TEST SCORE MEASUREMENT AND THE BLACK-WHITE TEST SCORE GAP

Jeffrey Penney*

Abstract—Research as to the size of the black-white test score gap often comes to contradictory conclusions. Recent literature has affirmed that the source of these contradictions and other controversies in education economics 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 monotonic transformations. Under fairly weak assumptions, this metric has interval properties and thus solves the ordinality problem. The measure can serve as a valuable robustness check to ensure that any results are not simply statistical artifacts from the choice of scale.

I. Introduction

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 on topics that have long seen considerable research effort. For the debate over the dynamics of the black-white test score gap, the current majority view is that there is a small gap at entry that quickly grows large by third grade (Fryer & Levitt, 2004, 2006) and that a substantial gap exists in the later grades (Clotfelter, Ladd, & Vigdor, 2009). However, some scholars have argued that the test score gap is moderate at kindergarten entry but shrinks after first grade (Murnane et al., 2006), or that it is large throughout schooling (Bond & Lang, 2013a).

Recent research has potentially found a source of some of these contradictions in the literature. Psychometricians have long accepted that test scores have only 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 & 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 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 ability score 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. Due to this desirable property, the proposed normalization can be employed for two reasons: to produce comparisons that are valid across different samples and different test score scales and to serve as a robustness check to ensure that any results obtained in the statistical analysis of test scores are not due to the choice of scale. I illustrate the use of the measure by examining the black-white test score gap. I find that the results at the mean reported in Fryer and Levitt (2004, 2006) and Clotfelter et al. (2009) are very similar to the results of this new measure.

This paper is organized as follows. Section II places the discussion in context by examining scaling issues related to test scores and then outlines the ordinality problem. The proposed test score normalization is outlined in section III. The application to the black-white test score gap is undertaken in section IV. The paper concludes with a brief discussion in section V.

II. Background

A. Scaling and Measurement Issues

Test scores are subject to rather complex issues relating to measurement. The majority of agreed-on scales that are used in the sciences have readily observable effects on the physical environment; for example, temperature until recently has been tied to the length of a column of mercury. These scales typically have interval properties: one can claim that the increase in temperature between 15 degrees and 20 degrees represents the same physical difference as between 25 and 30 degrees. By contrast, academic test scores do not have an effect on the physical environment and reference only the test from which they are measured, so what they represent is more difficult to quantify. Under the weakest assumptions, test scores have only ordinal properties: the difference between a test score of 40 and 50 may represent either more or less of a difference in ability than between a score of 80 and 90.1 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 toward more meaningful inference. These test scores allow for both relative and absolute performance measures; for example, a verbal score of 800 on the GRE verbal indicates a person with a wide vocabulary, while a score of 200 would signal a person as inarticulate. However, IRT is not a catholicon, as the scoring scales still refer to only the tests themselves (e.g., one cannot immediately compare an SAT verbal score to an ACT reading score). While IRT test scores are still on ordinal scales, they can be interpreted as

1 More information about the different types of numerical scales can be found in Stevens (1946).
interval scales with additional assumptions, but the assumptions needed to do so are rather strong. A brief introduction to the technical details of IRT can be found in section III of Bond and Lang (2013b).

The standard practice in analyses involving test scores in the education literature is to convert them into z-scores. Many of the problems associated with test scaling are at least partially addressed by using this test score normalization procedure. First, it provides some 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 or skill but are scored on different scales (such as when comparing SAT verbal scores to ACT reading scores).

### B. The Ordinality Problem

Nearly all research in economics and education that uses test scores as a dependent variable implicitly makes the following assumptions: (a) there exists an unobservable score \( A \) that represents ability in a subject; (b) \( A \) has interval properties; and (c) observable test scores \( T \) have interval properties since they are an affine transformation of an 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. 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 illustrates the ordinality problem. Three values of the latent true test score \( A \) are listed, and there is a constant difference in true ability 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.

<table>
<thead>
<tr>
<th>( A )</th>
<th>( \Delta A )</th>
<th>( T = \ln(A) )</th>
<th>( \Delta \ln(A) )</th>
<th>( T = e^A )</th>
<th>( \Delta e^A )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>–</td>
<td>0</td>
<td>–</td>
<td>2.7</td>
<td>–</td>
</tr>
<tr>
<td>2</td>
<td>1</td>
<td>0.7</td>
<td>0.7</td>
<td>7.4</td>
<td>4.7</td>
</tr>
<tr>
<td>3</td>
<td>1</td>
<td>1.1</td>
<td>0.4</td>
<td>20.1</td>
<td>12.7</td>
</tr>
</tbody>
</table>

Both sets of scores would fit the IRT model with the same likelihood.

### III. Methodology

The metric developed here employs the ordinary least squares variant of unconditional quantile regressions (UQR) as developed by Firpo, Fortin, and Lemieux (2009) to estimate the test score gap at the median and then normalizes the coefficients of 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.

The proposed measure solves the ordinality problem outlined in the previous section. Its invariance to monotonic transformations means that the same regression results will be obtained as if one had access to the “true” set of test scores. Therefore, investigators can verify whether their regression results are simply artifacts of scaling decisions or if they represent genuine findings.

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

\[
IF(y; q_t, F_y) = (\tau - 1[y \leq q_t])f_y(q_t) \equiv \tilde{y},
\]

where \( \tau \) is the quantile of interest, \( q_t \) is the value of \( y \) at the quantile \( \tau \), \( 1[\cdot] \) is an indicator function taking the value of 1 if the statement in the square brackets is true and 0 otherwise, \( f_y(q_t) \) is the density of \( y \) at \( q_t \), and \( IF(\cdot) \) denotes the influence function.4

**Lemma 1.** The term \( 1[y \leq q_t] \) is the same for any monotonic transformation of \( y \).

**Proof.** Define \( y^* \equiv g(y) \), and let \( g \) be a monotonic function. Then \( \tau = Pr[y^* \leq j^*] = Pr[g(y) \leq j^*] \equiv Pr[y \leq g^{-1}(j^*)] \). Thus, \( j^* = g(j) \), and therefore \( Pr[y^* \leq j^*] = Pr[y \leq g^{-1}(g(j))] = Pr[y \leq j] \). Hence, \( j = q_t \).

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 lagged 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.

---

2 Some claim that using the Rasch IRT model generates an interval scale (Salzberger, 2010).

3 I refer to this practice as z-normalization.

4 Note that the influence function used here is not recentered. The influence function rather than the recentered variant is used to considerably simplify the proof of theorem 1. Using the recentered influence function (RIF) in place of the influence function does not affect the results since \( q_t \) 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.
Recall that \( \tilde{y} \) is the transformed value of \( y \) by the IF. By lemma 1, \( 1[y \leq q_1] \) is invariant to monotonic transformations. What is left to show is that the ratio \( \hat{\beta}/s \) is invariant to changes in \( f_j(q_1) \). The value of the ratio is

\[
\frac{\hat{\beta}}{s} = \frac{1}{(X^\top X)^{-1}X^\top \tilde{y}/\sqrt{\tilde{y}^\top M\tilde{y}(n-k)^{-1}}},
\]

where \( n \) is the number of observations and \( k \) is the number of estimated parameters. Note that \( f_j(q_1) \) scales the values of \( \tilde{y} \) by a constant factor. Suppose a monotonic transformation of \( y \) takes place, \( y' \equiv g(y) \), where \( g \) is monotonic, and thus \( \tilde{y}' = \theta \tilde{y} \). The ratio is thus

\[
\frac{\hat{\beta}'}{s'} = \frac{1}{(X^\top X)^{-1}X^\top \tilde{y}'/\sqrt{\tilde{y}'^\top M\tilde{y}'(n-k)^{-1}}}
= \frac{1}{(X^\top X)^{-1}X^\top \theta \tilde{y}/\sqrt{\theta^2 \tilde{y}^\top M\theta \tilde{y}(n-k)^{-1}}}
= \frac{1}{(X^\top X)^{-1}X^\top \tilde{y}/\sqrt{\tilde{y}^\top M\tilde{y}(n-k)^{-1}}} = \frac{\hat{\beta}}{s},
\]

where the third equality holds since \( \theta \) 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 \( y \).

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 lagged dependent variable, the ratio of any nonlag 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

\[
y_1^* = \frac{\alpha_0 + \alpha_1 X + f(\alpha_2, y_1^*) + \epsilon_1}{\sigma_1},
\]

(3)

\[
y_2^* = \frac{\beta_0 + \beta_1 X + g(\beta_2, y_2^*) + \epsilon_2}{\sigma_2},
\]

(4)

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 \( g(\beta_2, y_2^*) = \beta_{21} y_2^* + \beta_{22} y_2^* + \ldots + \beta_{2k} y_2^k \). Recall that \( y_2^* \), the transformed value of \( y_2 \), is monotonically related to the transformed value of \( y_1 \) by \( y_2^* = A y_1^* \), and similarly, \( \sigma_2 = A \sigma_1 \) for some unknown linear scaling factor \( A \). Thus, estimating equation (4) is equivalent to estimating

\[
y_1^* = \frac{\beta_0 + \beta_1 X + g(\beta_2, y_2^*) + \epsilon_2}{\sigma_1}.
\]

(5)

Let \( M_y^1 \) be a projection matrix that projects off the space spanned by the polynomial expansion of \( y_1 \). By the Frisch-Waugh-Lovell (Frisch & Waugh, 1933; Lovell, 1963), the estimates of \( \alpha_1 \) of the regression

\[
M_y^1 \frac{y_1^*}{\sigma_1} = a_0 + a_1 M_y^1 X + u_1,
\]

will be identical to those in equation (3). Therefore, the estimate of \( \beta_1 \) in

\[
M_y^1 \frac{y_1^*}{\sigma_1} = \beta_0 + \beta_1 M_y^1 X + M_y^1 g(\beta_2, y_2^*) + u_2
\]

(7)

will be the same as the estimate of \( \alpha_1 \) in equation (3) if we have \( \sum M_y^1 x_i u_2 = 0 \); 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 latter, \( \hat{\beta}_1 \) in equation (4) is equal to \( \hat{\alpha}_1 \) in equation (3) asymptotically.

The interpretation of an estimated parameter under the proposed normalization is similar to the usual z-normalization case: a measure of the benefit of an intervention in terms of the (unexplained) randomness of the test score. It is important to note that parameter estimates obtained using unconditional quantile regression (and, thus, this proposed methodology) are effects at the median of the unconditional distribution; this is in contrast to ordinary least squares, which estimates the effect of the change in the proportion of those treated on the mean of the unconditional distribution (Firpo et al., 2009). In general, these two measures do not coincide in the presence of heterogeneous effects. One can obtain an apples-to-apples comparison of the effects of the test score scaling decisions by comparing the unconditional quantile regression results using the standard z-normalization with the proposed normalization.5

IV. Application: Black-White Test Score Gap

Bond and Lang (2013b) call the entire body of research on the black-white test score gap 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 growth-minimizing evolutions. These results lead the authors to claim that the dynamics of the gap in the literature likely reflect test score scaling decisions.

5 Of course, in addition to the median, one can examine other quantiles of interest when employing UQR on z-normalized scores to compare with this method since the above proofs cover the cases of any quantile of interest.
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 equation (2) in an attempt to uncover the true evolution of the test score gap.

A. Data

This research employs the Early Childhood Longitudinal Study Kindergarten Cohort (ECLS-K), a nationally representative survey of 21,260 children who entered kindergarten in autumn 1998. I use this data set to facilitate comparisons with the existing literature, which also uses this data set (e.g., Fryer & Levitt, 2004, 2006; Bond & Lang, 2013b). Due to space limitations, I provide only a limited overview of the data in this section; a detailed exposition of this data can be found in section III of Fryer and Levitt (2004).

The analysis of this chapter employs the same controls as Fryer and Levitt (2004): race, gender, age, birth weight, whether the mother of the child was a teenager when she first gave birth, whether the mother of the child was age 30 or over when she first gave birth, socioeconomic status (using a composite measure),6 WIC participation (a dummy variable indicating enrollment in the Special Supplemental Nutrition Program), and the number of children’s books in the home. The “white” racial category refers exclusively to non-Hispanic whites.

The ECLS-K data set includes several kinds of math and reading test scores. Those employed in this chapter are the longitudinal IRT test scores, which were designed to be fully comparable across grades. Because the scores are developed from IRT, we can employ the method outlined in this paper. To maintain comparability with the other literature on this subject, the test scores were converted into \( z \) scores.7

Students missing data on the Fryer-Levitt controls or who do not have at least one valid test score are dropped from the sample.8 For the variables that are time variant, the ones used in the analysis are those from the fall of the child’s kindergarten year. I do not use sample weights in this analysis because the results are not sensitive to their use.

B. Application

The regression equation estimated is

\[
T_i = \beta_0 + \rho \beta_1 + X_i \beta_2 + \epsilon_i,
\]

(8)

where \( T_i \) is the \( z \)-normalized test score of individual \( i \) at time \( t \), \( \rho \) is a vector of racial dummies (black, Hispanic, Asian, other), \( X_i \) is a vector containing the Fryer-Levitt controls, and \( \epsilon_i \) is the usual error term. Non-Hispanic whites are the baseline racial category.

Table 2 displays the coefficient estimates from estimating equation (8) using OLS, unconditional quantile regression at the median and the method outlined in this paper. Comparing the \( z \)-UQR estimates with those from the proposed method, not only are the evolution patterns almost identical, but the quantitative results are also largely consistent with each other despite the fact that they were constructed in radically different fashions. That similar results are obtained here using the proposed methodology and with UQR using the standard transformation points to the strong possibility that the ordinality problem is not a serious obstacle to inference in this case, possibly because of a well-constructed IRT scale.9 The OLS results, here for comparison purposes, also exhibit similar evolution patterns and magnitudes as those obtained from the other estimation methods.

The results for the normalized scores using the method outlined in this paper contrast favorably with the general consensus of the literature that there is a large gap between blacks and whites that persists across grades. This should allay concerns raised in Bond and Lang (2013b) that the black-white test score gap may be an artifact of test score scaling.

V. Discussion

Scaling issues are an important possible source for the different results in the black-white test score gap literature. However, there are additional possibilities relating to the issue of interstudy comparability. The first is whether

---

6 Details about the construction of the variable can be found starting on pages 7–8 of the ECLS-K User Guide.

7 By theorem 1, this normalization does not affect the analysis.

8 Analysis of the math and reading test scores in the fall of first grade is excluded because only a small portion of the sample took this particular test.

9 Recall that IRT test scores can have interval properties when they satisfy certain strong conditions (Bond & Lang, 2013b).
the tests used in different studies are measuring the same underlying factor (e.g., mathematical ability). For example, Murnane et al. (2006) employ the National Institute of Child Health and Human Development data set (NICHD) and show that the racial test score gap between blacks and whites holds steady for reading 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 incomparable. A second possibility concerns the issue of comparable samples. Reardon, Kalogrides, and Shores (2016) find considerable geographic variation in demographic test score gaps; given this, comparing black-white test score gaps using dissimilar samples may yield differences on this basis alone.

An alternative method to measure demographic test score gaps is to anchor them to adult outcomes (e.g., Cunha & Heckman, 2008; Bond & Lang, 2013a). The primary and significant advantage to anchoring is that the gaps are expressed in concrete units, such as completed years of education; these units have the highly desirable property of being ratio scales. Unfortunately, such an approach is not a panacea. The primary difficulty is that adult outcomes come with significant delays (Barlevy & Neil, 2012), which may limit the policy relevance of the results. Some results may be sensitive to distributional assumptions (Bond & Lang, 2013b) and heterogeneity in tastes may weaken the link between test scores and either earnings (e.g., Zafar, 2013) or schooling outcomes. The choice of whether to use scores anchored to adult outcomes or simply the raw scores themselves should depend on the context of the academic inquiry.

REFERENCES

Clotfelter, Charles T., Helen F. Ladd, and Jacob Vigdor, “The Academic Achievement Gap in Grades 3 to 8,” this review 91 (2009), 398–419.