Putting Test Scores on an Interval Scale

PIAAC Methodological Seminar – the Use of Test Scores in Secondary Analysis

Timothy N. Bond (Purdue University)

June 13, 2019

Source Material

This talk is based off of two papers co-authored with Kevin Lang of Boston University:

Bond, Timothy N. and Kevin Lang. 2013. "The evolution of the black-white test score gap in grades K-3: The fragility of results." *Review of Economics and Statistics.* 95 (5): 1468-1479.

Bond, Timothy N. and Kevin Lang. 2018. "The black-white education-scaled test-score gap in grades K-7." *Journal of Human Resources.* 53 (4): 891-917.

Ordinal versus Interval Scales

- Economists generally care about tests scores as a measure of early human capital
- E.g., the black-white test score gap tells us when human capital differences that impact labor market inequality first occur
- Published test scores measure human capital on an ordinal scale, not an interval scale
- There are an infinite number of potential cardinalizations of the ordinal scale, each of which could yield a different conclusion about the differences in human capital across groups, or the impacts of policy interventions

Goals of This Talk

1. Briefly outline why scale choice could matter in theory
2. Demonstrate the pliability of secondary analysis of test scores to alternative scalings using the evolution of the black-white test gap as an example
   - The black-white test gap can either rapidly grow, stay constant, or shrink with schooling depending on scale choice
3. Use adult outcomes to construct an interval scaled measure of test performance
   - This measure generates the opposite conclusion (shrinking gap) to what the published scales would generate on the same data

Example

- Suppose we have a 20 white students and 20 black students who take a test with three scores

Score	Black	White
1	10	5
2	4	10
3	6	5

- Typically economists would calculate the average test score by race to compute a racial test gap (or standard deviations thereof)
- Blacks: 1.8 Whites: 2.0, so test gap is .2

Example

- Approach implicitly assumes test score is interval with respect to policy outcome of interest
- No reason to think this is true
  - Score 1 indicates individual can recognize Latin alphabet
  - Score 2 indicates individual can read and write in English
  - Score 3 indicates individual can read and write in English and Latin
- "True size" of the test gap will depend on how we value the marginal return to fluency in English versus the marginal return to fluency in Latin

Scale Choice and Stochastic Dominance

- Ranking of the means will be independent of scale choice only if whites' scores first order stochastically dominate blacks' scores (or vice versa)
- Strong condition that is rarely satisfied in practice
- Magnitude of gap will always depend on scale choice
- Problem even more severe if concerned about changes in gap across time

Scale Choice in Practice

- Possible this is merely a theoretical concern
- Only "wild" or "implausible" transformations reverse gap found using published scales
- Explore this using black-white test gap in early years of schooling using two data sets
  - ECLS-K: Fryer and Levitt (2004, 2006) show that the black-white gap is modest at kindergarten entry and grows rapidly by third grade using these data
  - CNLSY: Peabody Individual Achievement Tests (PIAT) show similar pattern to Fryer and Levitt [though strikingly large gap at age 4 in the Peabody Picture Vocabulary Test (PPVT)]

Scale Bounds

- Attempt to bound the black-white test gap by search across all (continuous) monotonic transformations

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

- t is published test score
- $\beta_0 - \beta_6$ and c chosen to minimize or maximizing growth of gap from kindergarten through third grade
- Also search for transformation which maximizes correlation between initial and final reading score in ECLS-K; or pre-schooling (PPVT) and third grade (PIAT) tests in CNLSY

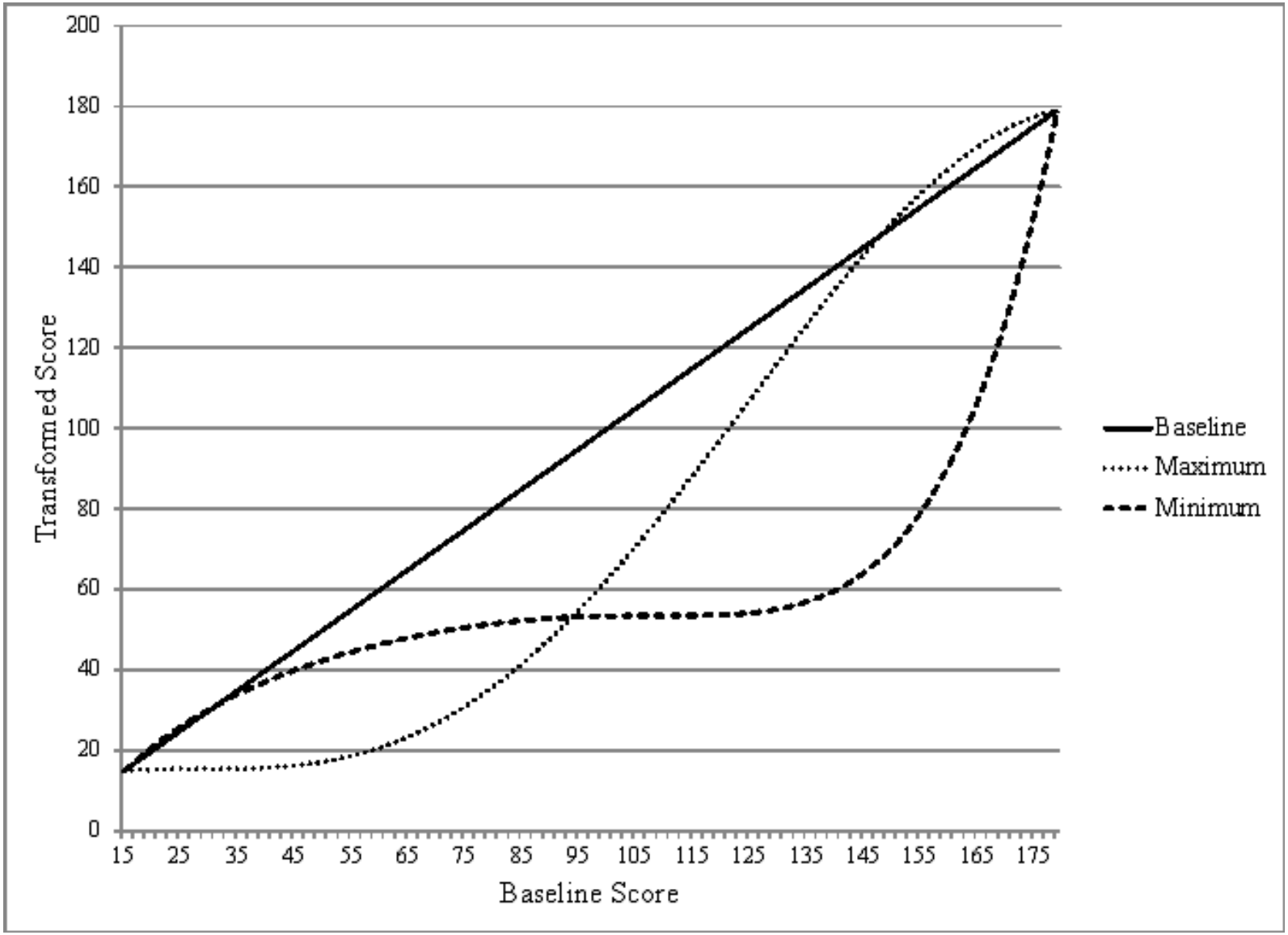
Table 1: Evolution of the black-white test gap under various transformations of the ECLS-K

	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)

Table 2: 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)

Are These Reasonable?



Scale Choice Matters

- Bounding exercise essentially uninformative of the nature of the black-white test gap
- Reasonable alternative scalings can produce anywhere from equality at school entry with blacks falling rapidly behind to large gap at school entry that is eliminated by third grade
- Any conclusion will depend on choice of scale
- Which scale should we choose?
- Proposal: "anchor" scale to outcomes policymakers care about

Theoretical Motivation

- Suppose we are interested in measuring some latent trait θ in grade g which predicts a future outcome S (e.g., adult wages, educational attainment)

$$S_i = \theta_{ig} + \nu_{ig} \tag{1}$$

- ν_{ig} is mean zero error term with $E[\theta_g \nu_{g'}] = 0 \forall g \leq g'$
- Assume θ_{ig} distributed normal with mean $\overline{\theta}_g$ and variance $\sigma^2_{\theta g}$
- Goal is to measure racial gap $\overline{\theta}_{wg} - \overline{\theta}_{bg}$

Published Scale vs. Interval Scale

- At best, a properly scaled test (for measuring θ) would yield

$$s_{ig} = \theta_{ig} + \eta_{ig} \tag{2}$$

- η_{ig} is classical test measurement error
- Instead, we observe a test score t_{ig} which measures θ, but not an an interval scale

$$t_{ig} = f_g(s_{ig}) \tag{3}$$

- f_g is function which transforms from published scale to outcome-based interval scale

Measurement Error

- We cannot observe f_g but can estimate it with $\mathbb{E}[S_{ig}|t_{ig}]$ (or empirical equivalent)
- What does that find?
- Applying Bayes' rule

$$\mathbb{E}[S_{ig}|t_{ig}] = \rho_g s_{ig} + (1 - \rho_g)\overline{\theta}_g \tag{4}$$

where p_g is test reliability

$$\rho_g = \frac{\sigma^2_{\theta g}}{\sigma^2_{\eta g} + \sigma^2_{\theta g}} \tag{5}$$

- s_{ig} projections of future educational attainment are implicitly and correctly shrunk to mean educational attainment because a single test score is a noisy measure of aptitude

Attenuation Bias in Racial Gaps

- Why is this a problem for estimating test gaps?

$$\begin{aligned} s_{wg} - s_{bg} &= [\rho_g \overline{\theta}_{wg} + (1-\rho_g)\overline{\theta}_g] - [\rho_g \overline{\theta}_{bg} + (1-\rho_g)\overline{\theta}_g] \\ &= \rho_g(\overline{\theta}_{wg} - \overline{\theta}_{bg}) \end{aligned} \tag{6}$$

- Measurement error biases achievement gap toward zero
- Intuition: One test is a very noisy measure of any individuals achievement, but lots of tests pretty good measure of of average achievement by race

Unshrinking Estimates

- If we can estimate bias directly, we can adjust our estimate of achievement gaps to find true magnitude of gap
- Consider estimating

$$s_{ig} = c + \beta_{1g}\theta_{ig} + \varepsilon_{ig} \tag{7}$$

- β_1 will be unbiased and consistent estimator of ρ_g
- In practice cannot observe θ_{ig}, but do observe S_i which is unbiased estimate of θ_{ig}

Data and Methodology

- Utilize CNSLY which has early adult outcomes as well as tests dating back to early childhood
- Calibrate scores based on ability to predict future education attainment (best available outcome at time of writing)

$$\hat{s_{ig}} = N_{t_{ig},g}^{-1} \sum_{j:t_{jg}=t_{ig}} S_j \tag{8}$$

- Use only whites to avoid concerns about anchoring lower scores to outcomes that are influenced by racial discrimination

Data and Methodology

- Estimate attenuation bias using

$$\hat{s_{ig}} = c + \beta_{1g} S_i + \varepsilon_{ig} \tag{9}$$

- Because $S_i = \theta_{ig} + \nu_{ig}$ estimate of β_{1g} will be biased downwards
- Construct leave-one-out estimate of lagged test score $\hat{s^*_{ig-1}}$ and use as instrument
- Lagged score uncorrelated with measurement error of next score, by definition (everything that is predictable about the future is embedded in lagged score)
- Measured achievement gap is then $\hat{\beta}_{1g}^{-1}(\hat{s_{wg}} - \hat{s_{bg}})$

Table 3: Raw Difference in Expected White Grade Completion Conditional on Test Score

	Math	Read-RR	Read-RC
	(1)	(2)	(3)
Pre-Age 5 PPVT		0.85	
		[0.55, 1.16]	
Kindergarten	0.57	0.21	0.24
	[0.34, 0.77]	[0.07, 0.38]	[-0.06, 0.50]
First Grade	0.52	0.33	0.38
	[0.35, 0.67]	[0.15, 0.45]	[0.17, 0.55]
Second Grade	0.75	0.60	0.43
	[0.55, 1.00]	[0.37, 0.80]	[0.21, 0.61]
Third Grade	0.70	0.62	0.62
	[0.52, 0.88]	[0.49, 0.81]	[0.43, 0.76]
Fourth Grade	0.67	0.53	0.62
	[0.52, 0.89]	[0.36, 0.70]	[0.42, 0.82]
Fifth Grade	0.71	0.48	0.47
	[0.54, 0.89]	[0.26, 0.61]	[0.29, 0.59]
Sixth Grade	0.63	0.57	0.58
	[0.48, 0.86]	[0.40, 0.74]	[0.36, 0.77]
Seventh Grade	0.72	0.56	0.60
	[0.54, 0.89]	[0.31, 0.70]	[0.43, 0.78]

Table 4: Measurement Error Difference in Ability in Units of Predicted White Education

	Math	Read-RR	Read-RC
	(1)	(2)	(3)
Kindergarten	1.14	0.62	1.26
	[0.58, 2.17]	[0.12, 2.27]	[-4.23, 8.81]
First Grade	1.01	0.82	0.74
	[0.56, 1.55]	[0.31, 1.32]	[0.28, 1.26]
Second Grade	1.07	0.87	0.79
	[0.57, 1.57]	[0.29, 1.53]	[0.35, 1.63]
Third Grade	0.97	0.64	0.66
	[0.52, 1.59]	[0.40, 0.99]	[0.24, 1.15]
Fourth Grade	1.12	0.53	0.72
	[0.69, 1.58]	[0.28, 0.76]	[0.02, 1.04]
Fifth Grade	0.79	0.55	0.63
	[0.51, 1.10]	[0.29, 0.75]	[0.33, 0.82]
Sixth Grade	0.84	0.74	0.75
	[0.54, 1.12]	[0.48, 1.01]	[0.43, 1.02]
Seventh Grade	0.87	0.66	0.68
	[0.51, 1.13]	[0.36, 0.93]	[0.38, 1.20]

Implications of Anchoring

- Anchored results give very different results than those from published test scales
- Much of these differences appear to be driven by a high degree of measurement error in early test grades
- Shortcoming: Results could differ depending on anchor
- We obtain similar results where we scale education by the mean annual earnings for that level of education, but one based on actual earnings may lead to different patterns
- But, different parts of the test distribution may be more important for different outcomes (e.g., criminality vs. obtaining a graduate degree)

Anchoring in Practice

- Anchoring appears very data intensive
- Could be possible to use old data to anchor current test scores (assume relationship between score and outcome is fixed across time)
- What if anchoring completely infeasible?
  - Plot the distribution of treatment and control – stochastic dominance unlikely but may be visual obvious that policy was effective
  - Learn about which types of students were affected by the policy
  - Are there any objective outcomes to verify test score results go in the same way (e.g., grade retention, high school completion?)