B | Economic Studies
at BROOKINGS

Evidence Speaks Reports, Vol 1, #25
August 11, 2016

Student test scores: How the sausage is made and why you should care

Brian A. Jacob

Executive Summary

Contrary to popular belief, modern cognitive assessments—including the new Common Core tests—produce test scores based on sophisticated statistical models rather than the simple percent of items a student answers correctly. While there are good reasons for this, it means that reported test scores depend on many decisions made by test designers, some of which have important implications for education policy. For example, all else equal, the shorter the length of the test, the greater the fraction of students placed in the top and bottom proficiency categories—an important metric for state accountability. On the other hand, some tests report "shrunken" measures of student ability, which pull particularly high- and low-scoring students closer to the average, leading one to understate the proportion of students in top and bottom proficiency categories. Shrunken test scores will also understate important policy metrics such as the black-white achievement gap—if black children score lower on average than white children, then scores of black students will be adjusted up while the opposite is true for white students.

The scaling of test scores is equally important. Despite common perceptions, a 5-point gain at the bottom of the test score distribution may not mean the same thing in terms of additional knowledge as a 5-point gain at the top of the distribution. This fact has important implications for the value-added based comparisons of teacher effectiveness as well as accountability rankings of schools. There are no easy solutions to these issues. Instead there must be greater transparency of the test creation process, and more robust discussion about the inherent tradeoffs about the creation of test scores, and more robust discussion about how different types of test scores are used for policymaking as well as research.

Testing is ubiquitous in education. From placement in specialized classes to college admissions, standardized exams play a large role in a child's educational career. The introduction of the federal *No Child Left Behind* (NCLB) legislation in 2001, which required states to test all students in grades 3-8 in reading and math, dramatically increased the prevalence and use of test scores for education policymaking.

Contrary to popular belief, all modern cognitive assessments—including the new Common Core tests—produce test scores based on sophisticated statistical models rather than the simple percent of items a student answers correctly. There are good reasons for this, as explained below. The downside is that what we see as consumers of test scores depends on decisions made by the designers of the tests about characteristics of those models and their implementation. These details are typically hidden in dense technical documentation, if publicly available at all.

In a recent paper, Jesse Rothstein and I review some of the peculiarities of modern assessment systems and discuss the implications of these design features for those who wish to use student test scores for research purposes.[i] While psychometricians (experts in the theory and methodology of psychological measurement) are familiar with these issues, in our experience most others are unaware of them and, as a result, frequently misuse test scores. Here I focus on two fundamental aspects of test scores—measurement and scaling—at a level meant to be accessible to readers who may not have a technical background but nevertheless have reasons to be concerned with how student test scores are used and interpreted.

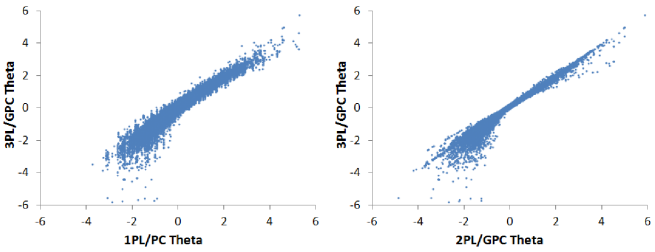
How is student ability measured?

The primary goal of a test is to obtain a measure of ability for an individual, which can then be aggregated up to a group level for various purposes.[ii] The simplest and most intuitive way to generate an estimate of a student's ability is to administer a set of items and measure the fraction of the items the student answers correctly. However, this approach has several limitations. First, if the test has a limited number of items, the estimate of a student's ability will be quite noisy, which means, among other things, that the student would get different scores on different occasions of testing without any change in underlying ability. Second, if there are different forms of the test, as is almost always the case, it will not be possible to compare raw scores across students because the items will differ in difficulty and substance across forms. Third, the simple fraction correct is not a very efficient measure because test items generally differ in their ability to discriminate between more and less able respondents (e.g., the items that almost every student gets correct or fails carry almost no information and yet contribute to the simple fraction to the same degree as items that differentiate between students with more or less knowledge and ability).

For these reasons, modern assessments utilize "item response models" to generate student ability measures.[iii] While this approach solves many problems, the choice of which model to use can have important implications. For example, a fundamental choice in the modern test development process is whether to use a one, two, or three "parameter" model. While the overall correlation of reported test scores across these three models is typically very high, the choice of model can make a sizeable difference for extremely high- or low-performing students.

Figure 1 below comes from a recent technical report on the new Smarter Balanced Common Core assessment. It shows a scatterplot of the student ability measures for 6th grade math scores based on a three parameter model shown on the vertical axis (called 3PL/GPC) versus a one or two parameter model shown on the horizontal axis (called 1PL/PC and 2PL/GPC, respectively). The distinguishing feature of the three parameter model is that it allows for the fact that students might correctly guess answers to test items. The fact that the points deviate from the 45-degree line at low values of student ability illustrates that the two models will assign substantially different scores to some students.

Figure 1. Scatterplot of grade 6 math scores from alternative assessment models



Source: Reproduced from page 272 in http://www.smarterbalanced.org/wp-content/uploads/2015/08/2013-14_Technical_Report.pdf

The length of the test also matters. The longer the test, the less measurement error there will be in student scores. Among other things, this means that shorter

tests will tend to produce more students in the top and bottom proficiency categories. The reason for this is that a student might get particularly lucky or unlucky on several questions, which will tend to have a greater influence on a short exam. Table 1 below shows the differences that arise using a test of 20 versus 41 items.

Table 1. Percent in proficiency level, by test length and score type

	Measured scores un-shrunken		Measured scores shrunken
	Long test (41 items)	Short test (20 items)	Long test (41 items)
Level 1 (low)	20.8	21.5	19.3
Level 2	35.9	35.9	38.7
Level 3	32.5	29.4	33.9
Level 4 (high)	10.8	13.1	8.2

Note: Reproduced from Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice*, 29 (3): 8-14.

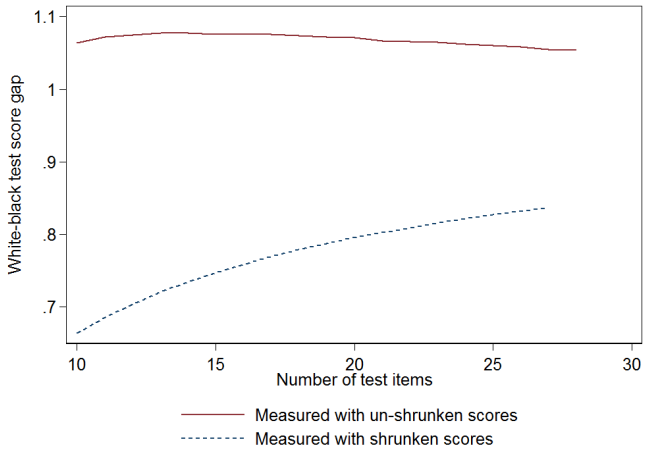
Even after the test designer has chosen the model and the length of the test, the way in which the scores are calculated also matters. Comparing two of the most common approaches to test scoring, one study found that roughly 12.5 percent of students would be classified into different performance levels depending on the technique chosen.[iv]

Another common method produces what psychometricians sometimes refer to as "shrunken" estimates of student ability. In this approach, instead of simply reporting a student's score based on the items he or she correctly answered, the test developer reports what can be thought of as a weighted average of the student's own score and the average score in the population. The reason for this is to account for the measurement error inherent in the student's own score which tends to be larger for scores that are more extreme. The logic is that if a student scores very high on a particular set of test items, most likely it is because she got a little bit lucky along with the fact that she probably has a high level of true ability. Conversely, a student who scores well below average on an assessment is more likely than the student with average scores to have had a bad day when taking the test and will have his or her score bumped up closer to the average.[v] As illustrated by Table 1, this approach to scoring will reduce the proportion of students classified into the top and bottom performance categories, as it pushes high and low scores toward the mean.

Similarly, the use of shrunken scores will lead one to underestimate group differences. Consider, for example, if one wanted to estimate the black-white test score gap using data from the large-scale, nationally representative, federal study of the progress of children through school (The Early Childhood Longitudinal Study of Kindergarteners, or ECLS-K). If black children score lower on average than white children, then the reported difference in the test scores between the groups based on the shrunken estimates will *understate* the black-white gap because, on average, scores of black students will be adjusted up, toward the population mean, while the opposite is true for white students.

Figure 2 illustrates several of these points. The data comes from a simulation that assumes an actual black-white test score gap of one.[vi] Note that the unshrunken estimates yield a gap of roughly one that does not depend much on the length of the test. In contrast, the shrunken scores significantly understate the gap, particularly on exams with fewer items.

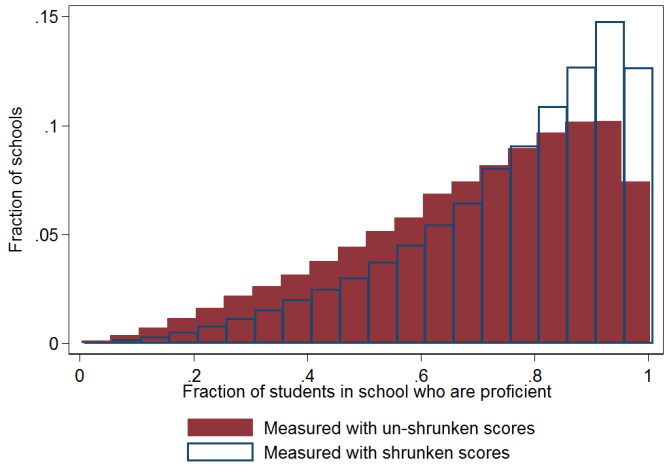
Figure 2. Simulated black-white test score gap



In the same way, the choice of scoring method influences school-level performance measures, which can impact accountability ratings. The use of shrunken scores will reduce the differences in performance across schools, just as the use of unshrunken scores tends to increase across-school differences. Figure 3 shows the distribution of school proficiency rates across schools under standard and shrunken scores, again based on simulated data. Relative to standard score estimates, the use of shrunken scores

produces fewer schools with very low proficiency rates and, perhaps surprisingly, more schools with very high proficiency rates. The reason is that the proficiency threshold is set below the mean of the whole population taking the test. Some low-performing students will have their scores pulled toward the mean and above the proficiency threshold. In contrast, the shrinkage of high-performing students toward the mean will never push these students below the proficiency threshold.

Figure 3. Distribution of school proficiency rates for shrunken versus unshrunken test scores



In an effort to increase the precision of estimated student ability measures, several well-known assessments incorporate student background characteristics as well as student responses to test items into test scores. As before, a student's final score is constructed as a weighted average of his test responses and the average score of other examinees. Instead of being shrunk toward the overall mean, a student's performance is shrunk toward the predicted performance of students with similar background characteristics. This is true for the National Assessment of Educational Progress (NAEP) and the Program on International Student Assessment (PISA) as well as several lesser-known assessments.

As a consequence, if a black and white student respond *identically* to questions on the NAEP assessment, the reported ability for the black student will be lower than for the white student, reflecting the lower average performance of black students on this assessment.

This does not bias the average black-white test score gap. The average score of all black students remains the same because the scores of high-performing black students are pushed down just as the scores of low-performing black students are pushed up, as is the case for white students. However, individual scores are affected, which can create important biases in more complex secondary analyses.[vii]

How is student ability reported?

Once test developers have generated a "raw" estimate of student ability, they must decide on what scale to report the scores. For example, the SAT college entrance exam is scaled so that each section has a mean of 500 and minimum (maximum) scores of 200 (800). The SAT's competitor, the ACT, uses integers between 1 and 36 for each of four subjects, with means around 21.

Most people view test scores as what statisticians refer to as "interval" measures. This means that a one unit change in the measure has the same meaning at any point on the scale. Take the case of temperature. We associate an increase of 5 degrees the same amount of additional warmth whether it involves going from 20 to 25 degrees or from 80 to 85 degrees.

Unfortunately, this is not the case with cognitive ability measures. Test scales are completely arbitrary. For example, there is no reason the College Board could not assign the lowest performing student on the SAT a score of 100, or have the highest score be 1000. And there is no reason to believe that the difference between a score of 300 and 350 reflects the same increase in knowledge as the difference between a score of 700 and 750.

At best, test scores are "ordinal" measures, meaning that they allow you to *order* students on a continuum from lowest to highest ability. We can confidently state that a student who scores 750 has more knowledge or skill than the student who scores 700. It is just not clear *how much* more.

This fact has important implications for virtually any way that policymakers seek to use test scores. Consider the case of the black-white test score gap. Past studies have found that white students typically score about one standard deviation higher than black students, before accounting for important socioeconomic factors such as family income. Moreover, many studies find that this difference actually grows over time.[viii] However, a recent study documents that the change in the black-white test score gap between kindergarten and third grade can be as small as zero or as large as 0.6 standard

deviations depending on how one chooses to scale the test.[ix]

Teacher evaluation is another good example. In recent years, many districts have started to use measures of teacher value-added as part of its determination of promotion, tenure, and even compensation. A teacher's "value-added" is based on the how much improvement his or her students make on standardized tests during the school year (sometimes adjusted for various demographic characteristics). A teacher whose students grew by, say, 15 points is considered more effective than a teacher whose students only grew 10 points. However, if the students in these classrooms started from a different baseline, then this type of comparison depends entirely on the scaling of the exam. For example, it might be the case that a teacher who raises the scores of low-achieving students by 10 points has provided the students more than her colleague who manages to raise the scores of higher-achieving students by 15 points.

What if scores are standardized? In policy evaluations, test scores are often presented in terms of "standardized" scores (also known as z-scores), which are calculated as the student's score minus the average score, and then divided by the standard deviation of scores. While these measures can be useful in some contexts, they are *not* a solution to the scaling issues described above.

Looking forward

The issues that arise in quantitative analysis of cognitive traits are only becoming more salient. The landscape of testing in US schools is changing rapidly, driven by the widespread adoption of the Common Core state standards for K-12 education. There is discussion of developing standardized assessments aimed at college students. Moreover, psychometric methods are spreading beyond cognitive skill assessment, and are now used to create measures of "non-cognitive" traits such as persistence, self-esteem, and socio-emotional regulation.

So, what should the conscientious analyst or policymaker do? There are no easy answers. With regard to the problem of scaling, one approach is to focus on measures that emphasize students' *rank* as opposed to their actual score, for example, by using percentile scores. Indeed, there are rank-based measures to characterize achievement gaps,[x] and a popular approach to calculating teacher value-added relies on student ranks rather than absolute scores.[xi] For researchers, we recommend a greater effort to test the robustness of their results to changes in the test score scale.

Perhaps most importantly, there must be greater transparency about how test scores are generated. Researchers, policy analysts and the public need to better understand the tradeoffs embedded in the various decisions underlying test scores. Only then can we have a productive discussion of the direction to take.

---

[i] Jacob, Brian and Rothstein, Jesse (2016). "The Measurement of Student Ability in Modern Assessment Systems." National Bureau of Economic Research, Working Paper #22434.
[ii] Here I use the terms ability, proficient and achievement interchangeably, though in the original article we discuss the distinctions between these terms more carefully.
[iii] The details of item response theory (IRT) are well beyond the scope of this essay. For a more complete discussion of item response theory models, see van der Linden, Wim J and Ronald K Hambleton, Handbook of Modern Item Response Theory, Springer, 1997; and Embretson, Susan E. and Steven P. Reise, Item Response Theory for Psychologists Multivariate Applications Series, Lawrence Erlbaum Associates, Inc., 2000.
[iv] Kolen, M. J., & Tong, Y. (2010). Psychometric properties of IRT proficiency estimates. *Educational Measurement: Issues and Practice*, 29 (3), 8-14.
[v] This is not common for state mandated assessments where scores are reported to parents, but is almost universal among assessments conducted by the Department of Education, including the widely used Early Childhood Longitudinal Study, the National Educational Longitudinal Study of 1988 (NELS:88), the Educational Longitudinal Study (ELS), and the High School Longitudinal Study (HSLS).
[vi] Both shrunk and unshrunk estimates are generated from a one-parameter (Rasch) model for simplicity.
[vii] Recent administrations of the NAEP use hundreds of student and school characteristics, including student demographics (like race, gender and age), family background characteristics (like parental employment and

parental education), school characteristics (including racial composition of the school and whether a school is in an urban location), student self-reports of study habits and school performance (including overall grades, expected educational attainment, time spent on homework), and teacher reports of aspects of the curriculum and of school policies. In most instances, the inclusion of such a large set of additional variables reduces the likelihood of bias. However, the NAEP does not include variables that are likely to be of interest for policy evaluations, such as whether the school offers performance pay to its teachers, whether the district has any type of school accountability policy or the type of school finance system in the state. This may mean that program evaluations using NAEP scores as outcomes will understate programs' true effects.

[viii] 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*, 86 (2): 447-464.

[ix] 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.

[x] A percentile-percentile plot comparing treatment and control groups would allow the researcher to fully characterize how the two distributions compare without relying on a particular scale. Likewise, one can calculate the probability that a randomly chosen black student will have a test score higher than a randomly chosen white student. See, for example, the following studies: Reardon, Sean, "Differential Growth in the Black-White Achievement Gap During Elementary School Among Initially High- And Low-Scoring Students," Institute for Research on Education Policy & Practice Working Paper, 2008, 7; Ho, Andrew D. and Edward H. Haertel, "Metric-Free Measures of Test Score Trends and Gaps with Policy-Relevant Examples (CSE Report 665)," Technical Report, Graduate School of Education & Information Studies University of California, Los Angeles 2006; Ho, Andrew Dean, "A Nonparametric Framework for Comparing Trends and Gaps Across Tests," *Journal of Educational and Behavioral Statistics*, June 2009, 34 (2): 201-228.

[xi] This approach is referred to as a student growth percentile model. However, use of teacher value-added measures from this model still requires some interval-like assumptions. For example, there is no assurance that a given increment to a teacher's median growth percentile is equally easy to achieve at all points in the teacher or student distribution.