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

Timothy N. Bond† Kevin Lang‡

October 2, 2012

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 of Youth, we examine how order-preserving scale transformations affect 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 reduce it in the ECLS-K during the early school years. All growth from entry through first grade and a nontrivial proportion from first to third grade probably reflects scaling decisions. JEL codes: I24, J15

*We are grateful to Henry Braun, Sandy Jencks, Dan Koretz, Sunny Ladd, Ed Lazear, Michael Manove, Dick Murnane, Sean Reardon, Bruce Spencer, an anonymous referee, the editor and participants in seminars and workshops at Baruch College, Boston University, Brown, CERGE-EI, Sciences Po and Stanford for helpful comments and suggestions. The usual caveat applies.

†Department of Economics, Krannert School of Management, Purdue University, 403 W. State Street, West Lafayette, IN 47907; email: tnbond@purdue.edu

‡Department of Economics, Boston University, 270 Bay State Road, Boston, MA 02215; 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 (2012) consider how changes in scaling affect estimates of the fade-out of teacher value-added. As discussed in greater detail in the conclusion, similar concerns arise in the happiness literature.

In this paper, we take the ordinality problem seriously 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) challenge the accepted view that a large black-white test score gap emerges in early childhood (Jencks and Phillips, 1998). They find that the gap in kindergarten is both modest and largely "explained" by a small number of socioeconomic characteristics but widens sharply in the early years of schooling.[1]

We perform a bounding exercise in which we examine how our estimates of the change in the black-white test score gap between kindergarten and third grade depend on choice of scale. Unfortunately, this puts only weak bounds on how the test score gap evolves. Without additional restrictions, we can conclude only that the change in the black-white test score gap between kindergarten and third grade is somewhere between 0 and .6 standard deviations.

One solution to the ordinality problem is to relate the test score to some outcome as in

Cunha and Heckman (2008) and Cunha, Heckman and Schennach (2010). When, we choose scales which maximize the ability of earlier scores to predict performance on later tests, we generally find 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.

The extent to which family background, environmental measures and parental behaviors can explain the test score gap is also controversial. This is in part because the influence of such factors varies among data sets[2] 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).

Despite these caveats, we also examine the relation between family background and the test score gap and ask whether it 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 and its growth.

In the next section, we show numerical examples of how scaling decisions can be important in interpreting the test gap. In section three, we provide a short primer on item response

theory scales which some researchers claim are interval scales. We then describe the data used for this study (section four) and present our approach (section five). Finally, we give our results in section six and then provide some concluding remarks in which we discuss the use of ordinal scales more broadly.

2 Scaling Issues

For ease of exposition, we begin with a scale that takes a small number of discrete values. We trust, however, that it will be apparent that the issues we raise apply to continuous scales. Suppose we have a perfect test of mastery of each of three progressively difficult skills. We assume that the skills are cumulative either because skills are learned in this order or skill 2 cannot be mastered before skill 1 (two-digit addition requires one-digit addition) and there is no partial mastery. The test therefore produces 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 are the number of skills mastered. But the marginal value of all three skills need not 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 population of two blacks and two whites. Initially, the blacks score a and c while the whites score b and c. If we use the naive scale, $0, 1, 2, 3$, then the difference in

the means is .5 or about .6 standard deviations. If instead we make the difference between a and b arbitrarily small relative to the difference between b and c, the test gap goes to 0. If we make the gap between b and c arbitrarily small relative to the gap between a and b, we get the maximum gap of about 1.15 standard deviations. Without some external reference for determining the proper scale, all we can say is that the test score gap is between 0 and 1.15 standard deviations.

The situation worsens when we ask how the gap evolves over time: a year later, we readminister our test and find that each student has progressed exactly one skill level (blacks score b and d; whites score c and d). Using the "natural" scale, the gap is unchanged. If we set b arbitrarily close to c, the gap *falls* from 1.15 to 0 standard deviations. On the other hand, if we choose $a \cong b$ and $c \cong d$, the gap *grows* from 0 to 1.15 standard deviations.

Like economists, noneconomists frequently ignore the complexities associated with working with interval scales. However, as early as 1983, Spencer pointed out that the performance of two groups can only by strictly ranked if the cumulative distribution functions (cdf) of their scores do not cross (one is higher than the other in the sense of first-order stochastic dominance) and suggested comparisons based on these cdfs.

Since percentiles are invariant to scale, a common solution is to use percentile ranks rather than test scores. The most prominent such measure is the percentile-percentile (PP) curve[3] which 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, the gap appears to be smaller for the comparison captured by the higher curve. Likewise, if the PP curves are

identical, it would appear that the gaps are also identical.[4]

Unfortunately the conclusions based on shifts (or lack thereof) of the PP curve can be misleading. In the above example, the PP curve does not shift between administrations of the test. But we saw that the gap might have increased or decreased. Moreover, the PP curve would also suggest no change if instead blacks scored a and d on the second administration. "No change" can only be right in both instances if $a \cong b \cong c$, in which case the first administration was completely uninformative.

3 Some Background on Educational Measurement

We have sometimes been asked how experts in educational measurement have responded to this paper. The simple answer is "with much kindness and patience for the gaps in our knowledge."The acknowledgements list several extremely helpful experts. One of them reported discussing the paper with others in the field and receiving reactions divided between "that is deeply disturbing" and "that cannot be important." A few have implied that we underappreciated the degree to which psychometricians are aware of the issues we raise, an error that we attempt to correct in this section. None has suggested that test scores are, in fact, measured on interval scales. A distinct minority of researchers in the field believe that certain scales are interval scales, but they either did not read our earlier drafts or did not bother to correspond with us.[5]

The concern that test scores are ordinal scales is long-standing. The earliest reference we found is Stevens (1946). Among the giants of the testing field, Thorndike (1966, p. 124) makes the point emphasized in this article: "... it is assumed that the numerals in

which the variables are expressed represent equal increments in some attribute. It is also recognized that this assumption is usually not well supported. But for 'rough and ready' studies of relationship, the violation of the assumption usually does not hurt much. However, when starting to deal with something as fragile as a change score, the violation of this basic assumption becomes a good deal more critical."

Here we provide a brief primer on item response theory (IRT) scaling so that the economists and others can understand why some researchers might believe that it generates interval scales. We again thank our tutors for their help and absolve them of responsibility for our errors. Readers familiar with the relevant literature or who do not care about it can skip this section. We refer readers who want more detail to Baker (2001).

IRT models are defined by their number of parameters. Across all models, questions are identified by a difficulty parameter b, while individuals vary by a parameter θ. The goal of IRT is to estimate θ for each test-taker.

In the most general (three-parameter) logistic IRT model, the probability that a student i will answer question q correctly is given by[6]

$$p_{iq} = c_q + (1 - c_q) \frac{1}{1 + e^{-a_q(\theta_i - b_q)}}. \tag{1}$$

θ, b and a measure student ability, question difficulty and how well the question distinguishes among students of different abilities. c is guessability or the probability that a student of extremely low ability will get the answer right. The two-parameter model assumes no guessability ($c = 0$).

It is universally recognized that the parameters are defined only up to a linear transfor-

mation. Adding the same constant to θ and b or multiplying θ and b and dividing a by the same constant does not change p. We can normalize the scale by, for example, setting the mean to be 100 and the standard deviation to be 15. Some authors claim that θ is *uniquely* identified given this normalization. The argument seems to be that given this normalization and adequate data, the θ_i are unique. However, as discussed by Lloyd (1975), one of the fathers of IRT, if $\theta' = f(\theta)$ where f is a strictly increasing function, then replacing θ with $f^{-1}(\theta')$ fits the data with the same likelihood. As noted, most psychometricians accept Lloyd's argument.

In the special case where all questions are equally informative, a can be normalized to 1, which gives the one parameter IRT model or Rasch scale. This has the property that regardless of question difficulty, a one unit increase in θ increases the log odds of getting the answer correct by 1. As a result, it is more common among individuals in the field to believe that Rasch scales can be interpreted as interval scales.[7]

4 Data

We use the Children of the National Longitudinal Survey of Youth (CNLSY) and the Early Childhood Longitudinal Study Kindergarten Class of 1998-1999 (ECLS-K).

4.1 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), a longitudinal survey that initially followed a sample of 12,686 youths who were between the ages of 14 and 21 as of December 1978. Beginning in

1986, the children of women in the NLSY79 sample were surveyed and assessed biennially. The assessments included a battery of tests of psychological, socioemotional, and cognitive ability, and questions on the home environment. 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.

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.

The primary advantage of the CNLSY is that it includes both the Peabody Individual Achievement Test (PIAT), which is similar to the test in the ECLS-K on which Fryer and Levitt find little difference in the performance of blacks and whites entering kindergarten, and the Peabody Picture Vocabulary Test (PPVT) which is similar to tests on which a test score gap is measured before kindergarten entry.

The PPVT 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. To avoid measuring differences that might reflect kindergarten quality, we analyze PPVT scores only for children who were less than four years old when they took the test. We further limit our sample to black and white youths. We have a total of 1,655 scores (1,072 white and 583 black).[8] There are no repeat exam takers. The highest

score attained by any child was 77. Two children scored 0, the lower bound. Based on the official scale, the test score gap is .97 standard deviations.

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. Sample size is fairly consistent across the grades we study. Average scores rise steadily from 17 in kindergarten to 39 by third grade. 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 gap between blacks' and whites' PIAT scores is quite modest in kindergarten (.25 standard deviations), but expands over the first four years of school, reaching .61 standard deviations in third grade. 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.

Recall that 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 conclusion 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.

4.2 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.[9]

In each survey year, the student's parents and teacher were interviewed about the child's background, home, and school environment. In addition, students took tests in reading, mathematics and general knowledge or science.[10] The material covered on the test was constant through first grade, but was modified in third grade to reflect students' growing knowledge. Children took a short "routing test" that directed them to a more comprehensive exam, the difficulty of which was determined by the routing test. According to the User's Guide, overall scores are calculated using IRT and represent the estimated number of questions the test taker would have answered correctly had she taken the entire test. In principle, a 112 on the kindergarten entry test represents the same level of accomplishment as a 112 on the third grade test. 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 released with the 5th grade data file.

We mimic Fryer and Levitt's sample construction methods. We focus on the reading scores because they show the most striking growth in the early years in the Fryer/Levitt study. We drop students who are missing a valid reading score from kindergarten through third grade or who do not have a valid entry for race. We use the sampling weights associated with grades kindergarten through three for child assessment studies and drop 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 baseline scale shows 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.

We note 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 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.

5 Methods

Following the literature, we define the test score gap as the difference between the mean test scores of whites and blacks divided by the standard deviation of test scores in that grade. Our approach, described more formally below, is to look for order-preserving transformations of the test score scale used in the original data set, and choose the ones that satisfy certain properties. Initially we look for the transformation that maximizes the growth in the scales and the one that minimizes the growth. When the number of points on the scale is small, we can literally transform every point on the scale, except the highest and lowest. When

the number of points is large, we use a very flexible order-preserving function. In addition to finding the bounds, we also choose the transformations that maximize the ability of an early test to predict performance on a later test.

We first search for the monotonic transformations of the original scale that maximize and minimize the growth of this gap. We impose the transformation

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

where t is the original scale, T is the transformed scale and a_{t+1} is a real number. Since the gap is unchanged by a linear transformation, we normalize $T(0)$ equal to 0 and $T(t_{\max})$ equal to $t_{\max}$ where $t_{\max}$ is the highest score observed in that grade.[11] Define G_g to be the test gap in grade g.

$$G_g = \frac{N_w^{-1} \sum_{i \in white} T(t_{ig}) - N_b^{-1} \sum_{i \in black} T(t_{ig})}{\sqrt{N^{-1} \sum \left(T(t_{ig}) - N^{-1} \sum T(t_{ig}) \right)^2}} \tag{3}$$

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 to minimize the objective function given by

$$D_{\min} = \min_a (G_3 - G_e) \tag{4}$$

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 scores are observed each year. We normalize a_{t+1} to 0 if no member of the sample in that grade has an initial test score of $t+1$.

This nonparametric approach gives bounds on the gap but produces implausible scales with only one or two steps. Additionally it cannot be used when the test score is a continuous variable, as the ECLS-K assessment approximately is. Therefore, we use a sixth-degree polynomial given by

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

where $\beta_0 - \beta_6$ and c are constants. This function is very flexible and can approximate a wide array of continuous functions. However, it need not be monotonic. Our algorithm checks for monotonicity and rejects parameters that violate this condition.[12] Of course, some monotonic functions may not 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 their standard deviation to equal 1 and choose β_0 so that their mean is 0. However, when showing the relation between the two scales, we fix the highest and lowest scores to be equal across scales.[13]

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 $G_{3.}$ Because the two test score distributions overlap, we cannot do the maximizations and minimizations separately.[14] Nevertheless, because there is not much overlap, the process of

selecting the transformations closely mimics this approach.

As we will see, in both data sets, the implications of $D_{\min}$ and $D_{\max}$ are very different. In the latter case, the black-white gap is trivial when children first enter school but grows substantially by the end of third grade. In contrast, in the former case, the black-white gap in the ECLS-K is modest but not trivial at school entry 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, we select among the possible transformations, including less extreme ones, by choosing the transformations with 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 kindergarten. For the ECLS-K, we maximize the correlation between the entry and third grade tests.

6 Results

6.1 Maximizing and Minimizing the Growth of the Gap

The first column of Table 3 shows the black-white test gaps in the PIAT, using the original scale. It increases significantly from a modest .25 standard deviations in kindergarten to .61 standard deviations by third grade. Under the growth-minimizing scale (column 2), the black-white test gap *shrinks* by .18 standard deviations during the first four years of education. The test gap in kindergarten is similar to that of the baseline at .24 standard deviations, but declines to .18 standard deviations in first grade and .08 standard deviations in second grade, ending at .06 standard deviations in third grade. The growth-maximizing

transformation reduces the gap at kindergarten to just .05 standard deviations, but at higher grades the gaps are similar to those in the baseline model, with blacks performing .63 standard deviations worse than whites in third grade, only slightly worse than they perform using the baseline scale. Thus, the gap grows by .58 standard deviations over the first four years of school.

The extreme transformations produce scales that differ noticeably from the baseline. They 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. If kindergartners differ only 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, this could be an appropriate scale at that grade. In fact, the PIAT reading test is designed somewhat like this. Students must pass a reading recognition test to advance to a reading comprehension test. The modal score in both our kindergarten and first grade sample is 18, the highest score at which students do not advance to the reading comprehension section.

Table 4 shows the achievement gaps on the ECLS-K reading assessment 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 transformation that minimizes the growth in the gap. The resulting growth is only .05. The gap at kindergarten is .46, only slightly higher than in the baseline,[15] but in third grade the gap is only .51 and thus noticeably less than in the baseline. Note that, in principle, minimizing growth between entry and third grade could still generate large swings in the first grade gap. However, there is no noticeable change in the gap between any pair of tests when this scale is applied.

Column (3) 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 .64 standard deviations in the racial test gap in the first four years of education, almost twice that using the baseline scale.[16]

Figure 1 shows the density function of test scores at kindergarten entry associated with each ECLS-K scale. Note that the baseline density is skewed with a long right tail. In contrast, visually, the density using 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 this distribution's weight is in a narrow band around its mode, with none substantially below. Nevertheless, this representation of the scores is not entirely counterintuitive. It is plausible that most children have few 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.[17]

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 densities of the three test score distributions. As for 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 very low scores. In contrast, the growth-maximizing transformation leaves the distribution of test scores looking similar to that associated with the baseline.

As discussed above, 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. The growth in the gap is minimized if 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, treating most differences among very low scores as uninformative. This would be appropriate if most children arrive in kindergarten knowing very little of the material covered by the ECLS so that throughout most of the distribution, differences in performance are relatively unimportant and only children with very high scores differ substantially from the mass of kindergarten entrants.

The results in this subsection bring out the fragility of any conclusion about how much 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 is a modest gap at entry and essentially no growth in the gap over this period." For the PIAT, there are scales that imply "black children moderately lag behind white children in achievement when they enter school, but match the achievement of their white peers by third grade."As is often the case with bounding exercises, the range of possible results is too large to be helpful.

Therefore, determining the right scale is important in establishing how the gap between blacks and whites evolves. We could attempt to choose scales that produce "aesthetically appealing" distributions of test scores, but 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 above, unintuitive distributions of ability may be plausible, both for young children and adults. In the next subsection we consider a more formal approach to choosing the appropriate transformation.

6.2 Selecting Transformations

We do not expect kindergarten or first grade scores to perfectly predict third grade scores. Performance on each test is random, and students' academic progress varies. 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 generally perform well on other tests. A reasonable criterion is to select the transformation that allows us to best predict future performance using information from previous tests.[18] We therefore choose transformations that maximize the correlation between test scores. If the tests measure a common underlying latent variable, this maximizes reliability, but we do not require this interpretation

For the CNLSY we construct a sample of 398 white and 253 black children who took both the PPVT before age 4 and the PIAT while in kindergarten. The racial test gaps (.97 standard deviation on the PPVT and .2 on the PIAT) for this subsample are very close to those for the full sample. The correlation between the untransformed test scores is .32.

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

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

Column (4) of Table 3 shows the evolution of the PIAT 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 is that roughly 18 percent of the third grade sample scored above the highest kindergarten score. This problem is less important 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 using the original scale.

In the ECLS-K can only maximize the correlation across reading assessments. First we maximize the correlation between the kindergarten entry and third grade tests. This new scale substantially increases the correlation between the two scores from .54 (R^2 = .29) at baseline to .62 (R^2 = .39), and produces gaps close to the potential maximums at kindergarten entry and in third grade.

As shown in column (4) of Table 4, the growth in the gap from kindergarten entry through third grade is .26 standard deviations, .09 smaller than the growth in the baseline. All the growth occurs between the end of first and the end of third grade. This differs from the baseline steady increase throughout the first four years of schooling.

We also explored the scales that maximize the R^2 from regression of the third and fifth grade scores on all prior scores. Both approaches yield results similar to those in column 4.

6.3 Controlling for Socioeconomic Factors

Surprisingly, Fryer and Levitt find that the modest black-white test score gap at school entry can be accounted for fully by a small number of socioeconomic characteristics (child'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 Table 5 we ask whether the

same is true for the scales we have developed. Strikingly the kindergarten entry results are robust to the choice of scale. 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 effect of the controls in third grade depends on the scale. For three of the four scales controlling for the socioeconomic factors reduces the gap from approximately .75 to about .3 standard deviations. This 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 significantly affects the magnitude of the increase of both the adjusted and unadjusted gap.[19]

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. We have already shown that, in contrast, much of the growth in the gap under the maximizing transformation can be explained by socioeconomic controls: the raw gap increases by .64 standard deviations, but 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 minimize the ratio of the growth in the controlled gap to the growth in the uncontrolled

gap. Using this transformation, the raw test gap grows by .59 standard deviations from kindergarten through third grade, while the controlled gap grows by only .28. 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 is similar.

6.4 Scale Sensitivity

The choice of scale limits the potential magnitude of between-group differences. An example may clarify this point. A researcher administers a test to 50 black and 50 white children to determine whether they are "learning to read" or "reading to learn." If he sets the scale so that 50 students are assigned to each group, if all the learning to read students are white and all the others are black, the test score gap is almost exactly two standard deviations. In contrast, if he sets the cutoff for reading to learn so that only 25 students earn this score, the highest possible gap is about 1.1 standard deviations. The gap can be bigger in the former case than in the latter.

The lower half of the scores in the ECLS-K kindergarten test are clustered within one standard deviation of the median. This characteristic of the test and scaling affects the potential for a large test score gap in kindergarten. To analyze the potential effect of such clumping 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.[20]

The first column of Table 7 shows the weighted test gaps and the upper bounds. While the observed black-white test gap increases over time, so does the bound. 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. Consequently, the observed racial test gap as a percentage of the possible test gap hardly changes over time, from 27 percent at the beginning of kindergarten to 33 percent 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 reflects changes in scale and test sensitivity rather than changes in the real achievement gap.

Columns (2) and (3) of Table 7 show the maximum test gap for the transformations that minimize and maximize the growth of the gap. The minimizing transformation yields a test gap that is 24 percent of the bound at kindergarten entry, but 53 percent at the end of third grade despite virtually no growth in its absolute size in standard deviations. In contrast, using the maximizing transformation the gap shrinks from 46 percent of the bound at the start of kindergarten to 35 percent at the end of third grade despite a nearly 700 percent increase in its absolute size. 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 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.

Column (4) shows the bounds 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.

An alternative approach is to look at the gap holding the test scale 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.[21] When we impose either the fall kindergarten or the spring third grade mapping, we see virtually no growth in the test score gap until the third grade, but substantial growth for this test.

Panel B instead supposes that we fix r while varying F_g (i.e. changing the scales across grades as occurs *de facto* in the ECLS-K.) Even if the rank order of students did not change, the test gap using the baseline scales in the ECLS-K would grow by .09 standard deviations 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 this gap occurs between the spring first grade and spring third grade tests, .05 standard deviations using entry rank and .07 using the third grade rank.

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.

7 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. Our message is not limited to the black-white test score gap but applies whenever test scores or other ordinal measures are used as dependent variables. There is increasing pressure in the United States and elsewhere to use "value-added measures" to determine teacher compensation and retention. Lang (2010) presents a simple example in which the ranking of teachers is highly sensitive to the choice of scale. Before linking important decisions mechanically to value-added measures, we should ensure that they are robust to arbitrary scaling decisions.

Cascio and Staiger (2012) consider a related issue. Economists frequently renorm scales to have mean zero and variance 1 each year. For any given year, this is simply a linear transformation. However, if the variance of the true scale grows as students progress through school, then we are using different linear transformations each year, and the scales are no longer strictly comparable. If variance grows quickly as students progress through school, whatever students learned earlier will seemingly become less important because the true scale is divided by a higher number later in school. Therefore, fade-out can be a mechanical result of the convention of normalizing test scores. They find evidence of such an effect but conclude that it is only of modest importance. However, Cascio and Staiger do not address

ordinality. Their approach depends on the ability to calculate correlations among test scores, which is possible only if they are measured on an interval scale. We believe it is unlikely that their result would hold for arbitrary monotonic transformations.

We are not the first to recognize the difficulties of working with ordinal test score scales. Cunha and Heckman (2008) and Cunha, Heckman and Schennach (2010) tie test scores to adult outcomes. This gives test scores a valid external reference but is not a panacea. The scale will differ if it is tied to wages or to log wages, and the choice between the two is essentially a welfare judgment. We do not know whether plausible variation in the choice of anchor would affect the outcome of their research.

Nor is our message limited to the education and child development literature. Similar problems arise in the happiness literature. A common, perhaps the central, finding in this literature is that the relation between happiness and income is clearly positive within countries but, at best, weak across countries. Happiness is typically measured on an ordinal scale of three to five points and rarely more than eleven. Researchers routinely average the answers to find the mean level of happiness or life satisfaction.[22] Without analyzing the underlying data, it is difficult to know how problematic this is.

In this paper, we have shown that scaling matters but differently for different tests. More broadly, our findings suggest that economists and other researchers should be much more circumspect in their use of test scores and other ordinal scales as dependent variables, particularly when comparing changes across groups. While many findings will be robust to scale changes, many will not be.

References

Baker, Frank, *The Basics of Item Response Theory* (College Park: ERIC Clearinghouse on Assessment and Evaluation, University of Maryland, 2001).

Braun, Henry I., "A New Approach to Avoiding Problems of Scale in Interpreting Trends in Mental Measurement Data," *Journal of Educational Measurement* 25 (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 (2008), 738-782.

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

Cascio, Elizabeth U. and Douglas O. Staiger, "Knowledge, Tests, and Fadeout in Educational Intervention," NBER working paper no. 18038 (2012).

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 (2009), 398-419.

Dickens, William T. and James R. Flynn, "Heritability Estimates versus Large Environmental Effects: The IQ Paradox Resolved," *Psychological Review* 108 (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 (2005), 35-54.

Dunn, Elizabeth and Norton, Michael, "Don't Indulge. Be Happy," *New York Times* Sunday Review, July 8, 2012.

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 (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 (2006), 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 (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 (2002), 3-17.

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

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

Kennickell, Arthur B., "Ponds and streams: Wealth and Income in the U.S., 1989 to 2007," Federal Reserve Board Discussion Paper 2009-13 (2009).

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

Kimeldorf, George and Allen R. Sampson, "Monotone Dependence," *The Annals of Sta-*

tistics 6 (1978), 895-903.

Lord, Frederic M., "The 'Ability' Scale in Item Characteristic Curve Theory," *Psychometrika* 40 (1975), 205-217.

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.

Reise, Steven P. and Waller, Niels G., "Item Response Theory and Clinical Measurement," *Annual Review of Clinical Psychology* 5 (2009), 27-48.

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

Reardon, Sean F., "Thirteen Ways of Looking at the Black-White Test Score Gap," Stanford University Institute for Research on Education Policy and Practice working paper no. 2008-08 (2008).

Spencer, Bruce D. "On Interpreting Test Scores as Social Indicators: Statistical Considerations," *Journal of Educational Measurement* 20 (1983), 317-333.

Stevens, S.S. "On the Theory of Scales of Measurement," *Science* 10 (1946), 677-680.

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

Notes

[1]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.

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

[3]The earliest reference appears to be Wilk and Gnanadesikan (1968). For examples of test gap measures based on the PP curve and extensions, see Braun (1988), Holland (2002), Ho and Haertel (2006) and Reardon (2008).

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

[5]See the discussion of the two views in Reise and Waller (2009).

[6]There is a similar model based on the normal distribution which, to conserve space, we will not discuss.

[7]See, for example, Reardon (2008) who concludes that the test he uses does not satisfy this condition.

[8]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.

[9]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.

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

[11]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 $t_{\max}$.

[12]This introduces a discontinuity into our objective function which creates many local minima (maximuma). We therefore searched from several different starting locations for each transformation and report that which produced the smallest (largest) gap.

[13]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.

[14]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.

[15]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 scale used in the ECLS-K comes close to maximizing that gap.

[16]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.

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

[18]Kimeldorf and Sampson (1978) refer to this as the monotone correlation.

[19]Given the arbitrariness of scale, there is no reason to believe that the socioeconomic factors should enter the equation linearly. We choose this specification following that of

Fryer and Levitt (2004,2006). The conclusions are robust to using a full set of interacted controls entering as a cubic polynomial.

[20]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.

[21]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.

[22]Two noted happiness researchers went further and treated the scale as a ratio scale: ".. our data showed that people who earned $55,000 were just 9 percent more satisfied than those making $25,000." (Dunn and Norton, 2012)

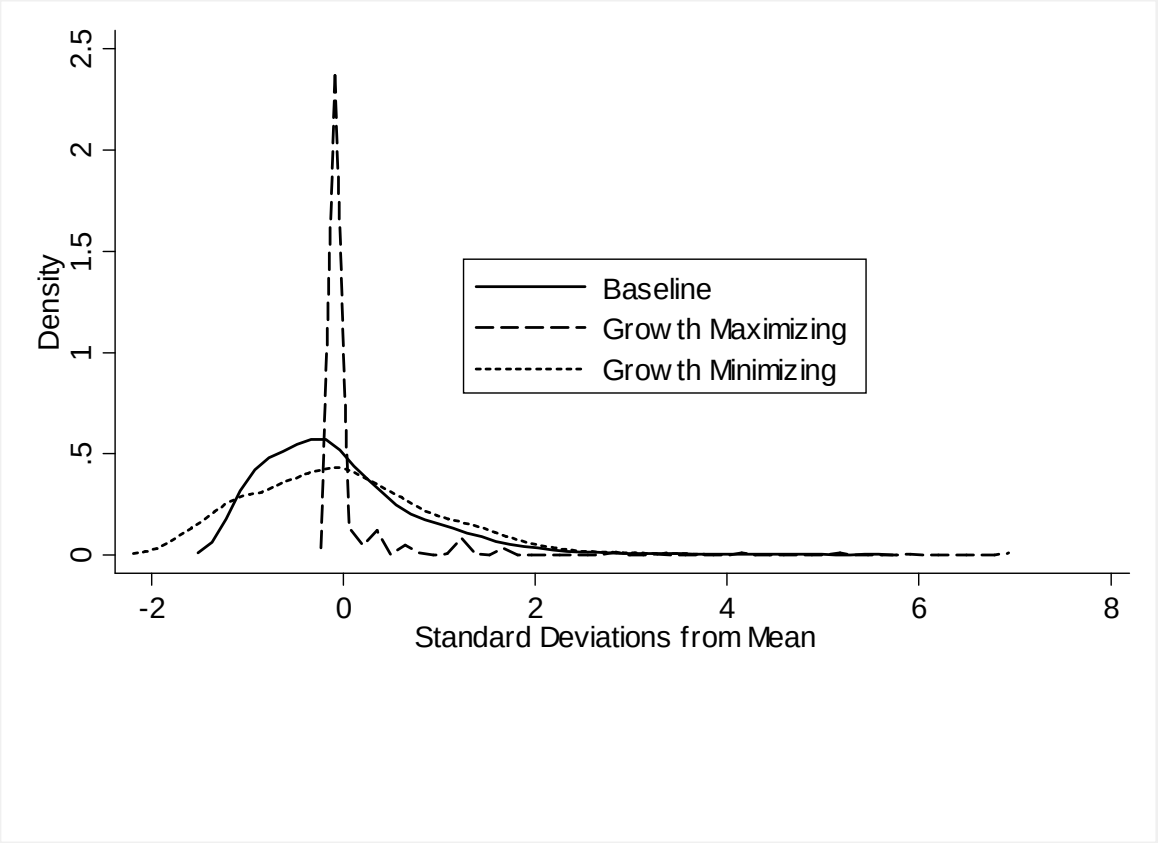

Figure displays densities of transformed test scores in kindergarten under transformations that minimize and maximize the growth of the test gap in the ECLS-K reading assessment. Outlying values beyond seven standard deviations above the mean are not displayed.

Figure 1

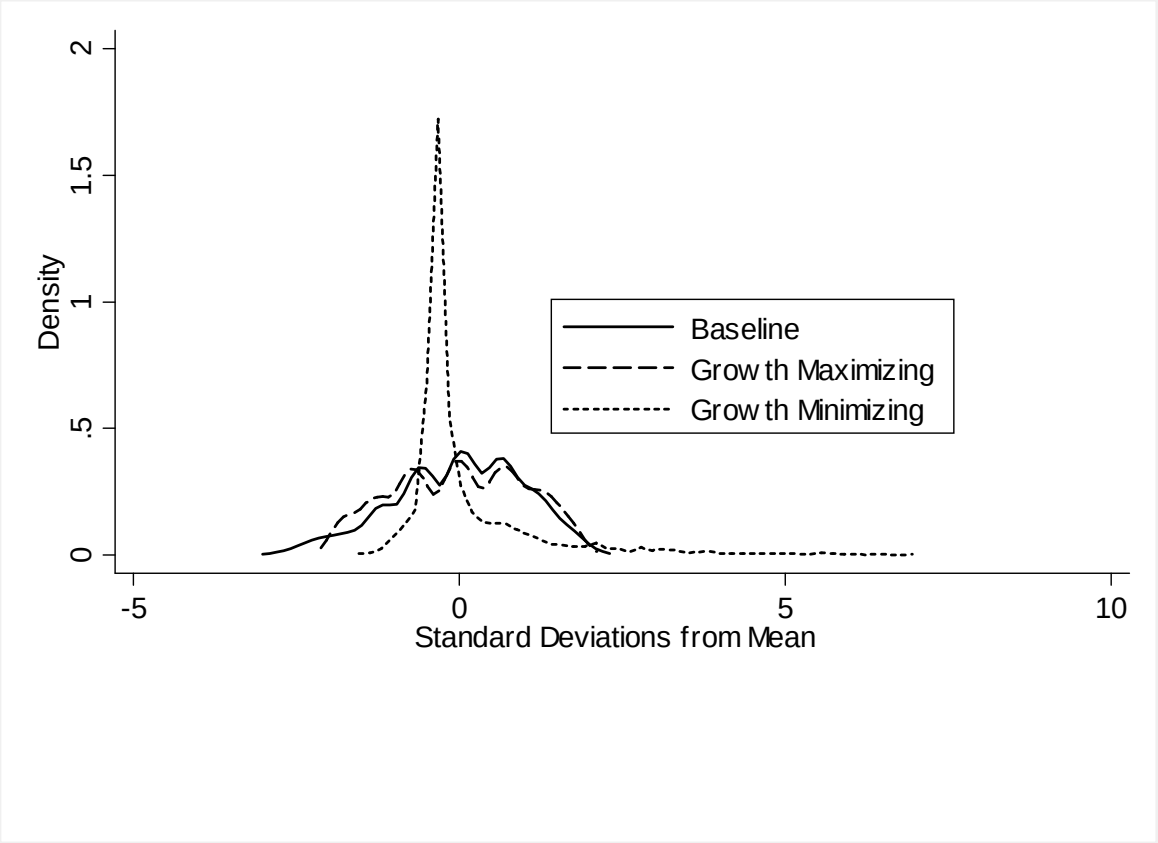

Figure displays densities of transformed test scores in third grade under transformations that minimize and maximize the growth of the test gap in the ECLS-K reading assessment. Outlying values beyond seven standard deviations above the mean are not displayed.

Figure 2

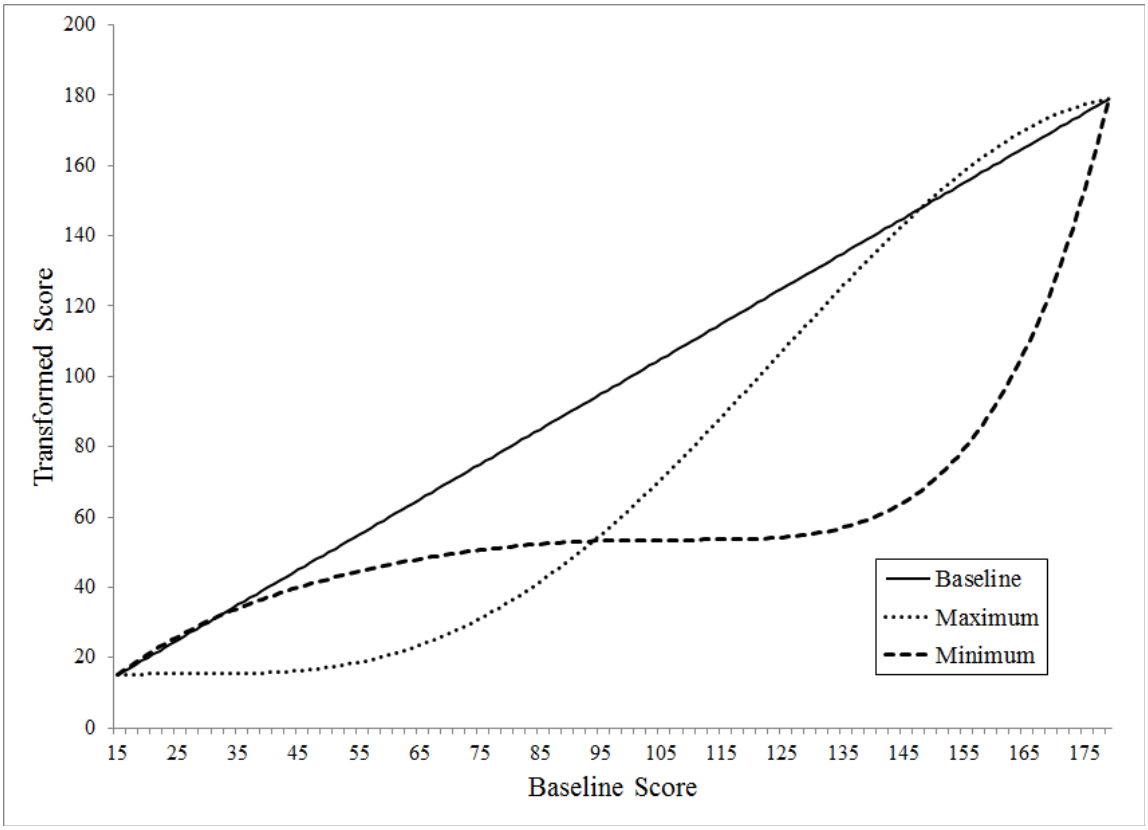

Figure displays transformation functions which minimize and maximize the growth of the test gap from fall kindergarten through spring third grade in the ECLS-K reading assessment. Transformations have been normalized to be over the same range as the original scales.

Figure 3

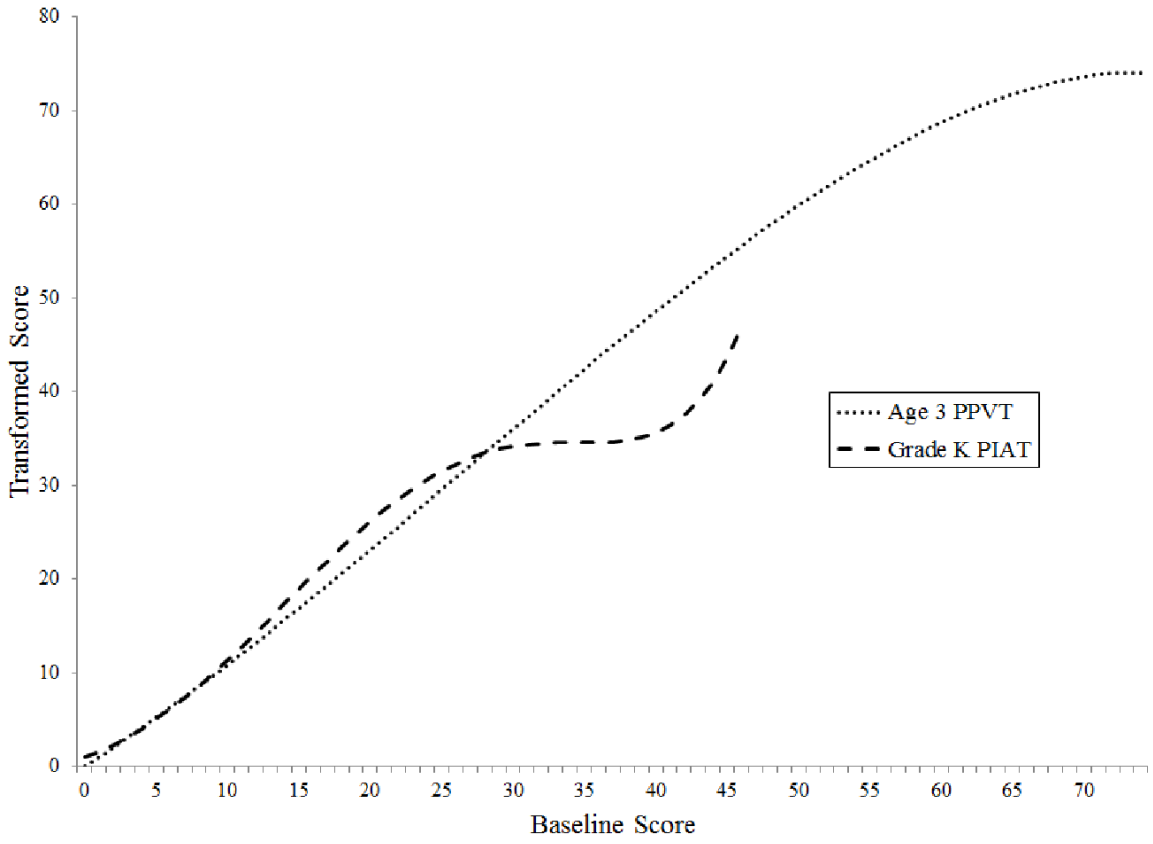

Figure displays transformation functions which maximize the correlation between the PPVT at age 3 and the kindergarten PIAT:RC. Transformations have been normalized to be over the same range as the original scales.

Figure 4

Table 1: CNLSY Descriptive Statistics

	(1)
Kindergarten	
Black	0.37
	(0.48)
Black-White Test Gap	0.25
	(0.04)
Observations	2771
First Grade	
Black	0.38
	(0.49)
Black-White Test Gap	0.42
	(0.04)
Observations	2765
Second Grade	
Black	0.40
	(0.49)
Black-White Test Gap	0.58
	(0.04)
Observations	2825
Third Grade	
Black	0.39
	(0.49)
Black-White Test Gap	0.61
	(0.04)
Observations	2832

Source: Children of the National Longitudinal Survey of Youth. Standard deviations are in parenthesis for variables. Test gaps are for PIAT: Reading Comprehension. Test gaps are measured in standard deviations and standard errors are in parenthesis.

Table 2: ECLS-K Descriptive Statistics

	Bond and Lang (1)	Fryer and Levitt (2)
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 are 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	Minimum	Maximum	Corr Max
	(1)	(2)	(3)	(4)
Kindergarten	0.25***	0.24***	0.05	0.24***
	(0.04)	(0.04)	(0.04)	(0.04)
First Grade	0.42***	0.18***	0.29***	0.29***
	(0.04)	(0.04)	(0.04)	(0.03)
Second Grade	0.58***	0.08***	0.52***	0.26***
	(0.04)	(0.04)	(0.04)	(0.03)
Third Grade	0.61***	0.06	0.63***	0.19***
	(0.04)	(0.04)	(0.04)	(0.03)

Gaps are average white score minus average black score on the PIAT-RC and are measured in standard deviations. Minimum and Maximum are scale transformations that minimize and maximize the growth of the test gap from kindergarten through third grade. "Corr Max" is 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	Minimum	Maximium	Corr Max
	(1)	(2)	(3)	(4)
Kindergarten - Fall	0.40***	0.46***	0.11***	0.47***
	(0.03)	(0.03)	(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 measured in standard deviations. Minimum and Maximum are scale transformations that minimize and maximize the growth of the test gap from fall kindergarten through spring third grade. "Corr Max" is the transformation that maximizes the correlation between the fall kindergarten and spring third grade test. 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	Minimum	Maximum	Corr Max
	(1)	(2)	(3)	(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, and are measured in standard deviations of the ECLS-K reading assessment. 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 first birth, and WIC recipiency. Minimum and Maximum are scale transformations that minimize and maximize the growth of the raw test gap from fall kindergarten through spring third grade. "Corr Max" maximizes the correlation between the fall kindergarten and spring third grade test. 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)	(0.02)	(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 and are measured in standard deviations on the ECLS-K reading assessment. Columns (1) and (2) maximize the percentage change in the ratio of the raw and unexplained test gap. Columns (2) and (4) maximize the change in the difference between the raw and unexplained test gaps. Columns (3) and (4) include controls for SES, number of books in the home, gender, birth weight, indicators for whether the mother was a teenager or over 30 at first birth and WIC recipiency. 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	Minimum	Maximum	Corr Max
	(1)	(2)	(3)	(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 Test 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 and are measured in standard deviations. Minimum and Maximum are scale transformation that minimize and maximize the growth of the test gap from fall kindergarten through spring third grade. "Corr Max" is transformation which maximizes the correlation between the fall kindergarten and spring third grade test. Maximum Test Gap is the test gap that would be observed if all the lowest scores belonged to blacks and all the highest scores belonged to whites.

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

	Fall-K	Spring-K	Spring-1	Spring-3
	(1)	(2)	(3)	(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

Panel A shows the test gaps given each grade's ranking, were that grades scores scaled as they were on the Fall-K and Spring-3 test. Panel B shows the test gaps given each grade's scale were they to have the same ranking as Fall-K and Spring-3. Gaps are average white score minus average black score on the ECLS-K reading assessment and are measured in standard deviations.