Test Score Measurement, Value-Added Models, and the Black-White Test Score Gap

Jeffrey Penney[1]

Department of Economics, Queen's University

penneyj@econ.queensu.ca

version: November 14, 2014

Research in the economics of education literature as to the effect of small class interventions, the usefulness of value-added models, and the size of the Black-White test score gap often comes to contradictory conclusions. Recent literature has affirmed that the source of these contradictions may be due to the fact that test scores contain only ordinal information. In this paper, I propose a normalization of test scores that is invariant to any monotonic transformation. Under fairly weak assumptions, this metric has interval properties and thus solves the ordinality problem. A heretofore unnoticed self-reference problem concerning test score normalization is also assuaged by using this measure. Because of these desirable properties, the measure can be employed to resolve longstanding debates in the economics of education.

Keywords: achievement, black-white test score gap, education, value-added model, race

JEL Classification: C10, H75, I21, J15

[1] Acknowledgements: I would like to thank seminar participants at Queen's University for their comments. I have greatly benefited from discussions with Sergio Firpo, Steven Lehrer, and Morten Nielsen. All remaining errors are my own.

1. Introduction

There are contradictory results in nearly all fields of empirical research, and the economics of education is no exception. In most any subfield in which a relationship between a variable of interest and test scores is examined, conflicting conclusions across different studies continue to arise even in topics that have long seen considerable research effort. For example, some research posits that smaller class sizes yield pedagogical benefits in later grades (Krueger, 1999; Krueger and Whitmore, 2001), while other research claims that the benefits fade very quickly (Hanushek, 1999; Ding and Lehrer, 2010). Another instance of disagreement is with value-added models (VAMs), which is an econometric methodology that is typically employed to obtain estimates of teacher quality proxied via their effect on student test scores. The validity of this research entered the public eye in 2014, when the landmark case Vergara v. California debated, inter alia, whether VAMs provided accurate estimates of teacher quality. Debate continues to rage in this area, with some research supporting the idea (Chetty et al., 2014a) and some claiming that VAMs suffer from serious flaws (Rothstein, 2009; Rothstein, 2014). An answer to the longstanding debate on the dynamics of the black-white test score gap also proves to be elusive. The current majority view is that there is a small gap at entry that quickly grows to be large by third grade (Fryer and Levitt, 2004; Fryer and Levitt, 2006), and that a substantial gap exists in the later grade (Clotfelter et al., 2009). However, some scholars have argued that the test score gap is moderate at kindergarten entry but instead shrinks after first grade (Murnane et al., 2006), or that it is large throughout (Bond and Lang, 2013a). There exist other persistent disagreements in subjects relating to test scores.

Recent research has potentially found a source of some of these contradictions in the literature. It has long been accepted by psychometricians that test scores only have ordinal properties, since these scores are monotonic transformations of some unobserved true measure of ability in a subject (Lord, 1975). Moreover, any monotonic transformation of a test score scale is also a valid scale (Cunha and Heckman, 2008). In light of these facts, Bond and Lang (2013b) perform a bounding exercise on the black-white test score gap. Using an algorithm to generate monotonic transformations of the original test score scale to maximize and then minimize the growth of the test score gap, they find that they find that the bounds they create are almost completely uninformative, and that the results of Fryer and Levitt (2004, 2006) of an increasing gap starting from kindergarten likely reflect scaling decisions.

In this paper, I outline a method to normalize test score scales that is invariant to any monotonic transformation. The primary strength of the proposed metric is that the same results will be obtained as if one had access to the latent true test scores themselves, since the observed test score is a monotonic transformation of the latent true scale. The measure has interval properties and therefore solves the ordinality problem. An additional benefit of this metric is that it solves a self-reference problem which caused difficulties when comparing coefficient magnitudes between different samples that has heretofore gone unnoticed; this issue was a mechanical side effect of employing the usual normalization of transforming test scores into z-scores for analysis. Due to these desirable properties, the proposed measure can be employed to produce comparisons that are valid across different samples and different test score scales. The metric has the additional benefit of being based on unconditional quantile regression (Firpo et al., 2009), which allows investigators to examine any quantile of interest. I illustrate the use of the measure with a value-added model and in examining the black-white test score gap. In the case of the latter, I find that the results at the mean in Fryer and Levitt (2004, 2006) and Clotfelter et al. (2009) to be very similar to the results of this new measure at the median; I additionally find substantial heterogeneity in the black-white test score gap across different quantiles in many of the different grades examined.

This paper is organized as follows. Section 2 places the discussion into context by examining scaling issues related to test scores, and then explains the ordinality and self-reference problems. The proposed measure is outlined in Section 3. The application to value-added models is contained in Section 4, while the application to the black-white test score gap is undertaken in Section 5. The paper concludes with a brief discussion in Section 6.

2. Background

2.1. Scaling and Measurement Issues

Briefly, numbers come in one of four types of scales (Stevens, 1946). The first is a nominal scale, which consists of numbers used to designate particular things, such as 1 for a bus and 2 for a train. The second type of scale is an ordinal scale. As its name implies, order matters. The magnitude of the differences between equal intervals is unknown: for example, consider a happiness scale where a person can respond with a number from 1 to 3 with 1 being "unhappy" and 3 designating "happy". The difference in happiness between 1 and 2 may be different than that of between 2 and 3; this is the case whether or not we assume happiness is a latent continuous variable and every respondent selected the same thresholds for each response. The third is an interval scale, wherein the distance between numbers matters: the difference between 5 and 10 is the same as the difference between 80 and 85. An example is temperature. One does not say that 30 degrees is twice as warm as 15 degrees; for a scale to have this property, it needs to be a ratio scale, the last of the four types of scales [2]. Each subsequent scale includes the listed properties of all the prior listed scales in the order given above; for example, an ordinal scale includes the properties of the nominal scale.

Test scores are subject to rather unique issues relating to measurement. Most agreed-upon scales that are used in the sciences have readily observable effects on the physical environment: weight is a function of force on an object due to gravity, time is expressed in terms of how long it takes for the earth to orbit the sun, and temperature has until recently been defined as the length of a column of mercury. By contradistinction, academic test scores do not have an effect on the physical environment and reference only the test they are measured from; therefore, what they represent is more difficult to quantify. The development of Item Response Theory (IRT) to scale tests, which produces scores that are estimates of the underlying true trait of interest (e.g. mathematical ability), was a step towards more meaningful inference. These test scores allow for both relative and absolute performance

[2] However, ratios of differences are possible with interval scales: for example, one can say that the difference in temperature between 10 degrees and 20 degrees is twice that of between 30 degrees and 35 degrees.

measures; for example, a verbal score of 800 on the GRE verbal indicates an eloquent individual, while a score of 200 would signal a person as inarticulate. However, IRT is not a catholicon, as the scoring scales still only refer to the tests themselves. Comparing scores on different scales which measure the same cognitive trait or ability presents an additional challenge: for example, consider comparing a verbal score of 400 on the SAT with a reading score of 9 on the ACT.

Many of the problems associated with test scaling are at least partially assuaged using test score normalization. The usual practice in analyses involving test scores is to convert them into z-scores (e.g. Fryer and Levitt 2004, 2006; Rivkin et al., 2005; Chetty et al., 2014a); this is done by subtracting the mean from each score and then dividing by the standard deviation. Normalizing test scores in this fashion accomplishes two things. First, it provides concreteness to the test scores since the coefficient estimates in a model using normalized scores describe magnitude in terms of their variability. Second, it allows the results to be compared across tests that measure the same underlying trait but are on a different scale.

It is important to note that transforming a dependent variable into a z-score produces the same parameter estimates (except for the intercept term) as dividing the coefficient estimates after the regression is run by the standard deviation of the dependent variable. The reason is that adding or subtracting a constant to y for every observation only affects the intercept term in a linear regression because the vector of ones for the constant is not correlated with any of the other explanatory variables, and dividing y by a constant simply scales down the coefficient estimates: this can easily be seen by replacing y in the formula for the OLS coefficient estimate of beta with y/k where k is a constant. Moreover, t-statistics (and thus levels of significance) are also the same despite this transformation since the they are also invariant to these changes in y for the same reasons[3]. Therefore, the step at which normalization occurs is immaterial for inference.

[3] Since the correct t-statistic and coefficient estimate can be obtained ex post, so can the standard error of the normalized coefficient estimate.

2.2. The Ordinality Problem

Nearly all research in economics and education that uses test scores as a dependent variable implicitly makes the following assumptions: (i) there exists an unobservable score A that represents ability in a subject; (ii) A has interval properties; (iii) observable test scores T have interval properties since they are an affine transformation of unobserved ability A: $T = mA+b$, where m and b are parameters. The first two assumptions are uncontroversial and are in agreement with the psychometric literature. However, the third is potentially problematic because psychometricians almost universally assume instead that T is a monotonic (rather than affine) transformation of A; that is, $T = f(A)$ for some unknown monotonic function f (Lord, 1975). This belief is partially based on the fact that the IRT test scores are not uniquely identified: for any set of estimated test scores T, any arbitrary monotonic transformation of these test scores $g(T)$ produces scores that fit the IRT model with the same likelihood; therefore, latent ability in a subject or skill A cannot be identified.

TABLE 1 here

Table 1 illustrates the ordinality problem. Three different values of the latent true test score A are listed, and there is a constant difference between each step. Suppose that a test is created to measure the underlying trait that A represents. Under one monotonic transformation, the first step is larger than the second; with the other, the second step is larger than the first. While each monotonic transformation preserves ordinality, the interval properties dissipate; even if the monotonic transformation f were affine, it would be impossible to determine whether this was the case. Both sets of scores would fit with the same likelihood.

2.3. The Self-reference Problem

The "self-reference problem" refers to the fact that the magnitude of a coefficient in a regression whose dependent variable is normalized by some function of the variance is affected by the variance of the explanatory variables (such as group membership dummies). Recall that in the data generating process (DGP) of a variable of interest, its unconditional variance can be decomposed into two separate factors: (i) the variance due to the characteristics of the observation, and (ii) the variance due to the randomness of the unexplained component. In terms of what could be estimated given the data, we could think of (i) as the set of observable characteristics, and (ii) as the unobserved factors and noise. When the unconditional variance is used to normalize a variable, one is implicitly making use of the sum of both of these components; in particular, the variance of the covariate of interest itself interacts with its magnitude in a way that can affect inference. I call this the "self-reference problem", which I illustrate below.

Take for example a demographic test score gap. Consider the model with two groups, P and R. Suppose the DGP is

$$testscore = \beta_0 + \beta_1 D + e \tag{1}$$

where D is a dummy variable indicating membership in group R, *testscore* is a scaled test score, and e is the usual error term that is uncorrelated with D. Therefore, the test score gap between groups P and R is β_1. Let there be two samples, I and II. In sample I, P is 95% of the sample membership and R is 5%. In sample II, P is 80% of the sample membership and R is 20%. If we express the test score gap using the scaled test score, both samples will produce the same OLS estimate for β_1 asymptotically. However, if both samples normalize the test score into a z-score, the coefficient for β_1 will be smaller in absolute value for sample II despite the scaled score gap being exactly the same; this is because the unconditional variance of the test score in sample II is larger. Of course, it would be misleading to claim that the test score gap between groups P and R is smaller in sample II simply due to differences in the proportion of each group.

3. Methodology

The metric proposed herein employs the ordinary least squares variation of unconditional quantile regressions[4] as developed by Firpo, Fortin, and Lemieux (2009) to estimate the test score gap at a given quantile (such as the median) and then normalizes the coefficients of

[4] There are three different methods to estimate it; Firpo, Fortin, and Lemieux (2009) provide evidence that they all yield substantially the same results. The OLS variation is the one typically used in practice.

interest by dividing them with the standard error of the regression; this is in contrast to the usual method, which instead normalizes the coefficients by dividing them by the standard deviation of the dependent variable. Unconditional quantile regressions possess a number of useful invariance properties: for example, t-statistics and R-squared values are invariant to monotonic transformations of the dependent variable.

This measure solves the ordinality and self-reference problems outlined in the previous section. By using the standard error of the regression to normalize the coefficients rather than the standard deviation of the test scores, the problem of coefficient magnitudes being potentially influenced by the variability in the explanatory variables is sidestepped. This measure's invariance to monotonic transformations means that the same regression results will be yielded as if one had access to the true set of test scores. Since this methodology produces estimates at given quantiles, it implicitly relaxes the assumption that the effect of interest is the same at every quantile of the distribution of test scores.[5]

I now provide a formal proof of the invariance property. Recall that an unconditional quantile regression transforms the response variable y as

$$IF(y; q_\tau, F_y) = (\tau - 1[y \leq q_\tau])/f_y(q_\tau) \equiv \tilde{y} \tag{2}$$

where τ is the quantile of interest, q_τ is the value of y at the quantile τ, $1[\cdot]$ is an indicator function taking the value of 1 if the statement in the square brackets is true and 0 otherwise, and $f_y(q_\tau)$ is the density of y at q_τ.[6] Observe that the key to many of the invariance properties of unconditional quantile regressions is the indicator function.

Lemma 1. *The term $1[y \leq q_\tau]$ is the same for any monotonic transformation of y.*

Proof. Define $y^* = g(y)$ and let g be monotonic. Then, $\tau = Pr[y^* \leq j^*] = Pr[g(y) \leq j^*] = Pr[y \leq g^{-1}(j^*)]$. Thus, $j^* = g(j)$, and therefore $Pr[y^* \leq j^*] = Pr[y \leq g^{-1}(g(j))] = Pr[y \leq j]$, hence $j = q_\tau$. □

[5] Of course, the mean effects estimated in the literature could also be interpreted as the average effect of the covariate of interest across the distribution.

[6] Note that the influence function used here is not recentered. The use of the influence function rather than the recentered variation is used to considerably simplify the proof of Theorem 1. Using the recentered influence function (RIF) in place of the influence function (IF) does not affect the results since q_τ is a constant element for all $\tilde{y}$; because of this, its only influence on the coefficient estimates is on the constant term, and thus it can be ignored.

With this in hand, the invariance property of the proposed measure can now be proven.

Theorem 1. *In an unconditional quantile regression model without a lag dependent variable, the ratio of any regression coefficient to the standard error of the regression is invariant to any monotonic transformation of the dependent variable.*

Proof. Recall that $\tilde{y}$ is the transformed value of y by the IF. By Lemma 1, $1[y \leq q_\tau]$ is invariant to monotonic transformations. What is left to show is that the ratio $\hat{\beta}/s$ is invariant to changes in $f_y(q_\tau)$. The value of the ratio is

$$\hat{\beta}/s = (X^\top X)^{-1} X^\top \tilde{y} / \sqrt{\tilde{y}^\top M_x \tilde{y} (n-k)^{-1}} \tag{3}$$

where n is the number of observations and k is the number of estimated parameters. Note that $f_y(q_\tau)$ scales the values of $\tilde{y}$ by a constant factor. Suppose a monotonic transformation of y takes place, $y^* = g(y)$ where g is monotonic, and thus $\tilde{y^*} = \theta \tilde{y}$. The ratio is thus

$$\begin{aligned}
\hat{\beta^*}/s^* &= (X^\top X)^{-1} X^\top \tilde{y^*} / \sqrt{\tilde{y^*}^\top M_x \tilde{y^*} (n-k)^{-1}} \\
&= (X^\top X)^{-1} X^\top \theta \tilde{y} / \sqrt{\theta \tilde{y}^\top M_x \theta \tilde{y} (n-k)^{-1}} \\
&= (X^\top X)^{-1} X^\top \tilde{y} / \sqrt{\tilde{y}^\top M_x \tilde{y} (n-k)^{-1}} = \hat{\beta}/s
\end{aligned}$$

where the third equality holds since θ is a scalar. Therefore, the ratio of the unconditional quantile regression coefficient estimate to the standard error of the regression is invariant to any monotonic transformation of y. □

Note that the invariance property holds for *any* quantile of interest. In the case where a lagged dependent variable is present on the right-hand side, the invariance property holds only asymptotically. Because of the unknown monotonic transformation of the response variable, it is necessary to include a polynomial expansion of the lagged response variable in order to approximate the unknown form of its transformation.

Theorem 2. *In an unconditional quantile regression model with a lag dependent variable,* *the ratio of any non-lag regression coefficient to the standard error of the regression is asymptotically invariant to any monotonic transformation of the dependent variable.*

Proof. Specify the untransformed and transformed unconditional quantile regressions as

$$\frac{y_1^*}{\sigma_1} = \alpha_0 + \alpha_1 X + f(\alpha_2, y_1') + \epsilon_1 \tag{4}$$

$$\frac{y_2^*}{\sigma_2} = \beta_0 + \beta_2 X + g(\beta_2, y_2') + \epsilon_2 \tag{5}$$

where f and g are polynomial expansions of the response variable. Recall that $y_2' = h(y_1')$ for some function h that is monotonic and continuous. Therefore, f can be well-approximated by the polynomial expansion $g(\beta_2, y_2') = \beta_{21} y_2' + \beta_{22} y_2^{2\prime} + \ldots + \beta_{2k} y_2^{k\prime}$. Recall that y_2^*, the transformed value of y_2 by the influence function, is related to the transformed value of y_1 by $y_2^* = Ay_1^*$ and similarly $\sigma_2 = A\sigma_1$ for some unknown linear scaling factor A. Thus, estimating (5) is equivalent to estimating

$$\frac{y_1^*}{\sigma_1} = \beta_0 + \beta_1 X + g(\beta_2, y_2') + \epsilon_2. \tag{6}$$

Let My_1' be a projection matrix that projects off the space spanned by the polynomial expansion of y_1. By the Frisch-Waugh-Lovell Theorem (1933, 1963), the estimates of α_1 of the regression

$$My_1' \frac{y_1^*}{\sigma_1} = \alpha_0 + \alpha_1 My_1' X + u_1. \tag{7}$$

will be identical to those in equation (4). Therefore, the estimate of β_1 in

$$My_1' \frac{y_1^*}{\sigma_1} = \beta_0 + \beta_1 My_1' X + My_1' g(\beta_2, y_2') + u_2 \tag{8}$$

will be the same as the estimate of α_1 in (4) if we have $\sum_n My_1' x_i j u_2 = 0$: that is, the error of approximation $f(\alpha_2, y_1') - g(\beta_2, y_2')$ is uncorrelated with the residuals from the linear relationship between y_1' and X. Since the former is a mathematical object that is asymptotically uncorrelated with the later, β_1 in (5) is equal to α_1 in (4) asymptotically. □

4. Application: Value-Added Models

The goal of valued-added models (VAMs) is to measure teacher quality by examining the test score gains of their students during their tutelage. It is somewhat controversial in that it is exclusively based on test scores, and many scholars criticize it on this front (e.g. Ewing, 2011; Koretz, 2008). Nonetheless, these models have allowed education economists to come to substantive policy conclusions. For example, using data from Texas public schools, Rivkin et al. (2005) find that a class size reduction of ten students yielded a smaller benefit than a one standard deviation increase in teacher quality. Evidence has begun to accumulate that students who learn under high quality teachers (as defined by having high value-added) experience more desirable outcomes in adulthood, such as earning higher salaries and being more likely to attend college (Chetty et al., 2014b). However, these findings have recently been challenged (Rothstein, 2014). Analysis concerning value-added has thus far been exclusively conducted under the assumption that test scores have interval properties.

The currently preferred model of student achievement is the VAM2 model using the terminology of Rothstein (2010). It is parameterized as follows:

$$T_{it} = \alpha + D_i\beta + \gamma T_{i,t-1} + \epsilon_{it} \tag{9}$$

where T_{it} is the student's test score in grade t, D_i is a J vector of dummies indicating which teacher a student was assigned to, and $T_{i,t-1}$ is the student's test score in the previous grade. The interpretation of the coefficient on β for the j^{th} teacher is "Controlling for previous inputs, how much did the average student of teacher j learn this year?", i.e. the value-added of that teacher. The VAM1 model drops the lagged term on test scores and replaces the response variable with $\Delta T_{it} \equiv T_{it} - T_{i,t-1}$, while the VAM3 model is identical to the VAM1 model but adds student fixed effects on the right-hand side.

The VAM2 model is preferred for several reasons. Including a lagged test score as a control can be thought of as a sufficient statistic for past inputs; that is, the number provides an accurate summary of what the student has experienced in the prior periods. In their measure of value-added, Chetty et al. (2014a) show that controlling for past scores is necessary in order to obtain unbiased estimates. The primary downside to this specification is that the effect of past inputs is assumed to decline at a common geometric rate (Todd and Wolpin, 2003). With reference to the methodology proposed herein, the VAM2 specification is the only option, since VAM1 and VAM3 models can change the ranking of teacher value-added (Penney, 2014).

4.1. Data

The data employed in this section come from a cohort of students that participated in Project STAR, an experiment that took place in Tennessee that ran from 1985 until 1989. The primary goal of the experiment, as its acronym (Student-Teacher Achievement Ratio) implies, was to determine the effect of class size on student achievement in primary education (Finn et. al., 2007). A cohort of over 6000 students from 79 schools took part in the experiment. To qualify for participation, schools required sufficient enrollment to support at least three different classes per grade. Both students and teachers were randomly assigned to three different class types, but this randomization took place within schools only. The class types were as follows: a small class (13 to 17 students), a regular class (22 to 25 students), or a regular class with a full-time teacher's aide. Compliance was nearly perfect in kindergarten, with approximately 0.3% of students enrolled in a class type that was not assigned to them. However, in first grade and beyond, there were some problems with noncompliance, with a number of students switching in or out of small classes. Noncompliance was primarily due to parental complaints or discipline problems. At the end of each academic year, all participating students were given a battery of academic and non-academic tests. More detailed overviews of Project STAR can be found in Krueger (1999) and Finn et. al. (2007). The Project STAR dataset was selected due to its experimental protocol, so that there would be no fears of bias due to non-random sorting in the teacher value-added estimates.

4.2. Application

In this section, I compare the results of a VAM2 model estimated by OLS using the z-score transformation with those of a VAM2 model that also includes polynomial terms on the lagged test score that is normalized using the proposed metric. The regression I will run for the later is

$$T_{i1} = \alpha + D_i\beta + \gamma_1 T_{i,k} + \gamma_2 T_{i,k}^2 + \gamma_3 T_{i,k}^3 + \gamma_4 T_{i,k}^4 + \epsilon_{it} \tag{10}$$

where again the vector β contains the estimates of teacher value-added. The test score T_{i1} is grade 1 test scores in mathematics. The fourth-order polynomial in kindergarten mathematics test score is employed as the control for past inputs.

The performance of three different teachers are evaluated and contrasted using both estimation methods on the test score data; in addition, regressions results are also be produced for two transformations of the original data: a square of the original test score, and a square root of the original test score. Tables 2 and 3 below display the results of this exercise.

TABLE 2 here

TABLE 3 here

Despite different transformations of the test score, all regressions give approximately the same value for the UQR case, while the OLS normalization produced estimates that are much more variable. The monotonic transformations in the example are rather tame – more extreme transformations would further display the robustness of the proposed metric relative to OLS.

5. Application: Black-White Test Score Gap

The subject of gaps in various outcomes between demographic groups remains a contentious issue that has attracted considerable attention from academics and policymakers for many decades. One of the most examined gaps in the education and economics literatures is the black-white test score gap (e.g. Bond and Lang, 2013a; Bond and Lang, 2013b; Clotfelter et al., 2009; Fryer and Levitt, 2004; Fryer and Levitt, 2006). It is widely believed that reducing this gap in test scores is an important step to promoting racial equity in outcomes such as crime, health, and family structure (Dee, 2005).

The seminal papers in this literature are the works of Fryer and Levitt (2004, 2006). Using a nationally representative dataset they find that, inter alia, the black-white test score gap is almost nonexistent once a small number of controls are taken into account; blacks were even found to have an advantage over whites in reading test scores at school entry. However, test score gaps began to emerge by the end of kindergarten, and by third grade, the gaps in both mathematics and reading were quite considerable even using controls. Clotfelter et al. (2009), using administrative data from North Carolina, show large persistent gaps between blacks and whites from third through eighth grade. While there were studies that found contradictory results (Murnane et al., 2006), most of the literature on the black-white test score gap finds it to exhibit these patterns.

Bond and Lang (2013b) have recently called this entire body of research into question. They argue that, once the ordinality of test scores is taken into account, the black-white test score gap can vary between "there is a small gap in kindergarten that declines thereafter" to "there is no gap in kindergarten but the gap grows to be significant". These results are obtained using a bounding exercise in which test scores are subject to various monotonic transformations in order to find their growth maximizing and minimizing evolutions. These results lead the authors to claim that the dynamics of the gap in the literature largely reflect test score scaling decisions.

The debate as to the evolution of the black-white test score gap, then, requires a metric that is invariant to monotonic transformations. In this section, I employ the proposed measure in an attempt to uncover the true evolution of the test score gap.

5.1. Data

This research employs the Early Childhood Longitudinal Study Kindergarten Cohort (herein "ECLS-K") is a nationally representative survey of 21,260 children who entered kindergarten in the autumn of 1998. This is the same dataset used by other prominent papers in the field of test score gaps (e.g. Bond and Lang, 2013b; Fryer and Levitt, 2004; Fryer and Levitt, 2006). It contains a wide breadth of information on topics ranging from the emotional health and wellbeing of children to various kinds of test scores. The ECLS-K surveyed students, parents, and educators such as teachers and school administrators. Data were collected in the spring and fall of kindergarten, the spring and fall of first grade, and in the springs of third, fifth, and eighth grade.

The analysis of this paper employs the Fryer-Levitt controls, which are as follows: race, socioeconomic status, gender, age, whether the mother of the child was a teenager when she first gave birth, whether the mother of the child was 30 or over when she first gave birth, WIC participation, and the number of children's books in the home. This set of controls was found to produce estimates of the corrected racial test score gaps that were largely similar with a much more fully specified model that included 94 control variables (Fryer and Levitt, 2004). The racial categories are broken down into white, black, Hispanic, Asian, and other. The "white" category refers exclusively to non-Hispanic whites, and the "other" category includes Native Americans, native Alaskans, and other racial backgrounds that do not fall into those previously listed. The socioeconomic status measure is continuous variable that is a function of the education levels of the parents or guardians, the occupations of the parents or guardians, and household income.[7] WIC participation is a dummy variable indicating enrollment in the Special Supplemental Nutrition Program for Women, Infants, and Children, which is a program targeted to low-income mothers and children.

TABLE 4 here

The summary statistics of the data are listed on Table 4. White children tend to be roughly a year older than children of other races when they first enter kindergarten; since test-at-age effects are quite strong early in a child's life, not correcting for age will overestimate the test score gap between whites and other minorities. Blacks and Hispanics are more than twice as likely as whites to participate in the WIC nutrition program. Black and Hispanic mothers tend to have their children 3 to 4 years earlier than whites on average; moreover, over half of black mothers in the sample began their families as teenage mothers, while almost 40% of hispanic mothers did the same. Whites tend to have many more children's books in the home than the other races.

The ECLS-K dataset includes several kinds of math and reading test scores. Those employed in this article are the longitudinal Item Response Theory (IRT) test scores, which

[7] Details about the construction of the variable can be found starting on page 7-8 of the ECLS-K User Guide.

were designed to be fully comparable across grades. These test scores were constructed using a Bayesian three-parameter IRT model. One of the primary benefits of the IRT approach is that it is able to yield precise estimates of the underlying latent trait of interest despite a small number of test questions. Moreover, use of the Bayesian variation minimizes possible issues related to test score shrinkage.[8] The testing procedure was adaptive in that routing items were used to give students tests that were commensurate with the likely range of their ability, which should minimize floor and ceiling effects as well as maximize the accuracy of the measurement. The test scores were found to have a very high level of reliability. The reliability score is defined as $1 - \psi/\theta$, where ψ is the within-person variance estimate of underlying ability and θ is the between-person estimate of the underlying ability (the total between-person variance of the posterior mean). A value of 1 indicates that latent ability is perfectly measured, while a value of zero signals that there is no information about the latent trait in the data. The various reliability scores for the math and reading tests for each year can be found below on Table 5. All but one of the values exceed 0.9 and some even exceed 0.95, which should assuage measurement error concerns.

TABLE 5 here

It is important to note that the IRT test scores were recalibrated in every round, thus estimates of the test score gaps across different studies that do not use the same iterations of the ECLS-K data may be dissimilar to a small degree. The recalibration is necessary since the IRT scale scores in the database represent estimates of the number of items children would have answered correctly at each point in time if they had answered all of the questions from each round: thus, the scaled score at time t is equal to the sum of the probabilities of a correct answer for every question in the database. Therefore, as new questions are added in successive waves, the scaled score for each child at every grade will increase if they are estimated to have a non-zero probability of answering at least one of the new questions correctly.

TABLE 6 here

[8] Test score shrinkage is occasionally a threat to the validity of results, for example, see Bond and Lang (2013a) who implement a procedure to correct for it.

The test score summary statistics are displayed on Table 6. To maintain comparability with the other literature on this subject, the test scores were converted into z scores.[9] At the population level, there exists an increasingly large gap between blacks and whites for both reading and math test scores from kindergarten to eighth grade.

The criteria for inclusion in the dataset for this analysis is as follows. Students missing data on the Fryer-Levitt controls or do not have at least one valid test score are dropped from the sample.[10] For the variables that are time-variant, the ones used in the analysis are those from the fall of the child's kindergarten year. Unlike Fryer and Levitt (2004, 2006), students who are missing some waves of test score information are not dropped from the sample; I do not follow the same practice because attrition in the 5th and 8th grade is quite high relative to the earlier years. Nonetheless, replication exercises give similar results, so the samples should be quite comparable.

The ECLS-K contains an extensive selection of sample weights, with the suggested weight varying depending on the unit of analysis; see section 4-3-1 of the ECLS-K User Guide for more information. I do not use sample weights in this analysis as the results are not sensitive to their use; this was also found to be the case in the previous research by Fryer and Levitt (2004, 2006) on this subject.

5.2. Application

The regression equation estimated is

$$T_{it} = \beta_0 + \rho_i \beta_1 + X_{it}\beta_2 + \epsilon_{it} \tag{11}$$

where T_{it} is the normalized test score of individual i at time t, ρ is a vector of racial dummies (black, Hispanic, Asian, other), X_{it} is a vector containing the Fryer-Levitt controls, and ϵ_{it} is the usual error term. Non-Hispanic whites are the baseline racial category, therefore all test score gaps are relative to their performance.

It is important to note that, because race cannot be changed, the β_1 coefficients require

[9] By Theorem 1 below, this normalization does not affect the analysis.

[10] Analysis of the math and reading test scores in the fall of first grade are excluded because only a small portion of the sample took this particular test.

a slightly different interpretation than usual. Chernozhukov et al. (2013) show that a dummy coefficient in an unconditional quantile regression is a first order approximation to the following expression:

$$inf_{y \in Y}[\int_{\chi_1} F_{Y_0|X_0}(y|x) dF_{X_1}(x) \leq \tau] - inf_{y \in Y}[\int_{\chi_0} F_{Y_0|X_0}(y|x) dF_{X_0}(x) \leq \tau] \qquad (12)$$

where y is an arbitrary value of the response variable, Y is the set of values y can take, Y_i is the set of values for group $i = 0, 1$, x is an arbitrary set of values for the set of control variables, X_i is the set of characteristics for group i with χ_i as its support, F is a cumulative distribution function, and τ is the quantile of interest. Thus, the first term of this difference is the counterfactual quantile τ for the group of interest, while the second term is the definition of the quantile of the baseline group. Defining $X_1 = X_0$ for all controls except the dummy, we can interpret the dummy variables (11) as the difference at quantile τ if the group of interest had the same distribution of coefficients as the baseline group. In the case of this paper, a coefficient β on black at the quantile τ indicates the difference between the blacks and whites at that quantile of their distributions of test scores if the former had the same distribution of control variables as the latter. Applications of Firpo et al. in the context of gender gaps can be found in Boudarbat and Connolly (2013) and Fortin et al. (forthcoming).

FIGURE 1 here

FIGURE 2 here

Figures 1 and 2 display graphs containing information on the unconditional quantile estimates of the proposed metric and their confidence intervals, as well as OLS estimates for comparison purposes. In order to ensure comparability with the extant literature, the test scores in the OLS regression have been converted to z-scores [11]

Comparing the OLS estimates with the quantile regression results at the median, there is a remarkable degree of consistency between the two measures, despite the fact that they were constructed in different fashions. The similarity of the results suggests that research

[11] Recall that for the unconditional quantile regressions, regression (11) will yield the same result regardless of how the test scores are monotonically transformed.

using the ECLS-K math and reading test score data may be more robust to scaling concerns than what was originally thought.

The black-white mathematics test score gap increases from kindergarten until third grade for most of the distribution, and then remains roughly constant afterwards: black students across the distribution do not regain much if any lost ground by eighth grade. However, those near the top of the distribution of test scores experience little if any growth of the gap in test scores at all, with it remaining approximately 0.1 standard errors throughout. For those near the bottom of the distribution, they fall significantly behind by third grade: those under the 4th decile experience a gap of almost 0.4 standard errors at this point. While the gap at different deciles is fairly even in kindergarten and first grade, by the spring of third grade, the gap is larger for those near the bottom and shrinks as we approach the top of the distribution.

Reading scores for blacks are actually superior to whites when entering school, and this is true across the distribution. By the spring of the first year, a small gap forms for about the bottom half of the distribution, while those at the top of the reading distribution experience no statistically significant gap associated with their race compared to whites. By fifth grade, those under the 5th decile experience larger gaps than those above it. Blacks at the top of the distribution only have a small gap by this point of about 0.1 standard errors, while those near the bottom have only fallen further behind with a gap of roughly 0.4 standard errors.

The results of this analysis are largely in agreement with other literature on the subject. Recall that Fryer and Levitt (2004) examines kindergarten and first grade, Fryer and Levitt (2006) covers kindergarten through third grade, and Clotfelter et al. investigate third through eighth grade.[12] Fryer and Levitt (2006) observed a widening of the black-white test score gap in mathematics in the first four years of school, which I also document here. The test score gap in mathematics appears to stabilize at third grade and remain the same all the way up until at least eighth grade, which is the same pattern that was observed in Clotfelter et al. (2009). I also document the initial advantage for blacks in reading that turns into a deficiency by third grade. The gap is approximately constant in third and fifth grade, but

[12] However, the latter use administrative data from North Carolina rather than a nationally representative sample such as the ECLS-K.

widens for the eighth grade test score, which is mostly in agreement with Clotfelter et al. (2009).

In conclusion, there is substantial distributional heterogeneity in the gaps over time. The dynamics of the gaps at the median mostly agree with those of the literature at the mean[13]. The gaps at the tops of the distributions in nearly all grades show a gap of zero or close to zero; this suggests that there is no racial test score gap between whites and blacks at the top of the ability distribution.

6. Discussion

In this paper, I proposed a method to normalize test scores that is invariant to monotonic transformations. In addition, the metric assuaged a problem with z-score normalization which caused the parameter estimates to be sensitive to the sample variability in the control variables. Because of these desirable properties, the metric could be used to compare results across different datasets. I applied the metric to a value-added model and to the debate surrounding the black-white test score gap. For the latter, I concluded that the results of the literature (Fryer and Levitt, 2004; Fryer and Levitt, 2006; Clotfelter et al., 2009) were largely correct despite not taking into account the ordinality problem of test scores. Substantial heterogeneity in the distribution of these gaps was found: most of the test score gap between blacks and whites appears to be driven by poor-performing black students at the bottom of the ability distribution. Blacks at the top of the ability distribution experience no economic nor statistically significant gap with whites in most cases. The metric developed in this paper should prove useful in resolving debates that are primarily based on test scores in the economics of education literature.

The use of this proposed metric comes with an important caveat. While it is an effective tool to compare results across different tests, the tests must be measuring the same underlying factor in order for comparisons to be meaningful. For example, Murnane et al. (2006) employ the National Institute of Child Health and Human Development dataset (NICHD) and show that the racial test score gap between blacks and whites holds steady for reading

[13] This analysis was repeated for the other races (asians and hispanics) and the results were also found to be consistent with the literature. A discussion of these results was omitted for reasons of space.

and decreases by almost half for mathematics from kindergarten to third grade, which are results that do not agree with Fryer and Levitt (2006). However, the tests used in the NICHD cover basic skills, while the ECLS-K focuses on subjects learned in school; therefore, the results are not comparable with each other. Discussion of demographic test score gaps, like any other subject of interest, necessitates that they be clearly defined.

An alternative method to measure gaps is Oaxaca-Blinder decomposition. This methodology is most commonly used to determine the extent of possible discrimination in the gender wage gap literature. The procedure decomposes an observed gap into two parts: an explained gap due to observable characteristics, and an unexplained gap that is present due to differences in the coefficients and in the unobservables. It has a desirable technical property of being "doubly robust", i.e. it is consistent if either the propensity score assumption or the model for outcomes is correct (Kline, 2011). The procedure outlined in this paper can be suitably modified for Oaxaca decomposition analysis. However, it has recently been argued that using a dummy variable to measure a gap between two groups is not only sufficient, but it may also possess advantages over Oaxaca-Blinder decomposition (Elder et al., 2010).

An alternative method to measure demographic test score gaps is to anchor them to adult outcomes (e.g. Bond and Lang 2013a, Cunha and Heckman, 2008; Cunha et al., 2010). Such an approach is not a panacea. Some results may be sensitive to distributional assumptions, such as whether to anchor test scores with either earnings or log earnings (Bond and Lang, 2013b). Adult outcomes come with significant delays (Barlevy and Neil, 2012), which may limit the policy relevance of the results, or may fail to materialize altogether, leading to possible selection bias. Moreover, the extent to which unobservable heterogeneity in tastes affects the relationship between test scores and adult outcomes is currently unknown: for example, educational attainment may be lower purely due to the desire of a person to pursue a trade such as electrician before completing high school despite having the cognitive skills necessary to complete higher education. Earnings may be lower due to occupational choice, such as the desire to become a social worker. In these two particular examples, the former yields high earnings and a low level of education, while the latter vice versa. Nonetheless, the primary and significant advantage to anchoring is that the gaps are expressed in concrete units, such as completed years of education; magnitudes expressed in such a metric are much more easily understood than the measures typically used in the test score gap literature. These units also have the very desirable property of being ratio scales. I view the anchoring literature as complementary to the approach advanced here for purposes of test score analysis.

References

[1] Barlevy, Gadi, and Derek Neal. (2012) "Pay for Percentile," *American Economic Review*, vol. 102: 1805-31.

[2] Bond, Timothy N., and Kevin Lang. (2013a) "The black-white education-scaled test-score gap in grades K-7," NBER working paper 19243. July.

[3] Bond, Timothy N. and Kevin Lang. (2013b) "The Evolution of the Black-White Test Score Gap in Grades K-3: The Fragility of Results," *Review of Economics and Statistics*, vol. 95: 1468-1479.

[4] Boudarbat, Brahim, and Marie Connolly. (2013) "The gender wage gap among recent post-secondary graduates in Canada: a distributional approach," *Canadian Journal of Economics*, vol. 46:1037-1065.

[5] Chernozhukov, Victor, Iván Fernández-Val, and Blaise Melly. (2013) "Inference on Counterfactual Distributions," *Econometrica*, vol. 81: 2205-2268.

[6] Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. (2014)"Measuring the Impacts of Teachers I: Evaluating Bias in Teacher Value-Added Estimates," *American Economic Review*, vol. 104: 2593-2632.

[7] Chetty, Raj, John N. Friedman, and Jonah E. Rockoff. (2014)"Measuring the Impacts of Teachers II: Teacher Value-Added and Student Outcomes in Adulthood," *American Economic Review*, vol. 104: 2633-2679.

[8] Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor. (2009) "The Academic Achievement Gap in Grades 3 to 8," *Review of Economics and Statistics*, vol. 91: 398-419.

[9] Cunha, Flavio, and James J. Heckman. (2008) "Formulating, Identifying and Estimating the Technology of Cognitive and Noncognitive Skill Formation," *Journal of Human Resources*, vol. 43: 738-782.

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

[11] Dee, Thomas S. (2005) "A Teacher Like Me: Does Race, Ethnicity, or Gender Matter?" *American Economic Review Papers and Proceedings*, vol. 95: 158-165.

[12] Ding, Weili, and Steven F. Lehrer. (2010) "Estimating Treatment Effects from Contaminated Multiperiod Education Experiments: The Dynamic Impacts of Class Size Reductions," *Review of Economics and Statistics*, vol. 92: 31-42.

[13] Elder, Todd E., John H. Goddeeris, Steven J. Haider. (2010) "Unexplained gaps and OaxacaBlinder decompositions," *Labour Economics*, vol. 17: 284-290.

[14] Ewing, John. Mathematical Intimidation: Driven by the Data. Notices of the AMS. May 2011: 667-673.

[15] Finn, Jeremy D., Jayne Boyd-Zaharias, Reva M. Fish, and Susan B. Gerber. Project STAR and Beyond: Database Users Guide," Lebanon: Heros, inc. 2007.

[16] Firpo, Sergio, Nicole M. Fortin, and Thomas Lemieux. (2009) "Unconditional quantile regressions,"*Econometrica*, vol. 77: 953-973.

[17] Fortin, Nicole M., Philip Oreopoulos, and Shelley Phipps. (forthcoming) "Leaving Boys Behind: Gender Disparities in High Academic Achievement," *Journal of Human Resources.*

[18] Frisch, Ragnar, and F. V. Waugh. (1933) "Partial time regressions as compared with individual trends." *Econometrica*, vol. 1: 387-401.

[19] Fryer, Roland G., and Steven D. Levitt. (2004) "Understanding the Black-White Test Score Gap in the First Two Years of School," *Review of Economics and Statistics*, vol. 86: 447-464.

[20] Fryer, Roland G., and Steven D. Levitt. (2006) "The Black-White Test Score Gap Through Third Grade," *American Law and Economics Review*, vol. 8: 249-281.

[21] Hanushek, Eric A. (1999) "Some Findings from an Independent Investigation of the Tennessee STAR Experiment and from Other Investigations of Class Size Effects," *Educational Evaluation and Policy Analysis*, vol. 21: 143-163.

[22] Kline, Patrick. (2011) "Regression, Reweighting, or Both: Oaxaca-Blinder as a Reweighting Estimator," *American Economic Review: Papers & Proceedings*, vol. 101: 532-537.

[23] Koretz, Daniel. Measuring Up: What Educational Testing Really Tells Us. Harvard University Press. Cambridge, Massachusetts, 2008.

[24] Krueger, Alan B. (1999) "Experimental Estimates Of Education Production Functions," *Quarterly Journal of Economics*, vol. 114: 497-532.

[25] Krueger, Alan B. and Diane M Whitmore. (2001) "The Effect of Attending a Small Class in the Early Grades on College-Test Taking and Middle School Test Results: Evidence from Project STAR," *Economic Journal*, vol. 111: 1-28.

[26] Lord, Frederic M. (1975) "The ability scale in item characteristic curve theory," *Psychometrica*, vol. 40: 205-217.

[27] Lovell, Michael C. 1963. "Seasonal adjustment of economic time series," *Journal of the American Statistical Association*, vol. 58: 993-1010.

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

[29] Penney, Jeffrey. On the specification of value-added models. 2014. Mimeo.

[30] Rivkin, Steven G., Eric A. Hanushek, and John F. Kain. (2005) “Teachers, Schools, and Academic Achievement,” *Econometrica*, vol. 73: 417-458.

[31] Rothstein, Jesse. (2009) “Student Sorting and Bias in Value-Added Estimation: Selection on Observables and Unobservables,” *Education Finance and Policy*, vol. 4: 537-571.

[32] Rothstein, Jesse. (2010) “Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement,” *Quarterly Journal of Economics*, vol. 125: 175-214.

[33] Rothstein, Jesse. Revisiting the Impacts of Teachers. October 2014. mimeo.

[34] Stevens, S. S. (1946) “On the Theory of Scales of Measurement,” *Science*, vol. 103: 677-680.

[35] Todd, Petra E., and Kenneth I. Wolpin. (2003) “On The Specification and Estimation of The Production Function for Cognitive Achievement,” *Economic Journal*, vol. 113: F3-F33.

Table 1: Example of test score ordinality

A	ΔA	$T = ln(A)$	Δln(A)	$T = e^A$	Δe^A
1	-	0	-	2.7	-
2	1	0.7	0.7	7.4	4.7
3	1	1.1	0.4	20.1	12.7

Table 2: UQR with standard error of regression normalization

teacher	$T = math$	$T = math^2$	$T = \sqrt{math}$
A	-1.47	-1.47	-1.47
B	-0.73	-0.72	-0.74
C	-0.48	-0.46	-0.48

Table 3: OLS with standard error of T normalization

teacher	$T = math$	$T = math^2$	$T = \sqrt{math}$
A	-1.13	-1.09	-1.15
B	0.05	0.11	0.02
C	-0.09	-0.03	-0.11

Table 4: Summary Statistics

	Total	White	Black	Hispanic	Asian
Sample proportion	1	0.58	0.14	0.18	0.05
Number of observations	17117	9890	2392	3027	868
Controls					
Female	0.49	0.49	0.51	0.5	0.5
	(0.50)	(0.50)	(0.50)	(0.50)	(0.50)
Age in fall kindergarten, months	65.5	65.93	65.12	64.78	64.55
	(4.28)	(4.25)	(4.23)	(4.28)	(4.03)
SES composite measure	0.03	0.23	-0.34	-0.39	0.29
	(0.79)	(0.73)	(0.75)	(0.71)	(0.87)
WIC participant	0.45	0.31	0.77	0.63	0.33
	(0.50)	(0.46)	(0.42)	(0.48)	(0.47)
Number of children's books in the home	73.79	94.62	40.15	42.4	48.47
	(59.64)	(59.36)	(40.17)	(46.16)	(49.98)
Mother's age at birth of first child	23.63	24.8	20.64	21.94	25.73
	(5.47)	(5.35)	(4.75)	(4.95)	(5.46)
First birth under age 20	0.27	0.19	0.51	0.37	0.15
	(0.44)	(0.39)	(0.50)	(0.48)	(0.35)
First birth at age 30 or over	0.16	0.2	0.06	0.09	0.25
	(0.36)	(0.40)	(0.24)	(0.29)	(0.43)

Author's calculations. Standard deviations are in parentheses. Statistics for time-variant variables are those for the fall of the child's kindergarten year. The first two rows do not add up to the total because the "other" racial category, which contains all the other races other than those listed on this table, is not listed here.

Table 5: Reliability Estimates

	Fall grade K	Spring grade K	Fall grade 1	Spring grade 1	Spring grade 3	Spring grade 5	Spring grade 8
Reading	0.92	0.95	0.96	0.96	0.94	0.93	0.87
Mathematics	0.91	0.93	0.94	0.94	0.95	0.95	0.92

Source: Taken from Table 3-10 in the Combined User's Manual for the ECLS-K Eighth-Grade and K8 Full Sample Data Files and Electronic Codebooks. A value of 1 denotes perfect measurement of the latent trait.

Table 6: Test Score Summary Statistics

	White	Black	Hispanic	Asian
Mathematics				
fall kindergarten	0.22	-0.39	-0.46	0.43
	(1.02)	(0.75)	(0.79)	(1.19)
spring kindergarten	0.22	-0.44	-0.41	0.33
	(1.00)	(0.80)	(0.85)	(1.12)
spring first grade	0.22	-0.51	-0.34	0.15
	(1.01)	(0.80)	(0.87)	(1.03)
spring third grade	0.24	-0.65	-0.33	0.24
	(0.94)	(0.90)	(0.94)	(1.01)
spring fifth grade	0.22	-0.75	-0.32	0.36
	(0.90)	(1.00)	(0.99)	(0.92)
spring eighth grade	0.2	-0.8	-0.33	0.41
	(0.88)	(1.07)	(1.05)	(0.91)
Reading				
fall kindergarten	0.11	-0.27	-0.27	0.49
	(1.00)	(0.78)	(0.88)	(1.45)
spring kindergarten	0.1	-0.29	-0.22	0.5
	(1.01)	(0.81)	(0.86)	(1.34)
spring first grade	0.15	-0.38	-0.29	0.36
	(1.00)	(0.87)	(0.89)	(1.08)
spring third grade	0.25	-0.56	-0.41	0.1
	(0.93)	(0.91)	(0.98)	(0.92)
spring fifth grade	0.25	-0.62	-0.41	0.12
	(0.90)	(1.01)	(0.98)	(0.93)
spring eighth grade	0.23	-0.78	-0.42	0.26
	(0.86)	(1.09)	(1.08)	(0.88)

Author's calculations. Test scores have been normalized to have a mean of 0 and a standard deviation of 1. Standard deviations are in parentheses.

Figure 1: Black-white Mathematics Test Score Gap

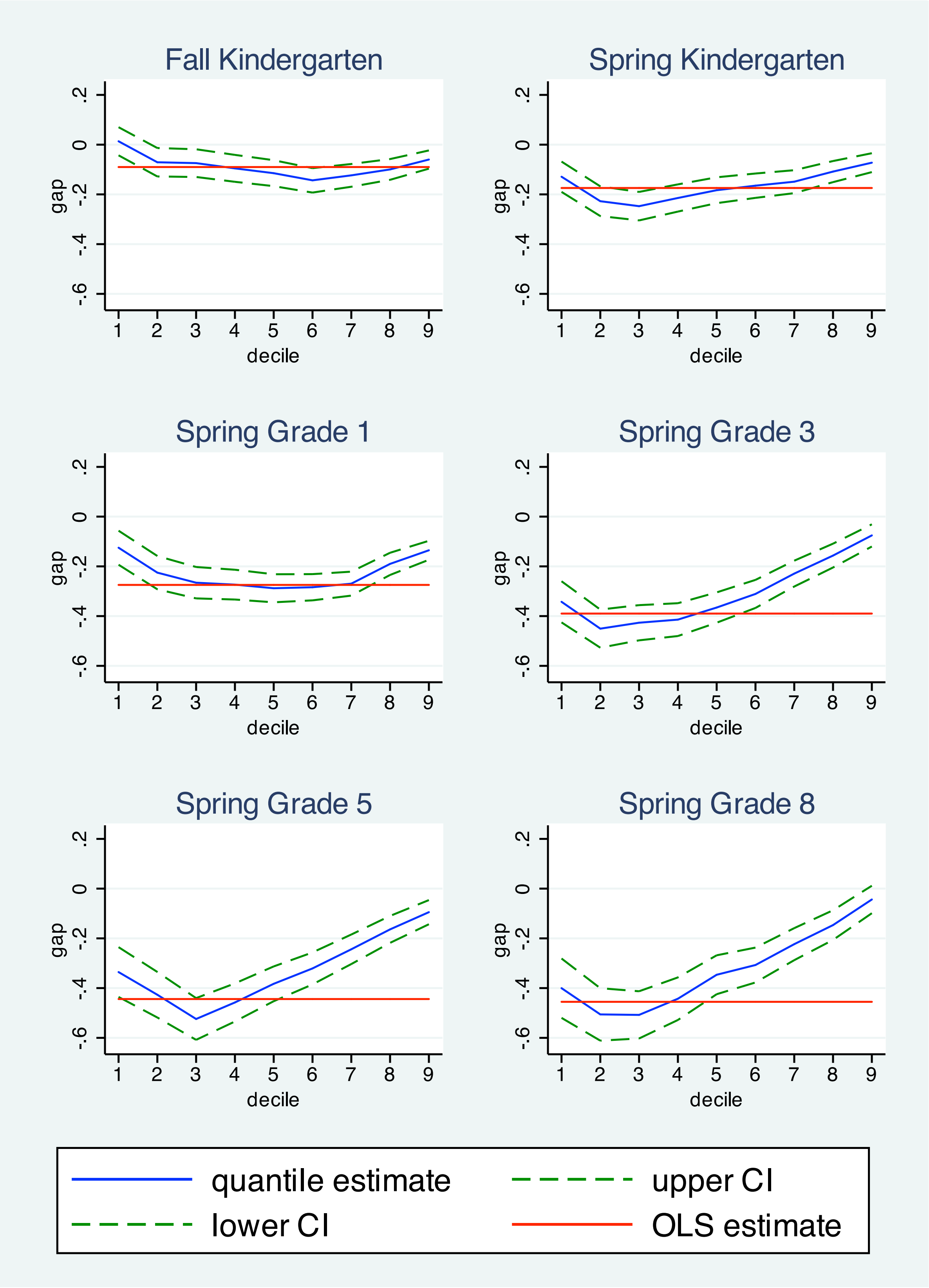
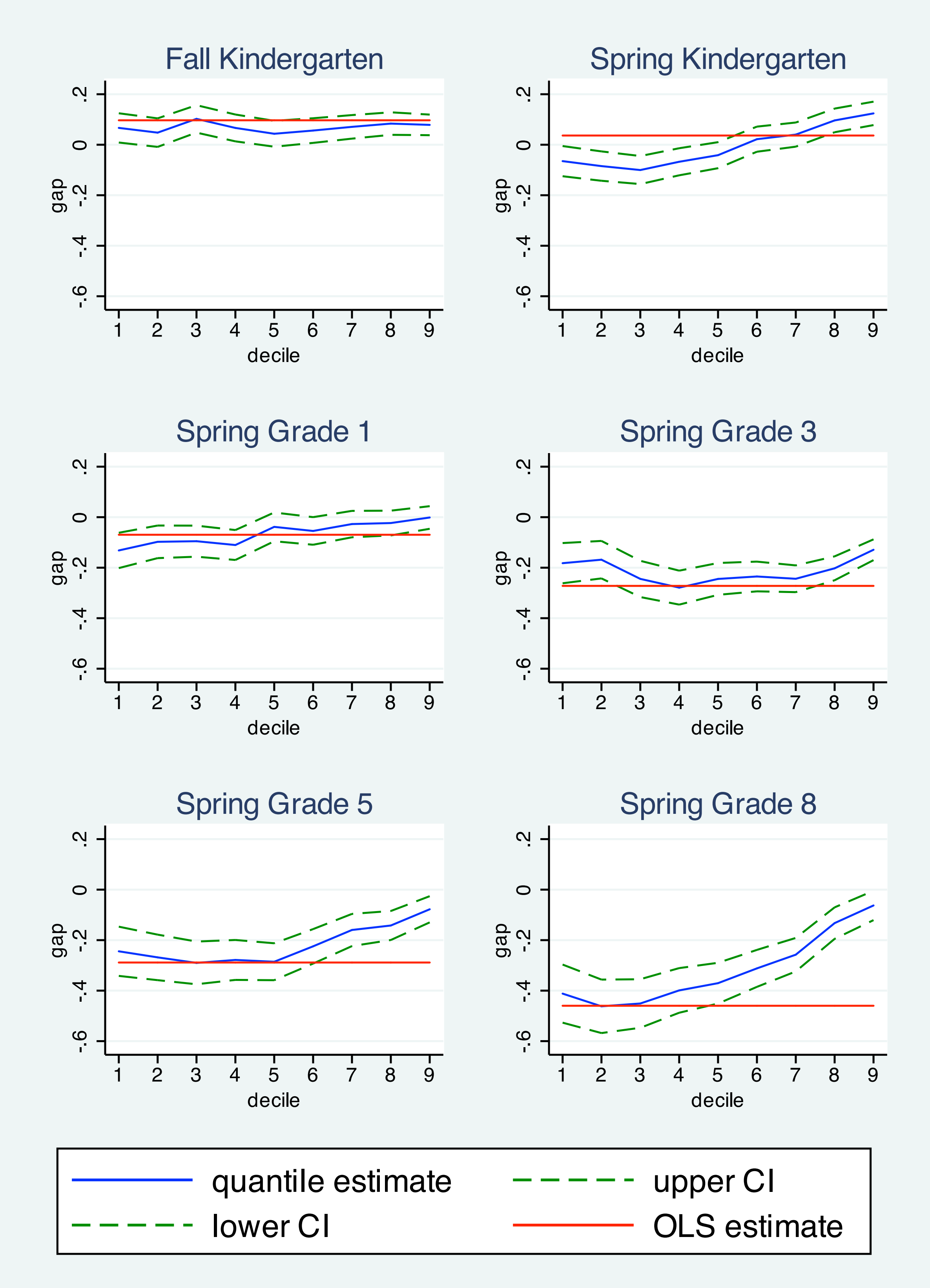


Figure 2: Black-white Reading Test Score Gap


Fall Kindergarten
Spring Kindergarten
Spring Grade 1
Spring Grade 3
Spring Grade 5
Spring Grade 8
gap
decile
quantile estimate
upper CI
lower CI
OLS estimate