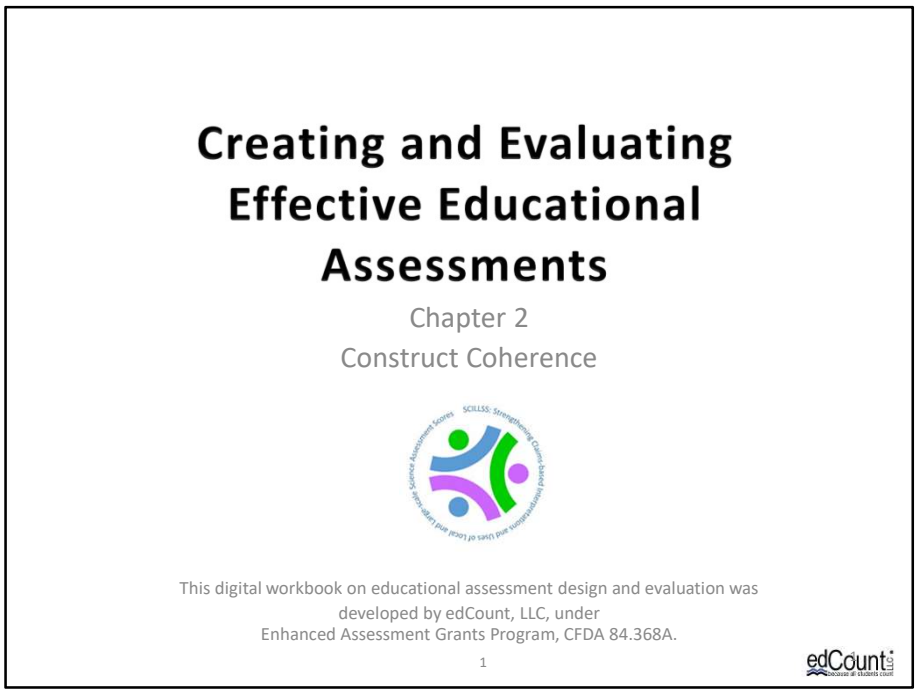
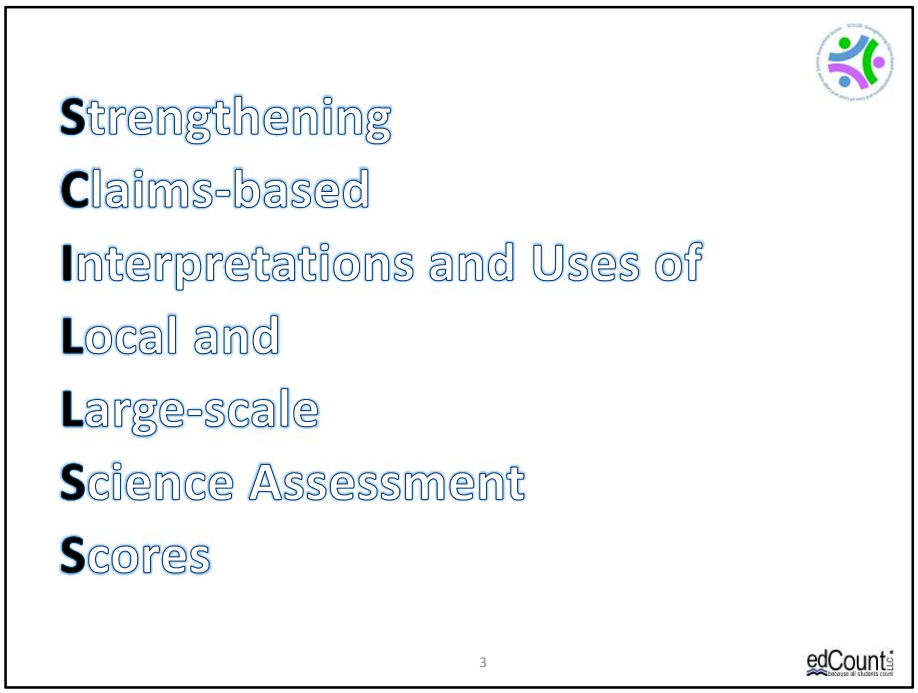


Chapter 2.0: Introduction

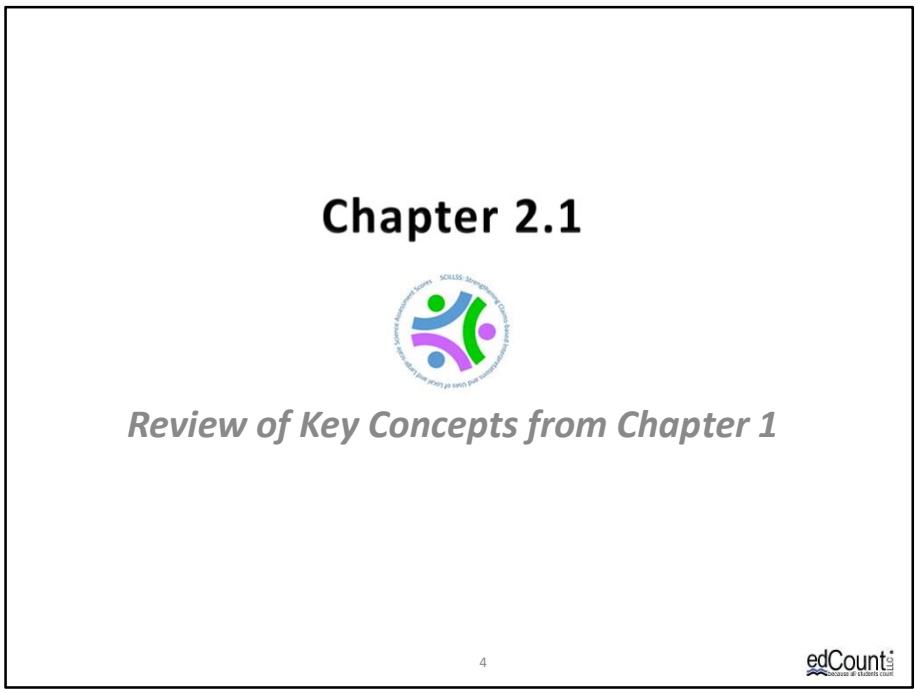
Welcome to the second of five chapters in a digital workbook on educational assessment design and evaluation. This workbook is intended to help educators ensure that the assessments they use provide meaningful information about what students know and can do.

This digital workbook was developed by edCount, LLC, under the US Department of Education's Enhanced Assessment Grants Program, CFDA 84.368A.

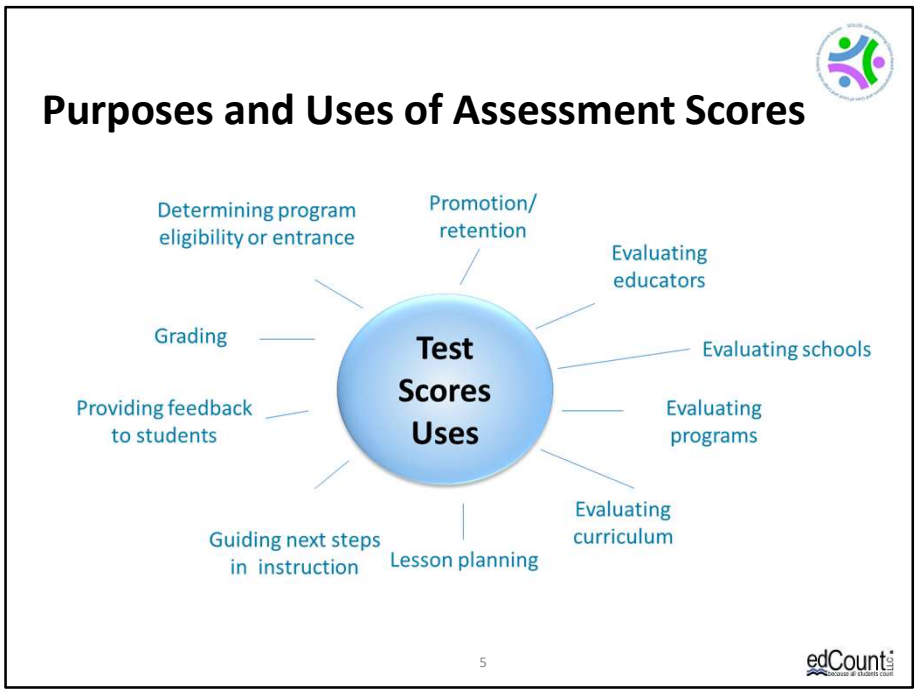
The grant project is titled the Strengthening Claims-based Interpretations and Uses of Local and Large-scale Science Assessment Scores…



or its acronym, “SCILLSS.”

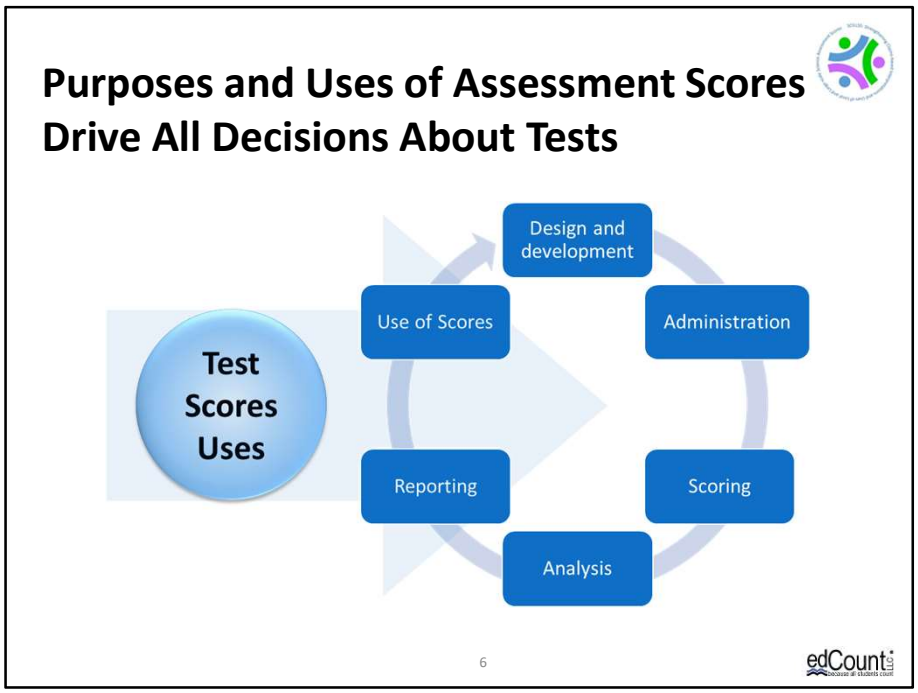


Chapter 2.1. Review of Key Concepts from Chapter 1

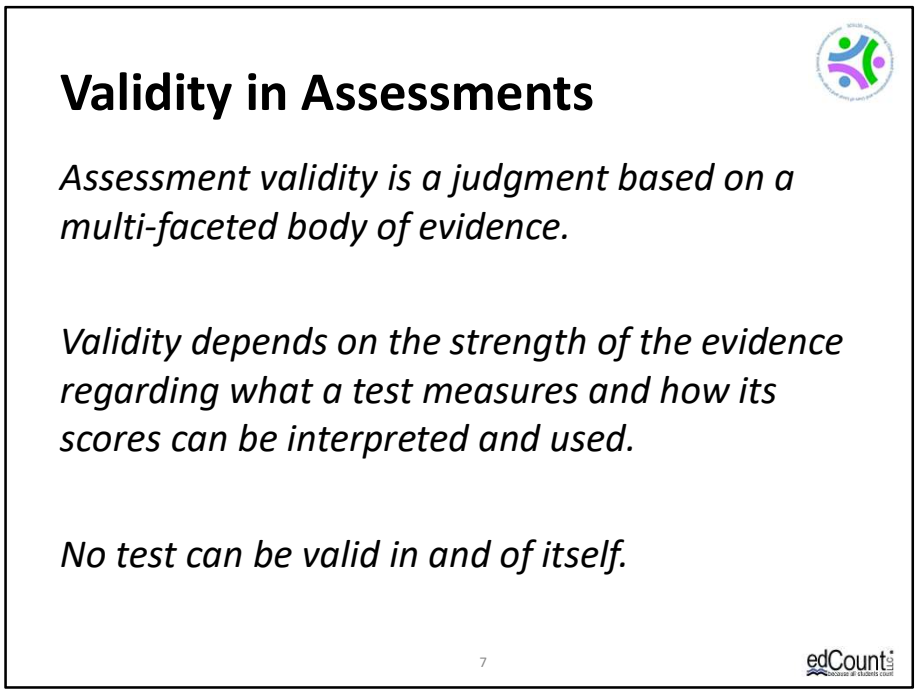


Let's begin with a brief recap of the key concepts covered in chapter 1 of this series.

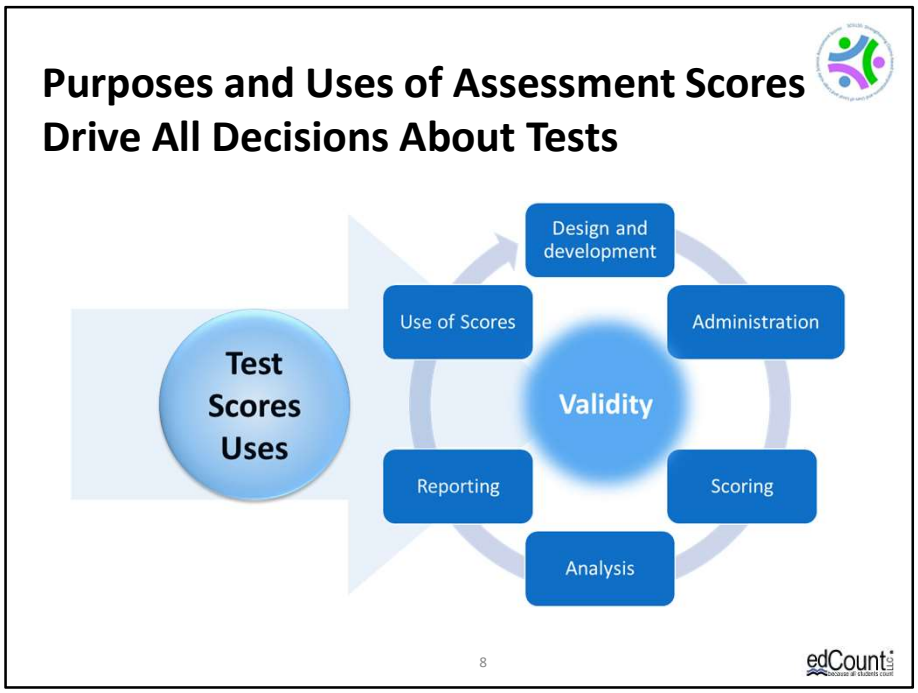
Chapter 1 focused on common reasons why we administer assessments of students' academic knowledge and skills and how we use those assessment scores.



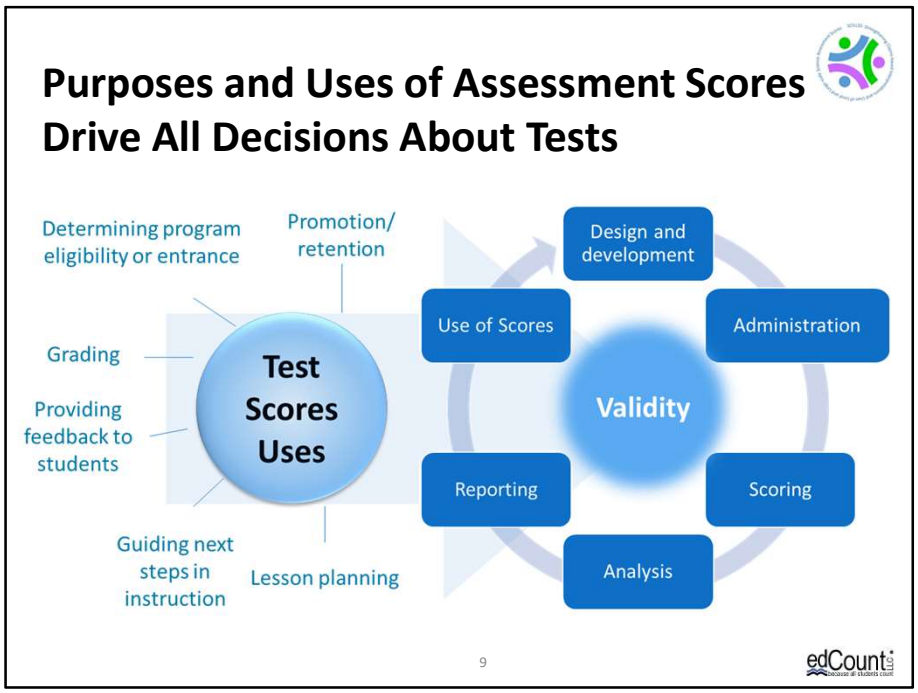
We learned that these purposes for administering assessments and the intended uses of assessment scores should drive all decisions about how assessments are designed, built, and evaluated.



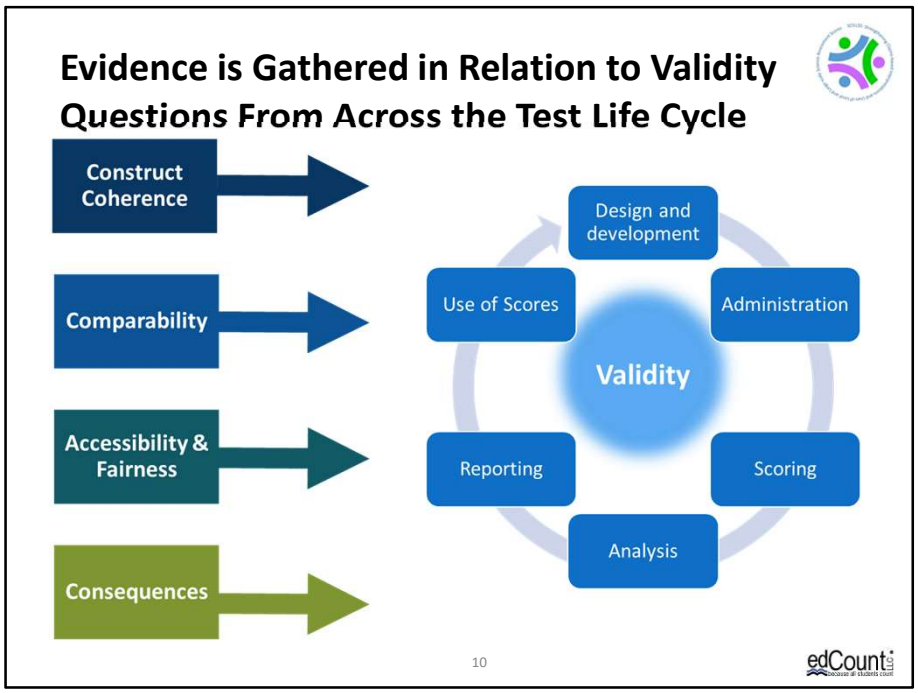
We learned in chapter 1 that validity relates to the interpretation and use of assessments scores and not to tests themselves. Validity is a judgment about the meaning of assessment scores and about how they are used.



We evaluate validity by gathering and judging evidence. This validity evidence is gathered from across the entire life cycle of a test from design and development through score use. Judgments about validity are based upon the adequacy and quality of this evidence in relation to assessment score interpretations and uses. Depending upon the nature of the evidence, score interpretations can be judged as valid or not. Likewise, particular uses of those scores may or may not be supported depending upon the degree and quality of the validity evidence.



For example, consider that some tests are meant to tell a teacher what his or her students know before or after a lesson or unit. The results of these assessments – which may be in the form of qualitative information or numerical scores or both – are intended to be used to inform decisions about upcoming instruction. To support those interpretations and uses of the scores, the teacher should have some evidence that the scores accurately reflect the knowledge and skills that are the instructional targets and that they are useful in guiding instructional decisions. Later in this chapter, and in the chapters that follow, we'll describe examples of what that evidence might look like.



Chapter 1 also included a brief overview of four fundamental validity questions that provide a framework for how to think about validity evidence. These four questions represent broad categories and each subsumes many other questions. The categories are: construct coherence, comparability, accessibility and fairness, and consequences.

The four validity questions are:

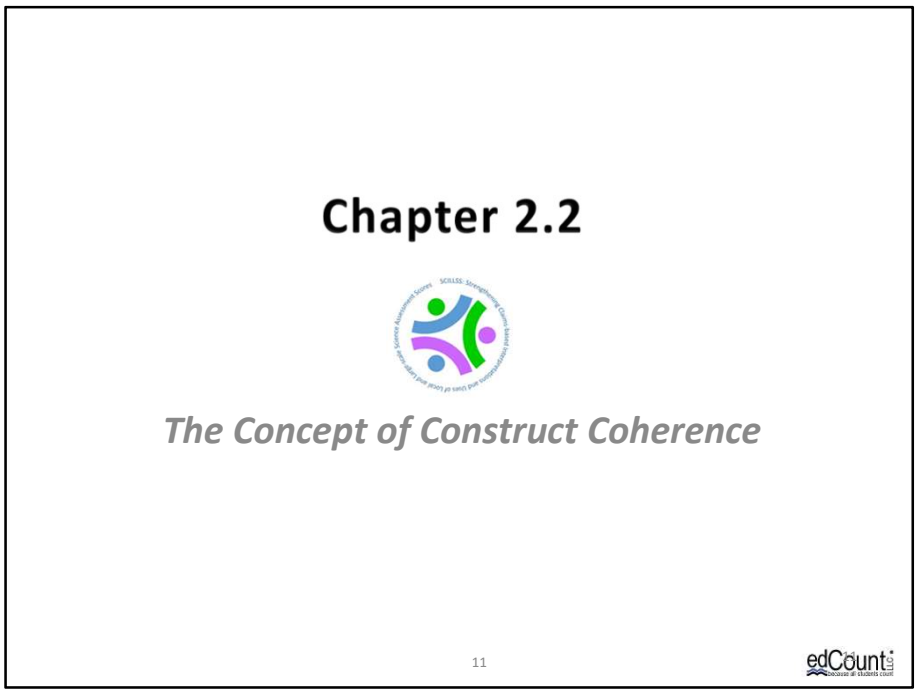
- To what extent do the test scores reflect the knowledge and skills we're intending to measure, for example, those defined in the academic content standards? This question addresses the concept of construct coherence.

- To what extent are the test scores reliable and consistent in meaning across all students, classes, schools, and time? This question addresses the concept of comparability.

- To what extent does the test allow all students to demonstrate what they know and can do? This question addresses the concept of accessibility and

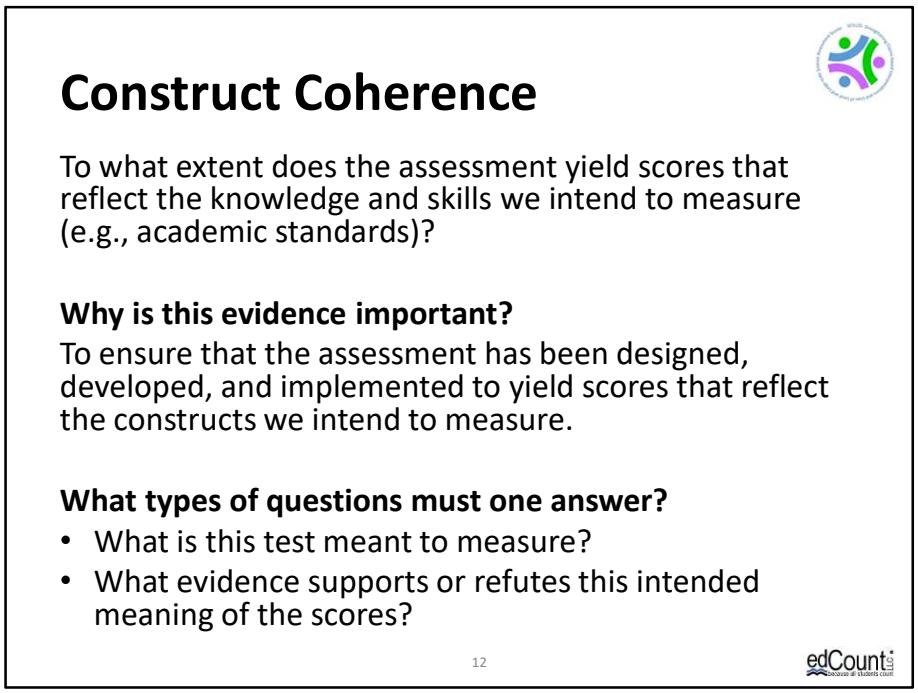
fairness. And

- To what extent are the test scores used appropriately to achieve specific goals? This question addresses the concept of consequences.

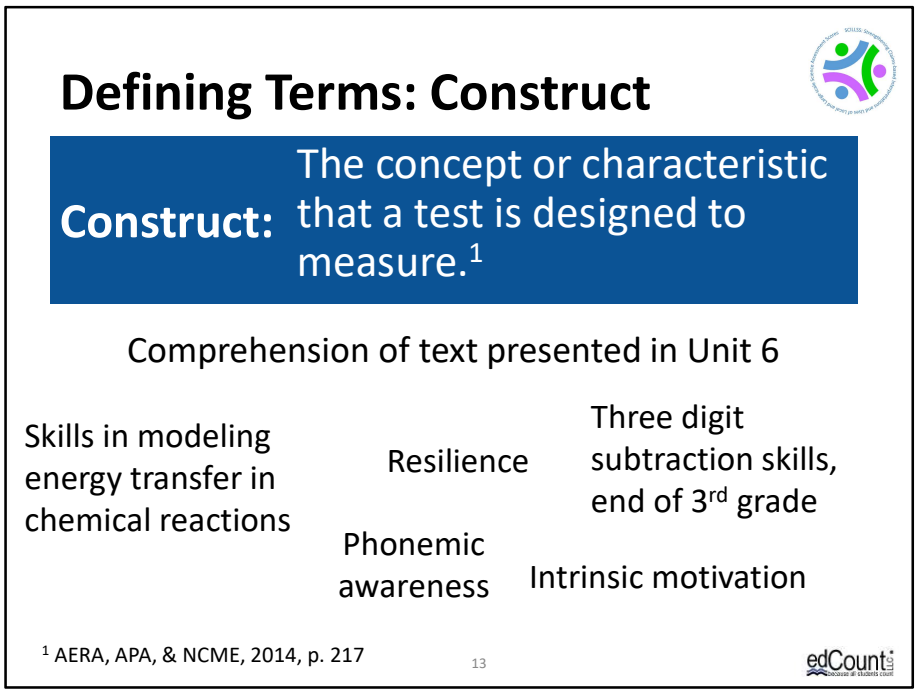


Chapter 2.2: The Concept of Construct Coherence

The purpose of this chapter in the five-chapter workbook series is to define the first category of validity questions, construct coherence, in greater detail and to provide examples of evidence related to these questions.



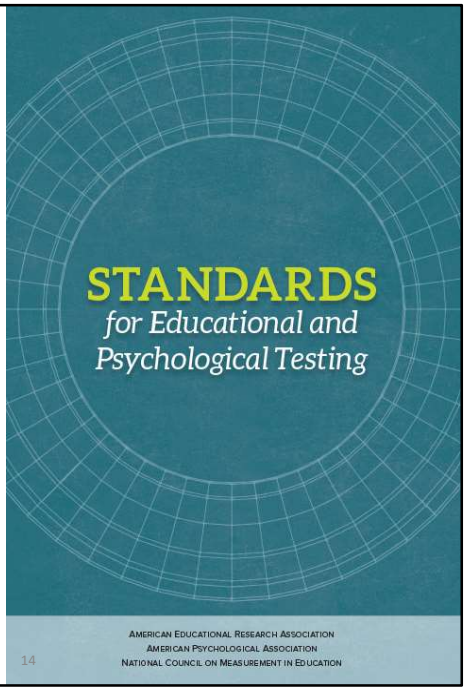
Construct coherence relates to the quality of evidence about what an assessment is meant to measure. This notion is clearly fundamental to the interpretation of assessment scores or, more simply, what test scores mean.



Recall from chapter 1 that a construct is the concept or characteristic that a test is designed to measure. In education settings, the constructs of most interest have to do with content knowledge and skills or personal or social characteristics that often relate to academic performance.

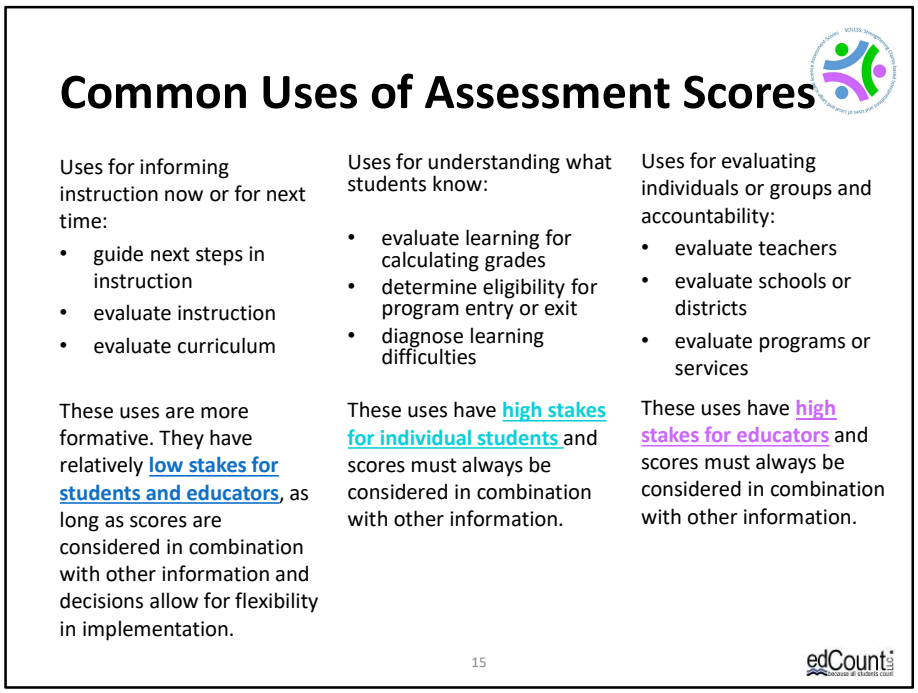
We cannot directly observe these constructs and must present students with opportunities – such as tests – when we can observe them demonstrate their knowledge and skills. If well-designed and well-implemented, tests can provide samples of performance that reflect the underlying constructs that are our real targets in education.

Standard 4.0: Tests and testing programs should be designed and developed in a way that supports valid interpretations of the test scores for their intended uses. Tests developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for the intended uses for individuals in the intended examinee population. (AERA, APA, & NCME, 2014, p. 85)



Anyone who plans to use an assessment, whether they plan to create that assessment themselves or adopt one built by others, must be clear about what the test scores are supposed to tell them and how they intend to use those scores. That is, every test user must establish a purpose for giving a test and identify the decisions that the test scores will inform. This notion is captured in the very first standard in the *Standards for Educational and Psychological Testing*, which guides professional practices in assessment, and reaffirmed in many other of these standards. For example:

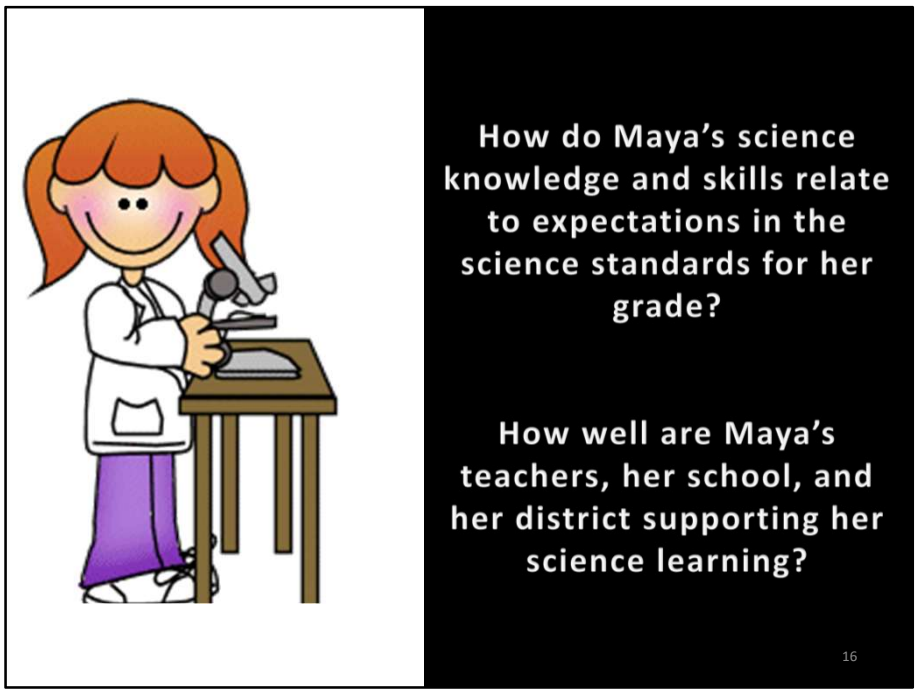
> Standard 4.0: Tests and testing programs should be designed and developed in a way that supports valid interpretations of the test scores for their intended uses. Tests developers and publishers should document steps taken during the design and development process to provide evidence of fairness, reliability, and validity for the intended uses for individuals in the intended examinee population.



Some common purposes for giving tests in educational settings include using scores to:

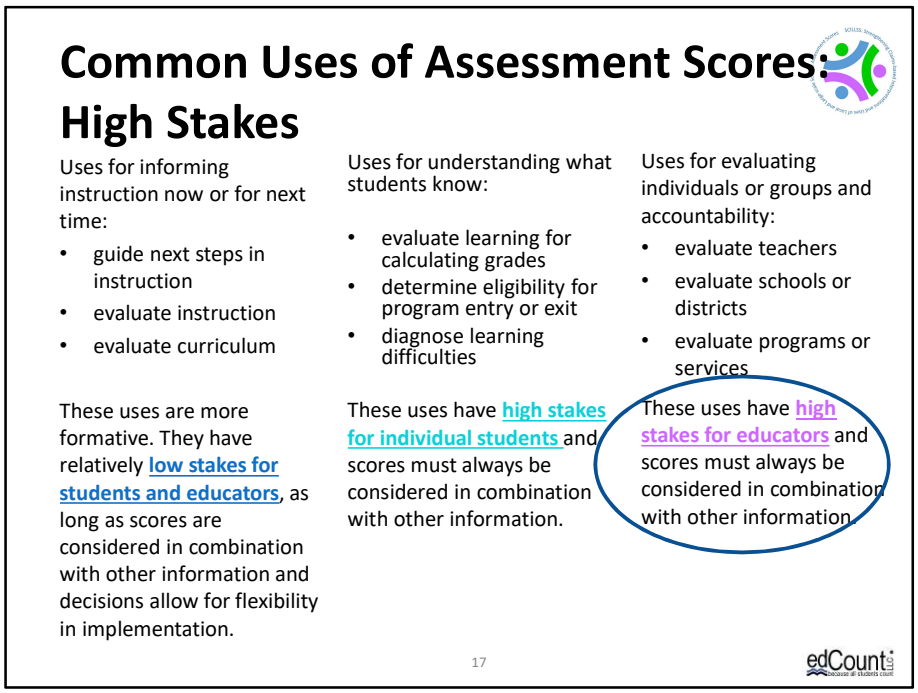
- guide next steps in instruction
- evaluate instruction
- evaluate curriculum
- evaluate learning for calculating grades
- determine eligibility for program entry or exit
- diagnose learning difficulties
- evaluate teachers
- evaluate schools or districts
- evaluate programs or services
- predict performance in a later setting

Anyone using test scores for any of these purposes must determine what test-based information would be appropriate and useful for that purpose.

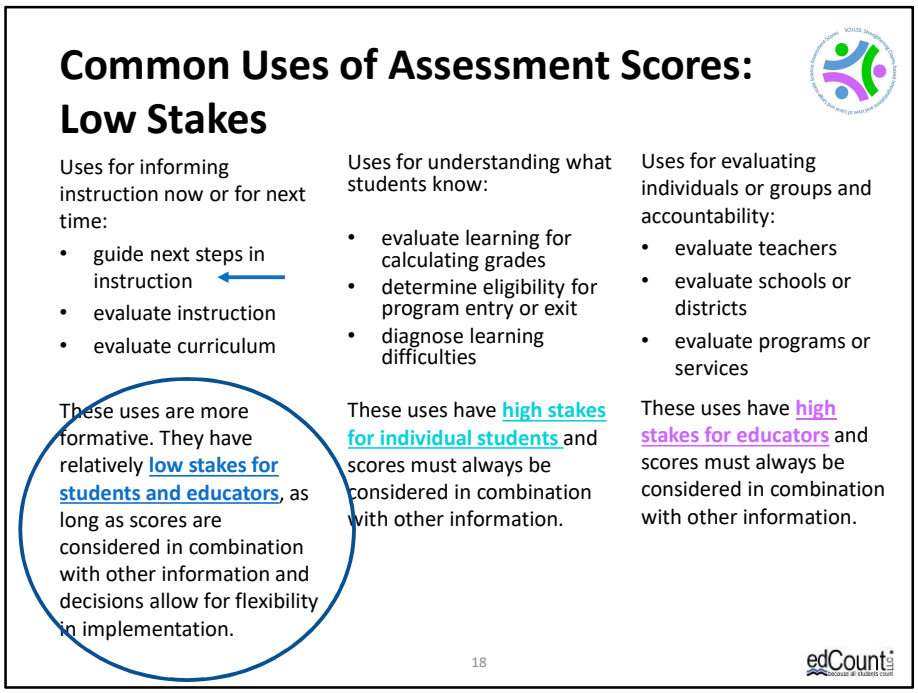


Consider the use of test scores to evaluate students' learning in relation to a state's academic content standards on an annual basis. Every public school, school district, and state in the U.S. uses test scores for this monitoring purpose.

Further, annual test scores inform accountability decisions for schools, school districts, and sometimes for individual educators. This means that the scores are interpreted not only as reflecting students' knowledge and skills, but also as indicators of how effective their teachers, schools, and school districts have been in teaching them.

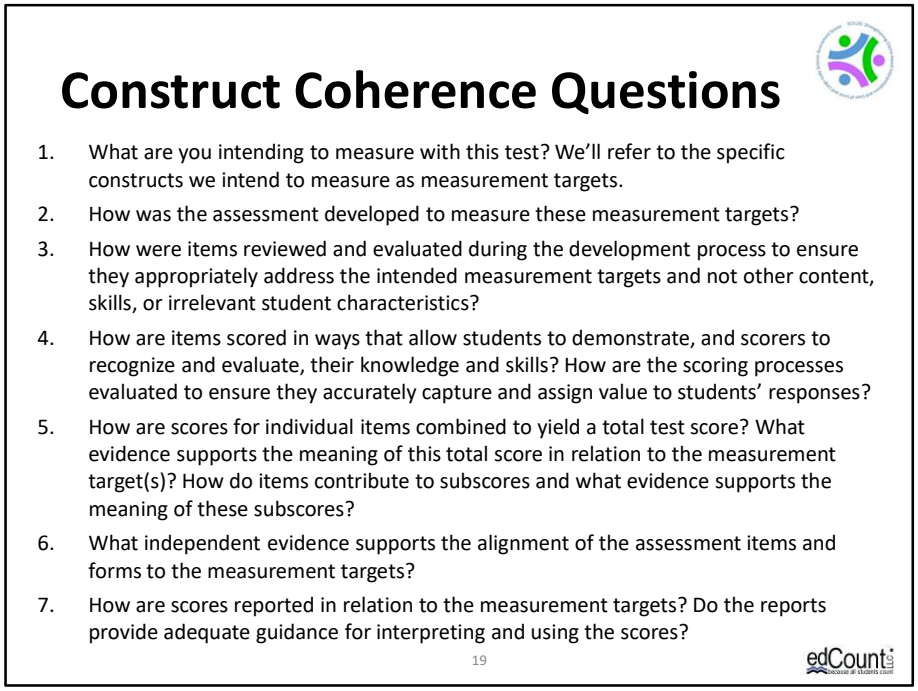


The uses of test scores for these evaluation and accountability purposes have high stakes associated with them. Therefore, the entity using the scores is obligated to establish validity evidence regarding these uses. The body of evidence should address all four validity questions and this chapter will address evidence related to construct coherence.



Another common use of test scores is to evaluate students' learning for the purpose of monitoring their progress across a school year. Here, the scores are also being interpreted as reflecting students' knowledge and skills in relation to academic expectations. But, the stakes associated with progress-monitoring are generally low as long as teachers or others close to the students have some flexibility in using the results.

Those requiring these assessments are also obligated to establish validity evidence. In this chapter, we'll consider the construct coherence aspects of this evidence and how evidence for these progress-monitoring uses differ in some ways from the evidence necessary for the high stakes accountability uses.



Evidence regarding our over-arching concept of construct coherence addresses the degree to which the test scores reflect the knowledge and skills we're intending to measure.

The types of questions that construct coherence questions include:

What are you intending to measure with this test? We'll refer to the specific constructs we intend to measure as measurement targets.

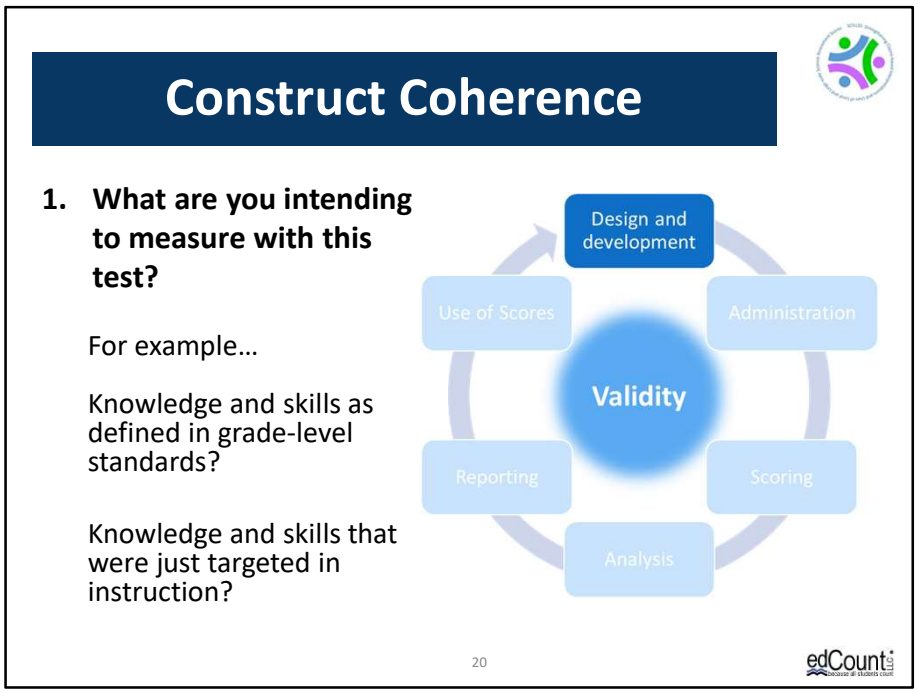
How was the assessment developed to measure these measurement targets?

How were items reviewed and evaluated during the development process to ensure they appropriately address the intended measurement targets and not other content, skills, or irrelevant student characteristics?

How are items scored in ways that allow students to demonstrate, and scorers to recognize and evaluate, their knowledge and skills? How are the scoring processes evaluated to ensure they accurately capture and assign value to students' responses?

How are scores for individual items combined to yield a total test score? What evidence supports the meaning of this total score in relation to the measurement target(s)? How do items contribute to subscores and what evidence supports the meaning of these subscores?

What independent evidence supports the alignment of the assessment items and forms to the measurement targets?
How are scores reported in relation to the measurement targets? Do the reports provide adequate guidance for interpreting and using the scores?
We'll describe evidence relevant to each of these questions next.

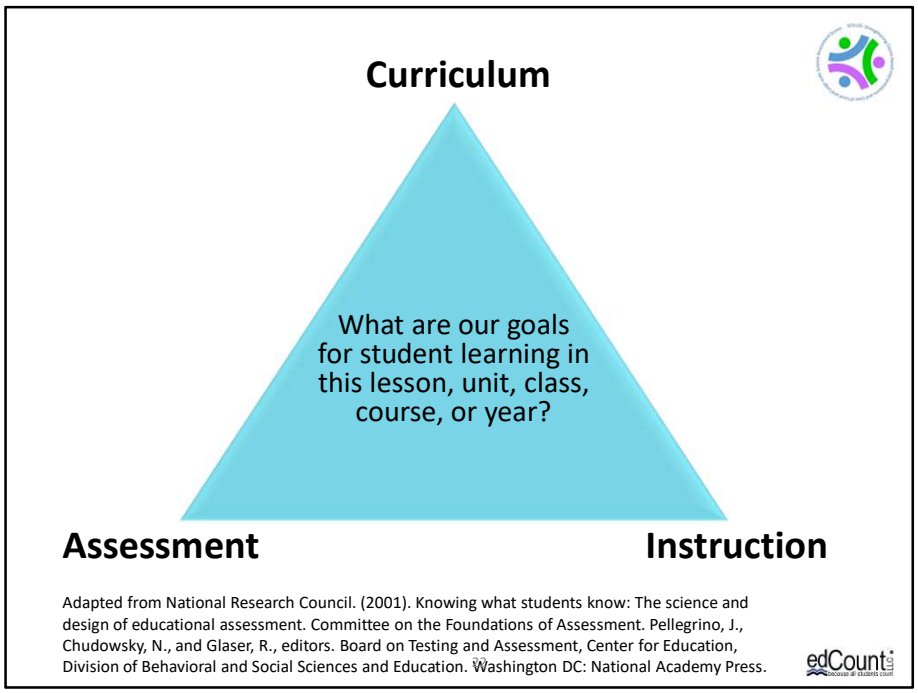


We'll start with the first of our construct coherence questions:
What are you intending to measure with this test?
Evidence to address this question comes from the Design and Development phase of the assessment life cycle.
Answering this question requires an articulation of the specific concepts and skills we want to know about. In both of the examples just described, the target knowledge and skills are those defined in a state's academic content standards, perhaps as specific in a district's curriculum.
For the statewide accountability test, the target knowledge and skills may be the comprehensive expectations for an entire academic year or course.
For the progress-monitoring test, the target knowledge and skills should be far more narrowly defined and relate to what students have just been taught.
In both cases, what appears on a test is just a sample of possible test content and we want to make inferences from the sample to the measurement target.
.

When selecting a test for one of these purposes, after clearly articulating what the scores are meant to mean, a teacher or administrator must evaluate what a potential test measures and how closely that aligns with the intended measurement targets.

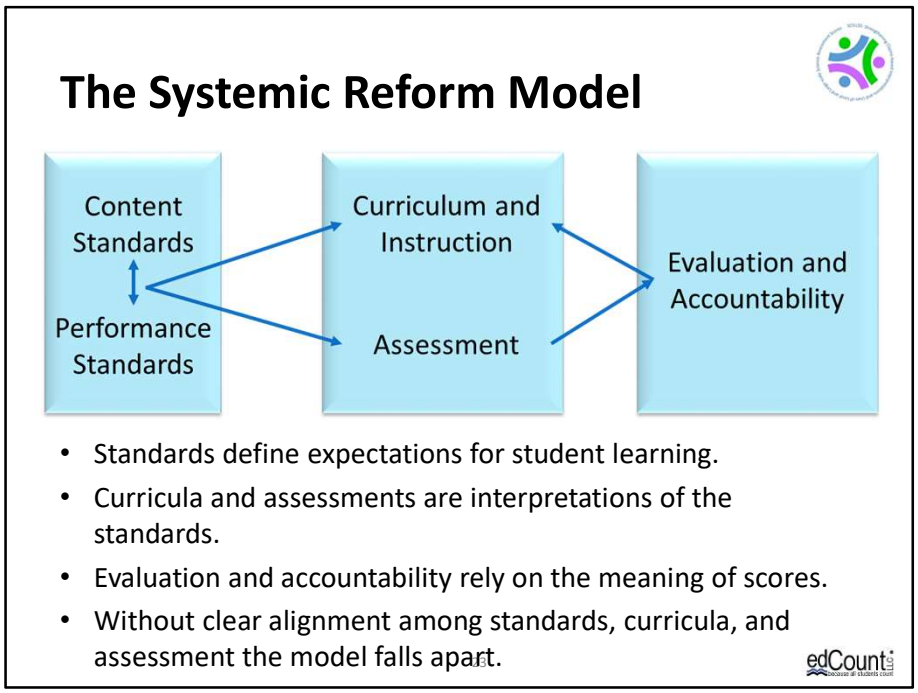
That is, what is the assessment meant to measure and are these measurement targets aligned with yours? The extent to which there is alignment affects the validity of score interpretations and uses.

Misalignment at this point would be like building a house without a foundation.



How would we know if an assessment was meant to measure our particular targets?

One way to help ensure alignment is to build a test specifically for this purpose. Most U.S. states have built their own assessments to assess students' knowledge and skills as defined in their own standards. If they are to interpret test scores in relation to their standards, the tests must demonstrably measure knowledge and skills defined in those standards. Further, the high stakes associated with the accountability uses of these tests further escalate the demand for high quality evidence that the test measures what schools, school districts, and educators are teaching.



It may be helpful to understand a bit more about this notion of alignment among a state's standards, its assessments, and what local educators teach. Federal and state education policies that underlie the use of standards-based tests and the use of such scores in accountability systems are based on a model called systemic reform. Systemic reform as an approach to school improvement asserts that standards define expectations for students' learning. Standards should drive the development of curricula and the delivery of instruction at the local level. To support standards-aligned curricula and instruction, assessment scores are meant to be used within a system of accountability that helps to identify where student learning is not meeting expectations and to direct additional resources to these schools and school districts.

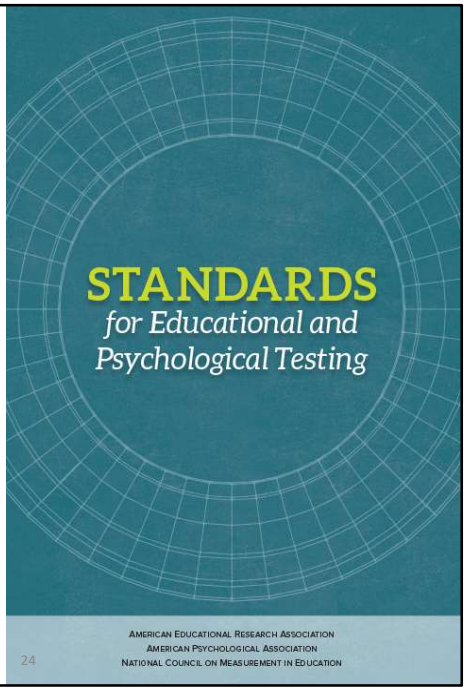
This model can work only if the assessment scores reflect student learning in relation to the same standards that guide curriculum and instruction. This is why alignment between the assessments and the standards is critical.

Standard 1.1: The test developer should set forth clearly how the test scores are intended to be interpreted and consequently used. The population(s) for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly. (p. 23)

Standard 4.1: Test specifications should describe the purpose(s) of the test, the definition of the construct of domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of the test results for the intended purpose(s). (p. 85)

Standard 7.1: The rationale for a test, recommended uses of the test, support for such uses, and information that assists in score interpretations should be documented. When particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified. (p. 125)

(AERA, APA, & NCME, 2014)



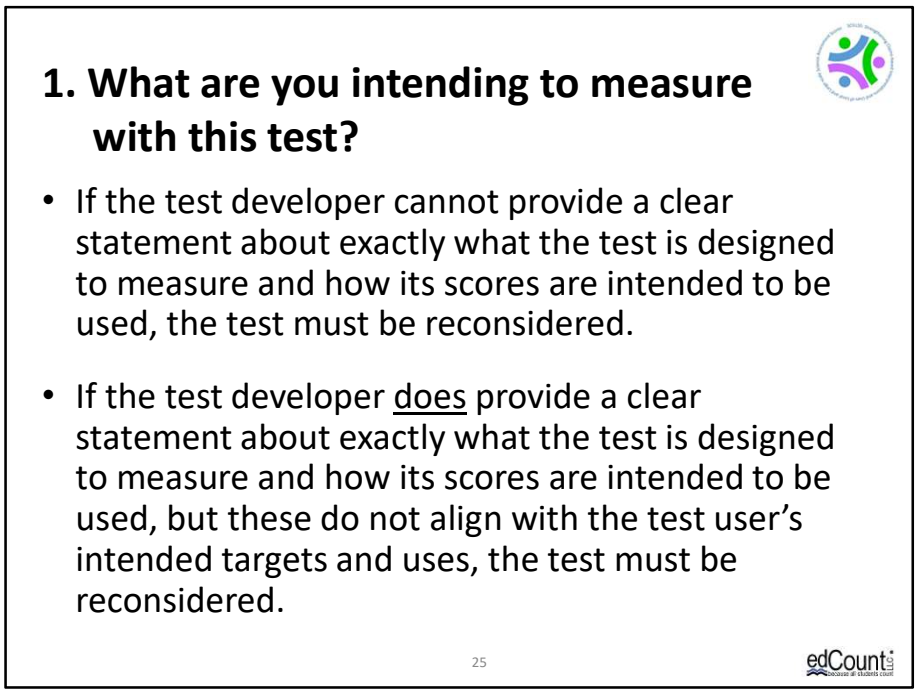
A test publisher must indicate what a test is designed to measure. This information would be found in the documentation that accompanies a test.

This obligation is referenced several times in the Standards for Educational and Psychological Testing. For example:

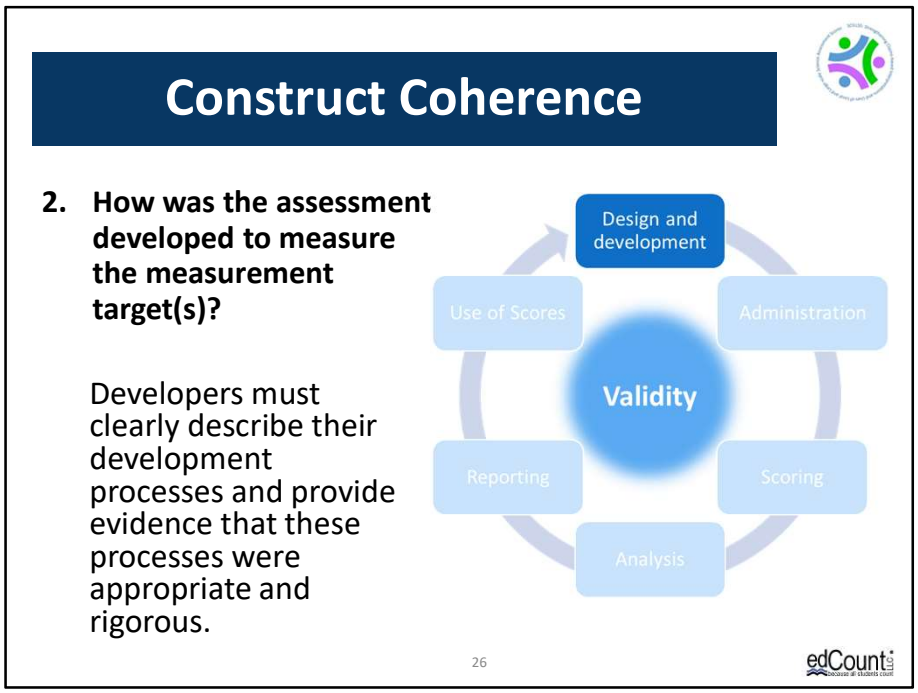
Standard 1.1: The test developer should set forth clearly how the test scores are intended to be interpreted and consequently used. The population(s) for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly.

Standard 4.1: Test specifications should describe the purpose(s) of the test, the definition of the construct of domain measured, the intended examinee population, and interpretations for intended uses. The specifications should include a rationale supporting the interpretations and uses of the test results for the intended purpose(s).

Standard 7.1: The rationale for a test, recommended uses of the test, support for such uses, and information that assists in score interpretations should be documented. When particular misuses of a test can be reasonably anticipated, cautions against such misuses should be specified.



The first construct coherence question presents a sort of "sudden death" situation for those deciding whether or not to use a test. If the test developer cannot provide a clear statement about exactly what the test is designed to measure and how its scores are intended to be used, the test must be reconsidered. If the test developer does provide a clear statement about exactly what the test is designed to measure and how its scores are intended to be used, but these do not align with the test user's intended targets and uses, the test must be reconsidered.



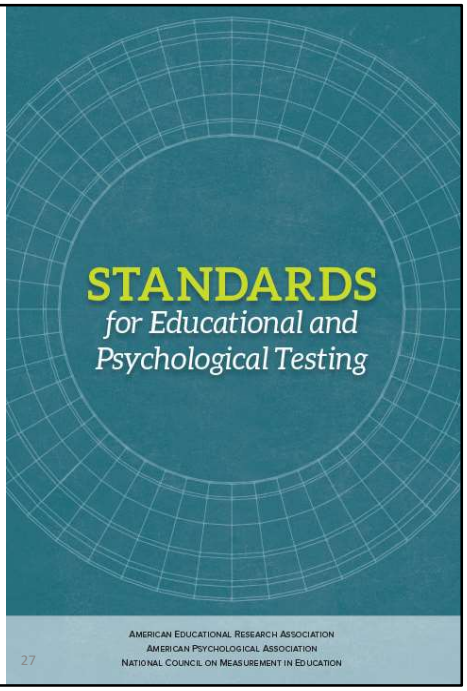
If these align, we move on to the second construct coherence question.
How was the assessment developed to measure the measurement target(s)?
Evidence to address this question comes from the Design and Development phase of the assessment life cycle.
It is not enough for a test publisher to simply say what they claim that a test measures. They are obligated to provide evidence of how the test was developed to support this claim. Developers must clearly describe their development processes and provide evidence that these processes were appropriate and rigorous.

Standard 1.11: When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. (p. 26)

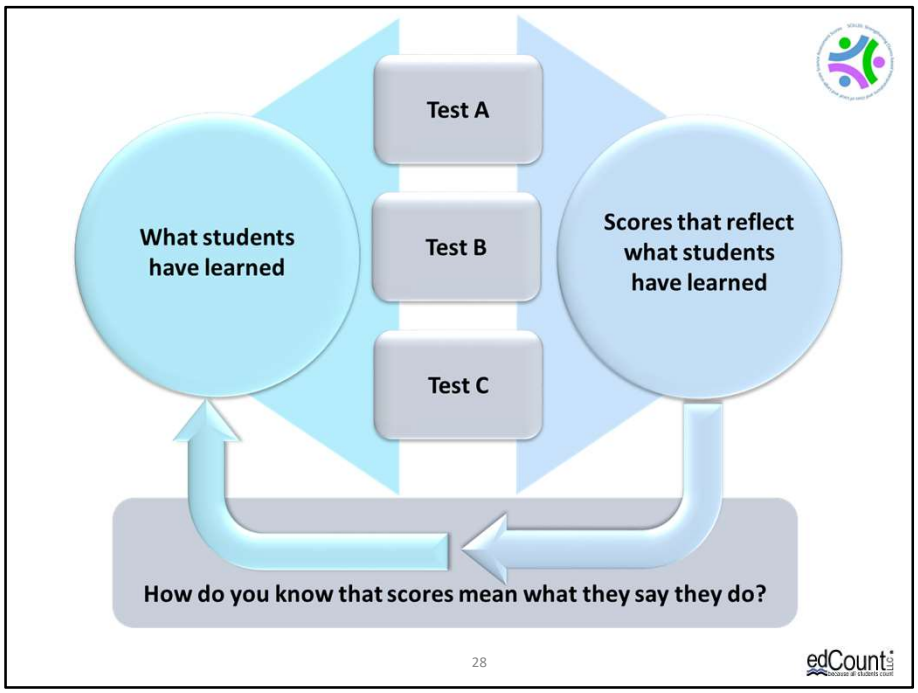
Standard 4.2: In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections. (p. 85)

Standard 4.12: Test developers should document the extent to which the content of a test represents the domain defined in the test specifications. (p. 89)

(AERA, APA, & NCME, 2014)



Our professional standards underscore these obligations. For example:
Standard 1.11: When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent.
Standard 4.2: In addition to describing intended uses of the test, the test specifications should define the content of the test, the proposed test length, the item formats, the desired psychometric properties of the test items and the test, and the ordering of items and sections.
Standard 4.12: Test developers should document the extent to which the content of a test represents the domain defined in the test specifications.



These standards mean that test developers must clearly describe the process they use to identify what is meant to be measured via tests so that scores from those tests can be interpreted appropriately. If test scores are meant to reflect what students have learned, the tests must be able to capture evidence of what students have learned. Relevant documentation encompasses descriptions about the development process and an evaluation of that process.

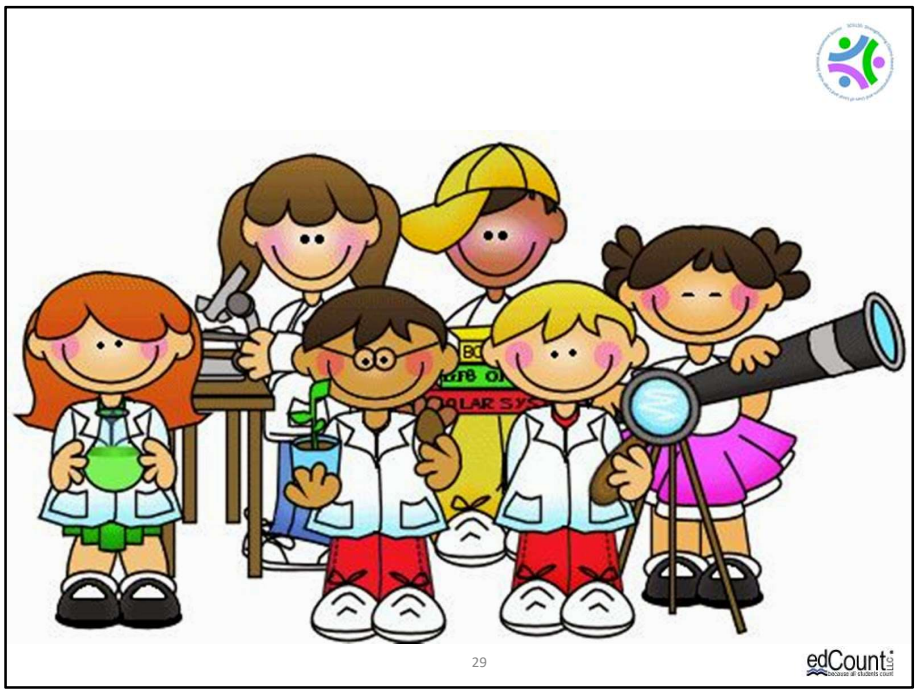
Developers should always provide a description of how they defined the domain and measurement targets.

Simply saying that a test measures "reading" or "3rd grade science" is unacceptable no matter what the scores are supposed to mean. How did the developer identify the specific samples and ranges of knowledge and skills the test would be developed to cover?

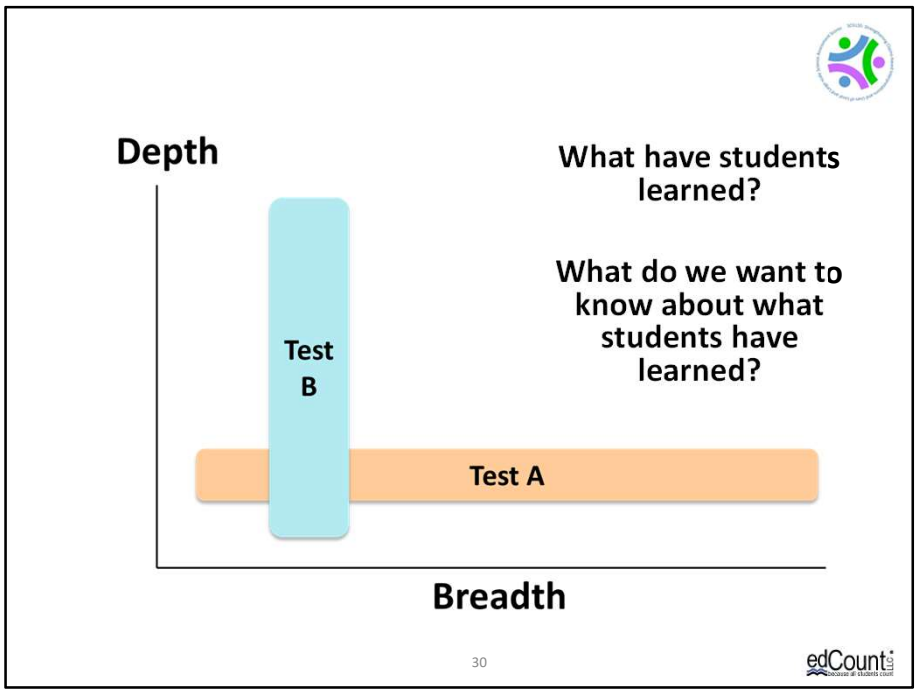
If the test scores are intended to reflect knowledge and skills that students are expected to have learned by a certain point in a school year or course, how did the test developer determine what the test should cover? Were state and local

educators involved in those decisions? Other knowledgeable and credible experts?

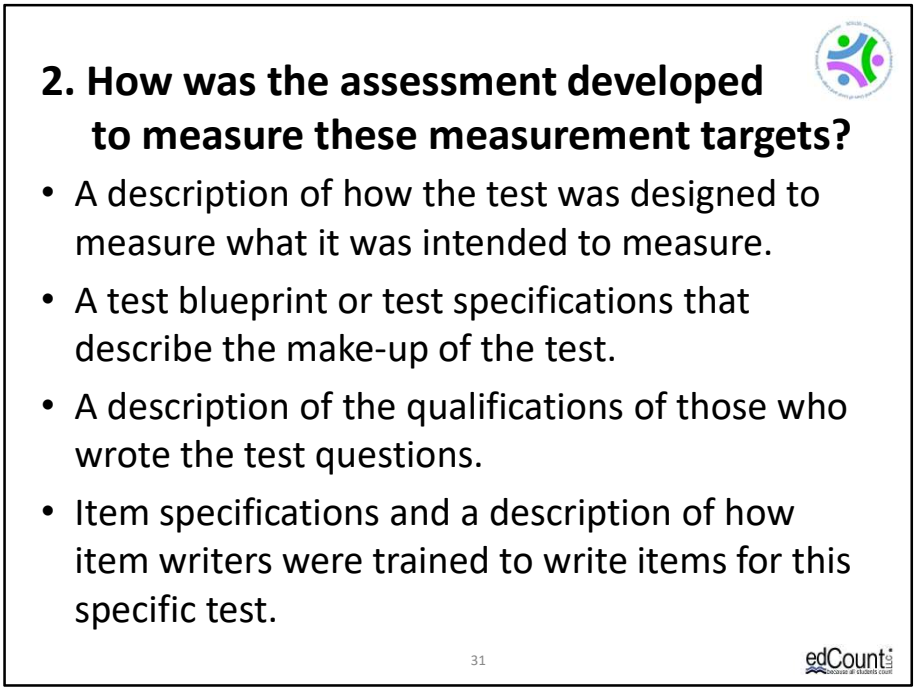
How did the test developer ensure that the test included an appropriate range of knowledge and skills at appropriate levels of difficulty and complexity? How challenging are the test questions and components such as reading passages and graphs? Were individuals who understand the diversity of the student population in this grade level involved in these decisions? In what ways?



In addition to knowing what the test was meant to measure and how developers made these decisions, a potential test user must consider whether the test is appropriate for all of his or her students. Can all students demonstrate what they know and can do on this test? We'll come back to this issue in chapter 4 of this series, which focuses on accessibility and fairness.



Consideration of a test should also include adequacy of coverage. Does the test provide students with enough opportunities to demonstrate their knowledge and skills across the breadth and depth of expectations the test is meant to measure? If a test is meant to address a relatively narrow set of skills, does it do so to an appropriate degree of depth and breadth and without tapping into other, non-target skills?



Note that all of these questions focus on concepts such as 'appropriate' and 'adequate.' Whether a test yields valid information depends on what it what intended to measure and judgments about validity are not black-and-white or based on a simple statistic.

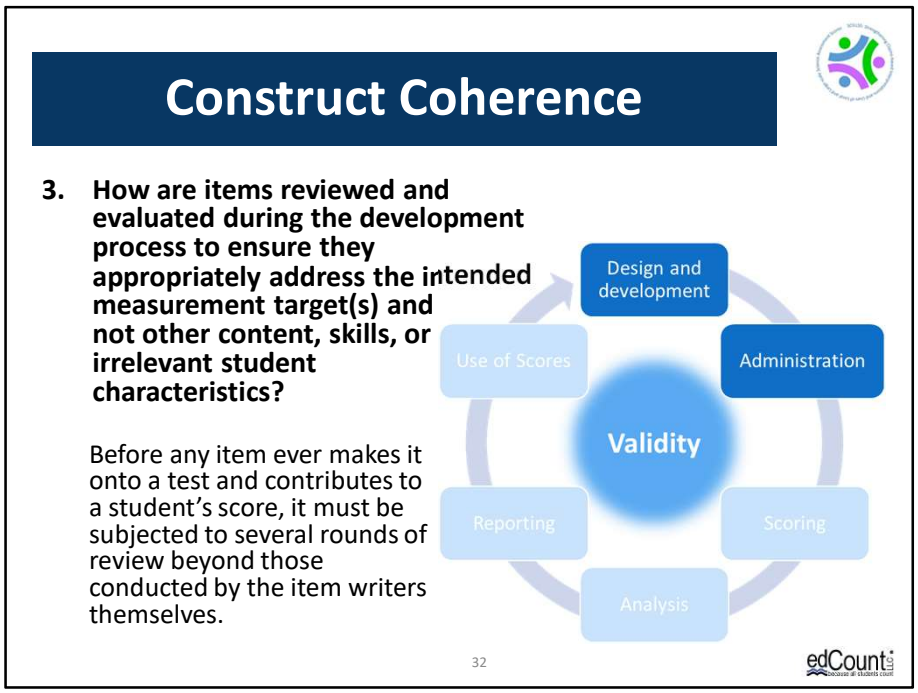
However, there are some specific pieces of evidence to look for when evaluating a development process. These include:

A description of how the test was designed to measure what it was intended to measure. How did the developers determine what questions would be on the test, what form these questions would take, and how many questions each student would take?

A test blueprint or test specifications that describe the make-up of the test. Is it clear how many items are on the test and what they measure? As we will see when we consider issues of comparability, any test that has more than one form – that is, has different versions with different items that are all meant to yield scores with the same meaning – must have a means for maintaining consistency across these forms.

A description of the qualifications of those who wrote the test questions. Items writers should have expertise in the subject matter, understand the population of students who will be taking the test, and experience writing similar items.

Item specifications and a description of how item writers were trained to write items for this specific test. Even experienced item writers need to understand the nature and purpose of the test on which their items will appear. In addition, those who develop tests for use in high stakes situations must provide evidence that the guidance and parameters for writing the items were sound and likely to support the development of good items (i.e., item and scoring specifications (rubrics, etc.)).
.



Our next question related to construct coherence is:
How are items reviewed and evaluated during the development process to ensure they appropriately address the intended measurement target(s) and not other content, skills, or irrelevant student characteristics?
Evidence to address this question comes from the Design and Development and Administration phases of the assessment life cycle. A test developer must take steps to ensure that the items don't require students to have particular outside knowledge that is not directly relevant to what the test is meant to measure. For example, students should not be expected to know anything about golf or beaches or salmon or growing wheat unless such knowledge is what the test is supposed to measure.

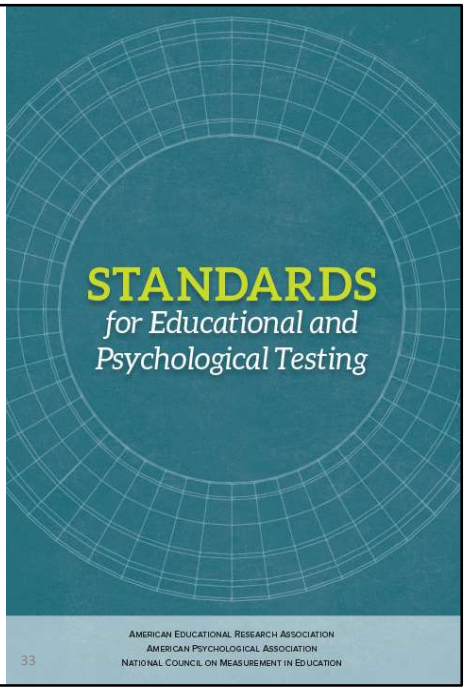
According to the principles of Universal Design for Learning (UDL), tests should be designed to facilitate and minimize construct-irrelevant barriers for all test takers in the target population. UDL seeks to make educational materials and assessments as accessible as possible to the widest variety of people while seeking to minimize separate-but-equal situations. Before any item ever makes it onto a test and contributes to a student's score, it must be subjected to several rounds of review

beyond those conducted by the item writers themselves. These reviews should address what the items are meant to measure as well as instances where fairness might be an issue.

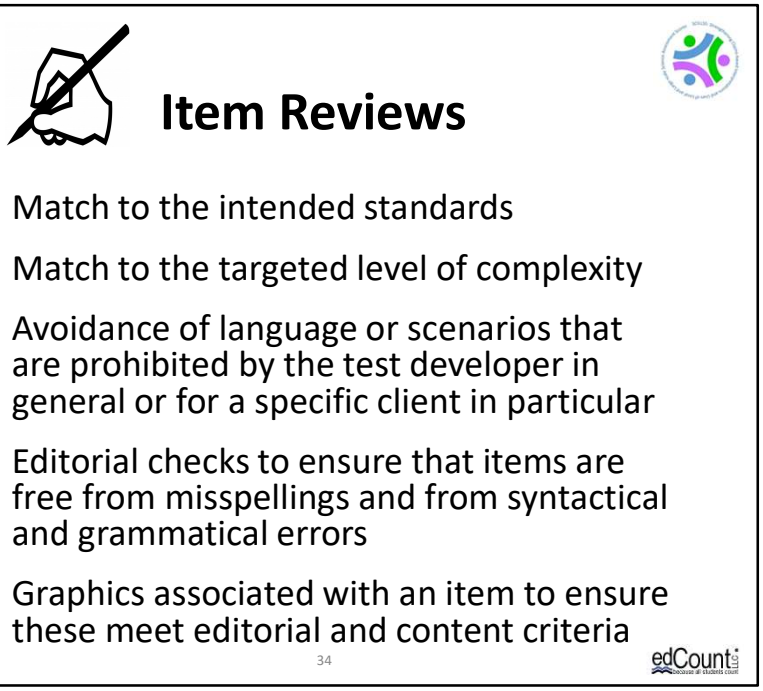
Standard 4.7: The procedures used to develop, review, and try-out items and to select items from the item pool should be documented. (p. 87)

Standard 4.8: The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive. (p. 88)

(AERA, APA, & NCME, 2014)



Professional standards related to a test developer's item review obligations include:
Standard 4.7: The procedures used to develop, review, and try-out items and to select items from the item pool should be documented.
Standard 4.8: The test review process should include empirical analyses and/or the use of expert judges to review items and scoring criteria. When expert judges are used, their qualifications, relevant experiences, and demographic characteristics should be documented, along with the instructions and training in the item review process that the judges receive.

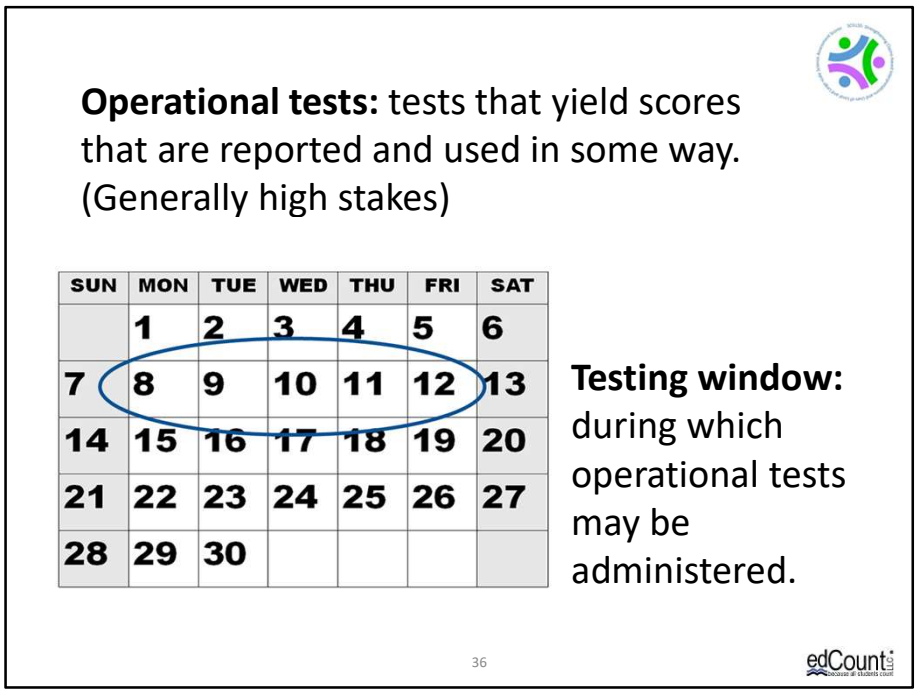


Those supervising item writers must conduct reviews to be sure each item meets criteria in the item development specifications. These may include, but are not limited to, the match to the intended standards, the targeted level of complexity, plausibility of distractors, and the avoidance of language or scenarios that are prohibited by the test developer in general or for a specific client in particular. For example, the game of golf is not familiar to many students in many areas so an item that is framed within a golf scenario may be confusing to them even if it might offer a great context for some math and science problems. Internal reviews of test items also include editorial checks to ensure that items are free from misspellings and from syntactical and grammatical errors. Reviewers also check any graphics associated with an item to ensure these meet editorial and content criteria.

It is common practice for developers of large-scale assessments to include rounds of review by educators in the areas where the tests are intended to be used. These reviews allow for those familiar with how standards are interpreted in local contexts to weigh in on how well the items correspond to the standards they are meant to measure.

Some of these local reviews may take place in the relatively early stages of item development; other local reviews are timed to occur after processes known as pilot-testing or field-testing.

The terms pilot-test and field test are sometimes considered synonyms. However, we make a distinction between pilot-tests, which we define as small-scale try-outs, and field-tests, which we define as a large-scale evaluation of items on the tests that students take.



Let's define some related terms to help with the notions of pilot testing and field testing.

Operational tests are tests that yield scores that are reported and used in some way. These tests "count" in one way or another.

Operational tests are administered during a testing window, which is the specific period of time when students are allowed to take the test. All high stakes tests have a clear testing window and those in charge of test administration are prohibited from giving students access to the test before or after this period.

Pilot tests and **field tests** are ways of trying out items before they count toward students' scores.

Pilot tests may include small collections of items rather than full test forms. Pilot tests often occur outside of the testing window and may only take place in a few districts or schools.

Field tests may involve small collections of items placed on operational test forms. Field tests often occur within the testing window and take place with a larger population of students.

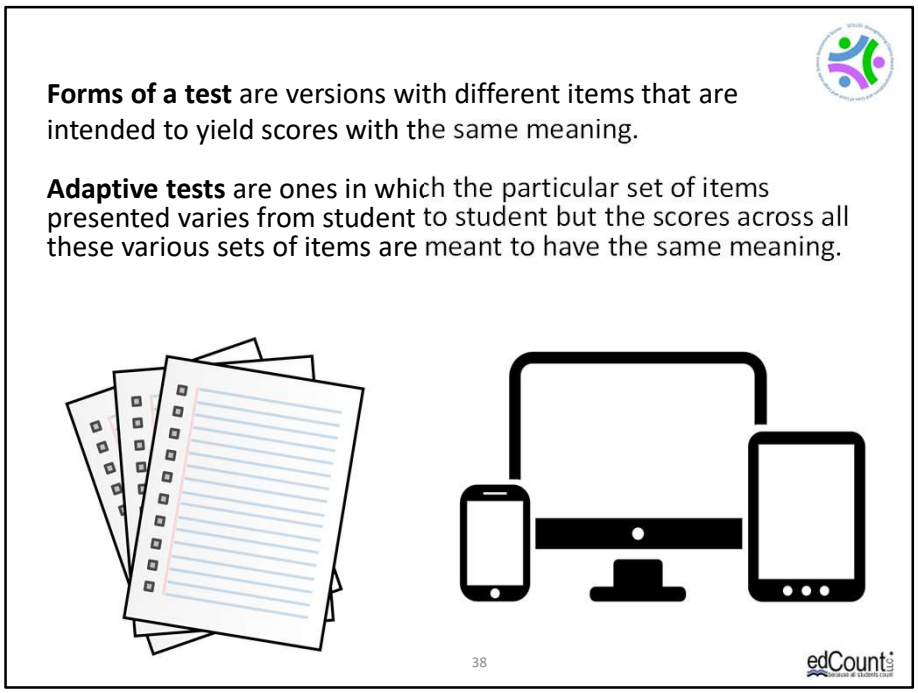
Scores on items from pilot tests and field tests do not count toward students' test scores, but the information gained from students' responses is used to evaluate the questions themselves.

37

edCount

Returning to the processes of pilot-testing and field-testing, pilot tests may include small collections of items rather than full test forms. Pilot tests often occur outside of the testing window and may only take place in a few districts or schools. The purpose of pilot testing is often to evaluate how students interact with new items that are unlike those that appear on current versions of a test.

The term field-testing is usually applied to larger collections of items administered more broadly. Virtually all large-scale, high stakes tests like those required annually by states include some field-test items. Think of a test form that includes a total of 50 questions. Perhaps 40 of those questions are "operational", meaning that students' answers to them count toward their test scores. The other ten questions would be field-test items. This means that they are being "tried out" and how a student answers them does not count toward his or her scores. Instead, the information gained from students' responses is used to evaluate the questions themselves.

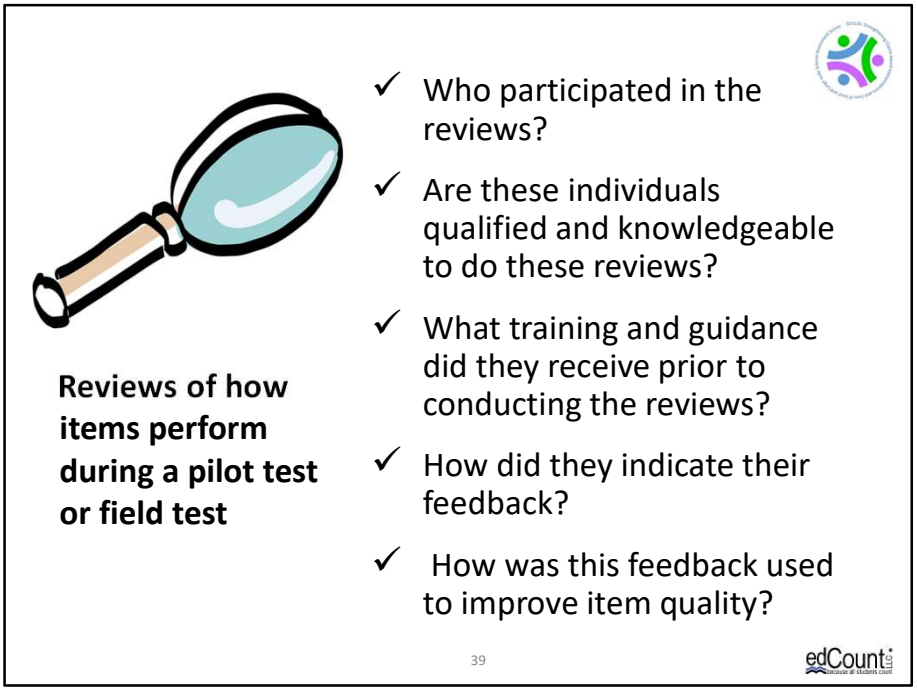


Large-scale, high stakes tests like those required annually by states, include several forms each year or are adaptive. In all cases, the items that appear on a test are drawn from what is called an item pool or an item bank, which is the repository in which items are stored.

Forms of a test are versions with different items that are intended to yield scores with the same meaning. The term "equivalent forms" indicates that it should not matter which form of a test a student took. The student should achieve the same score regardless of the form.

Adaptive tests are ones in which the particular set of items presented varies from student to student but the scores across all these various sets of items are meant to have the same meaning. In some adaptive tests, each item is selected based on how the student answered the prior item. In other adaptive tests, a student responds to a group of items and then the next group of items is selected based on the responses to that prior group. Both kinds of adaptive tests rely on a complex mathematical computation, or "testing algorithm", to identify the items a student sees. This algorithm often determines item choices based to some extent on item difficulty. If a student gets an item incorrect, the algorithm may select a somewhat easier item next. Likewise, if a student gets an item correct, the algorithm may

select a somewhat harder item next. Many other variables may be included in the algorithm, but the essential idea is that an adaptive test is, to some degree, tailored to each student while representing the test blueprint.

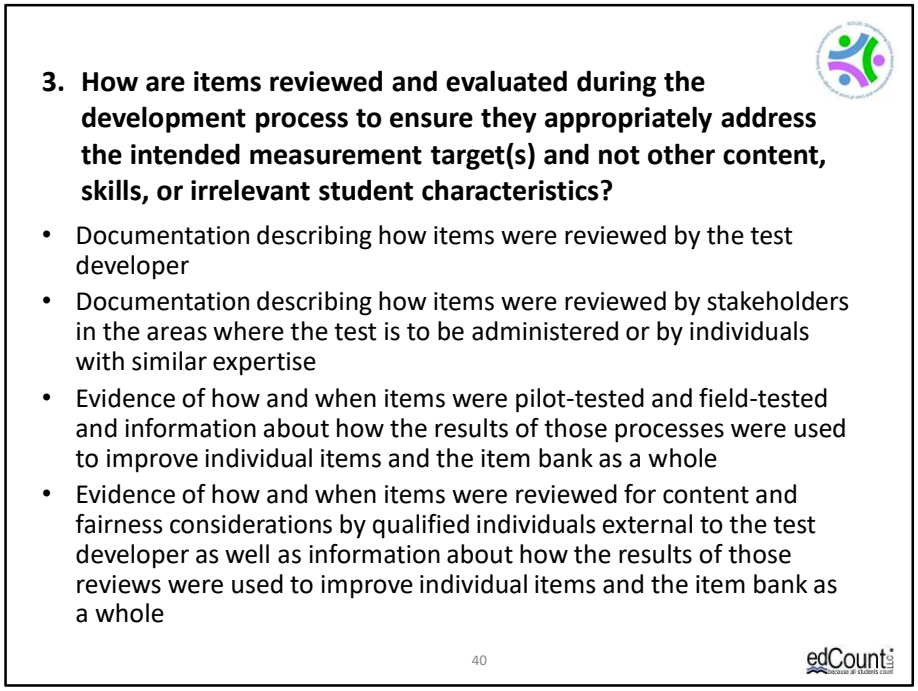


After a pilot-test or after an operational test that includes field-test items, a test developer will review statistics about the items to determine whether the items appear to be high quality enough to include as operational items on subsequent test forms. As noted earlier, local educators may be asked to judge these items from both a content perspective and from what is often called a “bias and sensitivity” perspective. Test developers do these reviews to ensure fairness, meaning all items are not inadvertently biasing against a subgroup of students.

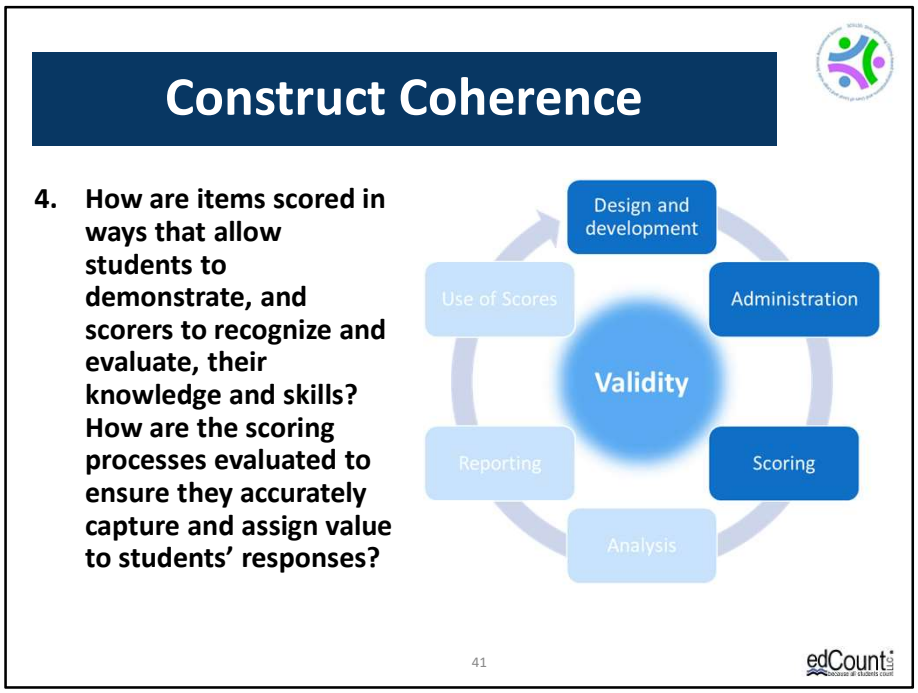
Here we are concerned with the quality of these reviews and the use of feedback to improve item quality.
Who participated in the reviews?
Are these individuals qualified and knowledgeable to do these reviews?
What training and guidance did they receive prior to conducting the reviews?
How did they indicate their feedback? How was this feedback used to improve item quality?

Such reviews are important to ensuring the quality of the items that eventually appear on tests and contribute to students’ scores. Failure to conduct such reviews

would be highly inappropriate on the part of an item developer for large-scale assessments.

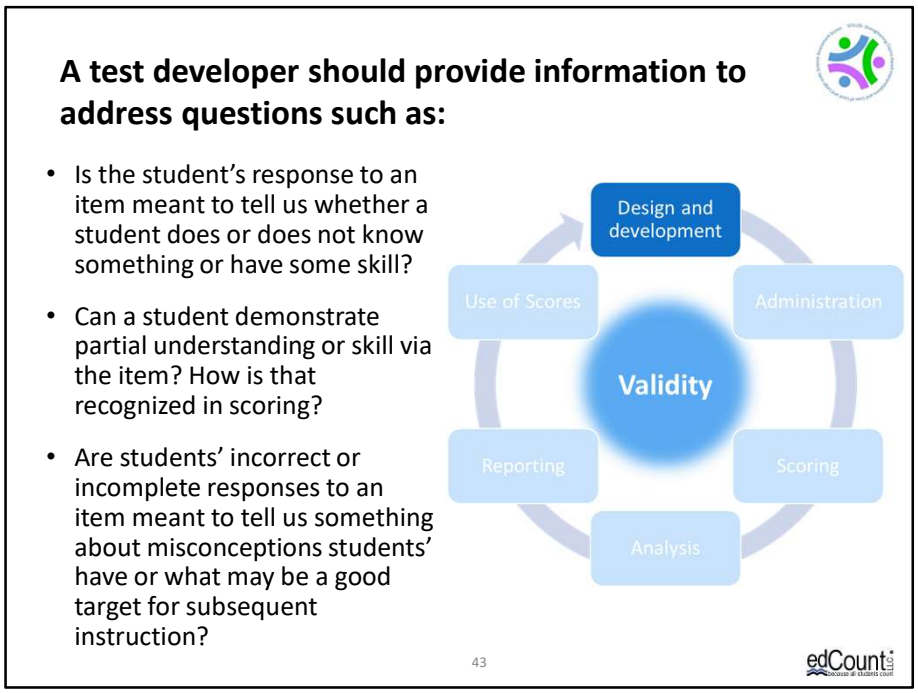


In summary, evidence related to our third construct coherence question should include:
Documentation describing how items were reviewed by the test developer
Documentation describing how items were reviewed by stakeholders in the areas where the test is to be administered or by individuals with similar expertise
Evidence of how and when items were pilot-tested and field-tested and information about how the results of those processes were used to improve individual items and the item bank as a whole
Evidence of how and when items were reviewed for content and fairness considerations by qualified individuals external to the test developer as well as information about how the results of those reviews were used to improve individual items and the item bank as a whole



Our next construct coherence question is:
How are items scored in ways that allow students to demonstrate, and scorers to recognize and evaluate, their knowledge and skills? How are the scoring processes evaluated to ensure they accurately capture and assign value to students' responses?
Evidence to address this question comes from the Design and Development, Administration, and Scoring phases of the assessment life cycle.

Sometimes, scoring may seem like the easiest part of assessment: is the answer right or is it wrong?
But, scoring can be complicated.



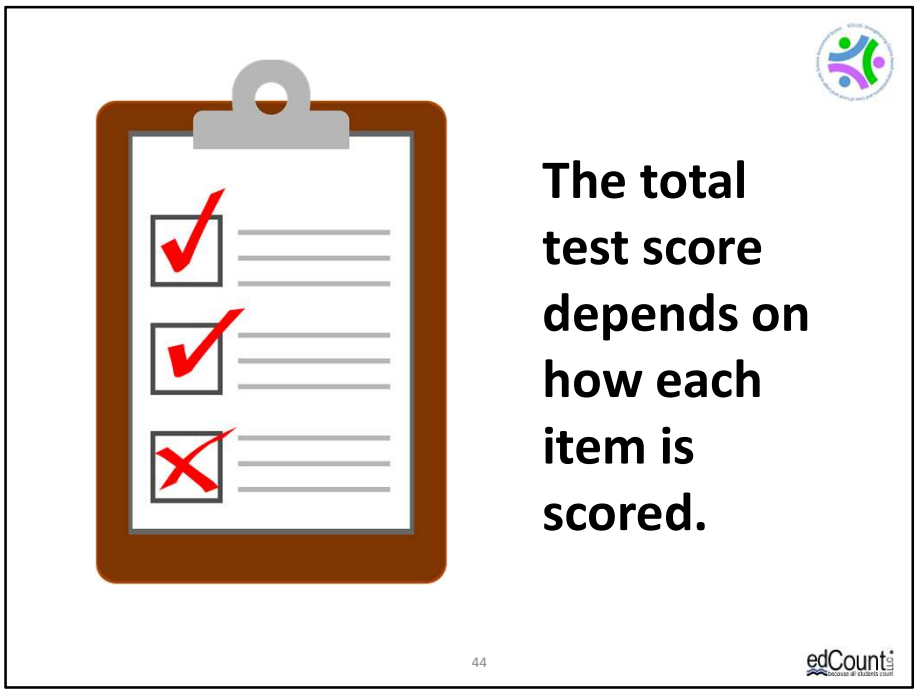
During the Design and Development phase, a test developer must determine what information an item is meant to elicit and how that information is to be represented in a score for that item.

Is the student's response to an item meant to tell us whether a student does or does not know something or have some skill?

Can a student demonstrate partial understanding or skill via the item? How is that recognized in scoring?

Are students' incorrect or incomplete responses to an item meant to tell us something about misconceptions students' have or what may be a good target for subsequent instruction?

We will raise a number of other questions about scoring in the chapters devoted to the comparability, fairness and accessibility, and consequences validity questions.

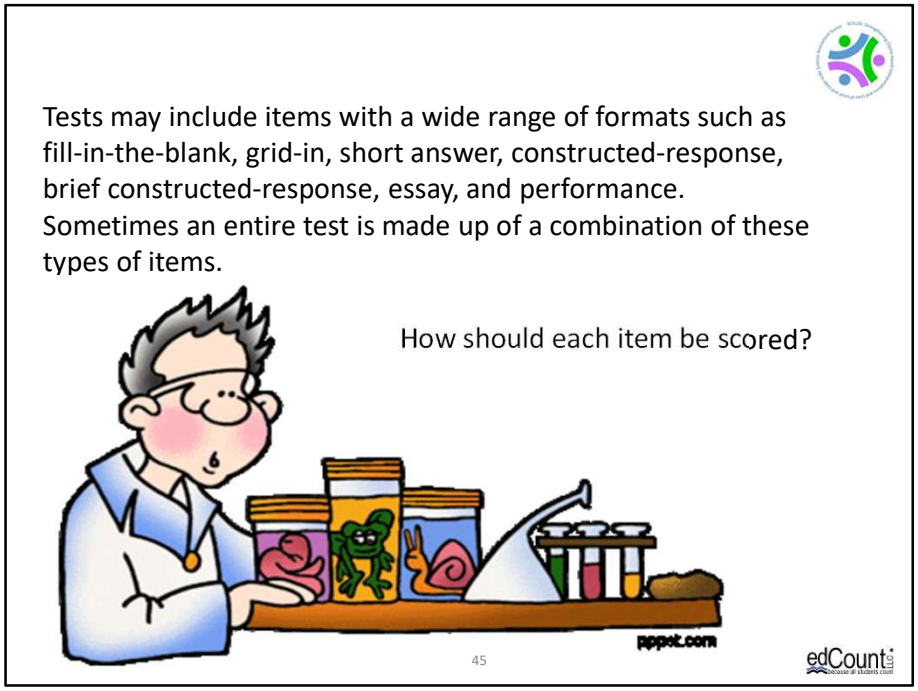


Consider that 50-item test in which 40 of the items are operational. Does that mean that a perfect raw score is 40?

A raw score is simply the sum of the scores for each of the items. If each item were worth one point, then a perfect raw score would be 40.

This is often the case when the test is made up of items known as “multiple-choice” or “selected-response”, which present a question or a statement and require a student to pick from a set of two or more options that include the correct answer and one or more “distractors.”

But, perhaps some of the items allow for partial credit. That is, a student could get half a point for a partly right answer and one point for a completely right answer. Or perhaps a completely right answer is worth two points and a partly right answer is worth only one point.



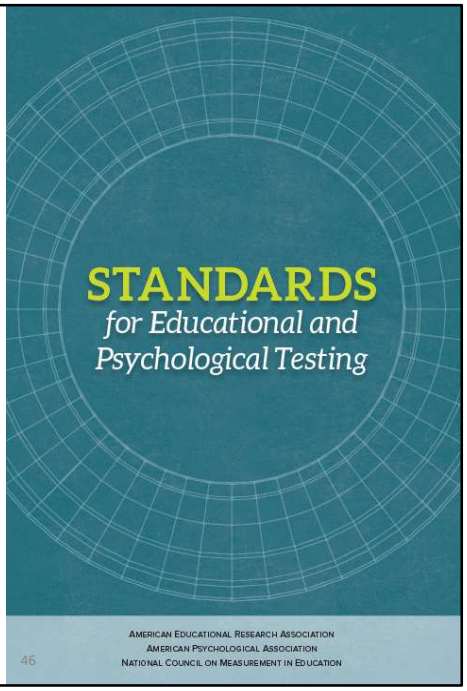
Perhaps some of the items require students to construct responses rather than select from the set of options presented to them. These types of items include a wide range of formats such as fill-in-the-blank, grid-in, short answer, constructed-response, brief constructed-response, essay, and performance. Sometimes an entire test is made up of a combination of these types of items.
Often, these types of items are worth more than selected-response items. But, how much more and how should this be determined?

Standard 4.18: Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear, the is especially critical for extended-response items such as performance tasks, portfolios, and essays. (p. 91)

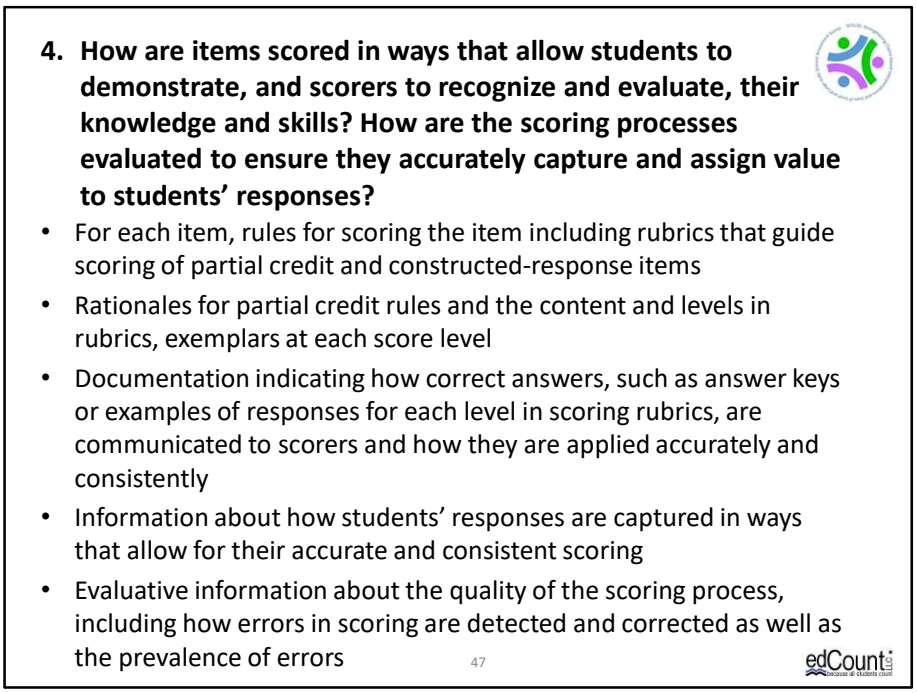
Standard 6.8: Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented. (p. 118)

Standard 6.9: Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected. (p. 118)

(AERA, APA, & NCME, 2014)



Rules for scoring items, including rubrics for scoring partial credit and constructed-response items, must be established as part of item design and development.
The professional standards for testing articulate several obligations related to the construct coherence aspect of scoring include:
Standard 4.18: Procedures for scoring and, if relevant, scoring criteria, should be presented by the test developer with sufficient detail and clarity to maximize the accuracy of scoring. Instructions for using rating scales or for deriving scores obtained by coding, scaling, or classifying constructed responses should be clear, the is especially critical for extended-response items such as performance tasks, portfolios, and essays.
Standard 6.8: Those responsible for test scoring should establish scoring protocols. Test scoring that involves human judgment should include rubrics, procedures, and criteria for scoring. When scoring of complex responses is done by computer, the accuracy of the algorithm and processes should be documented.
Standard 6.9: Those responsible for test scoring should establish and document quality control processes and criteria. Adequate training should be provided. The quality of scoring should be monitored and documented. Any systematic source of scoring errors should be documented and corrected.



Documentation that provides evidence related to the construct coherence aspect of item scoring should include:

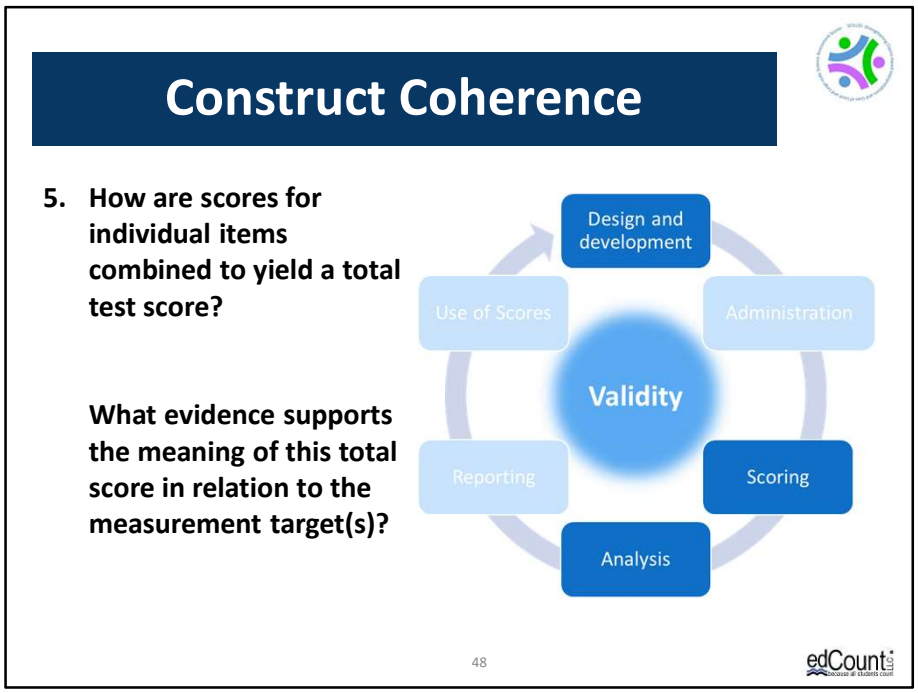
For each item, rules for scoring the item including rubrics that guide scoring of partial credit and constructed-response items

Rationales for partial credit rules and the content and levels in rubrics, exemplars at each score level

Documentation indicating how correct answers, such as answer keys or examples of responses for each level in scoring rubrics, are communicated to scorers and how they are applied accurately and consistently

Information about how students' responses are captured in ways that allow for their accurate and consistent scoring

Evaluative information about the quality of the scoring process, including how errors in scoring are detected and corrected as well as the prevalence of errors



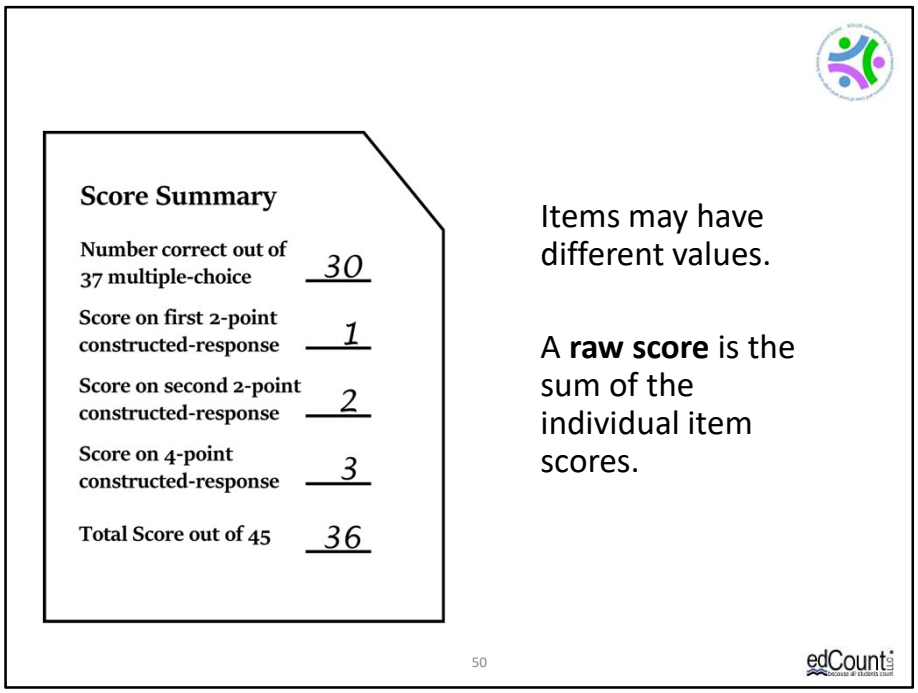
The fifth of our seven questions related to construct coherence is closely related to the fourth question:
How are scores for individual items combined to yield a total test score? What evidence supports the meaning of this total score in relation to the measurement target(s)?
Evidence to address this question comes from the Design and Development, Scoring, and Analysis phases of the assessment life cycle.

- How is information from students' responses to each item combined with information gained from other items?
- How are these combinations determined and do they represent the measurement target in a comprehensive and balanced manner?

Total score = 2?

49

As we've seen, those designing and developing a test must clarify how to score each item. That is, they must determine what constitutes the response or responses for which students get credit. Every test is made up of one or more items and a total test score, which is typically the key piece of information one seeks from a test, is based on the individual scores from the collection of items on the test. Therefore, test designers must address the questions:
How is information from students' responses to each item combined with information gained from other items?
How are these combinations determined and do they represent the measurement target in a comprehensive and balanced manner?

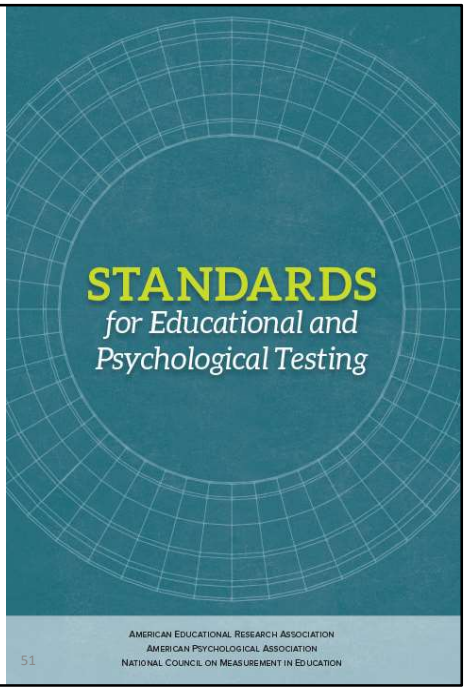


Recall our example of the 50-item test that includes 40 operational items – that is, items that count toward a student's score – and 10 field test items, which do not count toward any scores but contribute important information for future test construction.

We now know that the 40 items may be scored in a variety of ways. How they are scored individually tells us something about how they contribute to the total test score. For tests that teachers create for their own classrooms, and in some other cases, the total test score is usually a "raw" score, meaning that it is simply the sum of the scores for each of the items.

"Scale scores may aid interpretation by indicating how a given score compares with those of other test takers, by enhancing the comparability of scores obtained through different forms of a test, and by helping to prevent confusion with other scores."

(AERA, NCME, APA, 2014, p. 95)



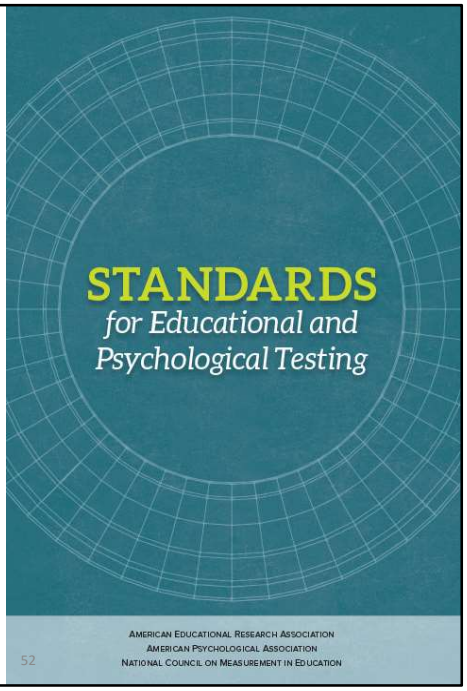
For nearly all commercial tests, "scale scores" are reported in addition to or instead of raw scores. Scale scores are statistical transformations of raw scores. We use scale scores when we are concerned with the comparability of score meaning across different forms of a test. For example, the score scale for the ACT ranges from one to 36. A student who takes the ACT twice can compare her scores and know if she really scored better the second time even though the questions on the test were different. The process of creating a score scale allows for individual items to be weighted in their contribution to the total test score by not only the raw score points each is worth, but by other characteristics such as item difficulty.

Standard 5.0: Test scores should be derived in a way that supports the intended interpretations of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use. (p. 102)

Standard 5.1: Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretations of scale scores, as well as their limitations. (p. 102)

Standard 5.4: When raw scores are intended to be directly interpretable, their meaning, intended interpretations, and limitations should be described and justified in the same manner as is done for scale scores. (p. 103)

(AERA, APA, & NCME, 2014)

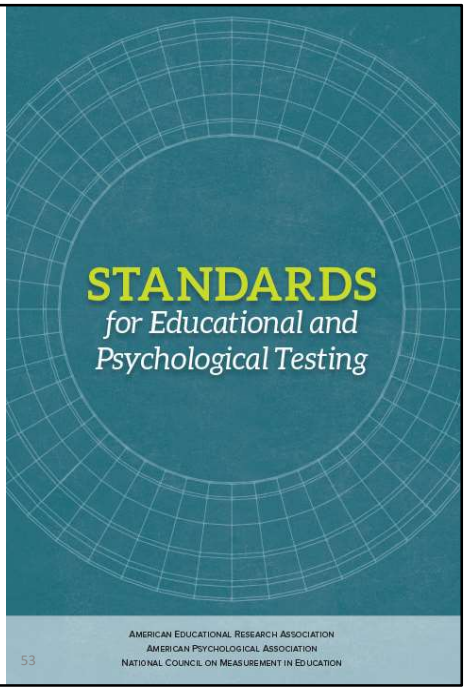


Our professional standards address obligations for both raw scores and scale scores.
Standard 5.0: Test scores should be derived in a way that supports the intended interpretations of test scores for the proposed uses of tests. Test developers and users should document evidence of fairness, reliability, and validity of test scores for their proposed use.
Standard 5.1: Test users should be provided with clear explanations of the characteristics, meaning, and intended interpretations of scale scores, as well as their limitations.
Standard 5.4: When raw scores are intended to be directly interpretable, their meaning, intended interpretations, and limitations should be described and justified in the same manner as is done for scale scores.

Our professional standards also address performance levels such as “proficient”, “on-track”, or “pass.” For example:

Standard 5.21: When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly. (p. 107)

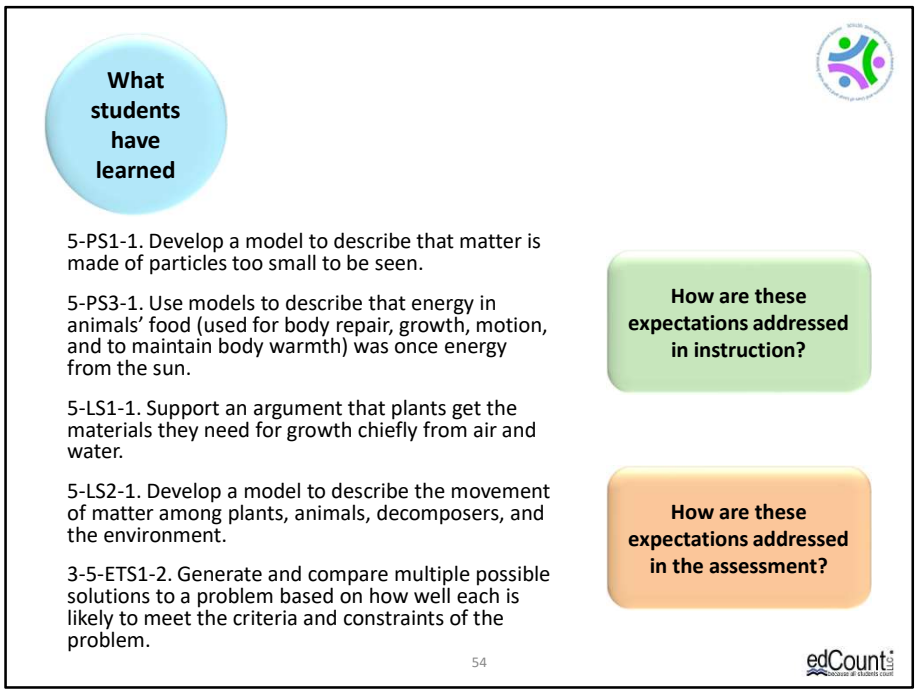
(AERA, APA, & NCME, 2014)



Our professional standards also address performance levels, which are always reported for statewide accountability tests and often for many other types of tests. Performance levels have names, such as “proficient” or “on-track”, as well as descriptions of the performance associated with test scores in each performance level range. The scores that differentiate between levels are called cut scores and these are established using a process called standard setting.

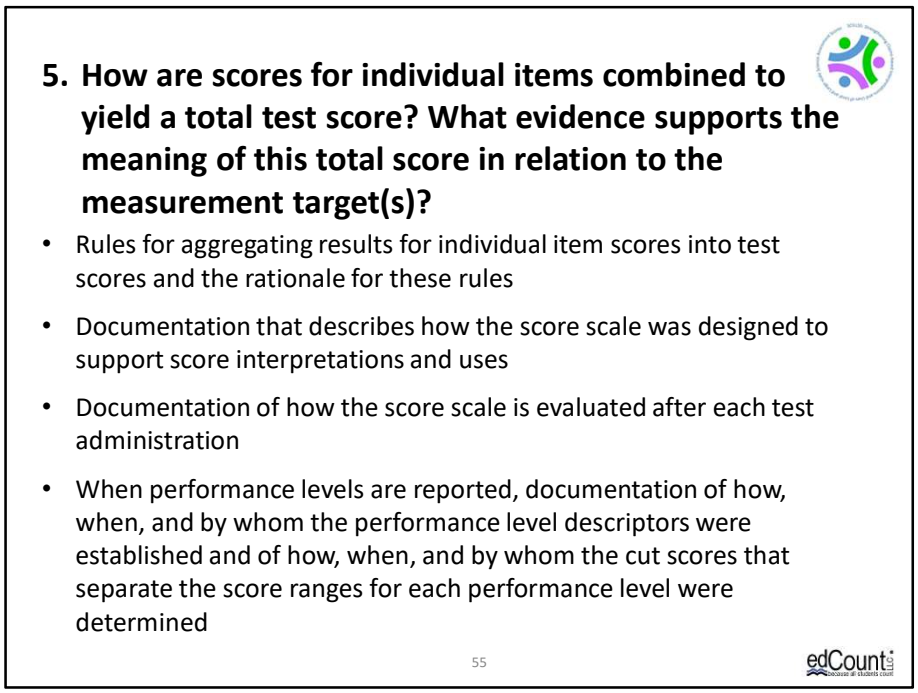
Standard 5.21: When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly.

Those using performance levels, such as “pass” or “fail” or such as “proficient” or “basic” need to understand how those levels were set and think carefully about the descriptions of performance or skill that accompany those labels. They should ask, “what evidence supports the claims in the performance level descriptions?”

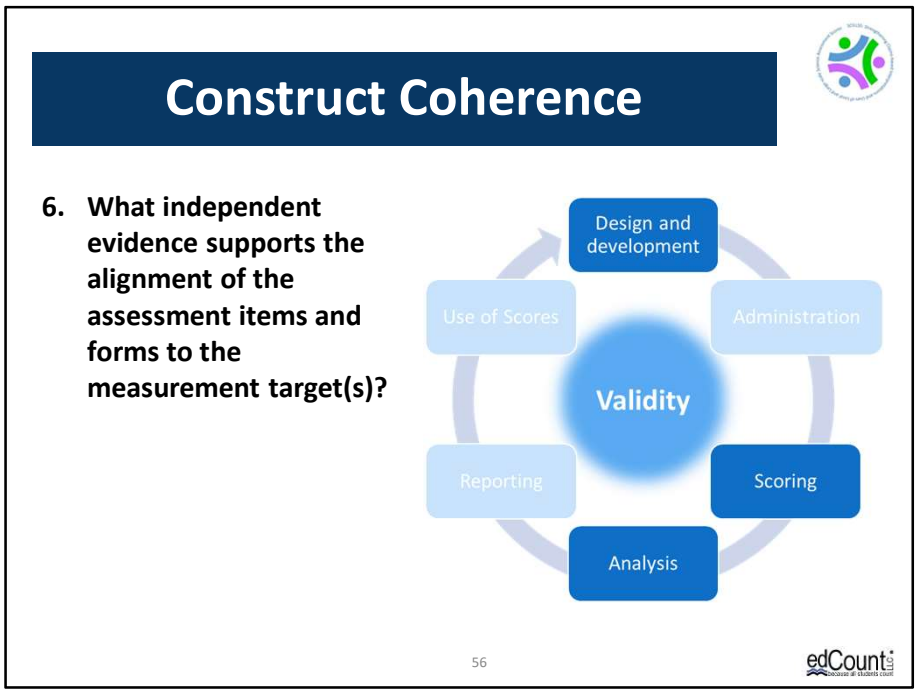


For all types of scores, test developers must explain how results from individual items are combined to create test scores. In addition, developers must provide evidence that the combination of items that contribute to test scores, and the way in which item results are combined, yields test scores that reflect what the test is supposed to measure.

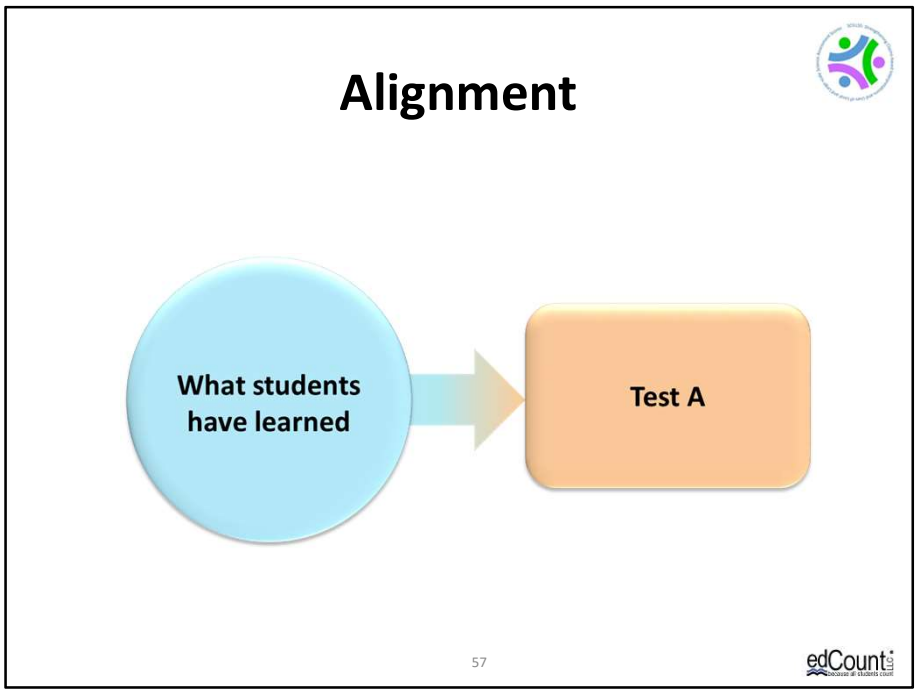
Let's say a test is meant to measure a student's science achievement at the end of fifth grade and includes items that adequately represent the knowledge and skills in the fifth-grade science standards. But, the test developer designs the scoring process such that some items are heavily weighted in their contributions to the total test score and others are not counted at all. Unless this pattern aligns with patterns in the science standards or in the teaching and learning related to them, this approach to calculating total test scores would not support the intended interpretations of the scores as reflecting students' science achievement.



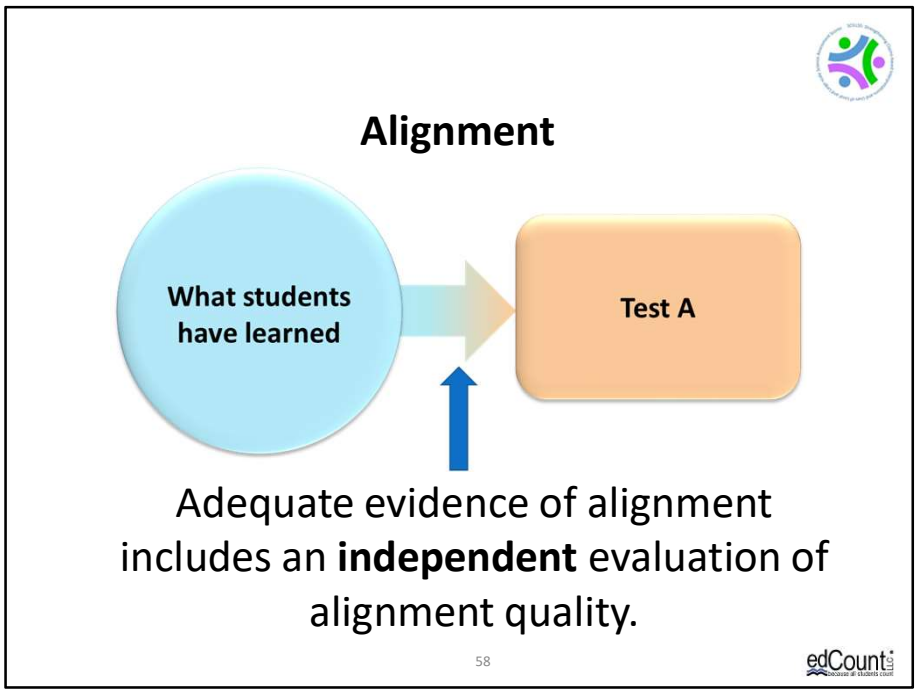
Documentation that provides evidence related to the construct coherence aspect of calculating test scores should include:
Rules for aggregating results for individual item scores into test scores and the rationale for these rules
Documentation that describes how the score scale was designed to support score interpretations and uses
Documentation of how the score scale is evaluated after each test administration
When performance levels are reported, documentation of how, when, and by whom the performance level descriptors were established and of how, when, and by whom the cut scores that separate the score ranges for each performance level were determined



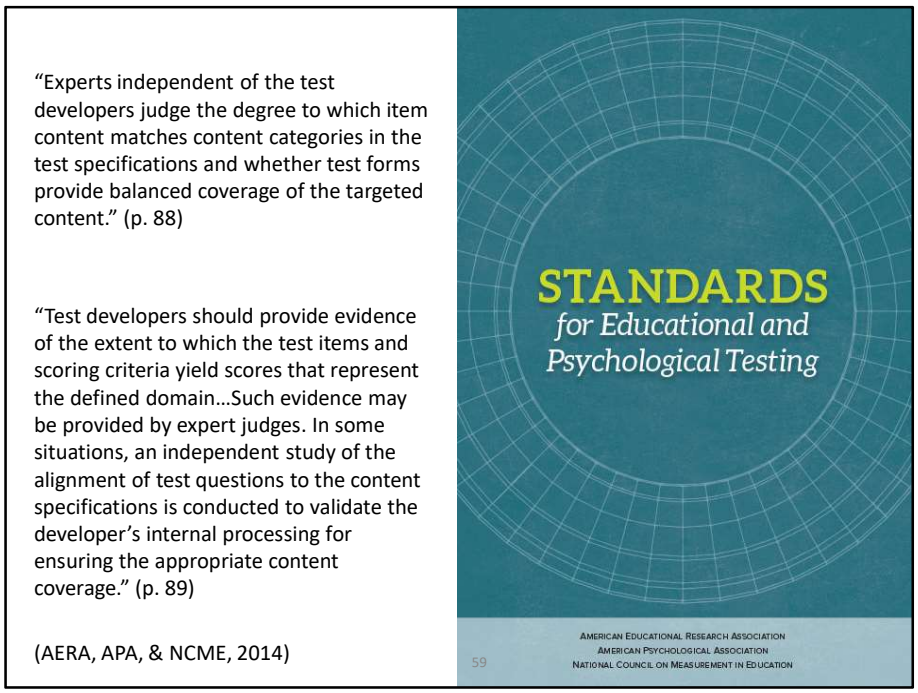
Our sixth question related to construct coherence is:
What independent evidence supports the alignment of the assessment items and forms to the measurement target(s)?
Evidence to address this question comes from the Design and Development, Scoring, and Analysis phases of the assessment life cycle.



We've used the words "align" and "alignment" a few times in this chapter. We've indicated that test developers must establish frameworks, blueprints, or specifications that define what a test is meant to measure and provide a rationale for how the test is designed to yield the intended information. In addition, developers must provide evidence of how test items are created and clear statements about what each is intended to measure. To further support the interpretation of test scores in relation to the constructs that tests are meant to measure, developers must create rules for how items are scored and how these results are combined into test scores and provide evidence that these rules yield meaningful test scores. All of these obligations relate to alignment.



Our sixth construct coherence question demands additional evidence of alignment, this time from those who were not involved in the test design and development process. Clearly, those who build tests have a vested interest in ensuring that scores from their tests mean what they say they mean. However, even with the most rigorous and diligent test development strategies, a test may miss its mark to some extent. As is the case for any credible evaluation, an independent perspective is necessary to protect against unfair results. A test user should never simply accept a commercial test developer's claim that the test is aligned. Evidence from an independent alignment evaluation is necessary to back up this claim.

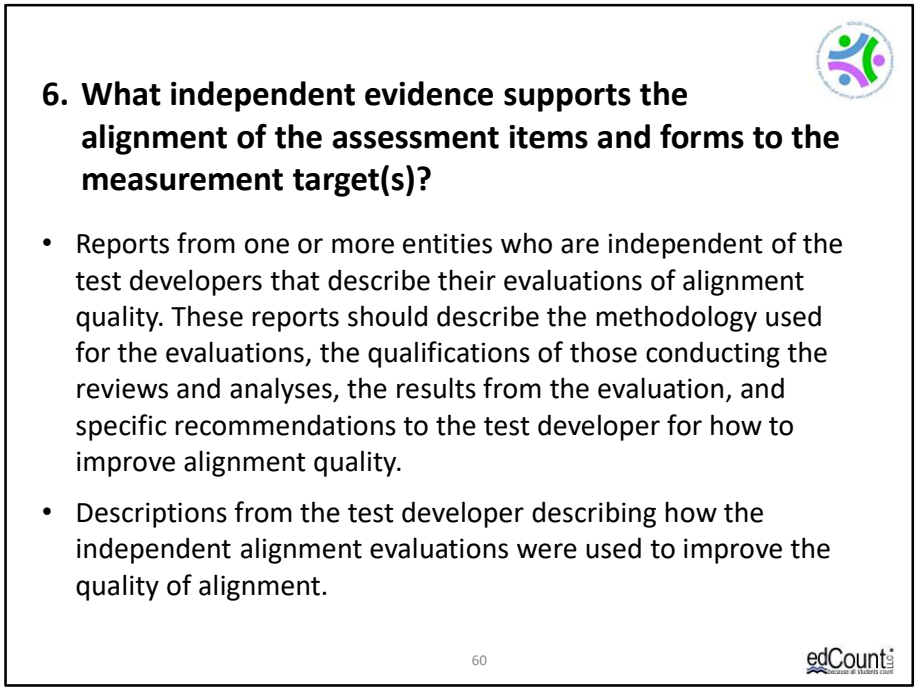


Our professional standards clearly call for this evidence. Under Standard 4.8, referenced previously in this chapter under our third construct coherence question, the explanatory text describes alignment evaluation.

"Experts independent of the test developers judge the degree to which item content matches content categories in the test specifications and whether test forms provide balanced coverage of the targeted content."

The explanation under Standard 4.12 also addresses the need for independent evaluation of alignment quality.

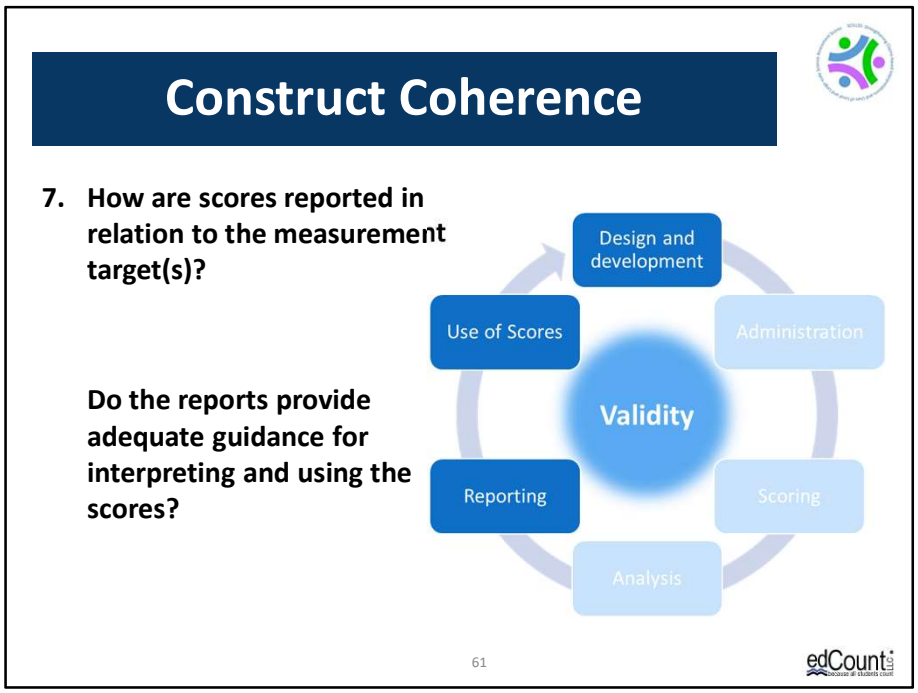
"Test developers should provide evidence of the extent to which the test items and scoring criteria yield scores that represent the defined domain…Such evidence may be provided by expert judges. In some situations, an independent study of the alignment of test questions to the content specifications is conducted to validate the developer's internal processing for ensuring the appropriate content coverage."



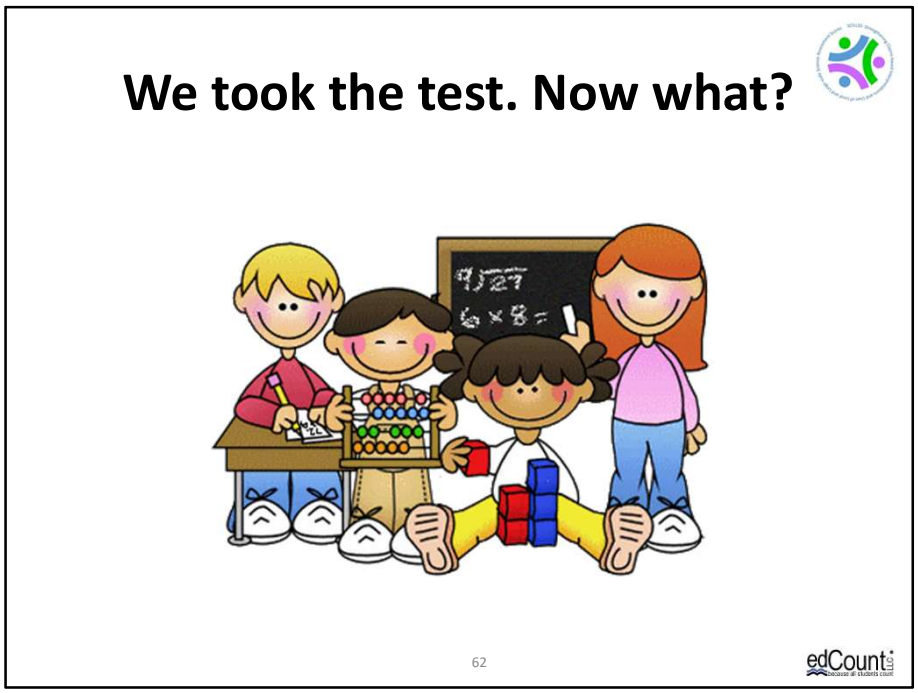
When interpretations of test scores are to an academic domain or part of that domain, those using the tests must take great care to consider the independent alignment evidence. It would be unwise to adopt a test when the developers cannot provide independent evidence to support their claims about what the test measures. Such evidence should include:

Reports from one or more entities who are independent of the test developers that describe their evaluations of alignment quality. These reports should describe the methodology used for the evaluations, the qualifications of those conducting the reviews and analyses, the results from the evaluation, and specific recommendations to the test developer for how to improve alignment quality.

Descriptions from the test developer describing how the independent alignment evaluations were used to improve the quality of alignment.



Our seventh and final construct coherence question is:
How are scores reported in relation to the measurement target(s)? Do the reports provide adequate guidance for interpreting and using the scores?
Evidence related to this question comes from the Design and Development, Reporting, and Score Use phases of the assessment life cycle.



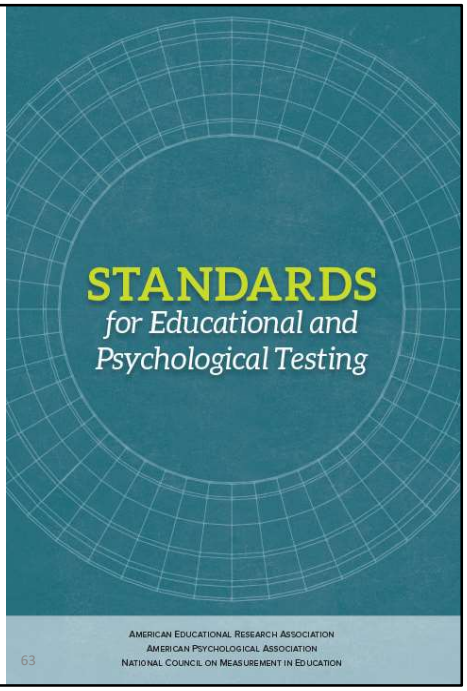
We may take as a truism that any test that does not yield scores that are meaningful and useful should not be given.
The entire purpose of this workbook is to help those who use tests, whether they develop them themselves or adopt ones that have been developed by others, ensure that the test scores and their subscores they use actually have their intended meaning. This particular validity question is meant to help test users consider how scores and subscores are reported in ways that support their appropriate interpretation and use.

Standard 6.10: When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what the scores represent, the precision/reliability of the scores, and how scores are intended to be used. (p. 119)

Standard 12.18: In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and be supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports. (p. 200)

Standard 12.19: In educational settings, when score reports include recommendations for instructional intervention or are linked to recommended plans or materials for instruction, a rationale for and evidence to support these recommendations should be provided. (p. 201)

(AERA, APA, & NCME, 2014)

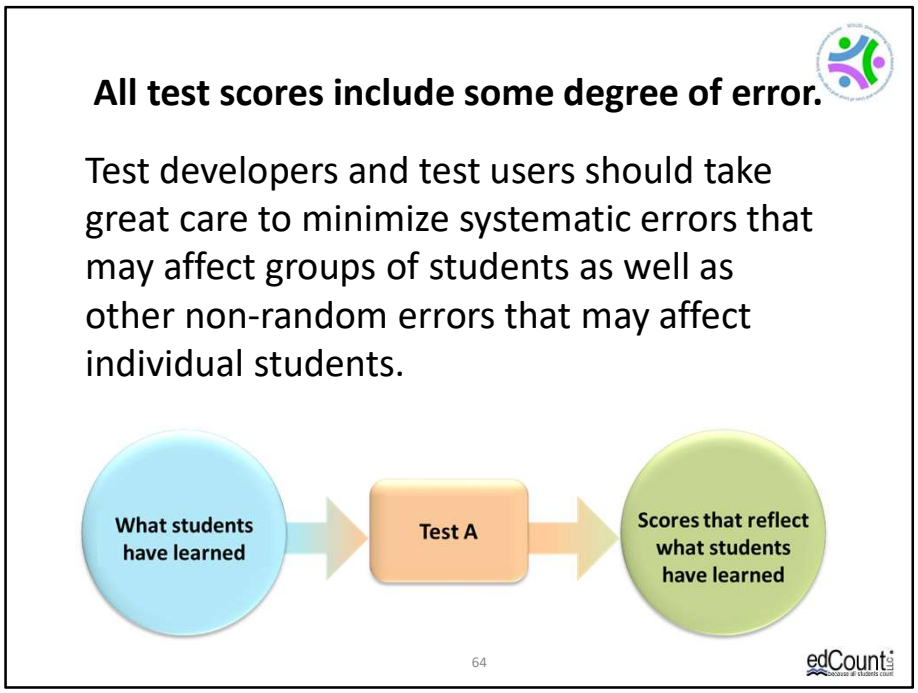


Our professional standards speak directly to the obligations of those reporting test scores. Relevant standards include:

Standard 6.10: When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what the scores represent, the precision/reliability of the scores, and how scores are intended to be used.

Standard 12.18: In educational settings, score reports should be accompanied by a clear presentation of information on how to interpret the scores, including the degree of measurement error associated with each score or classification level, and be supplementary information related to group summary scores. In addition, dates of test administration and relevant norming studies should be included in score reports.

Standard 12.19: In educational settings, when score reports include recommendations for instructional intervention or are linked to recommended plans or materials for instruction, a rationale for and evidence to support these recommendations should be provided.



These standards, as well as other standards related to reporting, call out the concept of error when they refer to precision/reliability, measurement error. Some degree of error is associated with all testing and we will address this issue in depth in chapter 3 of this workbook series.

For now, we note that test developers and test users should take great care to minimize systematic errors that may affect groups of students as well as other non-random errors that may affect individual students.

- ✓ What the scores mean in relation to what the test was meant to measure
- ✓ How the scores may be used appropriately

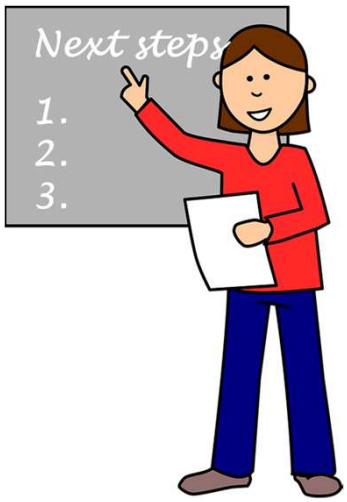
- ✓How scores should <u>not</u> be interpreted and used

65

edCount

Those who report scores are obligated to provide clear information about what scores mean in relation to what the test was meant to measure and how the scores may be used appropriately. Reports should also include information about how the scores should and should not be used. Reports should also provide information to help avoid uses that are not supported with validity evidence. All of this information must be clear and accessible for those who are meant to understand and use the scores; these individuals include teachers, administrators, parents, and students. Documentation that accompanies scores should describe the purpose of the test, what the scores mean, what evidence supports score meaning, and any cautions for score use.

If a test publisher offers recommendations for next steps in instruction, such as specific lessons for individual students or groups of students, that publisher should provide evidence that those recommendations are sound.



66

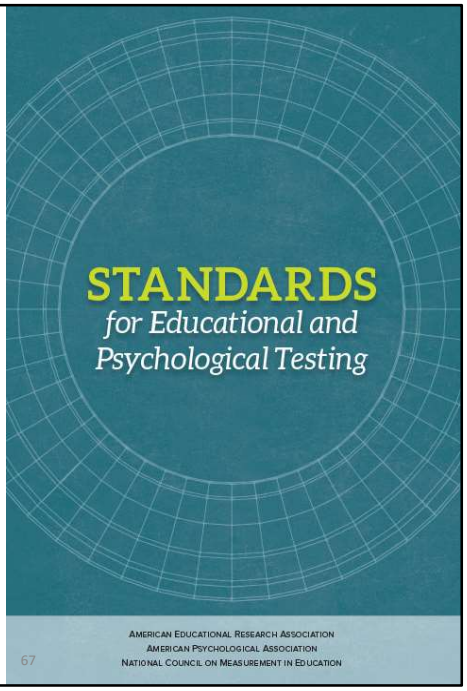
edCount

If those reporting scores further indicate a course of action based upon the scores, they are obligated to provide evidence to support that course of action. For example, if a test publisher offers recommendations for next steps in instruction, such as specific lessons for individual students or groups of students, that publisher should provide evidence that those recommendations are sound.

Standard 7.12: When test scores are used to make predictions about future behavior, the evidence supporting those predictions should be provided to the test user. (p. 129)

Standard 3.7: When criterion-related validity evidence is used as a basis for test score-based predictions of future performance and sample sizes are sufficient, test developers and/or users are responsible for evaluating the possibility of differential prediction for relevant subgroups for which there is prior evidence or theory suggesting differential prediction. (p. 66)

(AERA, APA, & NCME, 2014)



If a test publisher indicates that the test scores may be used to predict performance, say, in a college or career setting, then that publisher is obligated to provide evidence that these are appropriate interpretations and uses of the scores. In our professional standards, this obligation is expressed in general and in relation to how predictions may differ for individuals in different groups.

Standard 7.12: When test scores are used to make predictions about future behavior, the evidence supporting those predictions should be provided to the test user.

Standard 3.7: When criterion-related validity evidence is used as a basis for test score-based predictions of future performance and sample sizes are sufficient, test developers and/or users are responsible for evaluating the possibility of differential prediction for relevant subgroups for which there is prior evidence or theory suggesting differential prediction.

Those who use test scores to make predictions, and then to make decisions based upon those predictions such as admission to a program or school, must explore their data when groups differ – or could differ – in their predicted performance.

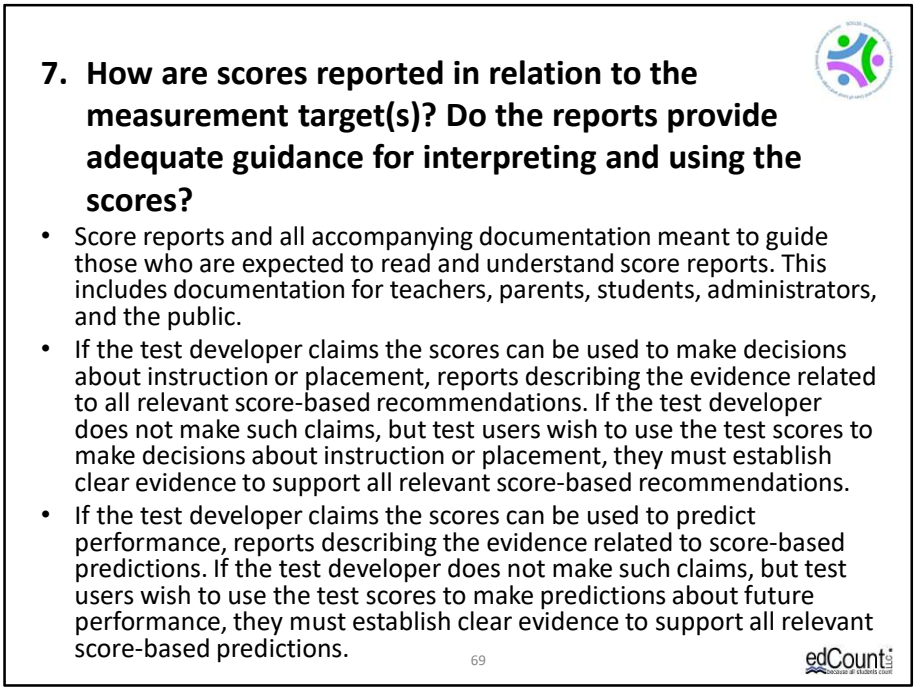
These groups could be male versus female, racial/ethnic groups, students with and without disabilities, and English learners versus students who are English proficient.

Those who use these scores are obligated to investigate the reasons for these differences and use that information in their decisions about whether and how to use the scores.

68

edCount

This latter point means that those using test scores to make predictions, and then to make decisions based upon those predictions such as admission to a program or school, must explore their data when groups differ – or could differ – in their predicted performance. These groups could be male versus female, racial/ethnic groups, students with and without disabilities, and English learners versus students who are English proficient. Those who use these scores are obligated to investigate the reasons for these differences and use that information in their decisions about whether and how to use the scores.

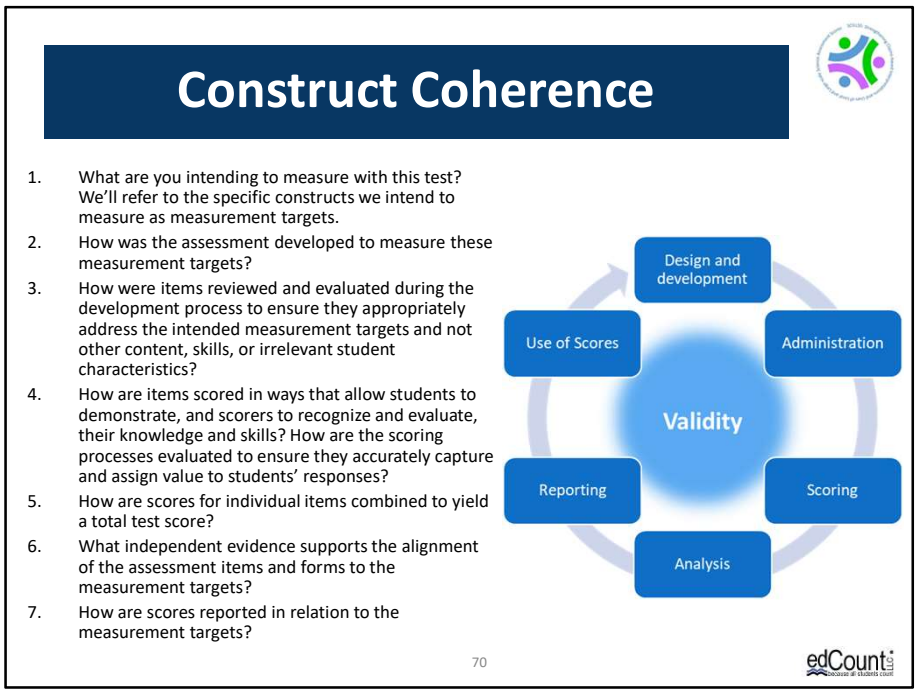


Developers should investigate the extent to which the reports are interpreted correctly by the relevant user through focus groups or review meetings for the purpose of collecting evidence related to our seventh construct coherence question which includes:

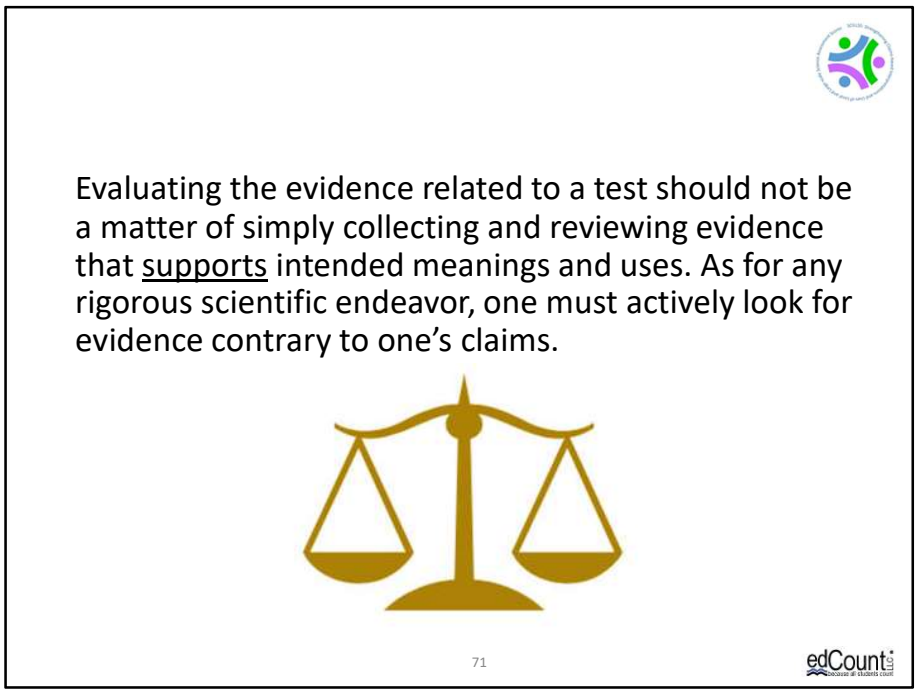
Score reports and all accompanying documentation meant to guide those who are expected to read and understand score reports. This includes documentation for teachers, parents, students, administrators, and the public.

If the test developer claims the scores can be used to make decisions about instruction or placement, reports describing the evidence related to all relevant score-based recommendations. If the test developer does not make such claims, but test users wish to use the test scores to make decisions about instruction or placement, they must establish clear evidence to support all relevant score-based recommendations.

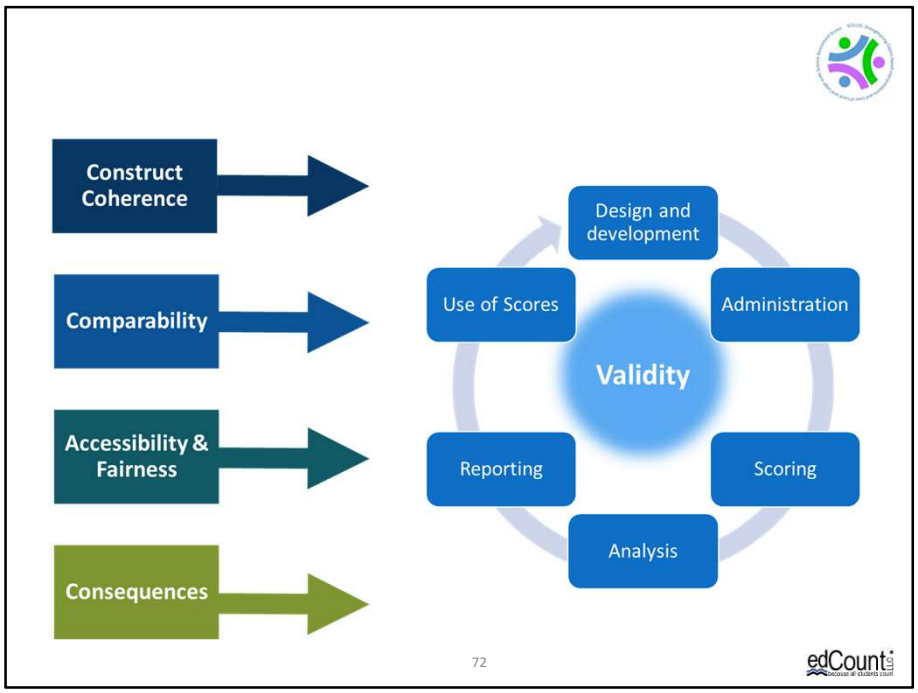
If the test developer claims the scores can be used to predict performance, reports describing the evidence related to score-based predictions. If the test developer does not make such claims, but test users wish to use the test scores to make predictions about future performance, they must establish clear evidence to support all relevant score-based predictions.



We have reached the end of our seven questions in this chapter. Clearly, test developers, or those using test scores for any purpose, are obligated to establish a great deal of evidence to support the ways they interpret and use test scores. With specific regard to construct coherence, this evidence must support those interpretations and uses in terms of what the tests are supposed to measure. What inferences about what students know and can do are we making based upon test scores? We need evidence that those inferences are appropriate.



Throughout this chapter, we've described evidence that test developers or users should establish throughout the testing life cycle. It's important to note that evaluating the evidence related to a test should not be a matter of simply collecting and reviewing evidence that supports intended meanings and uses. As for any rigorous scientific endeavor, one must actively look for evidence contrary to one's claims. Test developers, as well as those who use the tests and the scores, must consider how they would know if, for example, a test were not aligned as expected or if scores did not reflect students' knowledge and skills in 5th grade science. Commercial test publishers may highlight evidence that supports their claims about a test, but should also present information about how they have evaluated their practices and used evaluative information to improve their practice.



We have reached the end of the second chapter of the SCILLSS digital workbook on educational assessment design and evaluation and our focus on evidence related to construct coherence questions. We've focused on evidence for seven key construct coherence questions that test developers and those who require and use test scores must consider to support valid test score interpretations and uses. In the chapters that follow, we will address questions related to comparability, fairness and accessibility, and consequences of test use.

Resources and Additional Information

73

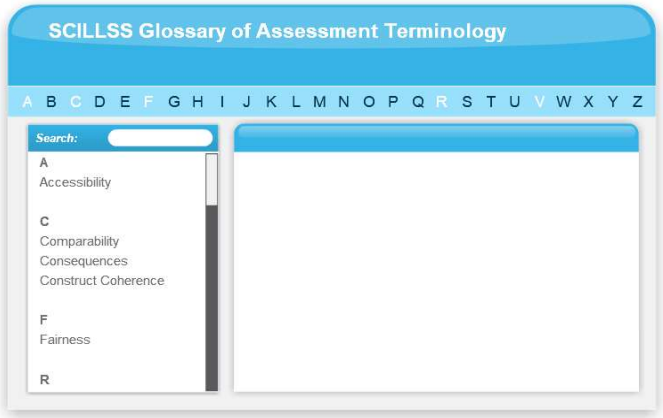
edCount

Finally, we offer additional resources that may be helpful to anyone interested in learning more about the concepts presented in this chapter. A glossary of terms and our reference list follow.

Thank you for your engagement in this second chapter of the SCILLSS digital workbook on educational assessment design and evaluation.

SCILLSS Glossary

Please refer to the SCILLSS Glossary for operational definitions of terms used.



74

Web links

In the web links pod, you can find the following resources.

- American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) Joint Committee on Standards for Educational and Psychological Testing. (2014). *Standards for educational and psychological testing*. Washington DC: American Educational Research Association.
- National Research Council. (2014). *Developing Assessments for the Next Generation Science Standards*. Washington, DC: The National Academies Press.
- SCILLSS Website

References

American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) Joint Committee on Standards for Educational and Psychological Testing. (2014). Standards for educational and psychological testing. Washington DC: American Educational Research Association.

Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. *Journal of Educational Measurement, 50*(1), 1–73.

National Research Council. (2001). Knowing what students know: The science and design of educational assessment. Committee on the Foundations of Assessment. Pellegrino, J., Chudowsky, N., and Glaser, R., editors. Board on Testing and Assessment, Center for Education, Division of Behavioral and Social Sciences and Education. Washington DC: National Academy Press.

Pellegrino, J. W., DiBello, L. V., & Goldman, S. R. (2016). A framework for defining and evaluating the validity of instructionally relevant assessments. *Educational Psychologist, 51*(1), 59-81.