# 8 reactions to Brian Jacob and Jesse Rothstein's "The measurement of student ability in modern assessment systems"… from a psychometrician

Andrew Ho, *Harvard Graduate School of Education*

August 17, 2016

A few weeks ago, Brian Jacob and Jesse Rothstein, economists at Michigan and Berkeley respectively, released an NBER working paper, "The measurement of student ability in modern assessment systems." Last Thursday, Brian published a Brookings report that summarized their joint work, "Student test scores: How the sausage is made and why you should care." As a measurement scholar committed to training interdisciplinary researchers in the judicious design and analysis of test scores,[1] I am grateful for their contributions. Here are eight quick reactions from my perspective as a "psychometrician."

1) **Duh.**

   Don't most educational researchers know all this already?

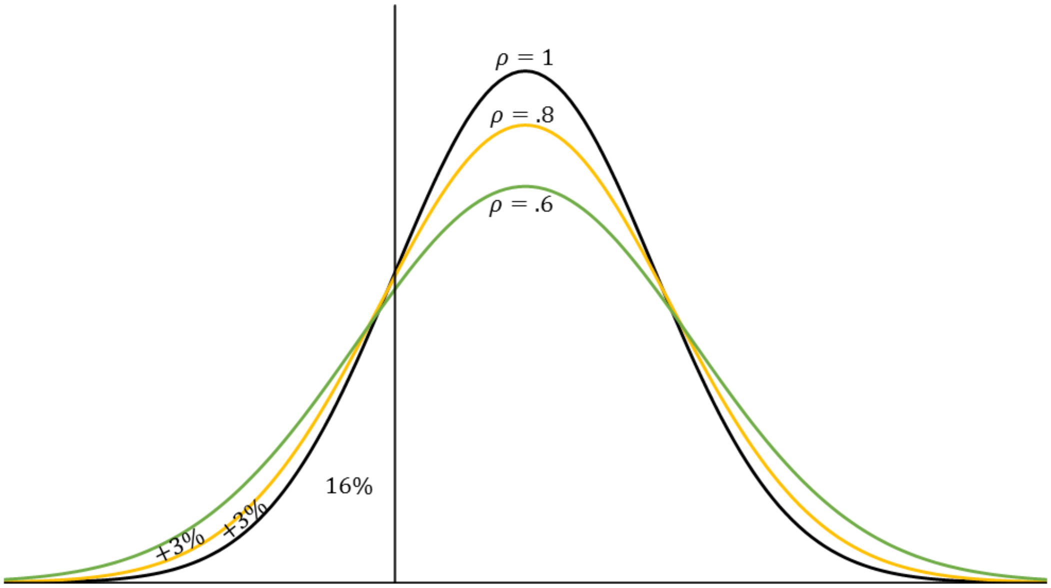
2) **Hmm, maybe not. We psychometricians should be writing more pieces like this.**

   Why didn't we write this? Our work is there in Brian and Jesse's citations, but we could do more to make it accessible and known beyond the measurement community. We should help to lead this effort. Test scores are powerful research tools. Those who know them best should be helping other researchers to use them wisely.

3) **Wait, I've got a riddle: When does a noisier test NOT lead to more extreme scores?**

   In his Brookings piece, Brian said, "all else equal, the shorter the length of the test, the greater the fraction of students placed in the top and bottom proficiency categories." I'll give an example of when this is true and when it might not be true. This is generally true because shorter tests have more extreme scores (just as you are far more likely to get 100% "heads" when you flip a coin 5 times than when you flip a coin 50 times). Test score distributions have variances proportional to their unreliability, $\sigma_x^2 \propto \frac{1}{\rho}$. Figure 1 illustrates this for tests with reliabilities of 1, 0.8, and 0.6. We can see that lowering reliability increases the percentage below a low cut score (-1$\sigma$) by 3 percentage points at a time, from 16% to 19% to 22%, as noise increases and the distribution spreads out.

[1] Their work was well-timed for two local reasons. First, my fall measurement course here at Harvard, S-061, begins in just two weeks. Second, Harvard and Michigan are both recent recipients of predoctoral research training grants from the Institute of Education Sciences. Measurement is a key component of this training. In this spirit, I'm visiting Brian, his colleagues, and their predoctoral cohort at Michigan in December to talk about many of the issues that Brian and Jesse have raised. (I also hope they don't mind me using their first names in this informal response.)



So how could the percentage of extreme scores stay constant? Because, in practice, the cutoff score that defines “extreme” might change, too. These cutoff scores are set judgmentally through a process called “standard setting,” and the panelists who set these standards consider the relative as well as absolute meaning of cutoffs. When scores are noisy, one could say the scores spread out, beyond the cutoffs, or one could say the scale spreads out, along with the cutoffs.

The upshot is that a short test (and its resulting imprecision) can be problematic, but it is particularly problematic when lengths/reliabilities change, such as when New York shortened its tests this past year, or when we make naïve comparisons between tests with different reliabilities, like state tests and NAEP.

### 4) Hold on, do shrunken scores really underestimate gaps?

Brian and Jesse correctly point out that many testing programs report “shrunken” score estimates. To the degree that scores are unreliable, these estimates are shrunken to the overall mean, improving their prediction. This raises the alarming possibility that these estimates will understate achievement gaps, as imprecision will cause all scores (and group differences) to shrink. But if a true gap is 1 arbitrary score point, and an estimate using shrunken scores is 0.8 arbitrary score points, is it really biased? In baseline IRT models, shrunken estimates are one-to-one functions of maximum likelihood estimates—individual ranks correlate perfectly. Although some applications may require anchoring of scores to specific locations on the original scale, for relative uses of scores, shrinkage rarely makes a difference.

If the underlying scale were concrete and interpretable, like dollars or counts, shrinkage would certainly be problematic—an eighth of a dollar is not a dollar. But the underlying scale for proficiency is latent and arbitrary. As long as scores are comparable on that scale, whether the scale spans 0 to 8 or 0 to 10 may make little interpretive difference. The goal becomes linear association with, not absolute recovery of, gaps and other parameters. Just as I argued in point #3 that the danger was less imprecision than changes in imprecision, here, for many score uses, the danger is less shrinkage than changes in shrinkage.

Of course, there are absolute interpretations of gaps that should be protected, such as effect sizes expressed in standard deviation units. These should be corrected for measurement error, as Brian and Jesse suggest. For these purposes, dividing effect sizes by the square root of classical reliability coefficients can help to get everyone on the same metric of "true-score" standard deviations. These corrections are imprecise, imperfect, and incomplete, but they get us close.[2] I would be interested to see Brian's Figure 2 from his Brookings report expressed in subgroup standard deviation units and corrected in this way.

### 5) Plausible values and policy variables, oh my!

My previous points are trivial clarifications compared to Brian and Jesse's important and underappreciated points about plausible values that use conditioning models. The sensitivity checks that they used to evaluate the robustness of their findings in Dee and Jacob (2011) and Lafortune et al. (2016) should be standard. I couldn't agree more that the impact of misalignment between conditioning variables and policy variables deserves further research.

### 6) Yes, please. More about the pliability of test-score scales

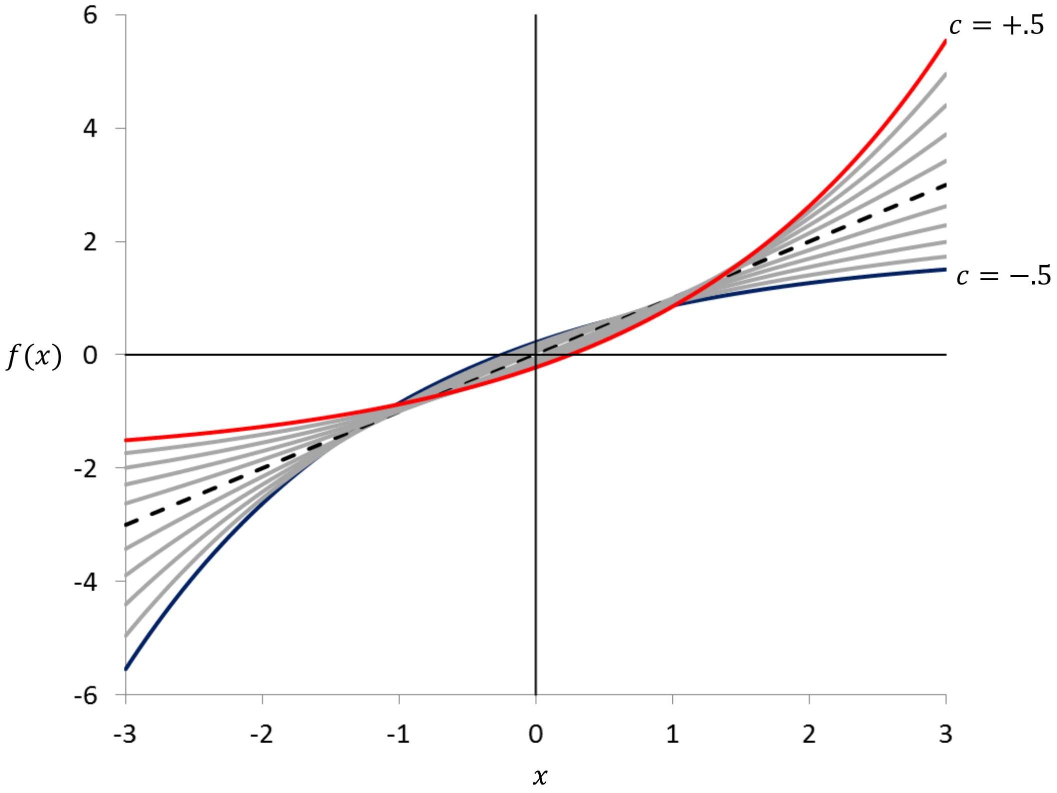
I've loved seeing the good work that economists have done to address the ordinality of test score scales, including but not limited to Ballou (2009), Barlevy and Neal (2012), Bond and Lang (2013), and Nielsen (2015). My own perspective is that scales are "pliable," bendable but not breakable, firm but flexible (Ho, 2009). The transformations proposed by some seem interval-nihilistic in their extremity, and I would also describe as fairly severe the exponential transformation that Brian and Jesse describe as "modest." An exponentiated standard normal distribution has a skewness of 6.2, far beyond observed score distributions in practice, whose skewnesses typically range from -2 to 2 (Ho & Yu, 2015). Observed score distributions need not define the limit but are a good reference point.

With these bounds in mind, Sean Reardon and I employed a generalized exponential transformation of the form, $f(x) = a + b * \exp(cx)$, where mean- and variance-preserving constraints lead to the transformation,

$$f(x) = -\frac{sgn(c)}{\sqrt{e^{c^2} - 1}}\left(1 - e^{cx - \frac{c^2}{2}}\right).$$

[2] Sean Reardon and I show that this gap correction holds up well under scale transformations, when reliabilities are what they are reported in state testing manuals (Reardon & Ho, 2015).

The skewness, $\gamma = sgn(c)\left(e^{c^2} + 2\right)\sqrt{e^{c^2} - 1}$, can range $\pm 2$ by setting $c$ to range $\pm.55$. I describe how to use these transformations to evaluate robustness to scale transformations in a 2014 workshop that SREE members can access here. The transformations look like this:



Broadly, this is in the spirit of Brian and Jesse's argument that we should define a family of plausible transformations within which we can agree or disagree productively.[3]

### 7) Why are we looking only where the light is? From reliability to generalizability.

As secondary researchers require more sophisticated corrections for unreliability, we should also become more sophisticated about our decomposition of measurement error—beyond item variance to variance due to raters, occasions, and domains. As Brian and Jesse allude to correcting for test-retest reliability, we should appreciate how rarely that is estimated and reported.

Proponents of Generalizability Theory have long demonstrated that item replications are but one facet of error. It's intimidating but necessary to overcome the practical (and political) limitations to estimating the real (and assuredly lower) reliability of scores across plausible replications. As performance assessments and their rater-dependent, task-dependent scores seem poised to rise again, we'll need greater vigilance in clarifying which reliability we mean.

[3] Other approaches include Briggs and Domingue (2013), Castellano and Ho (2015), Reardon and Raudenbush (2009), and Seltzer, Frank, and Bryk (1994).

### 8) I hope we keep these conversations going

As Brian and Jesse suggest, the issues they raise only begin to scratch the surface of a swamp of measurement issues that educational researchers face when using test scores. We haven't even touched linking, equating, discreteness, coarsening, accommodations, or test score inflation. These are just a few of the additional topics with implications for the secondary analysis of test scores.

As Brian and Jesse also make clear, there is thankfully a robust literature in measurement addressing these issues. Yet very little of it is oriented toward secondary researchers from other fields in education. I take Brian and Jesse's pieces as a helpful reminder of the work we still need to do to facilitate wiser research uses of test scores. It's work we can do together.

# References

Ballou, D. (2009). Test scaling and value-added measurement. *Education Finance and Policy, 4*, 351-383.

Barlevy, G., & Neal, D. (2012). Pay for percentile. *American Economic Review, 102*, 1805-1821.

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

Briggs, D. C., & Domingue, B. (2013). The gains from vertical scaling. *Journal of Educational and Behavioral Statistics, 38,* 551-576.

Castellano, K. E., & Ho, A. D. (2015). Practical differences among aggregate-level conditional status metrics: From median student growth percentiles to value-added models. *Journal of Educational and Behavioral Statistics, 40*, 35-68.

Dee, T. S., & Jacob, B. (2011). The impact of No Child Left Behind on student achievement. *Journal of Policy Analysis and Management, 30*, 418-446.

Ho, A. D. (2009). A nonparametric framework for comparing trends and gaps across tests. *Journal of Educational and Behavioral Statistics, 34*, 201-228.

Ho, A. D., & Yu, C. C. (2015). Descriptive statistics for modern test score distributions: Skewness, kurtosis, discreteness, and ceiling effects. *Educational and Psychological Measurement, 75,* 365-388.

Jacob, B., & Rothstein, J. (2016). The measurement of student ability in modern assessment systems (Working Paper No. 22434). Cambridge, MA: National Bureau of Economic Research.

Lafortune, J., Rothstein, J., & Schanzenbach, D. W. (2016). School finance reform and the distribution of student achievement (Working Paper No. 22011). Cambridge, MA: National Bureau of Economic Research.

Nielsen, E. R. (2015). Achievement gap estimates and deviations from cardinal comparability (Finance and Economics Discussion Series Paper 2015-040). Washington, DC: Board of Governors of the Federal Reserve System.

Reardon, S. F., & Ho, A. D. (2015). Practical issues in estimating achievement gaps from coarsened data. *Journal of Educational and Behavioral Statistics, 40*, 258-189.

Reardon, S. F., & Raudenbush, S. W. (2009). Assumptions of value-added models for estimating school effects. *Education Finance and Policy, 4,* 492-519.

Seltzer, M.H., Frank, K.A., and Bryk, A.S. (1994). The metric matters: The sensitivity of conclusions about growth in student achievement to the choice of metric. *Educational Evaluation and Policy Analysis*, 16(1):41-49.