

# The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results\*

Timothy N. Bond<sup>†</sup> and Kevin Lang<sup>‡</sup>

Department of Economics

Boston University

270 Bay State Road

Boston, MA 02215

October 8, 2011

## Abstract

Test scores are reported using ordinal scales. Using the Early Childhood Longitudinal Survey, we examine the effect of monotonic transformations of the scale on the evolution of the black-white reading test score gap from kindergarten entry through third grade. These transformations generate changes in the gap that range from effectively zero to twice that in the baseline scale with the most plausible showing modest reductions in growth relative to the original scale. In those scales showing a widening gap between first and third grades, much of the growth may be attributable to the greater potential for differences on the third grade test rather than to an actual change in relative performance.

---

\*We are grateful to participants in the empirical microeconomics workshop at Boston University for helpful comments and suggestions. The usual caveat applies.

<sup>†</sup>Boston University; email: timbond@bu.edu

<sup>‡</sup>Boston University, NBER and IZA; email: 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 suggested in (Carneiro and Heckman, 2003). The children in our sample are too young to permit us to base our scale on earnings, but we do choose the kindergarten and third grade scales to maximize the ability of kindergarten scores to predict third grades scores, and we also choose these scales to maximize our ability to predict fifth grade scores. Both approaches suggest little growth in the gap between kindergarten entry and the spring of first grade but very significant widening of the gap over the next two years.

Like Fryer and Levitt (2004, 2006), we use the Early Childhood Longitudinal Study (ECLS). We do so for two reasons. First, the ECLS is a longitudinal study of a large representative sample. Second, the Fryer/Levitt results are simultaneously very influential and controversial. They differ from earlier studies that found a large black-white test score gap emerged in early childhood (Jencks and Phillips, 1998). In contrast, Fryer and Levitt argue that in kindergarten the gap is both modest and largely “explained” by a small number of socioeconomic characteristics and that a racial gap only emerges in the early years of schooling.<sup>1</sup> To some extent, the different findings seem to reflect the use of different tests (Murnane et al, 2006), but the reason that different tests show different gaps may be related not only to content but also to scaling (Bond and Lang, 2011).

In this paper, we focus on the sensitivity to scaling decision of conclusions about

---

<sup>1</sup>Fryer (2010) finds that, depending on the measure used, the racial test gap either continues to expand through eighth grade or remains fairly constant from third through eighth grade.

how the gap evolves as students progress through school.<sup>2</sup> Hanushek and Rivkin (2006) find a widening gap in Texas while Clotfelter, Ladd and Vigdor (2009) find that 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.

Fryer and Levitt find that the gap at kindergarten entry is mostly or entirely explained by measures of family background but 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<sup>3</sup> 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. 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

---

<sup>2</sup>Our concerns are in some ways similar to those raised in Koretz and Kim (2007) who 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.

<sup>3</sup>See the summary in Rouse, Brooks-Gunn and McLanahan (2005) and the analysis in Duncan and Magnuson (2005).

different conclusions about the importance of socioeconomic factors in “accounting for” the racial test score gap. In the next section, we describe the data used for this study. We then present our approach (section three), followed by our results (section four) and concluding remarks.

## 2 Data

The Early Childhood Longitudinal Survey (ECLS-K) is a nationally representative longitudinal study 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 grade.<sup>4</sup>

The children were surveyed and assessed on a variety of different categories, such as school experience, motor skill development, height, weight, and direct cognitive assessments of reading and mathematical skill. Both the student’s parents and teacher were also interviewed in each survey in order to gather information on 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.<sup>5</sup> 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 takes into account the difficulty of the individual questions that are answered correctly, incorrectly, or left blank.<sup>6</sup> All scores are updated at each interview to expand

---

<sup>4</sup>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.

<sup>5</sup>Beginning in the third grade, the general knowledge test was replaced by a science test.

<sup>6</sup>Specifically, correct answers to difficult questions are discounted if the child answers easy questions in the same field incorrectly. It is assumed that their results on these difficult questions were due to guessing.

the range to account for improved performance with age, but the scale is the same for all tests. In principle, a 112 on the kindergarten entry test represents the same level of accomplishment as a 112 on the third grade test. 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.

To avoid concerns that differences in results reflect different sampling decisions, we construct our sample to mimic that of Fryer and Levitt. We focus on the reading scores because those are the ones that 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 1 shows descriptive statistics for our 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 (0.4 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. For purposes of comparison, the second column in Table 1 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. About 95 percent of students received scores on the entry test that were less than 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.

### 3 Methods

To explore the effects of monotonic transformations of the IRT scale derived for the ECLS, we consider transformations of the form

$$T(t) = \beta_0 + \beta_1(t - c) + \beta_2(t - c)^2 + \beta_3(t - c)^3 + \beta_4(t - c)^4 + \beta_5(t - c)^5 + \beta_6(t - c)^6 \quad (1)$$

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.

We define the test score gap in grade  $g$  as

$$G_g = \max_{\beta, c} \frac{N_w^{-1} \sum_{i \in white} T(t_{ig}) - N_b^{-1} \sum_{i \in black} T(t_{ig})}{\sqrt{N^{-1} \sum \left( T(t_{ig}) - N^{-1} \sum T(t_{ig}) \right)^2}} \quad (2)$$

where  $t_{ig}$  is the test score reported for individual  $i$  in grade  $g$ , and  $N$ ,  $N_w$  and  $N_b$  refer to the weighted size of the total, white and black samples.

Note that  $G$  is unchanged if we multiply  $\beta$  by a constant. Moreover,  $\beta_0$  does not affect the calculation of  $G$  since it does not change either the test score gap or the standard deviation. Therefore 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.<sup>7</sup>

---

<sup>7</sup>In practice, it was easier to do the estimation by setting the constant term to 0 and constraining

We define the maximum growth in the test score gap between kindergarten entry and third grade as

$$D_{\max} = \max_{\beta, c} (G_3 - G_e) \quad (3)$$

where  $\beta$  is the vector of the five components of  $\beta_0 \dots \beta_6$  that are not normalized.  $D_{\min}$  is defined analogously.

If the test score distributions on entry and in third grade were disjoint, then (subject to a minor caveat about the limits 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 maximizations and minimizations separately.<sup>8</sup> Nevertheless, because there is not much overlap, the process of selecting the transformations comes close to mimicking this approach.

As we will see, in our data, 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 is modest but not trivial when children enter school and changes little over the next four years.

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. We consider three transformations. The first uses only information from the entry test and the third grade test and maximizes the correlation between the two tests. The second uses information from all the tests through grade three. In this case we choose  $\beta$  and  $c$  to maximize the  $R^2$  from a regression of third grade test scores on kindergarten and first-grade test scores. The final set of transformations adds in the fifth-grade test scores and asks

---

the linear term and only subsequently transforming the estimated coefficients.

<sup>8</sup>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.

what transformation maximizes the  $R^2$  from a regression of fifth grade test scores on all the earlier test scores.

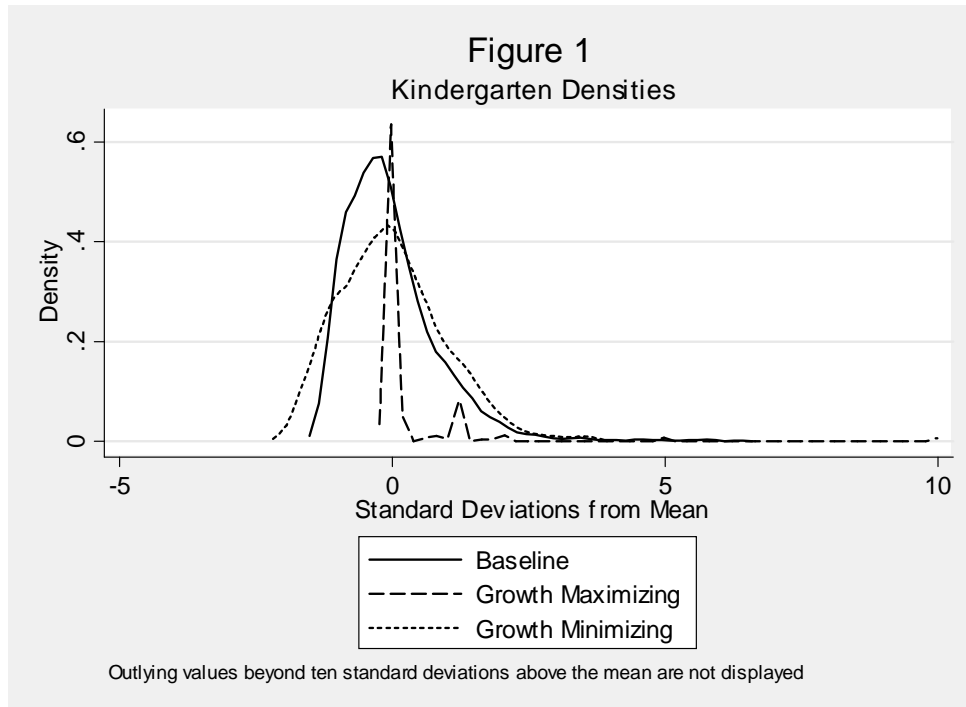
Ideally we would like to relate test scores to some future objective outcome. Unfortunately given the age of our sample, this is not possible. Instead, we ask how we can maximize the predictive power of current academic performance for future academic performance. In contrast with relating test scores to, for example, earnings where we treat the latter scale as a ratio scale, when relating test scores we have to ask what transformation of both the test scores used to predict the future scores and of the future scores themselves maximizes predictive power.

## 4 Results

### 4.1 Maximizing and Minimizing the Growth of the Gap

Table 2 shows how the test-score gap evolves from the beginning of kindergarten through the spring of third grade. The first column repeats the baseline pattern from Table 1. The second column shows the choice of transformations 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 come close to maximizing the entry gap. Thus it appears that the scaling system that is used in the ECLS is one that is not much below the one that maximizes that gap. The minimum possible growth in the gap is quite small. Using this scale, in third grade the gap is only .50 and thus noticeably less than the gap in the baseline. And the growth between entry and third grade is thus only .04. 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 2 shows the results of choosing the transformations that maximize 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 .09 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.

As noted above, using the test scale developed for the ECLS, the test score gap at kindergarten entry is .40 standard deviations. Figure 1 shows the density function of test scores associated with this choice of scale. Note that it is skewed with a long right tail.

In contrast, the value of the test score gap at kindergarten entry when  $T(t)$  is chosen to minimize the growth in that gap is 0.46, not much larger than the gap using the ECLS scale. As discussed earlier, while minimizing  $D$  does not literally maximize

$G_e$ , it comes close to doing this. Thus, subject to the caveat that some monotonic transformations will not be well approximated by our sixth-order polynomial, it is not possible to choose a transformation that dramatically increases the gap on kindergarten entry relative to the gap using the original scale. Visually, the resulting test score distribution, also shown in figure 1, more closely approximates a normal distribution than does the original.

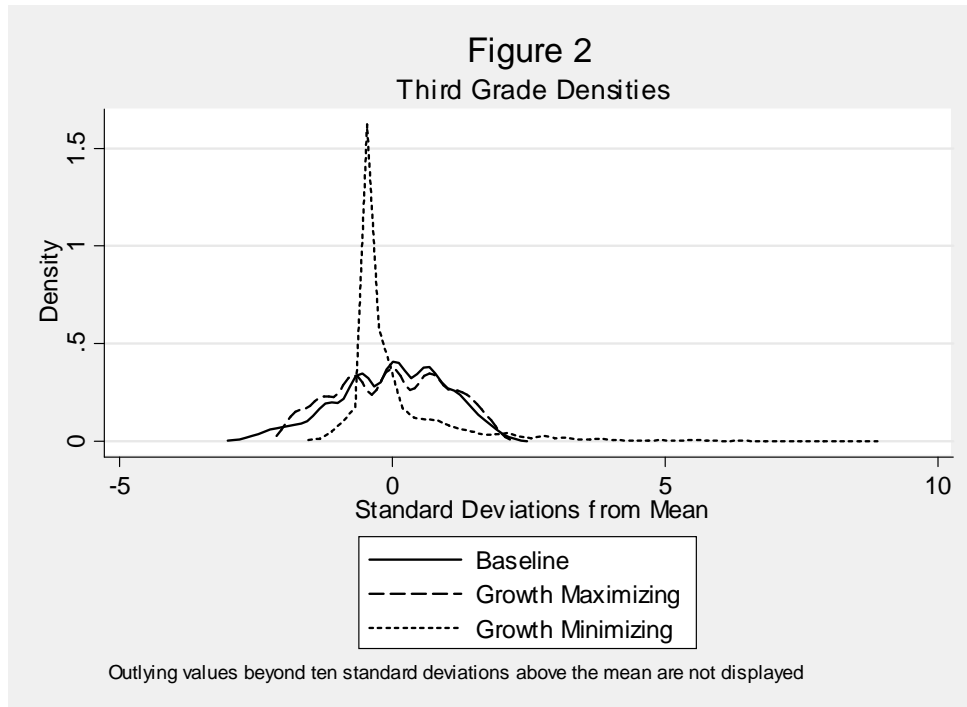
The third density in figure 1 is the one associated with the transformation that maximizes the growth in the test score gap. Using this objective, we estimate the gap at entry to be .11 standard deviations, which would generally be viewed as trivial. That said, we recognize that the test-score density associated with this transformation is 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.<sup>9</sup>

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 maximizing and minimizing transformations. As in the case of the kindergarten scores, the key to minimizing the third grade gap, and thus

---

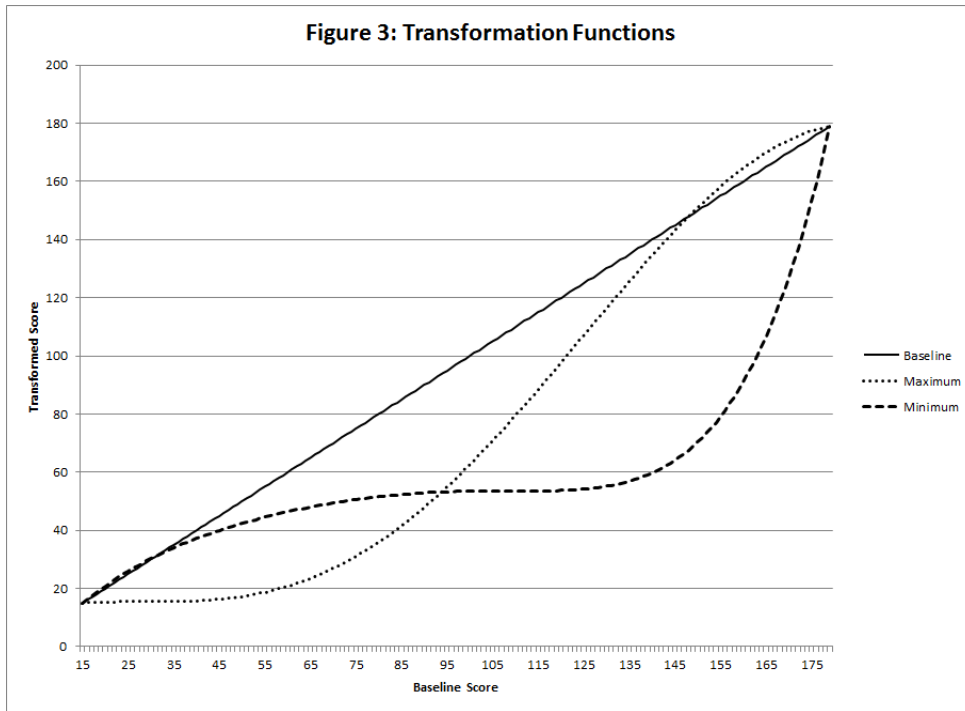
<sup>9</sup>This is based on our imputation from Kennickell's (2009) calculations based on the 1989-2007 Survey of Consumer Finances.



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. Again, 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 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. When described in this way, it is not obvious that the density shown in figure 1 is unreasonable.

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.” As is often the case with bounding exercises, the range of possible results is too large to be helpful. And choosing transformations

on aesthetic grounds or because the distributions of scores should not change “too much” is unsatisfactory. In the next subsection we consider a more formal approach to choosing the appropriate transformation.

## 4.2 Selecting Transformations

We would not expect kindergarten or first grade performance to perfectly predict third grade performance. There is randomness in performance on each of the tests. 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, the 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 predict best performance on the third grade test using only information from the kindergarten and first grade tests.

We begin by examining 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 Table 3, 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.

Our second approach to selecting among the potential transformations is to 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. The maximum  $R^2$  is .62 compared with .56 . Perhaps not surprisingly, adding the fall kindergarten and spring first grade tests does not change the pattern of the test score gap.

Finally, we add information from the fifth grade test and choose the scale that maximizes the  $R^2$  from a regression of the fifth grade score on the remaining test scores. As can be seen from the third column in table 3, the  $R^2$  when we use the baseline IRT scale is already quite high (.75). This is largely because the third and fifth grade scores are highly correlated. Changing the scale has little effect on the ability of the earlier tests to predict the fifth grade test score, raising the  $R^2$  by less than .01. Nevertheless, relative to the baseline scale, the revised scale shows less widening of the test score gap between entry and the spring of first grade. The subsequent widening of the gap between first and third grades is comparable.

### 4.3 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 4 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. Five of the six scales generate unadjusted test score gaps of approximated .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.

#### **4.4 Scale Sensitivity**

Why do some scales show larger test score gaps and different trends? And is the gap measured in standard deviations an appropriate metric? Our previous analysis suggests the hypothesis that the test score gap tends to be small when the scores are highly concentrated and only a relatively small fraction of the students have very different test scores from the mass of students. Therefore, in this subsection, we ask how large the test score gap could be given the fineness of the chosen scale and the sensitivity of the test.

Transforming raw test scores into a particular distribution limits the between-group differences we can observe.

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.

Concerned by the large test score gap but also somewhat perplexed by the different answers, the researchers collect an additional sample of 50 black children (but no more white children) whose distribution of test results turns out (miraculously) to be identical to that of the first group of black children. Obviously there is no need to ask the psychometricians to update their results, but a well-intentioned graduate student passes on the results and later informs the research team that both psychometricians now report a larger gap than they had previously.

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.

There are two related issues in our example. The first is that the composition of the sample affects the overall variance. Even though adding a new sample of fifty black children had no effect on the black-white test score gap, in both cases it lowered the overall standard deviation and therefore increased the gap relative to the standard deviation of test scores. At least equally important, the overall distribution of scores determines how large the gap can be.

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 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.<sup>10</sup>

Table 5 shows the weighted test gaps along with the hypothetical upper boundary for the gap on each test. 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.

Returning to our previous transformations, in Table 6 we look at the impact those transformations have on the boundary of the test gap. The first two columns show the maximum test gap for the transformations that simply try to minimize or maximize the growth of the gap in the first few 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

---

<sup>10</sup>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.

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 .94 standard deviations under the 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 3, 4, and 5, look at the boundaries for the test gap under the transformations that maximize the correlation across tests. The test gaps under these transformations as a percentage of the largest possible gap are similar to those in the baseline, though the transformation raises the upper bound of the test gap at kindergarten in each of the three cases.

## 5 Summary and Conclusion

The findings in this paper 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, 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.

Our preferred scales all tell a similar story. 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 in second 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.

Nevertheless, 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 rescaling is correct, this suggests one avenue to pursue in furthering our understanding of the gap.

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.

Carnerio, Pedro and James J. Heckman, "Human Capital Policy," *NBER Working Paper No. 9495*, 2003.

Cascio, Elizabeth U. and Douglas O. Staiger, Skill, Standardized Tests, and Fade-out 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.

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, Aruther B. 'Ponds and streams: Wealth and Income in the U.S., 1989 to 2007," *Finance and Economics Discussion Series 2009-13*, Federal Reserv 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.

Table 1 ECLS-K Descriptive Statistics		
	<u>Our Sample</u>	<u>Fryer &amp; Levitt</u>
Race		
White	0.617 (0.486)	0.554 (0.497)
Black	0.168 (0.374)	0.152 (0.359)
Hispanic	0.141 (0.348)	0.178 (0.382)
Asian	0.024 (0.153)	0.065 (0.246)
Female	0.488 (0.500)	0.489 (0.500)
Black-White Test Gap		
Kindergarten Fall	0.404 (0.030)	0.400 (0.029)
Kindergarten Spring	0.435 (0.032)	0.451 (0.029)
First Grade Spring	0.493 (0.034)	0.517 (0.030)
Third Grade Spring	0.746 (0.036)	0.771 (0.032)
Sociodemographic Controls		
Age (in months) fall Kindergarten	68.462	67.013
SES composite measure	0.022	0.005
Number of children's books in home	76.804	61.432
Mother's age at first birth	23.559	23.609
Child's birth weight (in ounces)	118.120	87.463
WIC participant	0.422	0.378
Observations	11414	10540

Table 2			
Evolution of the black-white test gap under various transformations			
	<u>Baseline</u>	<u>Minimum</u>	<u>Maximum</u>
	(1)	(2)	(3)
Kindergarten - Fall	0.4038 (0.0305)	0.4644 (0.0350)	0.1108 (0.0229)
Kindergarten - Spring	0.4355 (0.0322)	0.5049 (0.0373)	0.2054 (0.0216)
First Grade - Spring	0.4929 (0.0343)	0.4856 (0.0395)	0.4284 (0.0294)
Third Grade - Spring	0.7464 (0.0361)	0.5064 (0.0238)	0.7481 (0.0348)

Table 3 ECLS-K R-squared Maximizing Transformations			
Left Hand Side Variable Right Hand Side Variables	Third Grade Spring-K	Third Grade All Pre-3	Fifth Grade All Pre-5
Baseline R-squared	0.2940	0.5605	0.7483
Transformed R-squared	0.3881	0.6176	0.7541
Kindergarten - Fall	0.4694 (0.0354)	0.4746 (0.0364)	0.4591 (0.0344)
Kindergarten - Spring	0.5216 (0.0381)	0.5199 (0.0384)	0.5008 (0.0364)
First Grade - Spring	0.4947 (0.0375)	0.4914 (0.0397)	0.5035 (0.0374)
Third Grade - Spring	0.7325 (0.0325)	0.7289 (0.0385)	0.7425 (0.0369)

Table 4  
Evolution of the unexplained black-white test gap under various transformations

Transformation	Baseline (1)	Minimizing (2)	Maximizing (3)	Fall-K R2 (4)	All Pre-3 R2 (5)	All Pre-5 R2 (6)
Kindergarten - Fall	-0.0472 (0.0314)	-0.0346 (0.0361)	-0.0388 (0.0232)	-0.0339 (0.0366)	-0.0291 (0.0377)	-0.0371 (0.0355)
Kindergarten - Spring	0.0415 (0.0344)	0.0797** (0.0398)	-0.0111 (0.0226)	0.0959** (0.0407)	0.0959** (0.0410)	0.0771** (0.0389)
First Grade - Spring	0.0998*** (0.0378)	0.104** (0.0439)	0.0789** (0.0328)	0.100** (0.0424)	0.101** (0.0450)	0.104** (0.0417)
Third Grade - Spring	0.308*** (0.0397)	0.172*** (0.0262)	0.309*** (0.0384)	0.295*** (0.0358)	0.300*** (0.0420)	0.306*** (0.0405)

Standard errors are in parentheses.

\* significant at .1-level; \*\* significant at .05-level; \*\*\* significant at .01 level.

Controls for 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

Table 5 Black-White Test Gap as Percentage of Distribution Boundary				
	<u>Fall-K</u> (1)	<u>Spring-K</u> (2)	<u>Spring-1</u> (3)	<u>Spring-3</u> (4)
Observed Black-White Test Gap	0.4038	0.4355	0.4929	0.7464
Maximum Black-White Test Gap	1.4946	1.5979	1.9046	2.2319
% of Maximum Gap	27.02%	27.25%	25.88%	33.44%

Table 6 Black-White Test Gap as a Percentage of Boundary Under Various Transformations					
	<u>Minimizing</u> (1)	<u>Maximizing</u> (2)	<u>Fall-K R2</u> (3)	<u>All Pre-3 R2</u> (4)	<u>All Pre-5 R2</u> (5)
Fall-K Black-White Test Gap	0.4644	0.1108	0.4694	0.4746	0.4591
Fall-K Maximum Test Gap	1.9239	0.2393	1.9660	2.0475	1.8758
Fall-K % of Maximum Gap	24.14%	46.32%	23.87%	23.18%	24.48%
Spring-3 Black-White Test Gap	0.5064	0.7481	0.7325	0.7289	0.7425
Spring-3 Maximum Gap	0.9470	2.1576	1.9584	2.2655	2.2445
Spring-3 % of Maximum Gap	53.47%	34.67%	37.41%	32.17%	33.08%