What's Wrong With Inferences From Test Scores?


Edward Haertel
Stanford University

Educational testing and test reports are ubiquitous. By and large, people who hear reports that scores on state tests have increased since last year, or that the Black-White achievement gap is closing, or that 81 percent of Massachusetts fourth graders are proficient readers [editor--see page 192] think they understand pretty well what those reports mean. The logic of using test scores to evaluate schools, inform educational decisions, and influence curriculum and instruction is rarely questioned. The principal outcome of schooling is student learning; tests measure student learning; so test scores can show which students are meeting learning expectations and which schools are most effective. Test scores are objective, scientific, and easy to understand, so what's the problem?

It turns out that there are a lot of problems, and that we would do well to try and understand them better. *Measuring Up*, by Daniel Koretz, is an excellent place to start. This book is hard to classify. It is too sophisticated to be called a primer. There are no equations, so it can't be a measurement book. (Also, it is entertaining to read.) It says good things about testing and test use, and takes apart some arguments of testing opponents, so it can't be an anti-testing book. But, it raises profound challenges to the interpretation of score trends on high-stakes tests, to the meaning of achievement trend and gap reports in terms of "percent proficient," to the interpretation of cross-national achievement comparisons, and to popular assumptions about testing of students in special populations, including some assumptions written into law. So, it can't be a pro-testing book, either.

The book is based on a course Koretz has offered at Harvard to help nonspecialists become more intelligent consumers of tests and testing.[editor--page 3] He does a great service by clearly explaining measurement principles in the context of widespread testing uses and misuses. The book is rich with examples and helpful analogies, and in addition to the basics of reliability, validity, standard setting, scaling, and score equating, it nicely
explains the (sometimes flawed) chains of reasoning from examinees' test responses to various test score interpretations and educational policies.

Measurement is ubiquitous in the sciences, and many readers will be familiar with ideas like sampling error, measurement error, or scale conversions. But social measurements with stakes attached are a different matter. Achievement tests are used deliberately to influence schooling, as when reformers press for rich, complex "performance assessments" to focus instruction on "higher-order skills." [editor--page 62] Predictable responses to high-stakes testing, like a narrowed instructional focus on just the material tested, can lead to score inflation, undermining intended inferences from test scores. Some advocates of high-stakes testing as an educational reform tool have argued that even if there is some score inflation, tests cover material students should know, and so rising scores must signify some degree of real improvement. Koretz offers persuasive rebuttals, supported with empirical evidence.

The book is rich with concrete examples. Pointing out that even large score differences may show inconsistencies from one test to another, Koretz discusses the 2003 eighth-grade mathematics gap between the United States and Norway. That year, the TIMSS assessment place the U.S. far ahead of Norway, while the PISA assessment placed Norway ahead of the U.S. by a statistically significant margin.[editor--p. 319] Explaining selectivity bias, he describes an influential study that contributed to the popular idea of algebra as a "gatekeeper" course, showing how that study's conclusions were obviously unwarranted.[editor--p. 126] In clarifying within-group versus between-group variation in test scores, he presents the surprising statistic that if the average differences in test scores between the major racial and ethnic groups in the United States could somehow be eliminated, making each group's score distribution match the current distribution for nonhispanic white students, the total variability in scores would shrink by less than ten percent. (His estimates, using different subject-grade combinations, ranged from 0.5 percent to 9 percent.)[editor--pp. 141-142]

A major theme, reflecting Koretz's own research over the years, is score inflation on high-stakes tests. Optimistic reports of year-to-year improvement on state tests under NCLB should not be trusted. When the same score trend is estimated using a low-stakes "audit" test (e.g., the National Assessment of Educational Progress) and a high-stakes test (e.g., a state test), the audit test tends to show smaller gains over time, sometimes much smaller. When a school district changes from one test to another and then after several years the earlier test is readministered, it has been found that the performance gains shown over time on the new test were in fact accompanied by a corresponding decline in performance as measured by the old test.[editor--pages 245-246]

The final chapter offers reasoned advice on appropriate test use and cautious test interpretation, concluding as follows: "In all, educational testing is much like a powerful medication. If used carefully, it can be ... [a] powerful tool for changing education for the better. Used indiscriminately, it poses a risk of various and severe side effects. Unlike powerful medications, however, tests are used with little independent oversight. Let the buyer beware." [editor--pp. 331-332]