Spotlight on Assessment

Norm-Referenced and Criterion-Referenced Interpretations

Catherine Welch, Ph.D., Stephen Dunbar, Ph.D., and Ashleigh Crabtree, Ph.D.

The purposes of achievement testing are diverse and varied, but the common denominator to all purposes is information. Tests may be formative or summative, they may be based on defined domains or captured through a normative comparison, and they may be captured in daily classroom or in a more standardized testing environment.

But, underneath it all, effective achievement testing is designed to provide **sound, accurate** and **actionable information** for its users. An effective achievement test may successfully provide evidence to support one or more of its articulated purposes of testing. An effective achievement test must also be designed to support inferences about all articulated purposes and must strike a delicate balance among many complementary and sometimes competing and conflicting purposes and uses.

The purposes served by the ***Iowa Assessments***™ are diverse and varied. Figure 1 articulates many of the more common purposes. This document addresses four major purposes served by the ***Iowa Assessments***: tracking student readiness, measuring student-level outcomes, monitoring student growth, and making relative comparisons about a student's performance.

Achievement measures like the ***Iowa Assessments*** have been developed to inform instruction at the classroom level but can do much more than that. Indeed, the purposes supported by the ***Iowa Assessments*** cross with two primary types of interpretations. These two interpretations allow users to make decisions across classrooms, schools, or states and support claims for the validity of score interpretations.

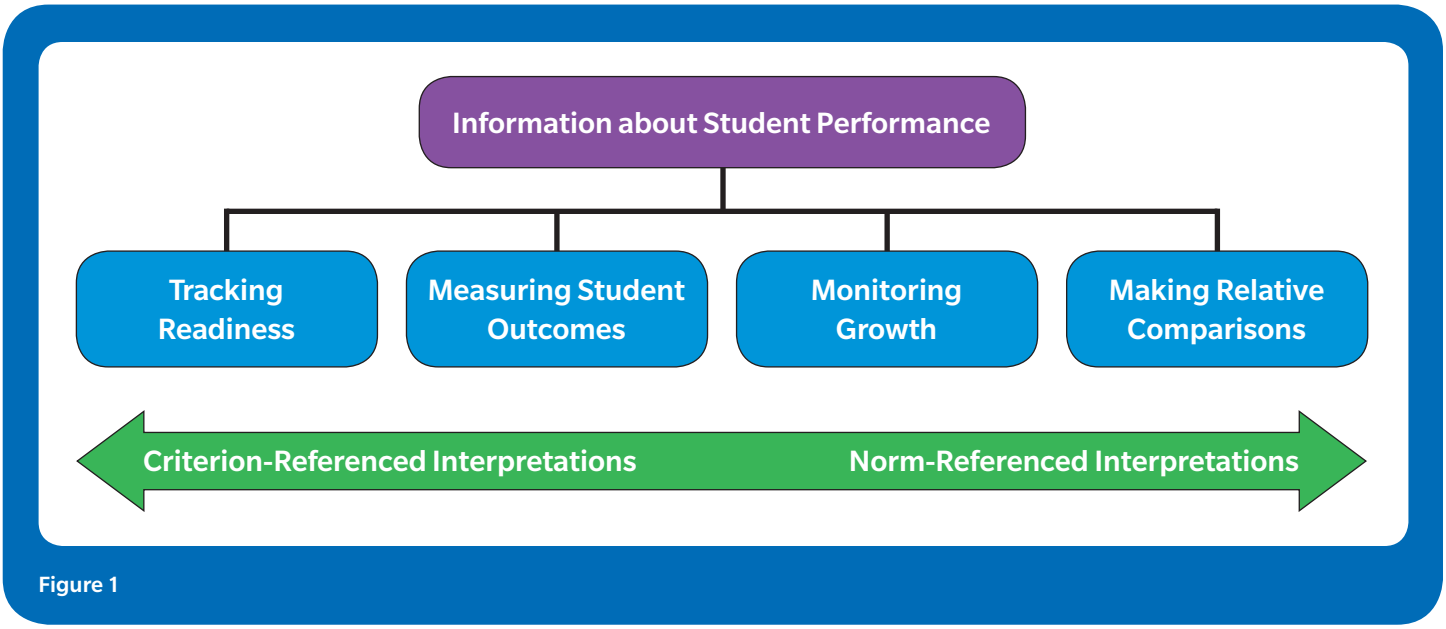

Figure 1

The Standards for Educational and Psychological Testing (2014) clearly state that assessments may be validated for various types of interpretations. Criterion-referenced interpretations are meaningful when a test taker's performance is referenced to a defined criterion domain. The interpretation indicates the level of performance for an individual student or group of students in relationship to a defined criterion domain. For example, the ***Iowa Assessments*** have defined domains of *proficient* and *not proficient*. The ***Iowa Assessments*** also have definitions of readiness for college-level coursework or not-yet ready for college-level coursework. Comparing a student's performance to these defined domains is a criterion-referenced interpretation.

Norm-referenced interpretations are meaningful when a test taker's performance is compared to the distribution of performance of other test takers within the same classroom, school, district, or state. This performance can be at the total test level, or at a finer level of detail such as subskill, content, or item level. For example, the ***Iowa Assessments*** have nationally representative, local or user-based points of reference to support norm-referenced interpretations.

Results from the ***Iowa Assessments*** are reported using a variety of scales designed to assist in score interpretation. Some scales allow a direct reference to the performance of other test takers while others allow an interpretation about the domain or level of performance of interest. Figure 2 provides a quick summary of the various scales available with the ***Iowa Assessments***.

Debunking the myths of Norm-referenced and Criterion-referenced Interpretations

Myth 1 – A test can be norm-referenced (NR) or criterion-referenced (CT), but not both.

False. NR and CR are different types of interpretations, and validity is always tied to interpretation. Collecting validity evidence according to test purpose and use permits varied interpretations about student knowledge and skills. An assessment may have adequate evidence to support both NR and CR interpretations.

This is the most pervasive myth surrounding norm- and criterion-referenced interpretations. However, when validity evidence exists for intended interpretations, it is possible for one test to allow for BOTH types of interpretations.

Myth 2 – A test with national norms only supports norm-referenced interpretations.

False. The existence of national norms does not preclude other interpretations if validity evidence exists to support them. A test that is designed, developed, and validated to support interpretations of readiness, growth, and student outcomes may also have national norms. On measures like the ***Iowa Assessments***, the collection of national norms is a separate research event that allows for the comparisons that go beyond the interpretations of growth and readiness. National norms offer interpretations in addition to interpretations of readiness, growth, and student outcomes.

Scales Designed to Assist in Score Interpretations

- National standard score thresholds for proficient or not proficient
- National standard score thresholds for on-track for college readiness or not yet on-track for college readiness
- Vertical scale to set growth goals and monitor growth toward proficiency or college readiness
- Relative performance (National percentile ranks, Local and state percentile ranks, Grade-equivalents, Stanines, and Normal Curve Equivalents)
- Vertical scale and growth norms to compare student growth to national growth

Figure 2

Myth 3 – A test with national norms is not concerned about content validity.

False. In the design of the ***Iowa Assessments***, content validity is the most fundamental consideration in developing and evaluating the items and the tests. Without content validity, the ***Iowa Assessments*** would be unable to make statements concerning what a student knows and is able to do. Any test that is not concerned with content validity fails at the most basic level of interpretation. Norm-referenced interpretations do not dismiss the importance of content validity.

Myth 4 – Only a criterion-referenced test can be used to inform classroom instruction.

False. Classroom instruction is informed by both criterion-referenced information as well as norm-referenced information. For example, knowing how many students in a particular classroom are on-track for college-level coursework is a criterion-referenced interpretation. However, comparing this same number of students in the classroom to the overall district is a norm-referenced interpretation.

Myth 5 – All items on an assessment with national norms are simply designed to differentiate between students.

False. Items are selected for inclusion on an assessment when it is determined that they match the content and cognitive specifications of the assessment. No item is selected for inclusion on an assessment based solely on its difficulty level or ability to differentiate. While it is important to have items that allow all students to demonstrate what they know and can do at both ends of the learning continuum, the content and clarity of the item are the first considerations when selecting items for an assessment.

For example, consider Figure 3:

Mathematics	
Domain	**Geometry**
Standard	Solve real-world and mathematical problems involving area volume and surface area of two- and three-dimensional objects composed of triangles, quadrilaterals, polygons, cubes, and right prims.

1. A candle company packages each candle in a box with dimensions 2 inches by 5 inches. The candle boxes will be placed in shipping boxes with dimensions 8 inches by 10 inches by 14 inches.

Candle Box Shipping Box

What is the greatest number of candle boxes that can fit in a shipping box?

A 14
B 23
C 40
D 26

Figure 3

This item would first be considered for selection on a test because it matches the domain and standard required by the test specifications for Grade 7 mathematics. In this case, the item aligns to the set of standards guiding the development of this assessment.

However, it is also possible to collect validity evidence that would allow users to make norm-referenced interpretations about this item. For instance, information collected after administering this item to a national sample of 7th graders and administering the item to a particular classroom suggested the following.

	In the Nation	In the Classroom
Percent of 7th grade students who answered the item correctly	64%	75%
Percent of high-performing 7th grade students who answered the item correctly	75%	85%
Percent of low-performing 7th grade students who answered the item correctly	32%	47%

Figure 4

Interpretations about how a particular group of students performed on this item relative to other students provides important information to the classroom teachers.

Conclusion

In summary, the foundation of measurement continues to rest with understanding the term validity. As test developers, we strive to continually provide evidence that supports the interpretation of test scores for various purposes. The responsibility for the accumulation and communication of this evidence must be assumed throughout all stages of design, development, and evaluation. With the ***Iowa Assessments***, this responsibility is a deliberate and purposeful part of the process resulting in an assessment system that can be appropriately and accurately used for various purposes.

To learn more about the *Iowa Assessments*, please go to hmhco.com/HMHAssessments to view author video clips and download informational brochures, scope and sequence resources, and additional white papers. Contact your Assessment Account Executive or call HMH Customer Experience–Assessments for a presentation.

Connect with us:

Houghton Mifflin Harcourt