EDUCATION WEEK
Published Online: March 11, 2014
Published in Print: March 12, 2014, as **Validity Counts: Let's Mend, Not End, Educational Testing**

COMMENTARY

Let's Mend, Not End, Educational Testing

By Madhabi Chatterji

The Common Core State Standards and accompanying K-12 assessments have recently sparked a fierce national backlash against testing. Sound educational testing and assessment are integral to good teaching and learning in classrooms and necessary for evaluating school performance and assuring quality in education. Rather than throw the baby out with the bathwater, I propose a more considered, "mend, not end" approach to testing, assessment, and accountability in America's schools, with validity at the forefront of the conversation.

Mending begins with understanding that most commercial standardized tests are designed to serve particular purposes well, for particular populations, and can support only particular decisions at best. To uphold validity principles in practice, it is worthwhile to ask: Are we using the test for the originally intended purpose, or for another purpose that taxes the tool beyond its technical limits? Multi-purposing a test indiscriminately is not a good idea from a validity standpoint, despite its efficiency.

Validity deals with the meaningfulness of test scores and reports. Technically, validity is determined by the built-in features of a test, including its overall content, the quality of test questions, the suitability of metrics for the domains tested, and the reliability of scores. In addition, how and where a test's results are applied, and the defensibility of inferences drawn, or actions taken, with test-based information affect the levels of validity we can claim from the scores and reports.

According to testing standards published by the American Educational Research Association, the National Council on Measurement in Education, and the American Psychological Association, once a validated test is taken out of its originally intended context, we may no longer be able to claim as much validity for a new population, purpose, or decisionmaking context, nor with as much certainty.

—Jori Bolton for Education Week

New proposed uses call for more tests of a test—a process called "validation." New evidence must be secured to support a new or different action. Too often, this basic guideline is overlooked, particularly under high-stakes accountability policies like the federal No Child Left Behind Act or the common core. Validity oversights also happen with relatively low-stakes international-assessment programs like the Program for International Student Assessment, or PISA.

No Child Left Behind, signed into law in 2002, mandated testing of all students in grades 3-8 to

measure progress of schools based on results of annually administered achievement tests. Variable state-set standards toward manifestly unattainable growth targets of "adequate yearly progress" and "universal proficiency" by 2014 stretched many school evaluation systems beyond their technical capabilities. NCLB's public rewards and sanctions based on school performance led to "teaching to the test," spuriously raising student test scores without lasting or replicable learning gains. This repercussion, in and of itself, undermined the validity of inferences from test scores, which no longer indicated clearly what students actually knew on tested domains.

Ripple effects of NCLB took hold in other school evaluation contexts, too, threatening validity in additional ways. Even the most enlightened and progressive of districts were pressured into missteps by high-stakes-testing requirements. In 2005, for example, Montgomery County, Md., sought to ratchet up performance and close achievement gaps districtwide by identifying its own model schools and school practices—a laudable goal. However, the county's selected measure of student achievement, aggregated to serve as an indicator of school performance in "value added" evaluation models, was the combined math and verbal SAT score of high school students.

Recent efforts have sought to **align the SAT** more with college-readiness and common-core standards, but at the time of the 2005 report, **"Value-Added Models in Education: Theory and Applications,"** the validity of the SAT as an indicator of school-level outcomes was questionable. A college-entrance exam, the SAT is designed to predict how well students will perform as college freshmen, with limited validity as a curriculum-based achievement test. Variability in the levels and kinds of coursework taken by students could significantly affect the meaning of the scores, weakening inferences about student achievement in K-12 scholastic programs.

Further, because students opt to take the SAT, test-takers are likelier to be stronger academically and inclined toward college, come from wealthier families, or have exposure to stronger schooling experiences. Self-selection biases schools' aggregate SAT scores, complicating interpretations of what caused them to rise or fall.

Neither the school district nor the SAT is at fault. Rather, punitive accountability measures tied to test results in the larger context of reforms may be called into question. The power of such accountability mandates influences decisions of even trained analysts, regardless of stakes tied to local actions.

In the current context of the common core, a parallel drama is playing out. The common-core tests now being developed have been criticized as too long, superficial or overly narrow, and out of alignment with the curriculum and common-core standards. Educators, parents, and local officials reasonably fear that, yet again, tests are serving as blunt policy instruments to drive top-down reforms with inadequate time and resources for designing deeper curriculum and assessments to match, with little or no professional development of teachers and school leaders and in neglect of critical supports that schools need to succeed.

With ill-prepared schools and students, what will the test results really tell us about student learning and the quality of schooling?

Yet, were the same tests implemented after standards were refined, teachers and schools readied, parents and students oriented, tests validated to measure what students actually learned better, and results freed from external rewards and sanctions, the results might be more meaningful. Further, the anti-testing backlash might well disappear.

No one was celebrating the recently released results on the 2012 PISA, ranking American 15-year-olds below their peers in many other industrialized countries, particularly in math and science. But how

meaningful and defensible are the intercountry comparative averages, given the differences in culture, educational opportunity, and backgrounds of 15-year-olds tested from different nations?

Despite popular claims, these sample survey statistics also cannot tell us much about whether particular regional reforms failed or succeeded. Interpreted carefully, PISA results yield useful benchmarks within particular nations, opening opportunities for education systems to improve.

Misinterpretation of PISA's intercountry rankings, however, reflects a larger syndrome of misuse of educational assessment results and hand-wringing about public education that could easily be avoided.

Most standardized instruments rest on a solid base of scientific knowledge that dates back to the first half of the 20th century. These tools have documented achievement gaps in ethnic, gender, and socioeconomic groups reliably, furnishing policymakers, educators, and our society at large with evidence for improving conditions.

But misuse and misinterpretation of standardized-test results is a pervasive problem in educational assessment that threatens levels of validity, especially in high-stakes testing contexts. Here's an area where scholars and practitioners; test-makers and test users; educators, parents, and students; and the media could work together to make a difference.

These and other issues will be open for debate and discussion in a time-limited blog hosted by edweek.org, to be launched next week and facilitated by James Harvey of the National Superintendents Roundtable and me. **Assessing the Assessments: K-12 Measurement and Accountability in the 21st Century** will feature expert commentary from scholars and practitioners, offering a variety of perspectives on today's critical assessment challenges.

Madhabi Chatterji is an associate professor of measurement-evaluation and education at Teachers College, Columbia University. She is the founding director of the ***Assessment and Evaluation Research Initiative at Teachers College****. Her recent writings include the book* Validity and Test Use: An International Dialogue on Educational Assessment, Accountability, and Equity *(Emerald, 2013).*

Vol. 33, Issue 24, Pages 30,36