Acad. Quest. (2009) 22:389–393
DOI 10.1007/s12129-009-9111-0

REVIEWS

Measuring Up: What Educational Testing Really Tells Us, by Daniel Koretz. Cambridge, MA: Harvard University Press, 2008, 368 pp., $29.95 hardbound.

Aca-metrics

Thomas Wood

Published online: 2 July 2009

Daniel Koretz is the Henry Lee Shattuck Professor of Education at Harvard. His research has focused mainly on educational assessment and its use as a tool of education policy. His principal area of concentration has been high-stakes testing, including the effect on score gains. He is best known in the field for documenting score inflation in state assessment exams in secondary schools.

Measuring Up: What Educational Testing Really Tells Us had its genesis in a course Koretz teaches at Harvard for master's students with little or no mathematical background. It has been a popular and successful course there, and one can see why. The author has a gift for simple, non-technical exposition. If you are looking for a good, non-technical discussion of topics in testing and measurement like validity, reliability, measurement error, bias, and the benefits and limitations of assessment and testing generally, this may be the book for you.

Thomas Wood is academic correspondent of the National Association of Scholars, One Airport Place, Suite 7 & 8, Princeton, NJ 08540–1532; nas@nas.org.

A Commonsense Approach to Testing

Koretz's approach to the issues is sane and sensible. As a teacher and parent as well as a professor specializing in the field, Koretz has been a strong proponent of accountability in education. He believes that testing is an essential part of educational accountability. However, while testing provides important information, one must also be aware of its inherent limitations and potential pitfalls.

The catalog of misuses and abuses of the "testing-accountability" movement in this book is long and sobering. "The shift from using tests for information to holding students or educators directly accountable for scores," he says, "is beyond a doubt the single most important change in testing in the past half century." Koretz believes that test scores should be treated as specialized information that supplements but

does not replace other information about students' performance. The most persistent meme in *Measuring Up* is that is a terribly bad idea to rely on testing to provide a single, simple measure of educational outcomes. Just say no, he says, to any proposal to use test scores as the sole or defining criterion to measure students and schools.

Koretz's focuses on K–12 and does not address the issue of outcomes assessment in higher education directly. However, his criticism of testing abuses in secondary education applies to higher education as well. Koretz would likely be a strong advocate of outcomes assessment in higher education, but would insist on its problems and pitfalls. In particular, those seeking a single, simple test to measure how well institutions of higher education are doing will find a great deal in *Measuring Up* to deflate their hopes.

E. F. Lindquist and Some Specific Principles

Koretz does not claim that his positions on the issues are new. In fact, he almost invites the reader to regard his views as little more than an extended, updated footnote to the 1951 paper, "Preliminary Considerations in Objective Test Construction," by E.F. Lindquist, a pioneer of educational testing.

Like Lindquist, Koretz insists that tests cannot measure many important factors. The goals of education are diverse, and only some are amenable to standardized testing. Tests, for example, cannot measure how well schools do in developing creativity or motivating students for life-long learning. Both are important goals for education, but we don't have tests for them, and likely never will.

A test contains only a finite number of items, and is therefore only a sample. At best, it can measure only part of what one wants to measure, since it is time-limited. Koretz is accordingly a strong proponent of holistic assessment, which he says the Graduate School of Education at Harvard, his institution, uses. At the same time, he favors the use of standardized testing as one part of an assessment program. Unlike grades, for example, standardized tests can be compared meaningfully across institutions and states.

Though I did not find mention of this in the book, it is clear from his statements of fundamental principles that the limitations of testing become more severe the less focused the curriculum. It should be easier to develop meaningful tests for accounting (and even easier to develop tests for subparts of accounting) than to develop tests for outcomes assessment in liberal arts education.

Bias in Testing and Performance Assessment

Since Koretz specializes in secondary education, his discussions are usually built around tests and other examples in this area. This may disappoint readers who want more examples from testing and measurement issues in higher education, but it does not really diminish the value of *Measuring Up* to those whose primary interest is higher education. Koretz's general principles apply just as well to examples taken from higher education, like the GRE or the SAT, as they do to K–12 testing.

Koretz does not ignore higher education completely. His discussion of bias in testing takes as its example what he calls the "Berkeley effect." Koretz does not attempt to show that the University of California at Berkeley's admissions process is free of racial bias; what he shows very clearly is that the disproportionate representation of groups at Berkeley can arise without any bias at all.

Koretz makes just two substantive assumptions: that a substantial difference in racial group scores exists, and that the distributions have most students grouped near their group's average. Both are well-established claims. The demonstration also uses cut-off scores. As Koretz acknowledges, this is a bit artificial, and is in fact an admissions policy that he does not recommend, because tests do not give consistent scores from one measurement to the next. A significant percentage of test-takers—those close to the cut-off score—will move between acceptance and rejection from one measurement to another even in the case of the best constructed tests. In testing and measurement, this unreliability is called measurement error. Nevertheless, cut-off scores can be useful in demonstrations because they make the mathematics particularly simple and compelling.

The mean scores of African Americans are so much lower than those of the rest of the applicant pool that the only selection process that has no adverse impact on them is a wholly unselective one (open admissions). If the cut-off score is set at the mean for the entire applicant pool, the representation of African American students falls sharply, from 15 percent to 6 percent. If the cut-off score is set at one standard deviation beyond the mean, the representation of African American students falls to 1 percent. One standard deviation above the average represents an SAT score of 633. This is not a very high level of selectivity according to the standards set by the most selective institutions. In 2006–2007, 72

percent of the students at the University of California at Berkeley had SAT verbal scores over 600, and 81 percent had math scores over 600.

Koretz also has some interesting things to say about performance assessment, and though his examples come from secondary education, the discussion is relevant to the Collegiate Learning Assessment (CLA), probably the most discussed and popular test for outcomes assessment in higher education today.

Performance assessment, sometimes called "authentic assessment," aims to avoid some of the pitfalls of other kinds of standardized testing. It has a fairly long history, going back to the late 1980s at least, and has been motivated by the widespread effort to replace the multiple-choice format with other kinds of tests, such as short-answer constructed-response items, items requiring longer written answers, hands-on performance (for example, using a scientific apparatus), portfolio assessments (defined as students putting together various kinds of work over the course of the term), and group work.

Because we don't put students in school simply in the hopes that they will do well while they are enrolled, educators have looked for tests that simulate the real world. Outside of school, problems are commonly complex—not artificially broken down into small pieces as they are in school. This limitation is inherent to the educational process. Teachers must necessarily focus in the classroom on what is manageable, in the hopes that students will transfer what they learn to real-life situations after and outside school.

While performance assessments, with their emphasis on problem-solving, reasoning, communication, and higher order skills, are useful and important, they have problems and limitations as well. Performance assessments tend to be expensive, because they take more time to complete and score than other kinds of standardized tests. It is often difficult to score them reliably, and harder to make the scores comparable from year to year and from school to school.

Another problem is the difficulty of determining what skills are being invoked in a performance task. Experience has shown that a casual examination of a test's content ("it seems valid on its face") is insufficient to determine its validity. For example, the skills that students will use in a performance assessment cannot simply be inferred from the "face" or surface of the test.

Koretz gives a real-life example of this problem. He arranged to meet three friends at a location in Manhattan

for brunch. His friends had trouble locating the venue, even though they were familiar enough with Manhattan to know the major avenues. They even knew when the avenues in the area and the venue's particular avenue reached zero. They were also able to figure out, with a little prompting, the rate of increase of the addresses on the avenues. Like Meno's slave, Koretz had the solution. Put mathematically, the solution could have been found by solving a simple linear equation in one variable: y=a+bx. His friends' difficulties were due to the fact that they were not in the habit of applying the elementary mathematics they had learned in school to real-life problems.

One cannot infer from this example that Koretz's friends didn't know the algebra. All three had, in fact, taken several semesters of college mathematics. In all likelihood, his friends did know the algebra; it's just that they could not or did not recall it under those circumstances. To test for how much algebra they still knew, a multiple-choice exam would probably be preferable to a performance assessment.

Recommendation

Measuring Up is a very useful, non-technical introduction to some problems and issues in educational testing. It does not aim to develop the level of expertise Charles Murray in *Real Education* has claimed America's elite must have in order to have informed opinions about policy issues. Nevertheless, it is a helpful, informative work for those who are not able to spend the time required to learn the mathematics. The book includes a comprehensive index, so the reader who wants to zero in on specific topics without reading the entire book can do so.