Conceptual Articles

Title of Article/Chapter	Student test scores: How the sausage is made and why you should care
Source Title (journal, book, etc.)	Economic Studies at Brookings. Evidence Speaks Reports Volume I, #25
Author(s)	Brian A Jacob
Year	2016
Participant Focus (teachers/students/etc.)	Student and teacher focused
Topic	Standardized assessments
Main Findings	This article discusses how the different types of test scores are used for policymaking as well as research. It focuses 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. The lack of transparency in the creation of standardized assessments, the logistics behind the statistical scoring and the effects it has on students and teachers. 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. The introduction of the federal No Child Left Behind, (NLCB) 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, aka accountability. The truth is that all modern assessments, produce test scores based on sophisticated statistical models rather than the simple percent of items a student answer correctly. For this reason, modern assessments utilize “item response models” to generate student ability measures. For example, a fundamental choice in the modern test development process is whether to use a one, two, or three “parameter” model. 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’s ability illustrates that the two models will assign substantially different scores to some students.

	The length of the test also matters. The longer the test the less measurement error there will be in student scores. 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. Another method refers to as "shrunken" estimates of student ability. 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. The use of shrunken scores will reduce the differences in performance across schools, just as the use of un-shrunken scores tends to increase across-school differences. 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. Instead of being shrunk toward the overall mean, a student's performance is shrunk toward the predicted performance of students with similar background characteristics. Additionally, these scales give us 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.