GUIDE

Not all assessment data is equal: Why validity and reliability matter

INTRODUCTION

Instructional time is critical. Teachers need time to teach, and students need time to learn. But instructional time isn't always protected. When it comes to assessment in particular, it's easy to argue that testing takes too much time and yields too few benefits.

It's tempting to define assessments merely by their duration: long or short. But time needed for completion isn't as important as whether an assessment has enough items to yield valuable data educators can use to improve student outcomes.

Shorter assessments tend not to ask enough questions and can cost more instructional time and resources in the long run than they save in the short term. Why? Because their data can misrepresent what students know and are ready to learn, leading to missteps in instruction and to reteaching. Students can also be identified for remediation when they don't truly need it, wasting the time of special ed teachers. Or they can miss opportunities for advanced material and become bored, requiring their teachers to spend time on reengagement.

Short or long, assessments also need to be intentionally built to measure the things they intend to measure. It sounds obvious, but if a test doesn't have well-designed test questions that are tightly aligned to subject matter, it won't yield useful information about what a student knows and is ready to learn in that content area. And if the assessment can't provide useful information to inform decision-making, why give it at all?

Assessments designed to provide accurate, actionable information about what students are ready to learn can protect instructional time and do more to help teachers and students set—and meet—academic goals. Assessments that do this focus on the validity and reliability of data. Without those two qualities, it is much harder to truly reach students where they are and give them what they need to succeed.

"Validity" and "reliability" defined

To grasp the importance of validity and reliability of data, understanding what each word truly means for assessment is crucial.

Validity asks a simple question: Is what is being purported to be measured what is actually being measured? Validity comes into play when, for example, a person gets on a scale and sees that they weigh 200 pounds. The purpose of the scale is to weigh people, and it fulfills its purpose. It does not display a temperature or blood pressure reading.

Reliability, on the other hand, is an indicator of precision. That bathroom scale would be unreliable if it displayed a different number after multiple steps onto it. It could very well be providing a valid measure—weight—but the measure would not be reliable if the scale read 200, 220, and 155 all within a matter of minutes.

No assessment can capture the knowledge of a student with the precision a bathroom scale has with weight; all assessments are an estimate of what a student knows and can do. But what high-quality assessments can do is lower the likelihood that the estimate of a student's ability is grossly off track. This is why reliability and standard error of measurement go hand in hand. The more reliable a data set is, the lower the standard error of measurement, that is, the less likely an estimate will be far off. The less reliable the data, the higher the standard error.

What validity and reliability look like in an assessment

When an assessment supplies data that is both valid and reliable, it does two things.

First, it asks questions that allow students to show their knowledge of a particular subject.

Just like the bathroom scale displaying weight instead of temperature, if an assessment is measuring addition it should ask the student for the sum of two and three, not the result of two times three. For a real-world example, consider a word problem in a math assessment. Test design experts must ask themselves if the question is really addressing math skills, or if it's more directly measuring reading

ability because passage length and complexity make it too difficult for some students to demonstrate they understand the actual math concepts.

Second, it asks enough well-targeted questions to provide a reasonable amount of confidence that the answers reflect an accurate estimate of what a student knows.

Just like three steps on the scale, all resulting in a reading of 200 pounds, would leave most people feeling certain about their weight, three correct answers to well-constructed addition questions would allow a teacher to be more sure of where a student was in developing that skill than just one question would.

Garnering quality data takes time. There is simply no way to remove 30% of the questions from an assessment without sacrificing confidence in the insights the assessment provides. And when an assessment is reduced from 50 items to 35 items solely for the purpose of taking less time to complete, it's not just insight into a particular subject—math, for example—that takes a hit. The reduction in items included in the assessment also robs educators of reliable data that pertains to specific instructional areas. For example, in math, that could include information in areas such as real and complex number systems, algebraic thinking, and statistics and probability. When you get down to the specific standard, there's no utility at all with so few questions.

How good data helps students

Valid and reliable data helps educators get the most from instructional time and paves the way for student success. It does this by improving confidence in placement decisions, making differentiation and goal setting more effective, and more accurately predicting performance on summative and college readiness exams.

Improves confidence in placement decisions

Assessments that don't provide valid, reliable data can increase—often dramatically—the chance that student needs may be misclassified. This can negatively impact all students, regardless of proficiency level. It could keep a student performing below grade level from getting the support needed to grow just as it could prevent a student performing ahead of peers from qualifying for a more rigorous program. Students are also more likely to be placed in programs they don't actually need, wasting the student's time and costing the school both time and money.

When educators can make instructional decisions based on higher-quality data, they can place students in remedial or advanced programs with greater confidence that the students belong in—and will benefit from—the programs into which they are placed. These are high-impact decisions that affect students' lives. They are too important to make informed by inferior data.

Makes differentiation and goal setting more effective

How can teachers meet the broad range of needs of every kid in their class? Through effective differentiation. How can students get invested in their learning? Through goal setting.

Valid assessment data gives teachers insight into student achievement relevant to the content they are expected to teach. Reliable data gives teachers the confidence to make instructional decisions based on those insights. High-quality assessments make collecting this data more efficient and effective.

This frees up teacher time to focus on differentiating instruction. Trustworthy assessment data shines the spotlight on student strengths and opportunities for improvement, taking the guesswork out of tailored instruction and allowing teachers to focus on exactly what needs attention so students can meet the learning objectives defined by the curriculum. The strengths and opportunities assessment reveals also lay the groundwork for goal setting. Together, teachers and students can chart an action plan likely to lead to success.

More accurately predicts performance on other tests

A valid and reliable assessment aligns with other tests that purport to measure the same thing, like year-end state assessments and the ACT® or SAT®. This is called concurrent validity. Concurrent validity is a measurement of the relationship between scores on two different tests at a specific point in time. If there is a high level of concurrent validity between two tests, when a student scores high on one, they are more likely to score high on another. So, for example, a student who scored well on an in-class assessment could tackle the SAT with less anxiety and more certainty that they'll do well if the two tests are correlated. Likewise a lower in-class assessment score would be a clear signal that more SAT prep was needed.

Predictive validity, how accurately an assessment at one point in time predicts performance on a future assessment, is especially beneficial in an interim assessment. When an interim assessment has strong predictive validity, you can have confidence in the inferences you're making about a student's status on year-end assessments throughout the year.

Because it's administered up to three times a year, an interim assessment allows teachers and students alike to

work toward the end-of-year state summative with more certainty. Low scores on the first test, early in the fall, can guide goal setting for the entire year. Testing again in winter or spring can show progress made since the start of the school year—while there's still time to course correct before the state test.

The power of good data at the school or district level

The quality of data provided by an assessment has ramifications for schools and districts, too. Valid, reliable data can help educators identify groups of students needing support—say, numerous first-graders struggling with reading—and plan for both remediation and resource allocation. Data from subsequent testing events can then prove efforts were well spent, supporting the development of new best practices, or it can show that a different approach is needed.

Without high-quality data, improvement efforts rely more on luck than reason. And when a school or district shows time and time again that it simply can't gain ground, funding is put in peril and student progress suffers.

In closing

An assessment shouldn't cut corners. Without valid, reliable data, educators simply don't have what they need to foster student success. The deeper understanding provided by high-quality assessments helps educators confidently make decisions that serve each student—decisions that can have a lifelong impact.

nwea

JUN20 | KAP5576