contrast, the growth-maximizing transformation leaves the distribution of test scores looking similar to that associated with the baseline.

As already discussed, we should not necessarily dismiss distributions that primarily distinguish the very high and very low performers from everyone else. While the large spike at the mode when using the growth-minimizing transformation initially appears problematic, the implied distribution is not obviously more implausible than the U.S. earnings, income and wealth distributions. However, it is perhaps more problematic that the growth-minimizing transformation requires this large spike to appear between school entry and third grade.

The relation between the original and transformed scales is shown in figure 3. We can see that the growth in the test score gap is minimized if we believe that differences in very low scores (roughly 15 to 40) and very high scores (roughly those

[13] This is based on our imputation from Kennickell's (2009) calculations based on the 1989-2007 Survey of Consumer Finances.

over 140) are very informative but those in between are relatively uninformative. The transformation that maximizes the growth of the test score gap does the opposite, at least at the bottom of the scale. It treats most differences among the very low scores as uninformative. This would be appropriate if we believed that most children arrive in kindergarten knowing very little of the material covered by the ECLS and that throughout most of the distribution differences in performance should be viewed as relatively unimportant and that only children with very high scores should be viewed as differing substantially from the mass of kindergarten entrants.

The results in this subsection bring out the fragility of any conclusion about the extent to which the test score gap increases between school entry and the end of third grade. The bounds permit conclusions ranging from “there is essentially no gap when students begin school and a very sizeable gap by the end of third grade” through “there are modest gaps at entry and at the end of the third grade and essentially no growth in the gap over this period.” There are even scales for the PIAT from which one could conclude "black children moderately lag behind white children in achievement when they enter school, but overtake them by third grade." As is often the case with bounding exercises, the range of possible results is too large to be helpful.

It is thus evident that determining the right scale is important in determining how the gap between blacks and whites evolves. We could attempt to choose scales that produce “aesthetically appealing” distributions of test scores, but this is unsatisfactory. There is no consensus on what the distribution of childhood ability should look like. Well accepted childhood tests, including the PIAT and the ECLS-K assessments, produce widely varying distributions of achievement. And as discussed before, there are reasons to think that unintuitive distributions of ability could be plausible, both for young children and adults. In the next subsection we consider a more formal approach to choosing the appropriate transformation.

## 5.2 Selecting Transformations

We would not expect kindergarten or first grade performance to perfectly predict third grade performance. There is randomness in performance on each test. Moreover, students make varying academic progress. Indeed the point of the current exercise is to ask whether blacks and whites progress academically at different rates during the first four years of school.

Nevertheless, tests measure related skills. Students who perform well on one test would generally be expected to perform well on the other tests. A reasonable criterion for selecting a transformation is to ask which transformation allows us to best predict future performance using information from previous tests. We therefore choose transformations which maximize the correlation between test scores. If the tests measure a common underlying latent variable, this approach maximizes reliability. If not, it merely maximizes the ability of an earlier test to predict performance on a later test.

For the CNLSY we have access to scores on an early childhood cognitive achievement test, the PPVT. We construct a sample of children who took both the PPVT before age 4 and the PIAT while in kindergarten. This sample consists of 398 white and 253 black children. The racial test gaps in this subsample are very similar to those of the full sample. Blacks perform .97 standard deviations worse on average on the PPVT than whites, and .2 standard deviations worse than whites on the PIAT. The correlation between the untransformed test scores is .32.

We use monotonic sixth degree polynomial transformations to find the set of scales which maximizes the correlation between individuals' PPVT and PIAT test scores. The resulting scales increase the correlation between these two tests only moderately, to .35 and do not noticeably alter the racial test gaps. The test gap on the PIAT falls to .24 and that on the PPVT increases to .98.

In Figure 4, we plot the correlation-maximizing transformations over the range of our sample, normalizing each scale to have the same range as the baseline. The transformed PPVT is similar to the original except that it compresses the highest scores. Interestingly, the PIAT transformation magnifies differences among the highest test scores, while compressing the scores somewhat below the highest. While this suggests that a high PIAT test score may be a more important predictor of performance than a high PPVT test score, it is important to remember that inferences on the highest range are based on only a few observations.

Column (4) of Table 3 shows the evolution of the PIAT test gap under this scale. Surprisingly given the modest transformation, the pattern differs substantially from the baseline. The kindergarten gaps are similar, but the gap drops to .19 in third grade for a *decrease* from kindergarten through third grade of approximately .05 standard deviations. One caveat for this result is that roughly 18% of the third grade sample scored above the highest kindergarten score. This is much less of a problem for the first and second grade tests. Yet, the gap using the transformed scale is essentially constant from kindergarten through second grade while it grows substantially from year to year using the original scale.

In the ECLS-K we do not have data on test scores outside of the cognitive assessments. We therefore maximize the correlation across the reading assessments. First we examine the transformation that maximizes the correlation between the tests taken at the beginning of kindergarten and the spring of third grade. This new scale substantially increases the correlation between the two scores. The correlation when using the transformation is .62 ($R^2 = .39$) compared with only .54 ($R^2 = .29$) using the baseline scores. This approach produces a kindergarten gap that is very close to the potential maximum gap at kindergarten entry and a third grade gap that is very close to the maximum gap at this point.

As noted earlier, the scale that maximizes the third grade gap is similar to the baseline scale and the one that maximizes the entry gap is only moderately different from the baseline. Therefore, the overall pattern of the racial test gap using the correlation-maximizing transformation does not differ dramatically from the baseline. As shown in column (4) of Table 4, the total growth in the gap from the beginning of kindergarten through the end of third grade is .26 standard deviations, .09 smaller than the growth seen in the baseline. All of the growth of the gap occurs between the end of first and the end of third grade. This differs from the story told by the baseline ECLS-K of a steady increase in the gap throughout the first four years of schooling.

We additionally choose the scale that maximizes the $R^2$ from a regression of the third grade score on the scores the student received on the first grade and two kindergarten tests, as well as from a regression of the fifth grade score on the third, first, and kindergarten tests. Both these approaches yield similar results to those of column 4.

## 5.3 Controlling for Socioeconomic Factors

One of the surprising results in Fryer and Levitt is that when students first enter kindergarten, the modest black-white test score gap can be accounted for fully by a small number of socioeconomic characteristics (children's age, child's birth weight, a socioeconomic status measure, WIC participation, mother's age at first birth, and number of children's books in the home). In this subsection we ask whether the same is true for the scales developed in the previous subsections. Of course, when there is no gap at entry, these characteristics cannot account for the gap, but it is possible that they could reverse it.

Table 5 shows the results of this exercise. Strikingly the kindergarten entry results are robust to the choice of scale. Regardless of whether the scale shows an unadjusted gap of .11 or .47, after controlling for this small number of factors, the remaining gap is actually reversed and favors blacks by between .03 and .05 standard deviations. In contrast, the importance of the controls in third grade depends on the choice of scale. Three of the four scales generate unadjusted test score gaps of approximately .75 standard deviations. After controlling for the socioeconomic factors, the gap falls to about .3 standard deviations but still indicates a very substantial deterioration in the relative performance of black children over the first three years of school. In contrast, the transformation that minimizes the growth of the unadjusted gap shows a noticeably more modest adjusted gap of .17. In this case two-thirds of the unadjusted gap is accounted for by the measured characteristics, a somewhat larger proportion than the little over half accounted for when the other scales are used.

Thus the choice of scale has a significant impact on the magnitude of the increase of the adjusted gap as well as of the unadjusted gap. We note that we have not chosen the scales on the basis of the adjusted gaps. We have, however, done some experimentation that suggests that maximizing and minimizing the adjusted gaps would not significantly alter the results.

Another surprising result in Fryer and Levitt is that the growth of the black-white test gap is virtually unaffected by whether or not socioeconomic controls are used. Both the controlled and uncontrolled test gaps grow by a similar magnitude between entry and the end of the third grade. We have already shown that this appears to be an artifact of the scale. Much of the growth in the gap under the maximizing transformation can be explained by socioeconomic controls. While the raw gap increases by .64 standard deviations over the first four years of education under this transformation, the controlled gap increases by only .35 standard deviations. Under the minimizing transformation, the socioeconomic controls actually have negative explanatory power. While the raw gap under this transformation grows by only .05 standard deviations, the adjusted gap increases by .2 standard deviations.

We further analyze the robustness of this result in Table 6. In the first two columns, we find the transformation that maximizes the percentage of the growth in the raw test gap that can be explained by the socioeconomic controls. That is, we minimize the ratio of the magnitude of the growth in the controlled gap over the magnitude of the growth in the uncontrolled gap. In this transformation, the controls explain only slightly more than in the growth-maximizing transformation. The raw test gap grows by .59 standard deviations from kindergarten through third grade, while the controlled gap grows by only .32. In columns 3 and 4, we instead maximize the difference between the growth of the raw test gap and the growth of the controlled gap. The pattern under this transformation looks similar to that of the maximizing transformation as well.

## 5.4 Scale Sensitivity

To some extent, the choice of scale limits the potential magnitude of between-group differences. An example may clarify this point. Suppose a group of researchers is interested in understanding early racial differences in reading. They administer a test to a group of 100 children, 50 of whom are black and 50 of whom are white. The performance of the children can be strictly ranked. They then give the results to two psychometricians with instructions to scale the results. The first reports that in this group the black-white test score gap is almost exactly two standard deviations. The second reports that it is about 1.1 standard deviations.

Further investigation reveals that both psychometricians believe that scales should reflect developmental milestones and that differences in performance on a given side of the milestone are insignificant. They also agree that the milestone is passed when children shift from "learning to read" to "reading to learn," but they differ about what test performance corresponds to this shift.

The first psychometrician set the milestone at a point for which half of the original sample was judged to be reading to learn. In contrast, the second psychometrician believes that only the 25 children with the highest scores merit this designation. Note that because each psychometrician uses a scale with only two points, their calculations are invariant to the issues we have addressed heretofore.

In our fictional example, we have allocated the fifty lowest scores to the "blacks" and the highest fifty scores to the "whites." In both cases, the reported test gaps are the largest consistent with the scales and distributions "chosen" by the psychometricians. Thus both gaps are at their maxima, but when the scale sets equal numbers of 0s and 1s,  the gap can be bigger than it can be when one one-fourth of the students receive 1s.

The lower half of the scores in the ECLS-K kindergarten test are all clustered within one standard deviation of the median. It is possible that this characteristic of the test and its scaling affects the potential for a large test score gap in kindergarten

To analyze the effect that such clumping may have on the evolution of the racial test gap in the ECLS-K, we calculate what the test gap would be in each grade if blacks had all the lowest scores and whites all the highest. Denoting $w_j$ as the weighted number of children with score $t_j$, where $j$ is the rank (from low to high) of the score, and $W_B$ is the weighted number of blacks in the sample, we create a set of weighted test scores $B = \{t_j | j \in [1, m]\}$ where $m$ solves the problem $\sum_{i=1}^{m} w_i = W_B$. We, likewise, assign all the highest scores to whites, based on their weighted proportion of the sample.[14]

The first column of Table 7 shows the weighted test gaps along with the hypothetical upper boundary for the gap on the ECLS-K assessments. While the observed black-white test gap increases over time, so does the boundary for that gap. The theoretical maximum gap based on the distribution at the beginning of kindergarten is 1.5 standard deviations. This rises to 2.2 standard deviations by the end of third grade. The result is that the observed racial test-gap, as measured as a percentage of the possible test gap, hardly changes over time. At the beginning of kindergarten, the achievement gap is 27% of the maximum possible achievement gap, given the scale, while the gap is 33% of the maximum gap at the end of third grade. This raises the concern that part of the large observed increase in the racial achievement gap in the ECLS-K may be attributable to changes in scale and test sensitivity, as opposed to changes in the real achievement gap.

The remaining columns of Table 7 look at the boundary of the test gap for our previously discussed transformations. Columns (2) and (3) show the maximum test gap for the transformations that simply try to minimize or maximize the growth of the gap in the first four years of education. Interestingly, these transformations have the opposite effect on the growth of the gap relative to the maximum test gap. The minimizing transformation yields a test gap 24% the size of the maximum gap at kindergarten entry, but one that is 53% at the end of third grade despite virtually no growth in the size of the test gap in terms of standard deviations over this period. Likewise, in the maximizing transformation the gap as a percentage of the maximum gap shrinks from 46% at the start of kindergarten to 35% at the end of third grade despite a nearly 700% increase in the size of the gap in terms of standard deviations over that same period. Our transformations appear to act mainly by changing the potential sensitivity of the scale to the racial test gap. The test gap at third grade can be no larger than .95 standard deviations under the

[14] The remaining middle scores are implicitly assigned to Hispanics, Asians, and others, though we do not look at their hypothetical test gaps in this situation.

minimizing transformation, compared to 2.23 standard deviations in the baseline. The maximizing transformations can have a gap no larger than .24 at kindergarten, which is not only lower than the maximum in the baseline of 1.5, but lower also than the actual observed test gap in the baseline of .4 standard deviations. Columns (4) looks at the boundaries for the test gap under the transformation that maximizes the correlation across tests. Strikingly, the maximum gap is almost identical at kindergarten entry and third grade. The increase in the estimated gap as a proportion of the maximum gap therefore reflects changes in the former rather than the latter. Recall, however, that the increase in the estimated gap with this scale is smaller than with the base scale.

Rather than look at the boundaries of the test score gap, an alternative approach is to look at what the gap would look like if the test scale remained constant. Denote $F_g(r)$ as the function that maps a child's performance rank to a test score. Panel A of Table 8 shows the evolution of the test gap if $F_g(r)$ did not vary with $g$. That is, we choose an initial grade and then take a child's rank on each grade's exam and reassign to him or her the score given to the child who was at that rank on the initially chosen exam.[15] Both when we impose the fall kindergarten and spring third grade mapping, we see virtually no growth in the test score gap until the third grade test, but substantial growth in the test gap at third grade.

Panel B instead supposes that we fix $r$ while varying $F_g$ (i.e. changing the scales across grades as they do *de facto* in the ECLS-K.) Even if the rank ordering of students did not change, we would still see growth in the test gap using the baseline scales in the ECLS-K. If the rank order of children remained what it was in kindergarten throughout the first four years, we would observe a .09 standard deviation increase in the test gap from entry to third grade due simply to changes in the spacing between ranks over time. Likewise, using the third grade rank order we would observe a .13 standard deviation increase in the test gap. Most of the increase in

[15] Since we are using weights, we cannot map rank in one grade directly to rank in the other grade. Instead we view rank as a continuum and look at masses at each score. This results in some children receiving a weighted average of two consecutive scores. The results are sensitive to the way in which ties at scores are broken, but since there are very few ties given the quasi-continuous nature of the scoring system, this sensitivity is only beyond the fourth decimal point.

this gap occurs between the spring first grade and spring third grade test. Using the entry rankings, this increase during this time span is .05 standard deviations, and using the third grade rankings it is .07.

Table 8 strongly suggests that the growth in the test gap from kindergarten through first grade reflects scales and not achievement. Moreover, a significant portion of the growth from first to third grade also reflects scaling decisions. Taken together, tables 7 and 8 suggest that when we use the base scale, something on the order of 8 to 13 percentage points of the growth in the gap between entry and third grade reflects scale sensitivity.

# 6 Summary and Conclusion

Our findings suggest that we should exercise great caution when using test scores to determine when a black-white test score gap first emerges and whether it widens in the early school years. By choosing the scale appropriately, we can make the initial gap in the ECLS-K, at kindergarten entry, in reading anywhere from a trivial one-ninth of a standard deviation to almost half a standard deviation. Similarly, the third grade gap varies between half and three-quarters of a standard deviation. Equally significantly, whether the gap widens after school entry depends on our choice of scale. Some scales show a decrease in the test gap in the CNLSY.

We use similar methods to choose appropriate scales but find strikingly different results across data sets. In the ECLS-K, the gap at kindergarten entry is somewhat but not dramatically larger than suggested by the untransformed scale and the growth in the gap through third grade is correspondingly smaller. Almost all of this growth occurs between spring of the first and third grades. But even this result is suspect because the third grade test is capable of generating a larger gap than are the earlier tests. Given this concern, we do not wish to place excessive emphasis on the finding that the gap widens between first and third grades. In the CNLSY, the gap in kindergarten is virtually identical in the untransformed and our preferred transformed scales. However while the untransformed scale shows a dramatic increase in the test gap over the first four years, our preferred scale actually implies a decline in magnitude by third grade.

We note that it has become something of a mantra in education circles that third grade is when students begin the transition from “learning to read” to “reading to learn.” If the timing implied by our ECLS-K rescaling is correct, this suggests one avenue to pursue in furthering our understanding of the gap. If the pattern in our CNLSY rescaling is correct, pre-enrollment interventions may be more important in reducing the test gap than those that are post-enrollment.

More broadly, our findings suggest that economists and other researchers should be much more circumspect in their use of test scores. While many findings will be robust to scale changes, many will not be.

# References

Bond, Timothy N. and Kevin Lang, “Test Choice, Scale Choice and the Black-White Test Score Gap,” unpublished, 2011.

Braun, Henry I., “A New Approach to Avoiding Problems of Scale in Interpreting Trends in Mental Measurement Data,” *Journal of Educational Measurement*, 25:3 (Autumn 1988): 171-191.

Cunha, Flavio and James J. Heckman, “Formulating, Identifying, and Estimating the Technology of Cognitive and Noncognitive Skill Formation,” *Journal of Human Resources*, 43:4 (Fall 2008): 738-782.

Cunha, Flavio, James J. Heckman, and Susanne M. Schennach, "Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Econometrica*, 78:3 (May 2010): 883-931.

Cascio, Elizabeth U. and Douglas O. Staiger, “Skill, Standardized Tests, and Fadeout in Educational Intervention,” unpublished, 2011.

Clotfelter, Charles T , Helen F Ladd and Jacob L Vigdor, “The Academic Achievement Gap in Grades 3 to 8,” *Review of Economics and Statistics*, 91:2 (May 2009): 398-419.

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

Duncan, Greg J. and Katherine A. Magnuson, “Can Family Socioeconomic Resources Account for Racial and Ethnic Test Score Gaps?” *The Future of Children*, 15:1 (2005): 35–54.

Fryer, Roland G., Jr. “The Importance of Segregation, Discrimination, Peer Dynamics, and Identity in Explaining Trends in the Racial Achievement Gap,” *NBER Working Paper No. 16257*, 2010.

Fryer, Roland G., Jr. and Steven D. Levitt, “Understanding the Black-White Test Score Gap in the First Two Years of School,” *Review of Economics and Statistics*, 86:2 (May 2004): 447-64.

Fryer, Roland G., Jr. and Steven D. Levitt, “The Black-White Test Score Gap Through Third Grade,” *American Law and Economics Review*, 8:2 (2206): 249-81.

Hanushek, Eric A. and Steven G. Rivkin, S. G. "School Quality and the Black–White Achievement Gap," *NBER Working Paper No. 12651*, 2006.

Ho, Andrew D., "A Nonparametric Framework for Comparing Trends and Gaps Across Tests," *Journal of Educational and Behavioral Statistics*, 34 (June 2009): 201-228.

Ho, Andrew D. and Edward H. Haertel. "Metric-Free Measures of Test Score Trends and Gaps with Policy-Relevant Examples," *CSE Report No. 665*, 2006

Holland, Paul W. "Two Measures of Change in the Gaps Between the CDFs of Test-Score Distributions," *Journal of Education and Behavioral Statistics*, 27:1 (Spring 2002): 3-17

Jencks, Christopher and Meredith Phillips, "The Black-White Test Score Gap: An Introduction," in Jencks, Christopher and Meredith Phillips, eds. *The Black-White Test Score Gap*, Washington, DC: Brookings Institution Press, 1998.

Jensen, Arthur R., 1969, "How Much Can We Boost IQ and Scholastic Achievement?" *Harvard Educational Review*, 39(1): 1-123.

Kennickell, Arthur B. "Ponds and streams: Wealth and Income in the U.S., 1989 to 2007," *Finance and Economics Discussion Series 2009-13*, Federal Reserve Board, 2009.

Koretz, Daniel and Young-Suk Kim,"Changes in the Black-White Test Score Gap in the Elementary School Grades," 2007.

Lang, Kevin, "Measurement Matters: Perspectives on Education Policy from an Economist and School Board Member," *Journal of Economic Perspectives*, 24:3 (Summer 2010): 167-181.

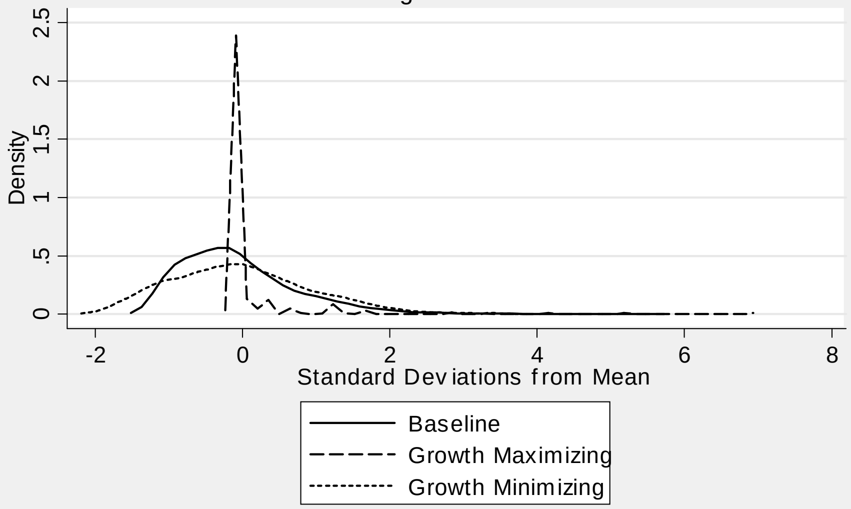
Murnane, Richard J., John B. Willett, Kristen L. Bub and Kathleen McCartney, "Understanding Trends in the Black-White Achievement Gaps during the First Years of School," *Brookings-Wharton Papers on Urban Affairs*, (2006): 97-135.

Rouse, Cecilia E., Jeanne Brooks-Gunn and Sara McLanahan, "Introducing the Issue, School Readiness: Closing Racial and Ethnic Gaps," *The Future of Children*, 15:1 (2005): 5-14.

Wilk, M. B. and R. Gnanadesikan, "Probability Plotting Methods for the Analysis of Data," *Biometrika* , 55 (March 1968): 1-17.

Figure 1

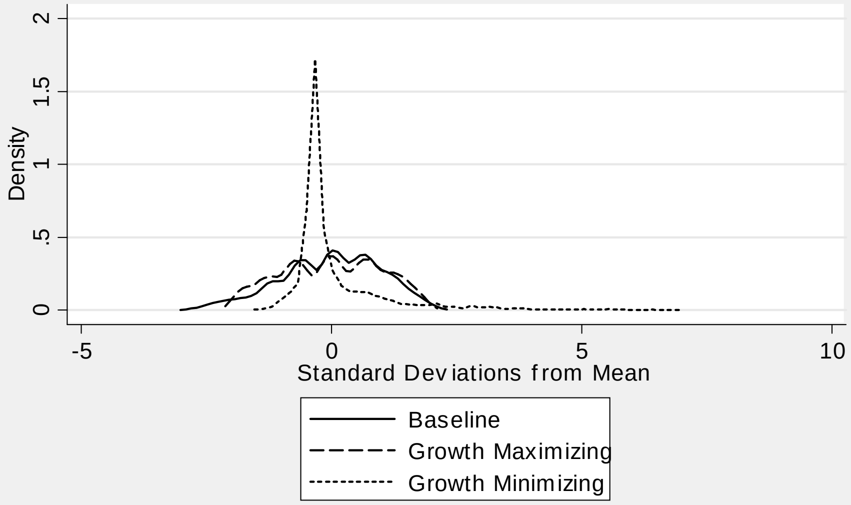
Kindergarten Densities



Outlying values beyond seven standard deviations above the mean are not displayed

Figure 2

Third Grade Densities



Outlying values beyond seven standard deviations above the mean are not displayed

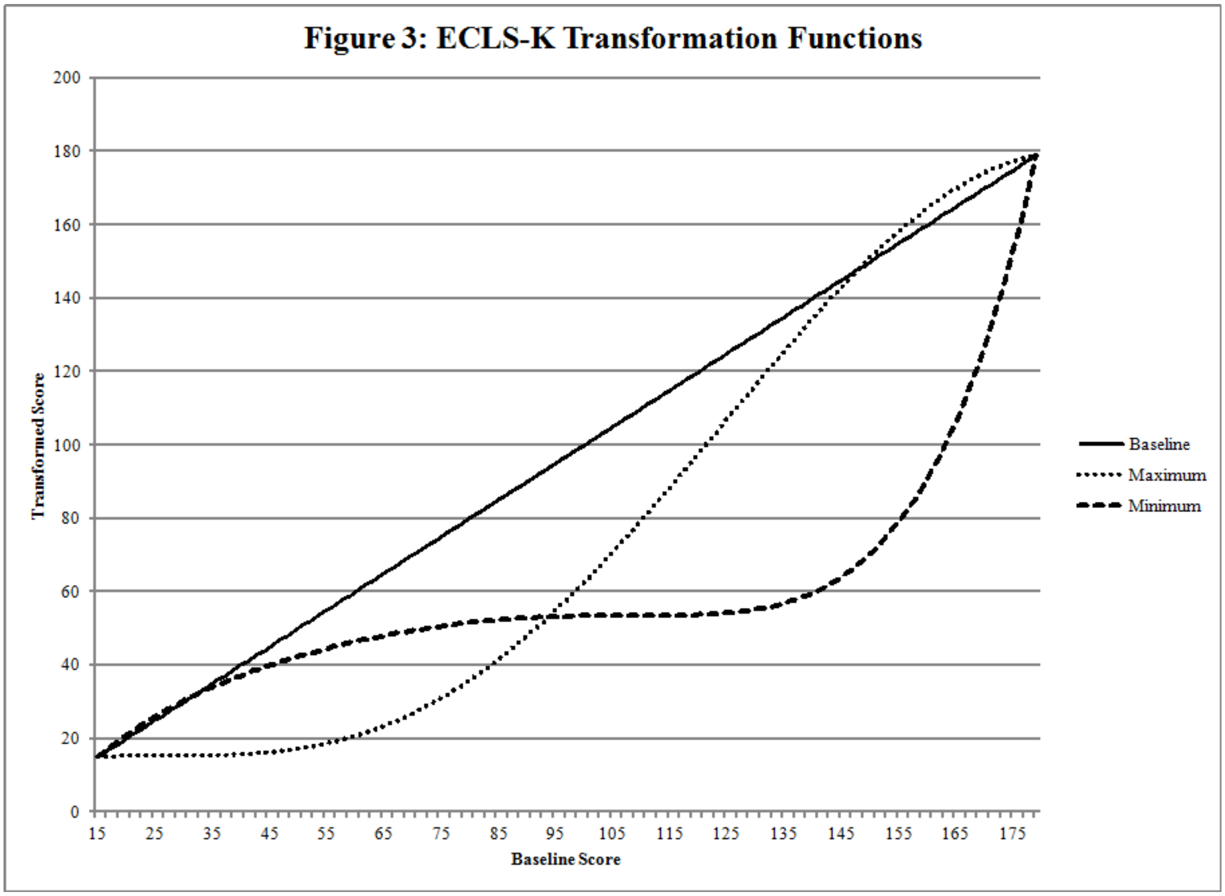


Figure 3: ECLS-K Transformation Functions

**Figure 4: Correlation Maximizing Transformations for CNLSY**

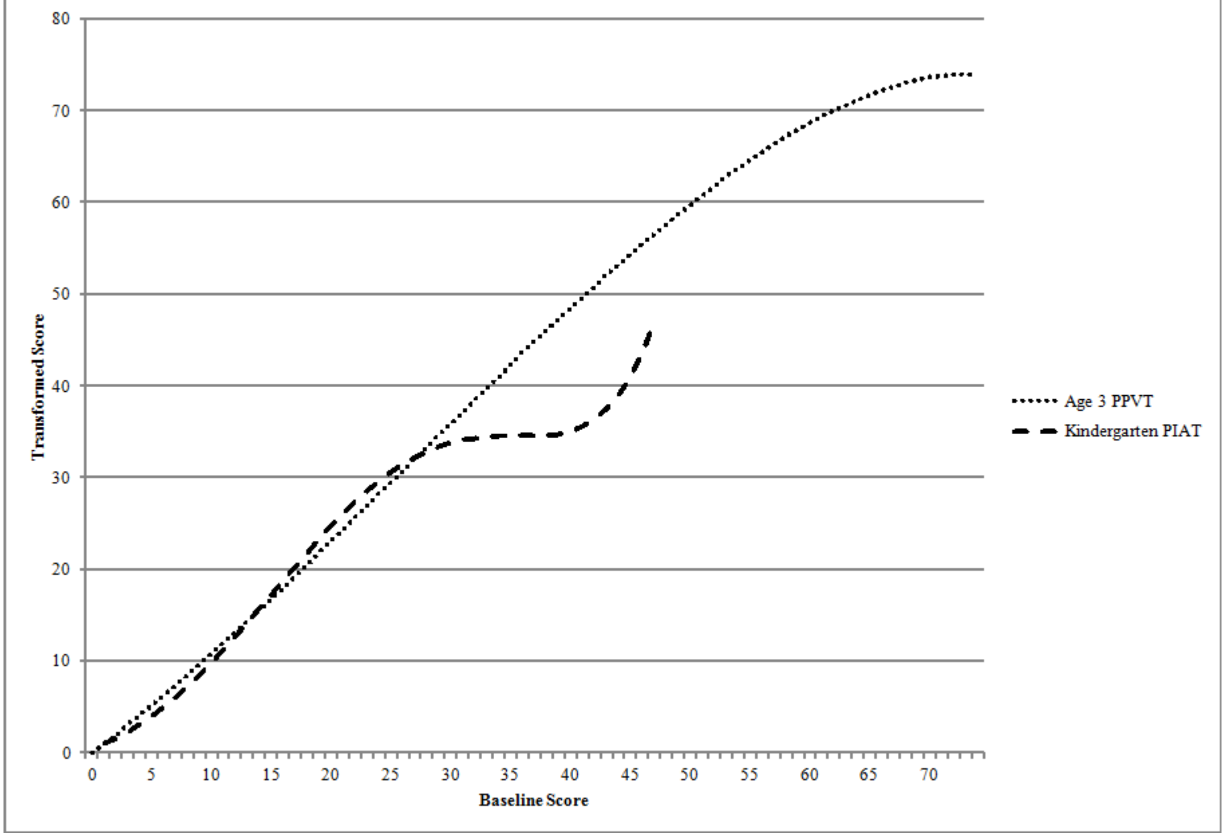


Table 1: CNLSY Descriptive Statistics

| | Total | Black | White |
|---|---|---|---|
| Kindergarten | | | |
| Mean | 17.06 | 16.21 | 17.55 |
| | (5.33) | (4.95) | (5.48) |
| Min | 0 | 0 | 0 |
| Max | 56 | 46 | 56 |
| N | 2943 | 1081 | 1862 |
| First Grade | | | |
| Mean | 24.96 | 22.89 | 26.24 |
| | (7.94) | (6.41) | (8.50) |
| Min | 0 | 0 | 0 |
| Max | 64 | 50 | 64 |
| N | 2761 | 1055 | 1706 |
| Second Grade | | | |
| Mean | 32.89 | 29.47 | 35.15 |
| | (9.73) | (8.57) | (9.79) |
| Min | 0 | 8 | 0 |
| Max | 68 | 59 | 68 |
| N | 2822 | 1125 | 1697 |
| Third Grade | | | |
| Mean | 38.64 | 35.04 | 40.98 |
| | (9.71) | (9.33) | (9.24) |
| Min | 0 | 0 | 2 |
| Max | 81 | 68 | 81 |
| N | 2833 | 1057 | 1716 |

Source: Children of the National Longitudinal Survey of Youth. Standard deviations in parenthesis.

Table 2: ECLS-K Descriptive Statistics

| | Bond and Lang | Fryer and Levitt |
|---|---|---|
| Race | | |
| White | 0.62 | 0.55 |
| | (0.49) | (0.50) |
| Black | 0.17 | 0.15 |
| | (0.37) | (0.36) |
| Hispanic | 0.14 | 0.18 |
| | (0.35) | (0.38) |
| Asian | 0.02 | 0.07 |
| | (0.15) | (0.25) |
| Female | 0.49 | 0.49 |
| | (0.50) | (0.50) |
| Black-White Test Gap | | |
| Kindergarten Fall | 0.40 | 0.40 |
| | (0.03) | (0.03) |
| Kindergarten Spring | 0.44 | 0.45 |
| | (0.03) | (0.03) |
| First Grade Spring | 0.49 | 0.52 |
| | (0.03) | (0.03) |
| Third Grade Spring | 0.75 | 0.77 |
| | (0.04) | (0.03) |
| Sociodemographic Controls | | |
| Age (in months) fall Kindergarten | 68.5 | 67.0 |
| SES composite measure | 0.022 | 0.005 |
| Number of children's books in home | 76.8 | 61.4 |
| Mother's age at first birth | 23.6 | 23.6 |
| Child's birth weight (in ounces) | 118.1 | 87.5 |
| WIC participant | 0.42 | 0.38 |
| Observations | 11414 | 10540 |

Source: Early Childhood Longitudinal Study Kindergarten Class of 1998-1999. Standard deviations are in paranthesis for variables. Test gaps measured in standard deviations and standard errors are in parenthesis.

Table 3: Evolution of the black-white test gap under various transformations of the PIAT

| | Baseline<br>(1) | Minimum<br>(2) | Maximum<br>(3) | Corr Max<br>(4) |
|---|---|---|---|---|
| Kindergarten | 0.25*** | 0.24*** | 0.12*** | 0.24*** |
| | (0.03) | (0.04) | (0.03) | (0.04) |
| First Grade | 0.42*** | 0.17*** | 0.38*** | 0.29*** |
| | (0.04) | (0.04) | (0.03) | (0.04) |
| Second Grade | 0.58*** | 0.07* | 0.57*** | 0.26*** |
| | (0.04) | (0.04) | (0.04) | (0.04) |
| Third Grade | 0.61*** | 0.06 | 0.64*** | 0.19*** |
| | (0.04) | (0.04) | (0.04) | (0.04) |

Gaps are average white score minus average black score on the PIAT-RC. Column 4 represents the transformation that maximizes the correlation between the PIAT-RC at kindergarten and the PPVT at age 3. Standard errors are in paranthesis. *p<.1 **p<.05 ***p<.01

Table 4: Evolution of the black-white test gap under various transformations of the ECLS-K

| | Baseline<br>(1) | Minimum<br>(2) | Maximium<br>(3) | Corr Max<br>(4) |
|---|---|---|---|---|
| Kindergarten - Fall | 0.40*** | 0.46*** | 0.11*** | 0.50*** |
| | (0.03) | (0.04) | (0.02) | (0.04) |
| Kindergarten - Spring | 0.44*** | 0.50*** | 0.21*** | 0.52*** |
| | (0.03) | (0.04) | (0.02) | (0.04) |
| First Grade - Spring | 0.49*** | 0.49*** | 0.43*** | 0.49*** |
| | (0.03) | (0.04) | (0.03) | (0.04) |
| Third Grade - Spring | 0.75*** | 0.51*** | 0.75*** | 0.73*** |
| | (0.04) | (0.02) | (0.03) | (0.03) |

Gaps are average white score minus average black score on the ECLS-K reading assessment. Standard errors are in parenthesis. *p<.1 **p<.05 ***p<.01

Table 5: Evolution of the unexplained black-white test gap under various transformations

| Transformation | Baseline<br>(1) | Minimum<br>(2) | Maximum<br>(3) | Corr Max<br>(4) |
|---|---|---|---|---|
| Kindergarten - Fall | -0.05 | -0.03 | -0.04* | -0.03 |
| | (0.03) | (0.04) | (0.02) | (0.04) |
| Kindergarten - Spring | 0.04 | 0.08** | -0.01 | 0.10** |
| | (0.03) | (0.04) | (0.02) | (0.04) |
| First Grade - Spring | 0.10*** | 0.10** | 0.08** | 0.10** |
| | (0.04) | (0.04) | (0.03) | (0.04) |
| Third Grade - Spring | 0.31*** | 0.17*** | 0.31*** | 0.30*** |
| | (0.04) | (0.03) | (0.04) | (0.04) |

Gaps are the coefficient on a white indicator variable with black as the excluded variable. Each regression controls for SES, number of books in the home, gender, birth weight, indicators for whether the mother was a teenager or over 30 at birth, and WIC recipiency. Standard errors in parenthesis. *p<.1 **p<.05 ***p<.01

Table 6: Scales which maximize the explanatary power of controls

| | Percent Difference | | Raw Difference | |
|---|---|---|---|---|
| | No Controls | Controls | No Controls | Controls |
| | (1) | (2) | (3) | (4) |
| Kindergarten-Fall | 0.07*** | -0.03 | 0.07*** | -0.04 |
| | (0.02) | (0.03) | (0.02) | (0.02) |
| Kindergarten-Spring | 0.14*** | -0.00 | 0.15*** | -0.00 |
| | (0.02) | (0.01) | (0.02) | (0.02) |
| First Grade - Spring | 0.32*** | 0.05** | 0.34*** | 0.06** |
| | (0.02) | (0.03) | (0.02) | (0.03) |
| Third Grade - Spring | 0.66*** | 0.25*** | 0.67*** | 0.26*** |
| | (0.03) | (0.03) | (0.03) | (0.03) |

Gaps are the coefficient on a white indicator variable with black as the excluded variable. Standard errors are in paranthesis. *p<.1 **p<.05 ***p<.01

Table 7: Black-White Test Gap as a Percentage of Boundary Under Various Transformations

| | Baseline<br>(1) | Minimizing<br>(2) | Maximizing<br>(3) | Corr Max<br>(4) |
|---|---|---|---|---|
| Fall-K Black-White Test Gap | 0.40 | 0.46 | 0.11 | 0.47 |
| Fall-K Maximum Test Gap | 1.49 | 1.92 | 0.24 | 1.97 |
| Fall-K % of Maximum Gap | 27.0% | 24.1% | 46.3% | 23.9% |
| Spring-3 Black-White Test Gap | 0.75 | 0.51 | 0.75 | 0.73 |
| Spring-3 Maximum Gap | 2.23 | 0.95 | 2.16 | 1.96 |
| Spring-3 % of Maximum Gap | 33.4% | 53.5% | 34.7% | 37.4% |

Gaps are average white score minus average black score on the ECLS-K reading assessment.

Table 8: Evolution of Black-White Test Gap Under Fixed Distribution

| | Fall-K<br>(1) | Spring-K<br>(2) | Spring-1<br>(3) | Spring-3<br>(4) |
|---|---|---|---|---|
| Panel A: Fixed Scale, Varied Rank | | | | |
| Fall-K Distribution | 0.40 | 0.42 | 0.44 | 0.62 |
| Spring-3 Distribution | 0.49 | 0.53 | 0.51 | 0.75 |
| Panel B: Varied Scale, Fixed Rank | | | | |
| Fall-K Rank | 0.40 | 0.42 | 0.47 | 0.49 |
| Spring-3 Rank | 0.62 | 0.65 | 0.72 | 0.75 |

Gaps are average white score minus average black score on the ECLS-K reading assessment.