NEPC
NATIONAL EDUCATION POLICY CENTER

Thursday, April 2, 2026

Newsletter

Testing and Assessment: 10 Terms to Know Now

As testing season kicks into high gear across the United States, you might be hearing a series of words like "construct," "reliability," and "validity." These seemingly familiar phrases often have specialized interpretations in the world of psychometrics (a.k.a. the study and design of assessments). In a textbook recently published by The Guilford Press, NEPC Fellow Madhabi Chatterji, a professor emerita at Teachers College, Columbia University, defines these and other terms as she explains an approach to testing that is adaptable to different uses, users, test takers, and contexts.

The book takes a multidisciplinary approach, which means it includes a glossary of testing vocabulary that can be applied across fields. Here are some of the terms she defines and further elaborates upon, since the text itself is aimed at graduate students and other advanced users.

Constructs: Constructs are hypothetical concepts measured by assessments. You can't see them and they cannot be directly measured. For example, weight, which can be directly observed, is not in itself a con-

struct. However, let's say the construct you want to measure is sadness—an abstraction that cannot be measured directly with a scale or a ruler. Body weight might be one of multiple pieces of information you collect to measure your construct because your evidence suggests that both overeating and undereating are associated with sadness. Other pieces of information might include number of hours per night slept, or self-reported emotions. These directly observable sets of indicators are known as "domains."

Criterion-referenced tests: These exams are designed to provide information about the degree to which test takers have met certain standards. Results are provided relative to these standards rather than relative to the results of other test takers (which is called "norm referencing"). In K-12 education, an example of criterion referencing is the National Assessment of Educational Progress, which reports scores in three categories: basic, proficient, and advanced. By contrast, a norm-referenced score might provide results as percentiles, with a higher percentile indicating that a student has obtained a higher score than other test takers and a lower percentile implying the reverse. On a criterion-referenced test, it is theoretically possible for every student to earn the top (or bottom) score, because the goal is to measure the degree to which test takers have met a standard. On a norm-referenced test, scores should always show a broad distribution since they indicate how a test taker performed relative to others who took the exam. This is sometimes referred to as scoring "on a curve," which is a reference to the normal curve shape that the scores form when plotted on a graph. (In reality, testing companies may not re-norm scores annually, which means students may be scored on a normal curve established several years earlier.)

Formative assessments: Distinct from summative assessments—which measure final outcomes—formative tests aim to provide feedback that leads to improvement. In a classroom setting, they can help teachers better understand what students do and do know at different points in their learning trajectories. The teachers can then use these results to shift instruction to better meet the needs of learners. Students can also use the results to identify areas where they need more practice. A common example of a formative assessment is an "exit ticket" in which a lesson ends with a brief assessment (sometimes even just a show of raised

hands) that provides feedback to the teacher on what students have and have not understood. Teachers can use that information as they plan the next lesson.

Opportunity-to-learn bias: Opportunity-to-learn biases are systematic and measurable differences in certain test questions or sets of test questions—among students who have the same overall test scores. For instance, if a district superintendent looks at all the students who earn high scores on a math test and notices that top scorers from School A all got a question about fractions wrong, it is possible that School A's teachers did not cover this topic, creating an opportunity-to-learn bias.

Portfolio-based assessments: Portfolio-based assessments use multiple methods and incorporate observations collected over time. In a K-12 setting, a portfolio-based reading assessment might draw upon essays, spelling test scores, oral presentations, and reading diaries collected during an entire semester and collected into a separate and unique portfolio for each student.

Reliability: A reliable test is consistent across test takers and conditions. For example, if I take an assessment that measures my political views three different times, and each time the score generates wildly different results—despite the fact that my views have remained the same—then the test itself is probably not very reliable.

Test equating: These are statistical methods used to generate equivalent results from different forms of the same test. This method is often used by large, standardized testing programs that attempt to limit cheating by varying the questions that appear on tests administered at different times of the year. For instance, equating is used to ensure that a student who takes a college admissions exam in the fall does not have a higher or lower chance of earning a given result relative to a student who takes the exam in the spring.

User-centered assessment design: The focus of Chatterji's book, user-centered design—as its name implies—centers on real-world users, uses, and useability. Chatterji describes user-centered assessments as clearly defining what is to be measured, who will be tested, why the measurement was designed, how the test is to be administered, interpreted,

and scored on a numeric scale.

Validity: Validity has many different aspects but, broadly speaking, it is the degree to which tests actually measure what they purport to measure without introducing unrelated statistical error or "noise" that distorts the results. Validity may vary by use, which is why it is important to design tests with their end uses in mind and to use tests for their designated purposes. One real world example is the use of student reading and math test scores to evaluate teachers. Critics of this approach argue that teacher evaluation is not a valid use of these exams because (among other reasons) they were not designed with this objective in mind.

Vertical scaling/linking: In K-12 education, vertical scaling is used to attempt (with varying degrees of effort and success) to link results from tests administered at one grade level to tests administered at another grade level, even when they address different aspects of a subject. For instance, vertical scaling might be used to compare fourth grade reading scores with third grade scores, even though the test questions may assess different aspects of literacy that students are expected to master at different grade levels.

User-Centered Assessment Design: An Integrated Methodology for Diverse Populations, by Prof. Madhabi Chatterji, was published in 2025.

NEPC Resources on Assessment

This newsletter is made possible in part by support provided by the Great Lakes Center for Education Research and Practice: http://www.greatlakescenter.org

The National Education Policy Center (NEPC), a university research center housed at the University of Colorado Boulder School of Education, sponsors research, produces policy briefs, and publishes expert third-party reviews of think tank reports. NEPC publications are written in accessible language and are intended for a broad audience that includes academic experts, policymakers, the media, and the general public. Our mission is to provide high-quality information in support of democratic deliberation about education policy. We are guided by the belief that the democratic governance of public education is strengthened when policies are based on sound evidence and support a multiracial society that is inclusive, kind, and just. Visit us at: http://nepc.colorado.edu