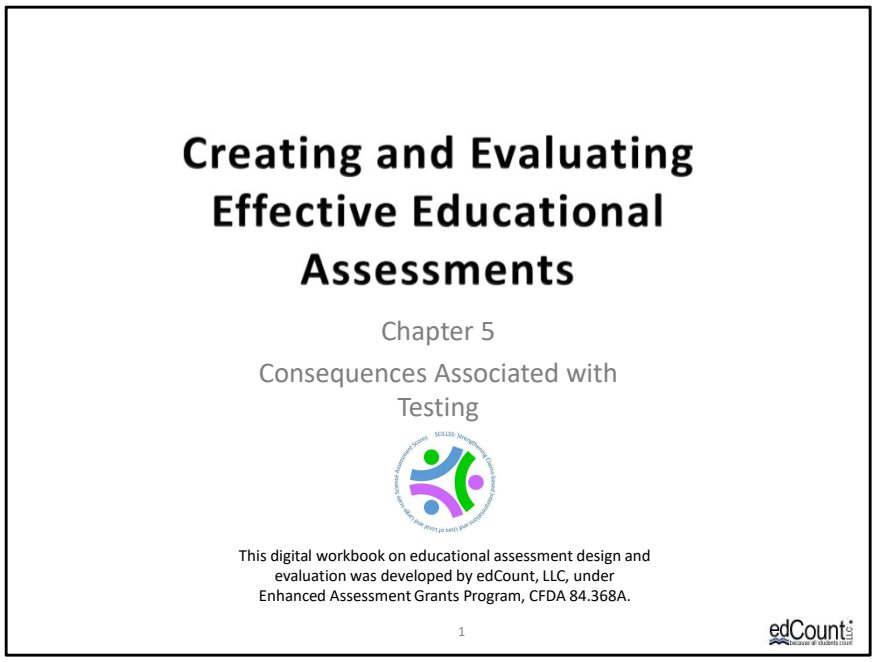
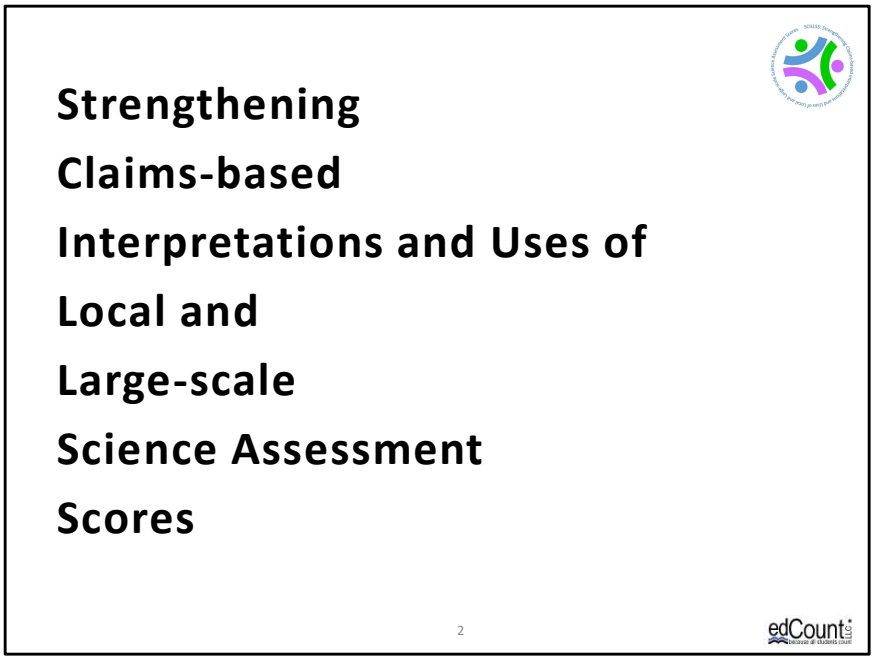


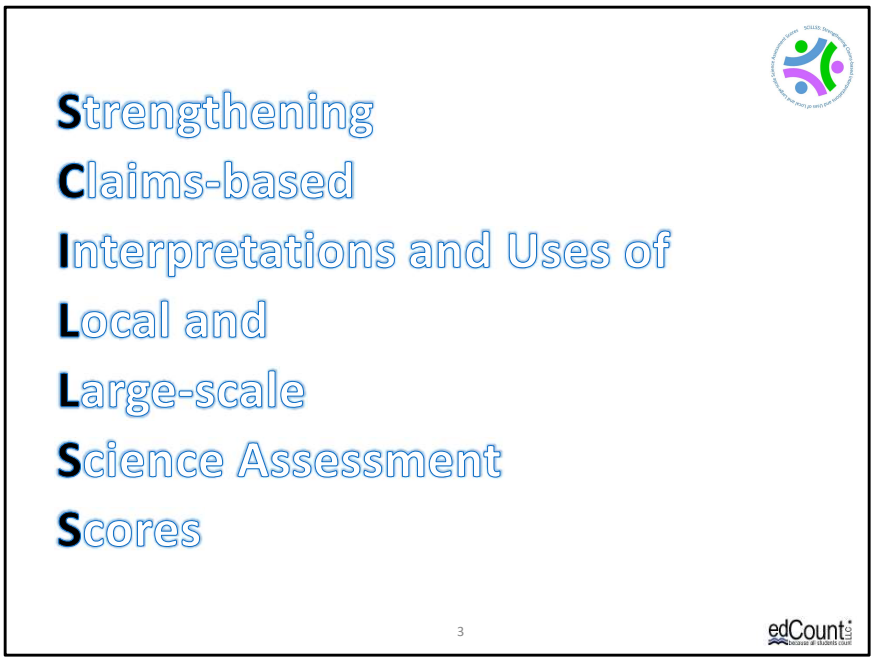
Chapter 5: Consequences

Welcome to the fifth 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.

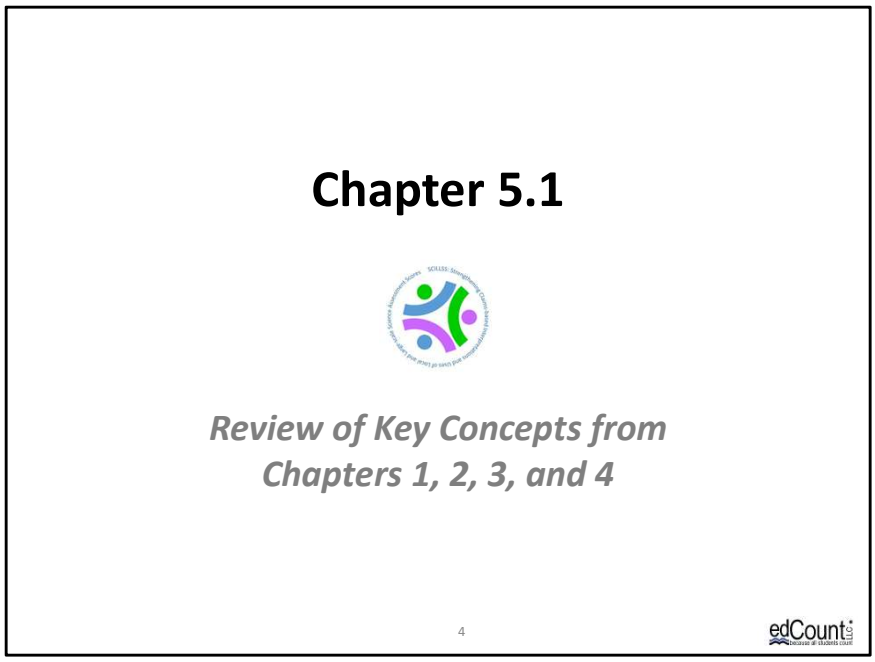
This digital workbook was developed by edCount, LLC, under the US Department of Education's Enhanced Assessment Grants Program, CFDA 84.368A.



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

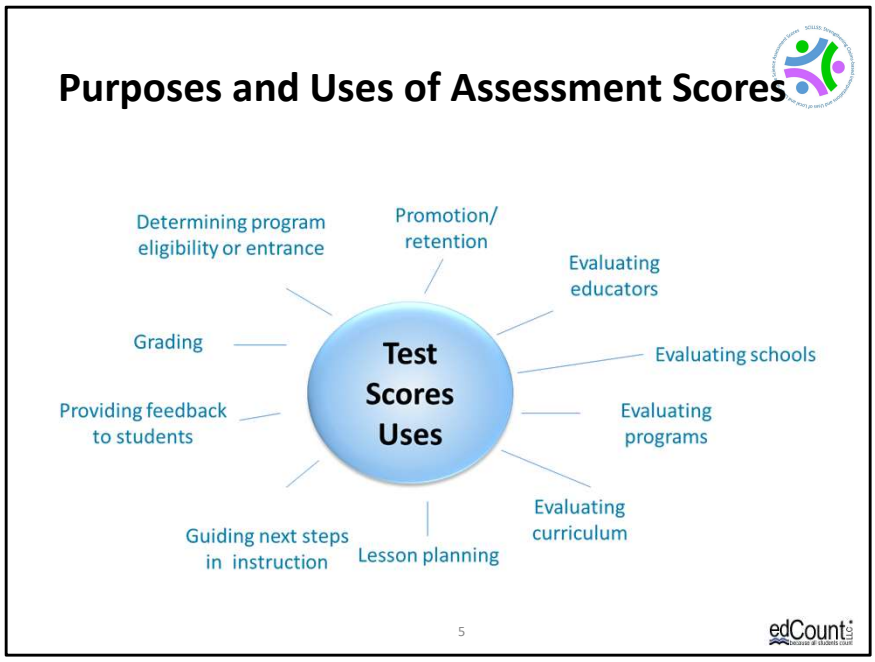


or its acronym, “SCILLSS.”



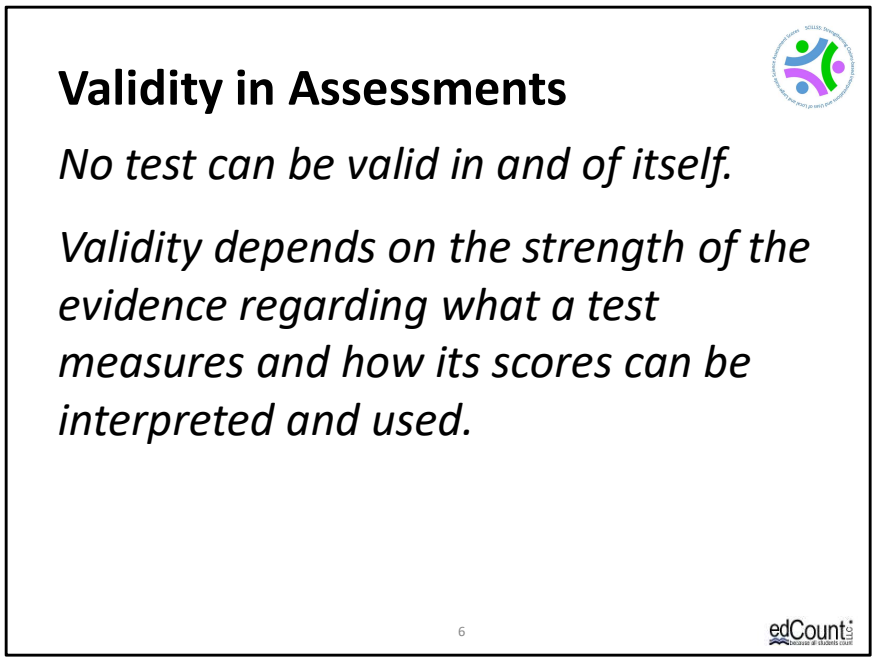
Chapter 5.1.

Review of Key Concepts from Chapters 1, 2, 3, and 4

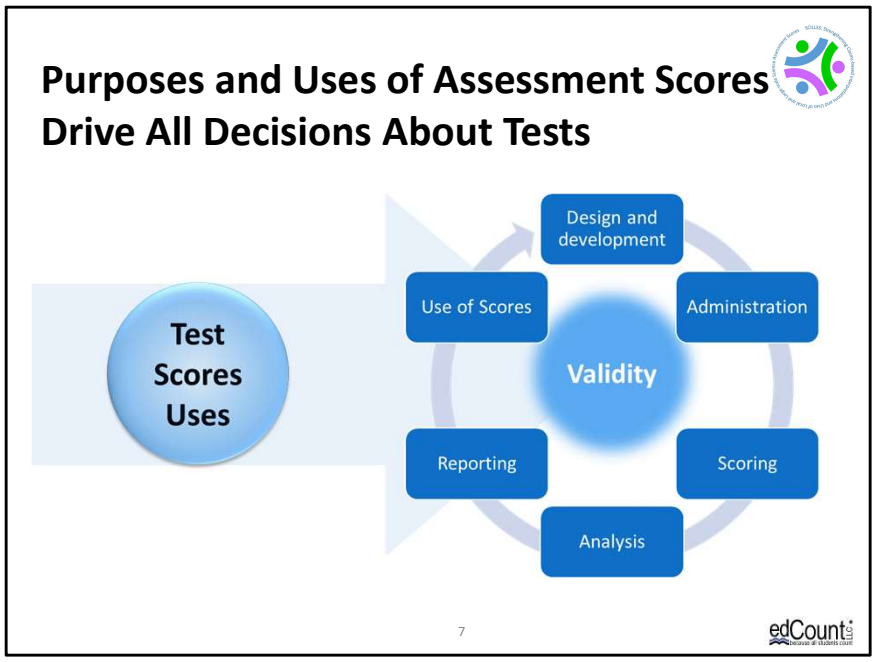


Let's begin with a brief recap of the key concepts covered in the first four chapters of this series.

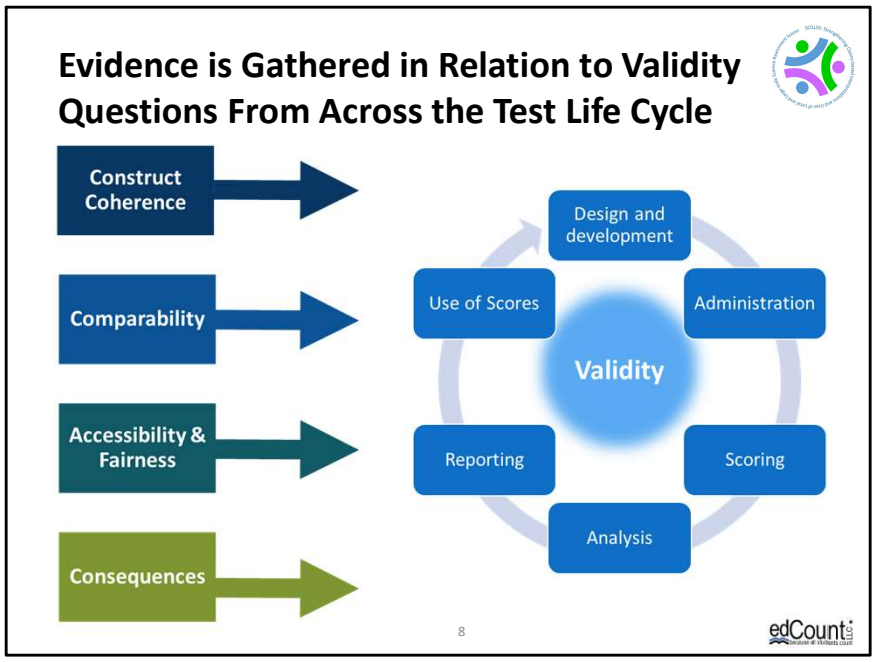
Chapter 1 focused on common reasons why we administer assessments of students' academic knowledge and skills and how we use those assessment scores. We learned that these purposes for administering assessments and the intended uses of assessment scores should drive all decisions about how assessments are designed, built, and evaluated.



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



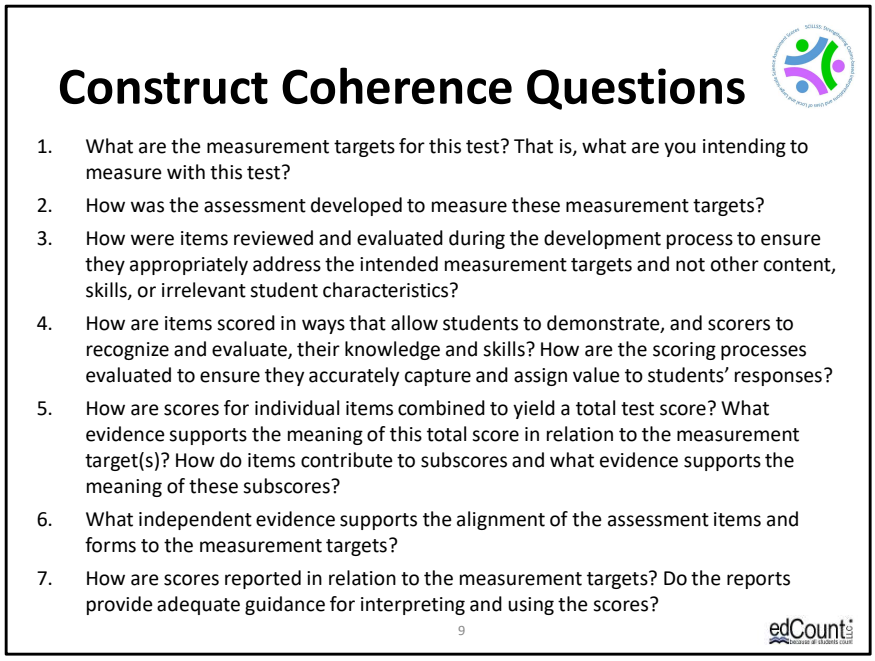
We evaluate validity by gathering and judging evidence. This validity evidence is gathered from across the entire life cycle of a test from design and development through score use. Judgments about validity are based upon the quality and adequacy of this evidence in relation to assessment score interpretations 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.



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 four validity question categories are:

- Construct coherence: 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?
- Comparability: To what extent are the test scores reliable and consistent in meaning across all students, classes, schools, and time?
- Accessibility and fairness: To what extent does the test allow all students to demonstrate what they know and can do? And
- Consequences: To what extent are the test scores used appropriately to achieve specific goals?

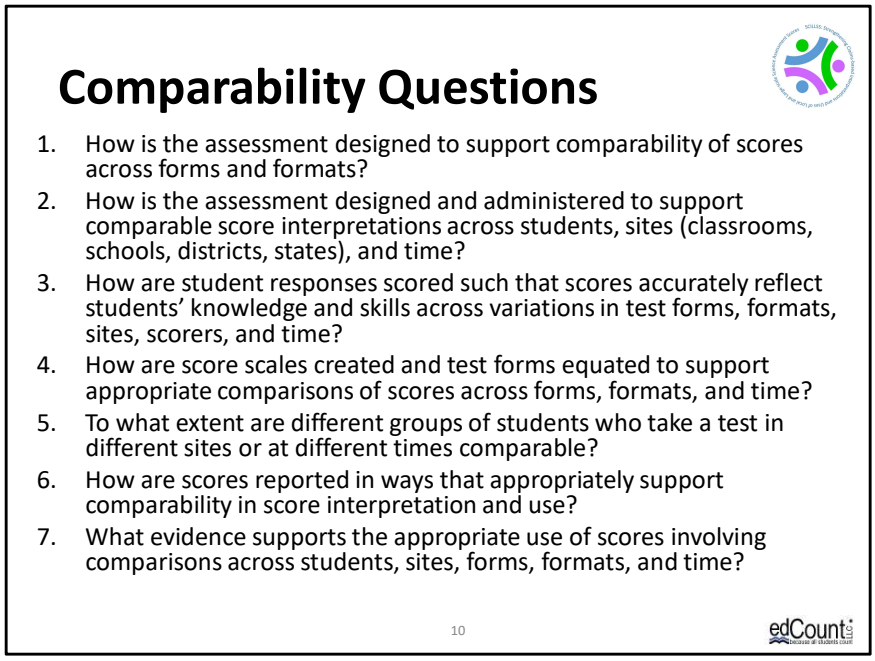


Chapter 2 of this digital workbook focused on the first set of these questions, construct coherence. We addressed the types of evidence that relate to seven key construct coherence questions.

1. What are the measurement targets for this test? That is, what are you intending to measure with this test?
2. How was the assessment developed to measure these measurement targets?
3. 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?
4. How are items scored in ways that allowed 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?
5. 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?
6. What independent evidence supports the alignment of the assessment items and

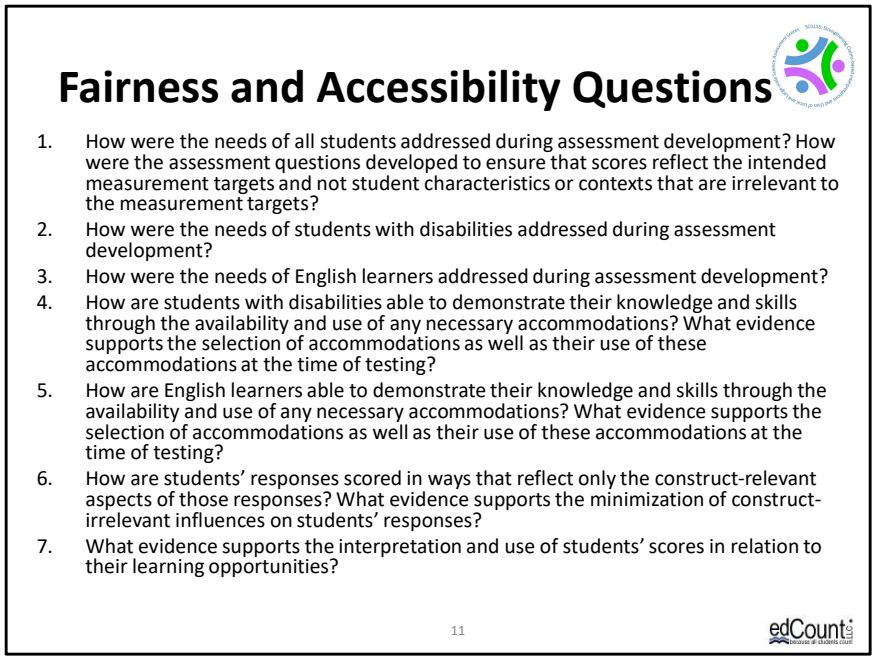
forms to the measurement targets? And,

7. How are scores reported in relation to the measurement targets? Do the reports provide adequate guidance for interpreting and using the scores?



Chapter 3 focused on the second set of these questions, which relate to comparability and reliability/precision. These questions are:

1. How is the assessment designed to support comparability of scores across forms and formats?
2. How is the assessment designed and administered to support comparable score interpretations across students, sites (classrooms, schools, districts, states), and time?
3. How are student responses scored such that scores accurately reflect students' knowledge and skills across variations in test forms, formats, sites, scorers, and time?
4. How are score scales created and test forms equated to support appropriate comparisons of scores across forms, formats, and time?
5. To what extent are different groups of students who take a test in different sites or at different times comparable?
6. How are scores reported in ways that appropriately support comparability in score interpretation and use?
7. What evidence supports the appropriate use of scores involving comparisons across students, sites, forms, formats, and time?



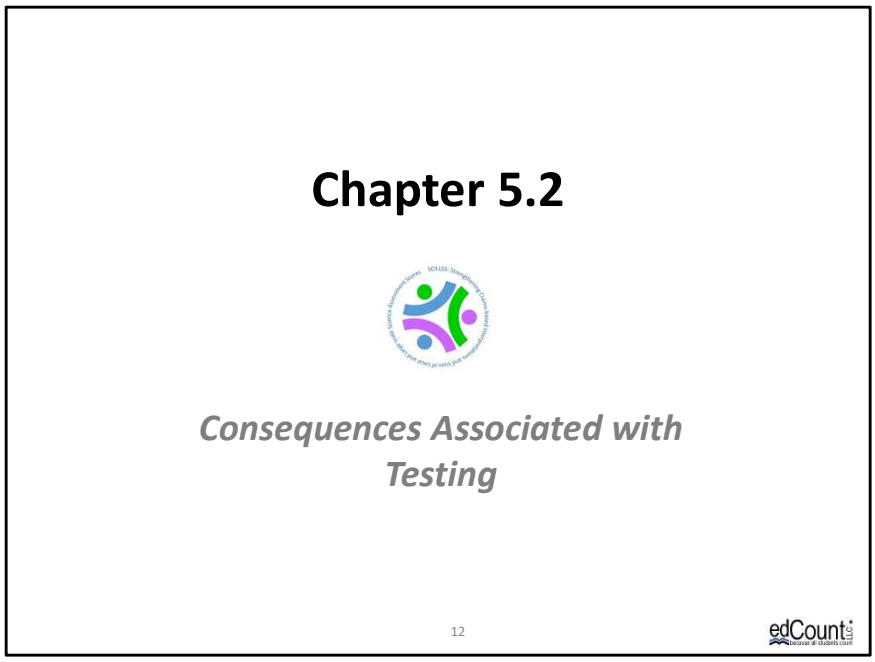
Chapter 4 focused on the third set of these questions, which relate to fairness and accessibility. These questions are:

1. How were the needs of all students addressed during assessment development? How were the assessment questions developed to ensure that scores reflect the intended measurement targets and not student characteristics or contexts that are irrelevant to the measurement targets?
2. How were the needs of students with disabilities addressed during assessment development?
3. How were the needs of English learners addressed during assessment development?
4. How are students with disabilities able to demonstrate their knowledge and skills through the availability and use of any necessary accommodations? What evidence supports the selection of accommodations as well as their use of these accommodations at the time of testing?
5. How are English learners able to demonstrate their knowledge and skills through the availability and use of any necessary accommodations? What evidence supports the selection of accommodations as well as their use of these accommodations at the time of testing?
6. How are students' responses scored in ways that reflect only the construct-

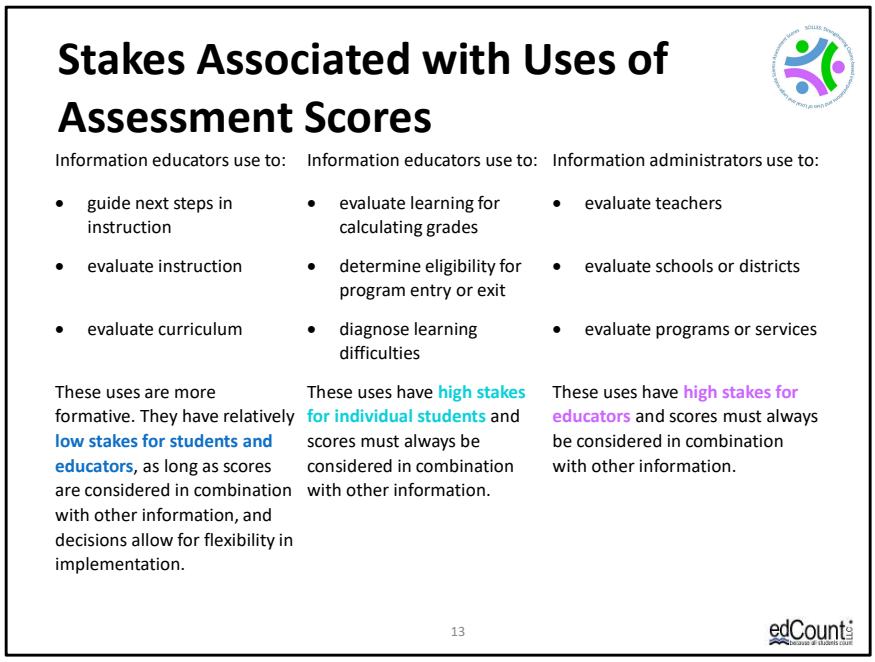
relevant aspects of those responses? What evidence supports the minimization of construct-irrelevant influences on students' responses?

7. What evidence supports the interpretation and use of students' scores in relation to their learning opportunities?

Now, in this chapter, we turn our attention to the fourth set of validity questions, which relate to the notion of consequences.



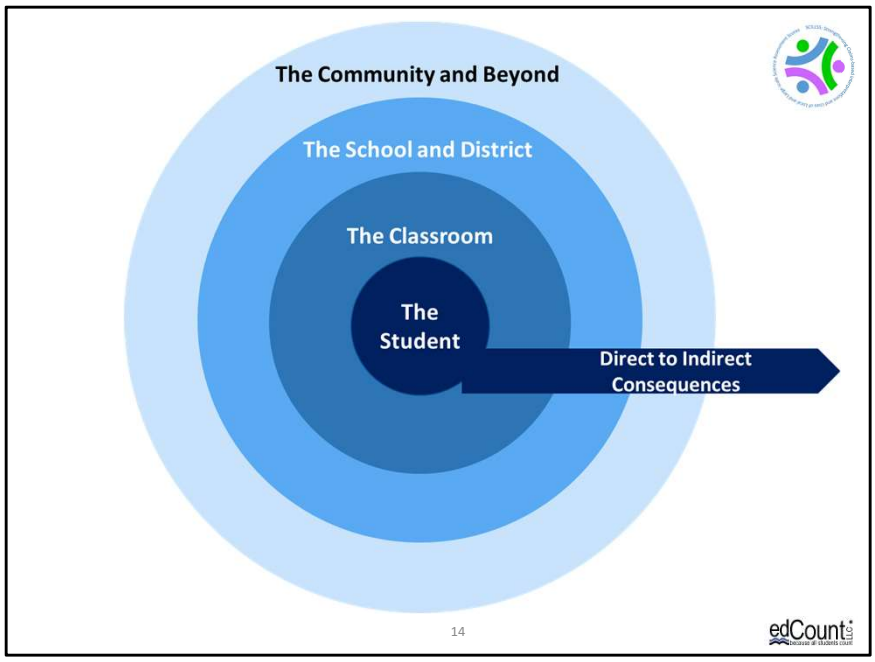
Chapter 5.2: Consequences Associated with Testing



All tests have consequences, whether for the students who take them or for other students or systems or policies. In this chapter, we will focus on types of consequences and how to evaluate evidence of the consequences of testing.

In education, we give tests for many reasons, such as to gauge what students have learned or need to learn or how well programs are working. Decisions associated with these kinds of reasons for testing can be about what to teach next or re-teach or what curriculum resources to invest in or who gets selected for a particular program or service. These are all consequences related to testing.

In prior chapters, we also learned that each purpose for giving a test and each decision we make with test scores is associated with stakes, some higher and some lower. Low stakes means that the decisions are not permanent or harmful to the student and can be modified easily and quickly as more information becomes available. High stakes can mean significant long-term impact that is not easily modified. Under no circumstances should high stakes decisions be made solely on the basis of a test score.

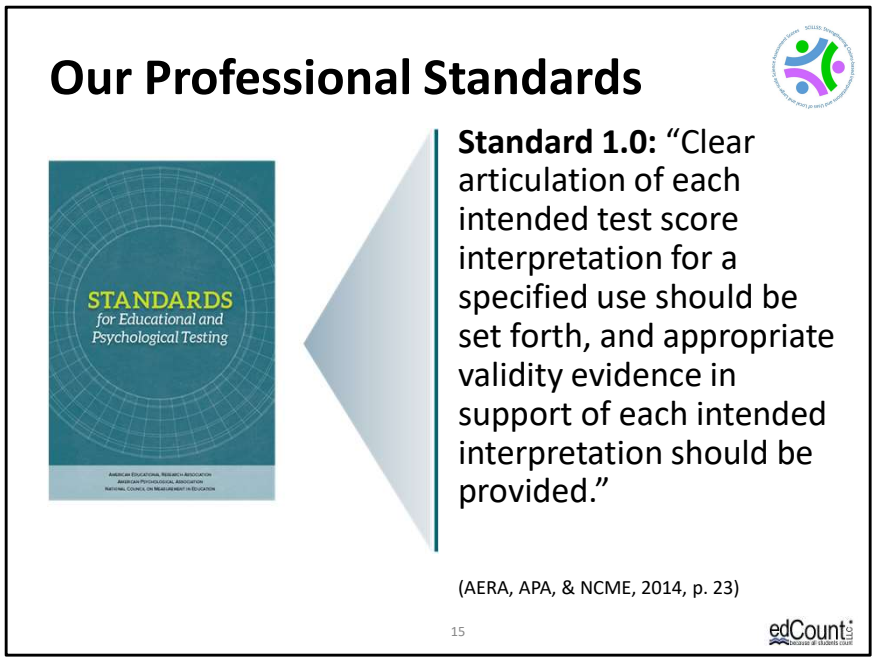


In addition to varying in terms of stakes, consequences vary in terms of their direct impact on or proximity to the students who take the test. Some consequences are very proximal to the student, such as when scores are used to guide instruction, for grading, and for making decisions about promotion, selection, or admission. These score interpretations are based on the information a test score provides about what a student knows and can do and these uses relate a student's future experiences, from the instruction a teacher delivers after reviewing testing information to post-secondary opportunities, to his or her performance on the test. Of course, we must remind ourselves of the caveat against basing decisions solely on the scores from a test.

Other consequences are more distal and indirect. These reflect the use of scores to help make decisions about curriculum revision or program adoption, teacher or school effectiveness, or other policy issues. These uses of scores as policy levers may ultimately affect what a student experiences or affect what other students experience, but generally not immediately.

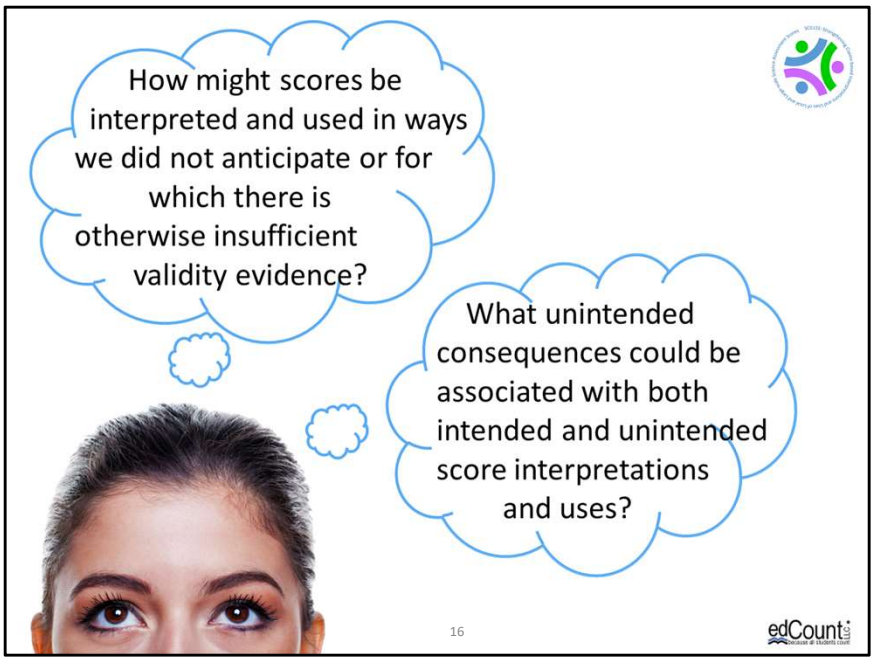
Direct, proximate consequences and indirect, distal consequences may be intentional, meaning that they reflect the score interpretations and uses that the test was

designed to support. They may also be unintentional. Some intentional and unintentional consequences may be positive and some may be negative.

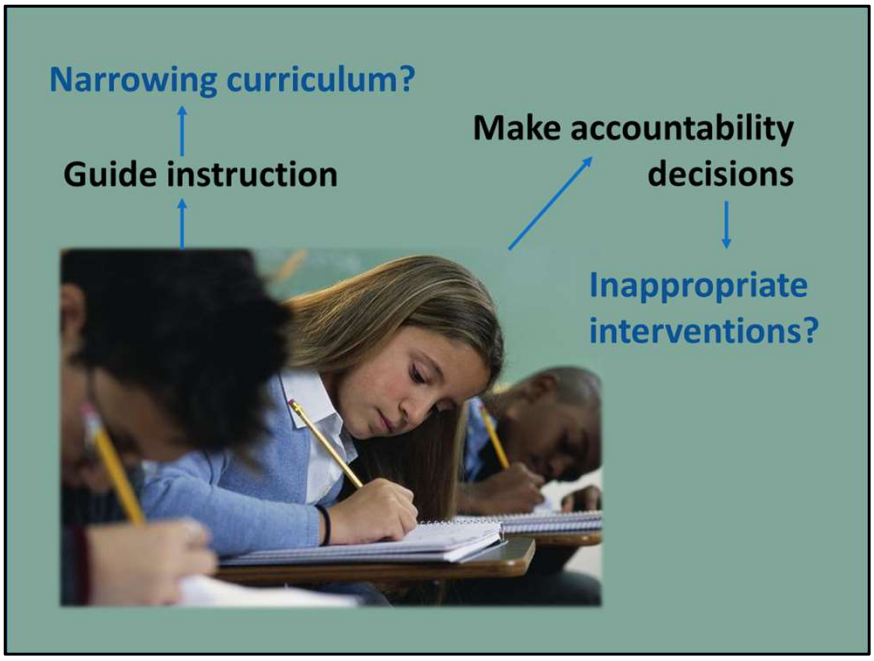


Sometimes scores from the same test are used for more than one purpose or more than one type of purpose. As we have learned in the other four chapters in this series, those using test scores are obligated to gather and evaluate validity evidence for each intended test score interpretation and use. This concept is so important in educational testing that it is addressed in the very first standard in the *Standards for Educational and Psychological Testing*, the book that defines expectations for quality and rigor in assessment practices.

Standard 1.0: Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided. (p. 23)
AERA, APA, NCME, 2014



In terms of consequences, this means that those who use test scores must consider consequences associated with intended interpretations and uses as well as consequences associated with interpretations and uses that may be unintended. That requires not only evidence related to intended interpretations and uses, but also asking questions such as, “how might scores be interpreted and used in ways we did not anticipate or for which there is otherwise insufficient validity evidence?” and “what unintended consequences could be associated with both intended and unintended score interpretations and uses?”



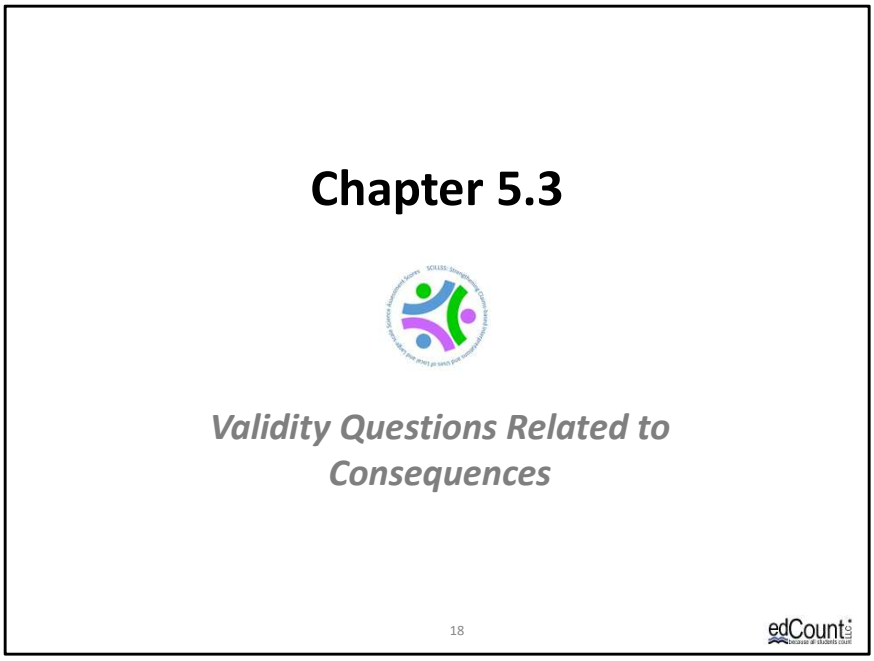
Consider a test that was designed and developed to yield scores for teachers to use in guiding instruction. The body of validity evidence for that test must include evidence to support any guidance that the test publisher provides linking students' scores to ensuing instruction, including evidence that the recommended instruction is effective. The intended, positive consequences of testing here would be for a student to get the instruction he or she needs. It would be possible, however, for a student to experience unintended, negative consequences. Even if the scores accurately reflected his or her knowledge and skills, she could not get the appropriate instruction because the guidance was wrong or the instruction was not delivered well or the delivery of that instruction meant the student missed out on other important opportunities.

On a larger scale, consider a test that is administered to all students in a district or state and yields scores used for school accountability. That is, the scores are interpreted as reflecting the effectiveness of schools in supporting student achievement and used to identify schools for awards and for various interventions. In addition to the validity evidence necessary to support such intended interpretations and uses of scores, those using the scores in this way are obligated to consider unintended consequences, particularly negative ones. These can include the

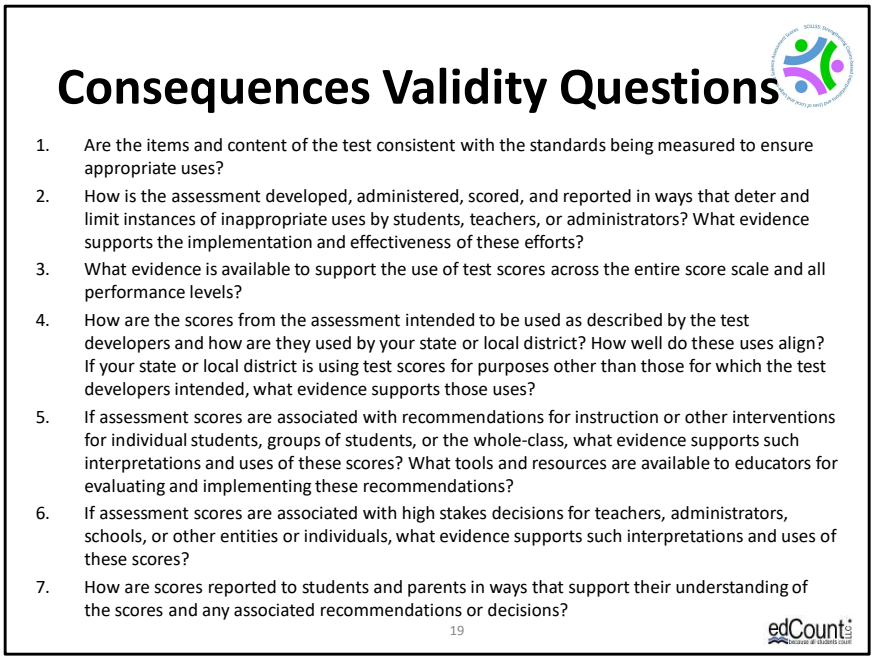
narrowing of curriculum to focus mostly or only on what is known to be on the test to the exclusion of instruction in other areas, so students miss out on the full range of learning opportunities. (Stecher, 2002)

Further, the uses of scores in this intended way could incentivize inappropriate decisions about instruction for individual students. For example, administrators could identify students who scored just below the level considered to be “on grade” or “proficient” and demand that teachers work intensively with these students. This is inappropriate because such instructional decisions are not based on what individual students actually need; the targeted students as well and the non-targeted students may miss the learning opportunities that are actually necessary and effective for them. (Booher-Jennings, 2005)

These issues will play out as we work through the validity questions in this chapter.



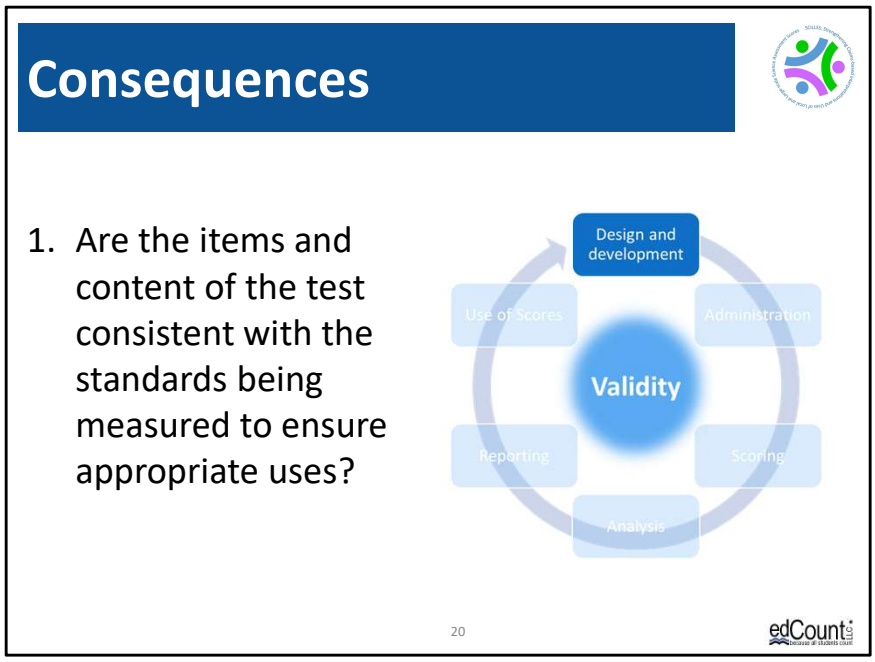
Chapter 5.3: Validity Questions Related to Consequences



The seven validity questions we will consider in this chapter are:

1. Are the items and content of the test consistent with the standards being measured to ensure appropriate uses?
2. How is the assessment developed, administered, scored, and reported in ways that deter and limit instances of inappropriate uses by students, teachers, or administrators? What evidence supports the implementation and effectiveness of these efforts?
3. What evidence is available to support the use of test scores across the entire score scale and all performance levels?
4. How are the scores from the assessment intended to be used as described by the test developers and how are they used by your state or local district? How well do these uses align? If your state or local district is using test scores for purposes other than those for which the test developers intended, what evidence supports those uses?
5. If assessment scores are associated with recommendations for instruction or other interventions for individual students, groups of students, or the whole-class, what evidence supports such interpretations and uses of these scores? What tools and resources are available to educators for evaluating and implementing these recommendations?

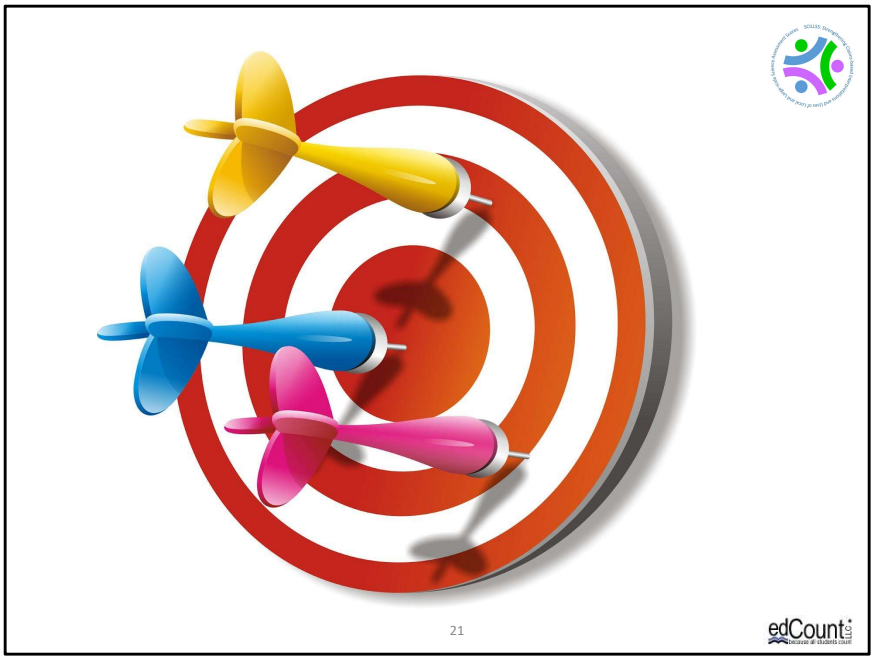
6. If assessment scores are associated with high stakes decisions for teachers, administrators, schools, or other entities or individuals, what evidence supports such interpretations and uses of these scores?
7. How are scores reported to students and parents in ways that support their understanding of the scores and any associated recommendations or decisions?



Now let's consider the validity evidence related to our first question in this chapter, which is:

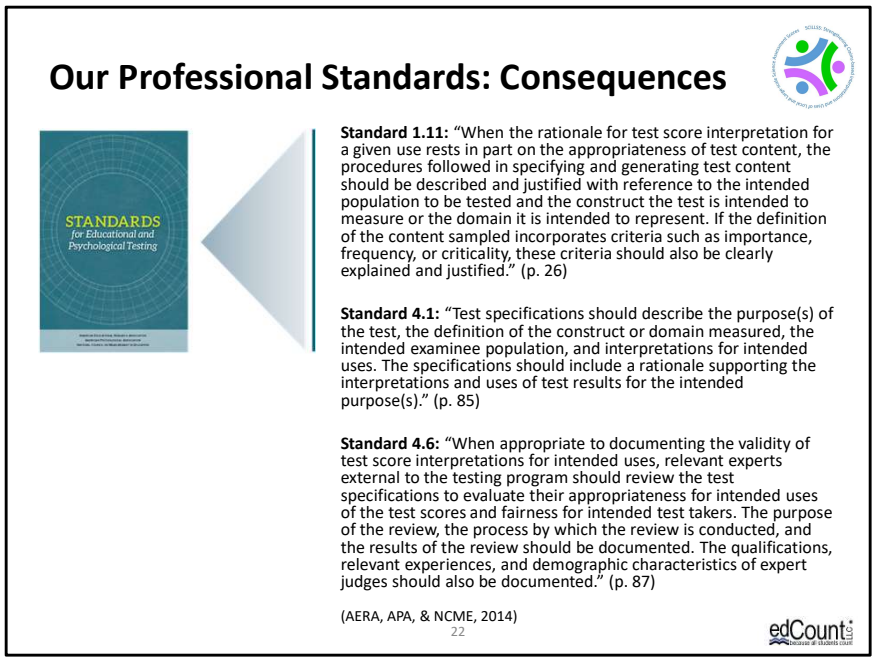
1. Are the items and content of the test consistent with the standards being measured to ensure appropriate uses?

Evidence related to this question comes from the Design and Development phase of the assessment life cycle.



This first question represents a sort of “do not pass go” concept for testing. There is no possibility that the interpretations and uses of test scores can be valid if the test has not been carefully designed and developed to measure what it is intended to measure.

As we learned in the second chapter of this series, which focuses on construct coherence validity questions, careful design and development requires a clear definition of what one is intending to measure and a series of appropriate, well-implemented processes to create a test that aligns with that intent. A test should neither under-represent the intended content or construct nor include content that is beyond or outside of what the test is intended to measure. If it does, the scores cannot be interpreted back to the intended measurement targets and cannot be used for purposes associated with those targets.



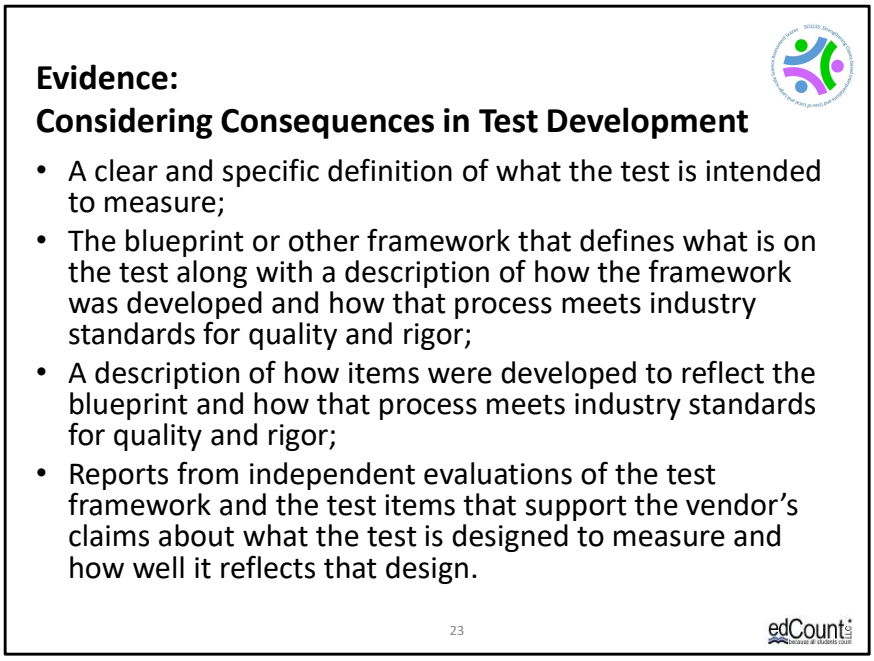
Our professional standards clarify several obligations relevant to this first question. For example:

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. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified. (p. 26)

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

Standard 4.6: When appropriate to documenting the validity of test score interpretations for intended uses, relevant experts external to the testing program

should review the test specifications to evaluate their appropriateness for intended uses of the test scores and fairness for intended test takers. The purpose of the review, the process by which the review is conducted, and the results of the review should be documented. The qualifications, relevant experiences, and demographic characteristics of expert judges should also be documented. (p. 87)



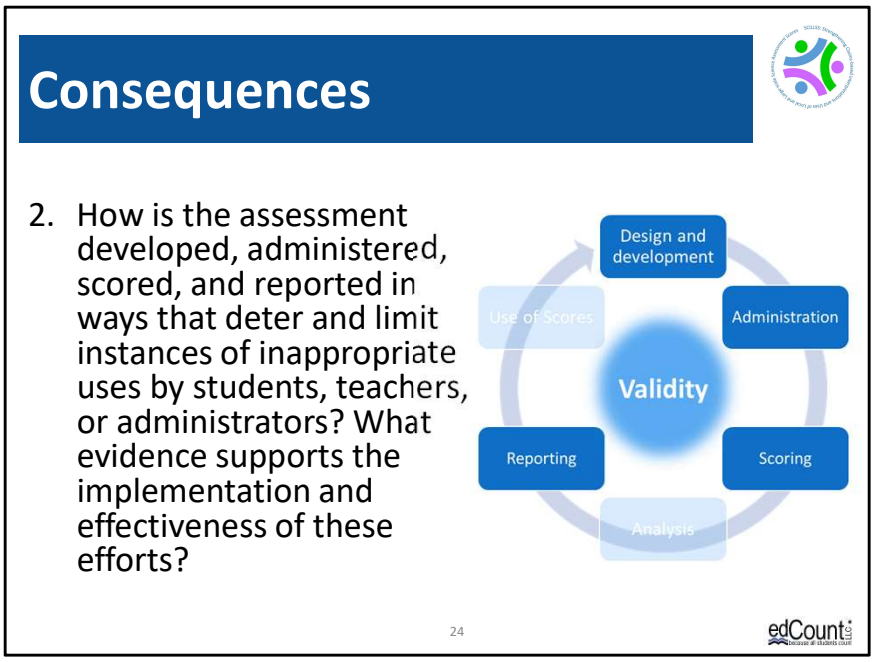
Evidence for this validity question should be supplied by the testing vendor or publisher and should include:

A clear and specific definition of what the test is intended to measure;

The blueprint or other framework that defines what is on the test along with a description of how the framework was developed and how that process meets industry standards for quality and rigor;

A description of how items were developed to reflect the blueprint and how that process meets industry standards for quality and rigor;

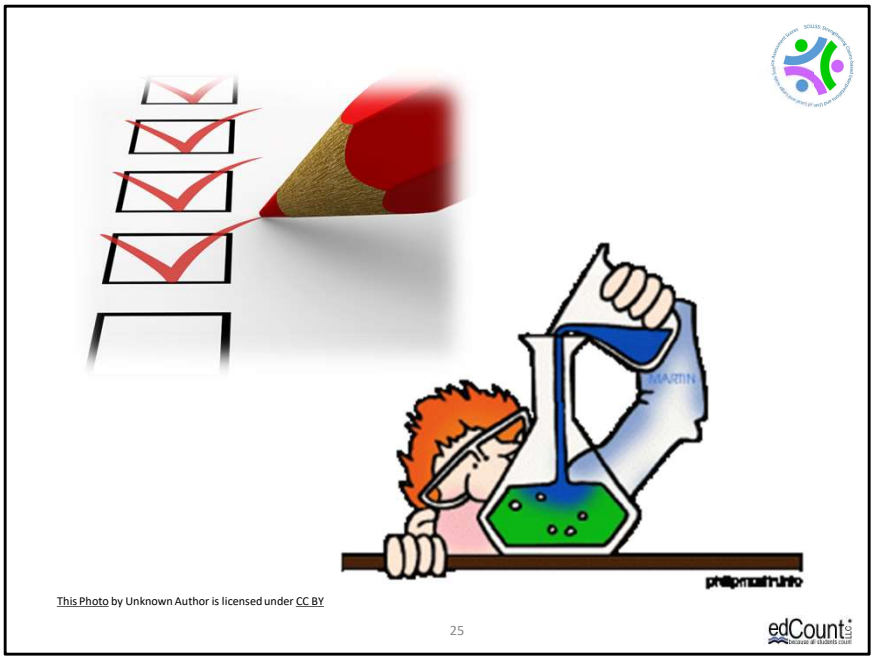
Reports from independent evaluations of the test framework and the test items that support the vendor's claims about what the test is designed to measure and how well it reflects that design. Such evaluations are independent only if carried out by companies or individuals who were not involved in test development and are not employed by the test users. Some use the term "third-party" to refer to these independent parties.



Our second validity question in this chapter is:

2. How is the assessment developed, administered, scored, and reported in ways that deter and limit instances of inappropriate uses by students, teachers, or administrators? What evidence supports the implementation and effectiveness of these efforts?

Evidence for this question comes from the Design and Development, Administration, Scoring, and Reporting phases of the testing life cycle.

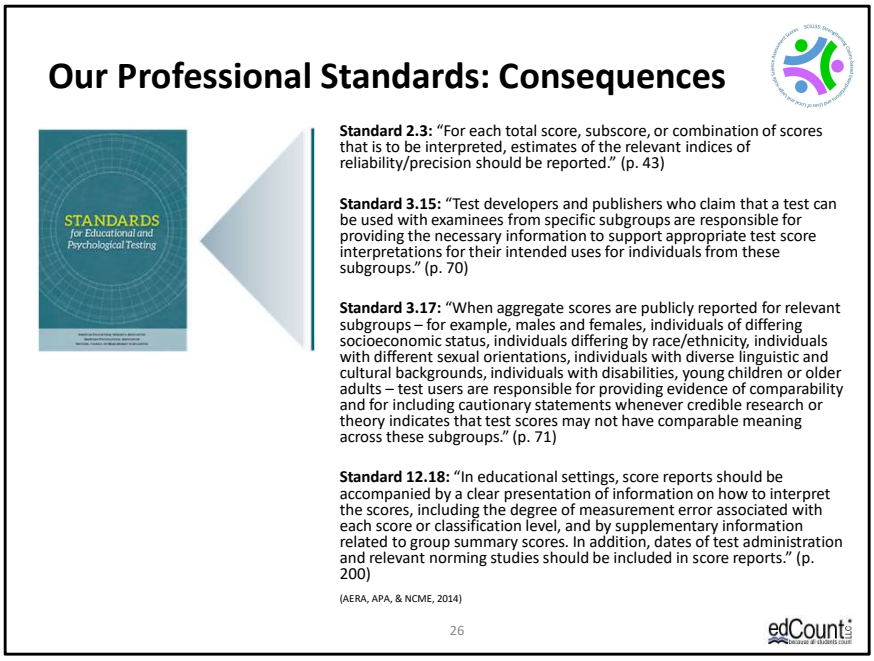


This question covers a lot of territory. Following from our first question, which addresses evidence related to how a test is designed and developed to measure the intended content or construct, this question is meant to focus on evidence that test scores do not reflect student characteristics or contexts or conditions that are irrelevant to what the test is meant to measure. Further, this question encompasses the form in which scores are reported and the information that accompanies test scores.

As we learned in our fourth chapter of this series, which focused on fairness and accessibility issues in testing, a test must be designed and developed carefully so that all students can legitimately interact with the questions and provide responses that reflect their knowledge and skills. This can be done by using the principles of Universal Design and involving individuals with expertise in the needs of students with disabilities, English learners, and students with diverse racial and cultural backgrounds in the design and development processes for the test items and for the materials and procedures for test administration and scoring.

During test administration, teachers, administrators, and others serving as proctors, must follow the test administration guidelines the publisher provides and remove any

conditions that may be obstacles for students during testing. This includes creating a testing environment that is clean, well-lit, and free from visual or auditory distractions. It includes ensuring that students have the space, surfaces, and tools they need before testing starts and that the decisions that IEP and 504 teams and the teams who make decisions for how English learners participate in testing, are honored.



When scores are reported, they must be accompanied by information that guides appropriate interpretations of the scores for specific uses and also strongly discourages interpretations and uses for which there is no validity evidence. Our standards speak to these obligations several times. For example:

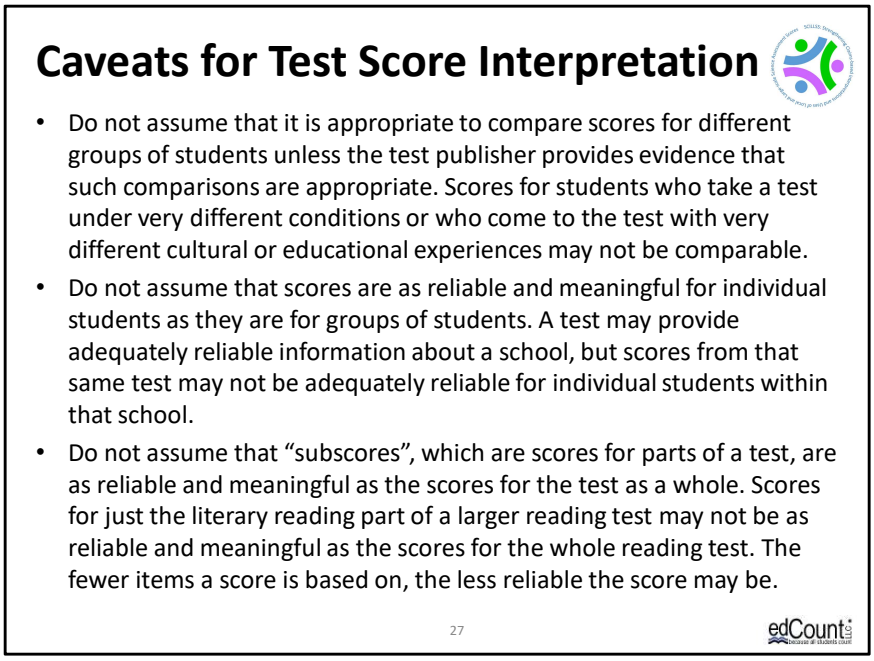
Standard 2.3: For each total score, subscore, or combination of scores that is to be interpreted, estimates of the relevant indices of reliability/precision should be reported. (p. 43)

Standard 3.15: Test developers and publishers who claim that a test can be used with examinees from specific subgroups are responsible for providing the necessary information to support appropriate test score interpretations for their intended uses for individuals from these subgroups. (p. 70)

Standard 3.17: When aggregate scores are publicly reported for relevant subgroups – for example, males and females, individuals of differing socioeconomic status, individuals differing by race/ethnicity, individuals with different sexual orientations, individuals with diverse linguistic and cultural backgrounds, individuals with disabilities, young children or older adults – test users are responsible for providing

evidence of comparability and for including cautionary statements whenever credible research or theory indicates that test scores may not have comparable meaning across these subgroups. (p. 71)

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 by 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)

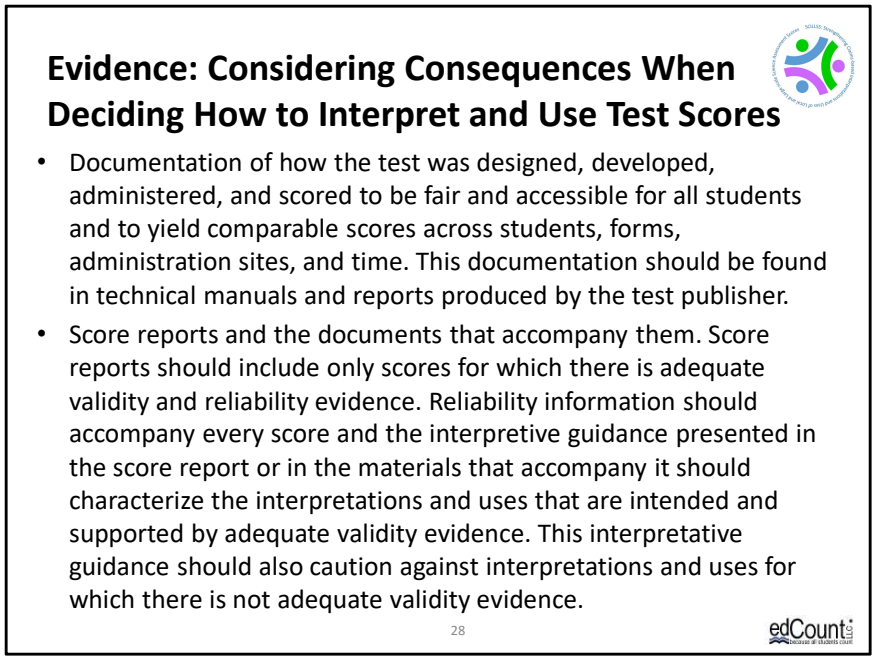


A number of caveats for test score interpretation are embedded within these professional standards, such as:

Do not assume that it is appropriate to compare scores for different groups of students unless the test publisher provides evidence that such comparisons are appropriate. Scores for students who take a test under very different conditions or who come to the test with very different cultural or educational experiences may not be comparable.

Do not assume that scores are as reliable and meaningful for individual students as they are for groups of students. A test may provide adequately reliable information about a school, but scores from that same test may not be adequately reliable for individual students within that school.

Similarly, do not assume that "subscores", which are scores for parts of a test, are as reliable and meaningful as the scores for the test as a whole. Scores for just the literary reading part of a larger reading test may not be as reliable and meaningful as the scores for the whole reading test. The fewer items a score is based on, the less reliable the score may be.



These caveats about score interpretations lead to cautions about score uses. As we have said several times throughout the five chapters in this series, no decision for a student or a group of students should be based solely on a single test score.

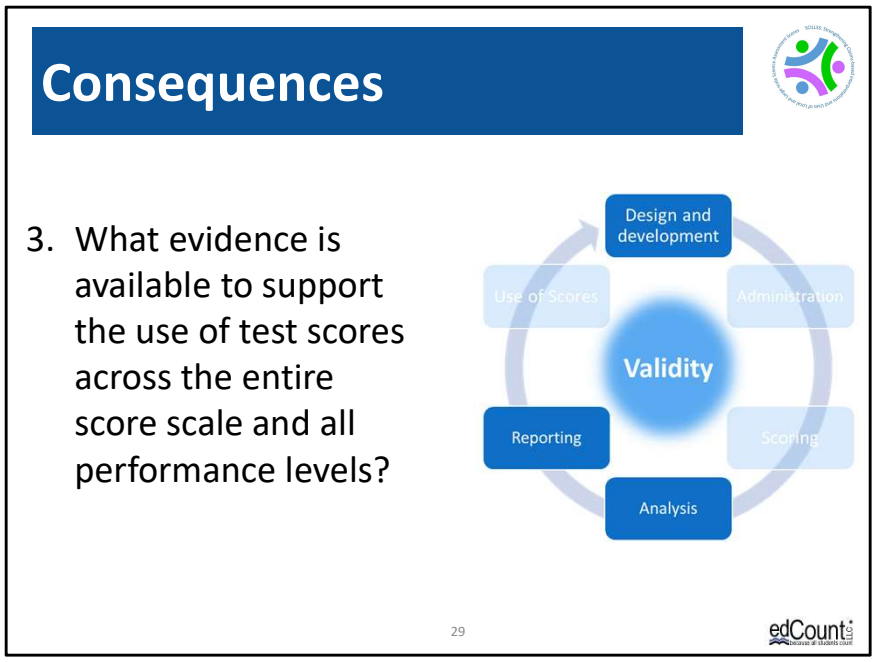
Evidence for this question would include:

Documentation of how the test was designed, developed, administered, and scored to be fair and accessible for all students and to yield comparable scores across students, forms, administration sites, and time. This documentation should be found in technical manuals and reports produced by the test publisher. Technical reports include summary information about scores, but they are distinct from the score reports that are distributed to students, parents, teachers, and schools.

Score reports and the documents that accompany them. Score reports should include only scores for which there is adequate validity and reliability evidence. Reliability information should accompany every score and the interpretive guidance presented in the score report or in the materials that accompany it should characterize the interpretations and uses that are intended and supported by adequate validity evidence and also cautions against interpretations and uses for which there is not

adequate validity evidence.

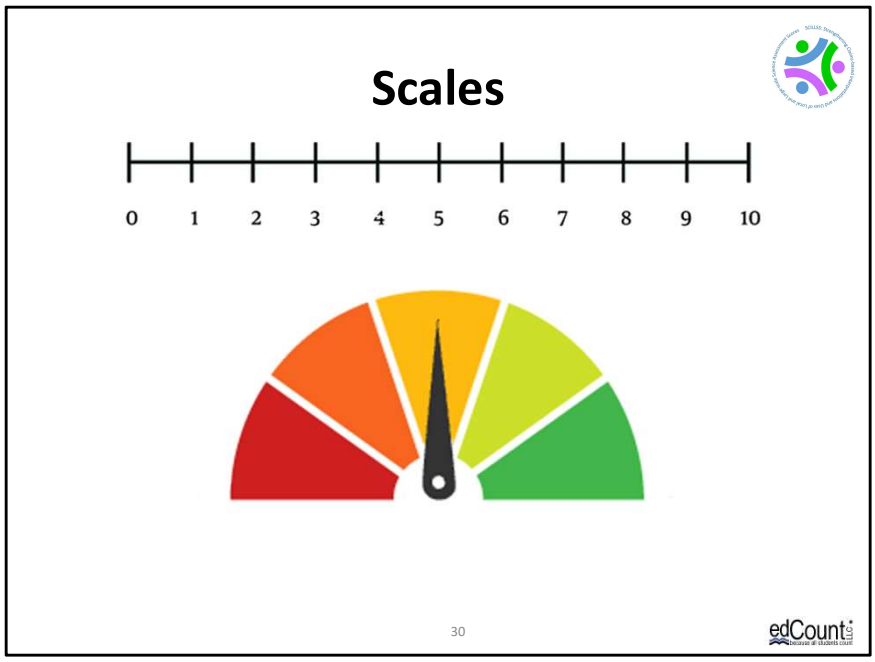
In addition, test users should always ask themselves whether they are interpreting and using test scores appropriately. It is wise to take a step back and reflect on whether one is reading too much into test scores or applying them to decisions that are unwarranted.



Our third question in this chapter is:

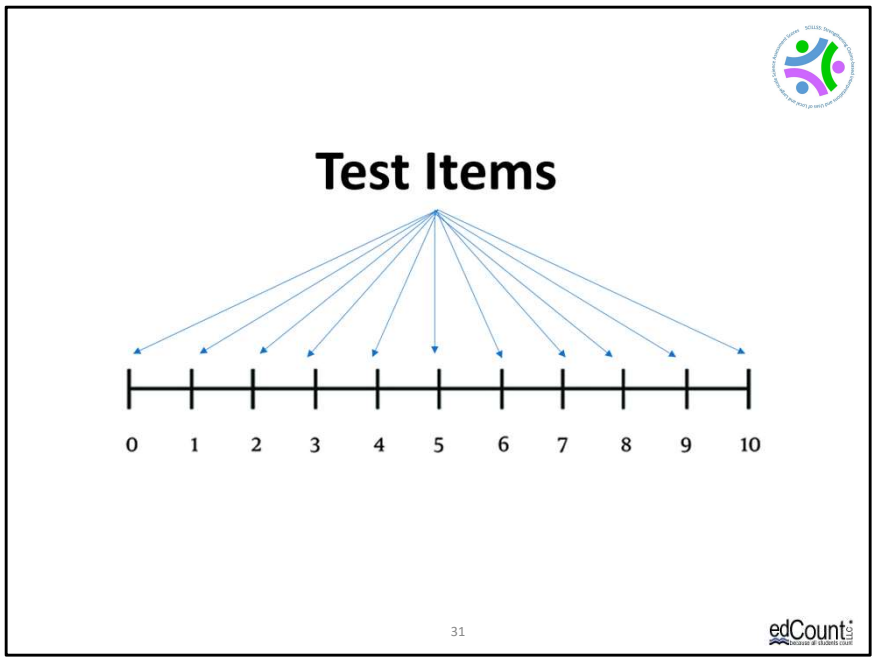
3. What evidence is available to support the use of test scores across the entire score scale and all performance levels?

Evidence related to this question can be found in the design and development, analysis, and reporting phases of the testing life cycle.

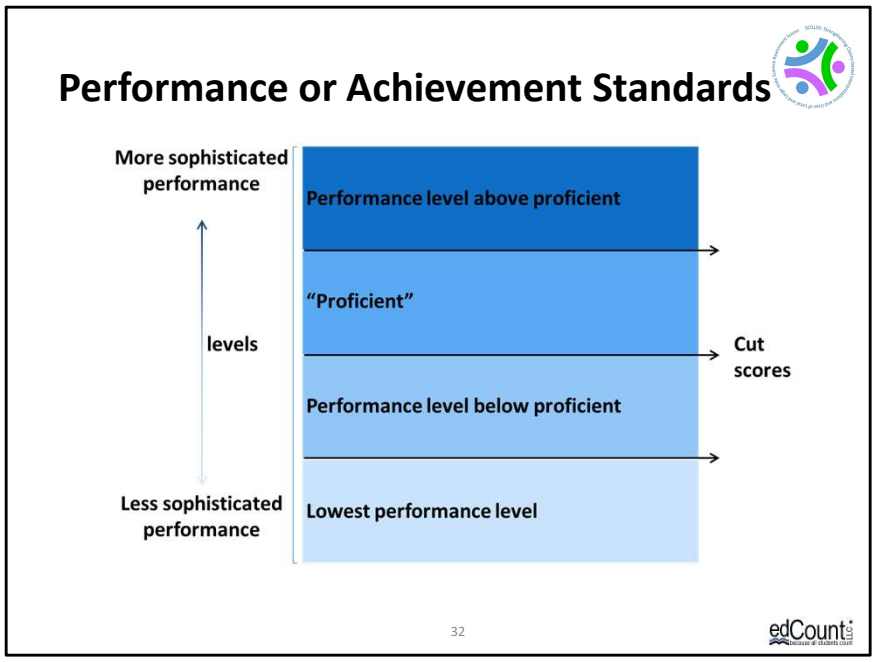


Test scores are points or ranges on score scales. Even when the scores are “raw scores”, meaning that that they are simply the number or percent correct and haven’t been scaled psychometrically, part of their meaning comes from where they are on the scale.

While technical aspects of score scales and how they are created is beyond the scope of this chapter series, it is important to understand that psychometricians create score scales to account for characteristics of test items. Items vary in characteristics such as difficulty and this has implications for how items contribute to scores for any given form and for comparisons of scores across forms.

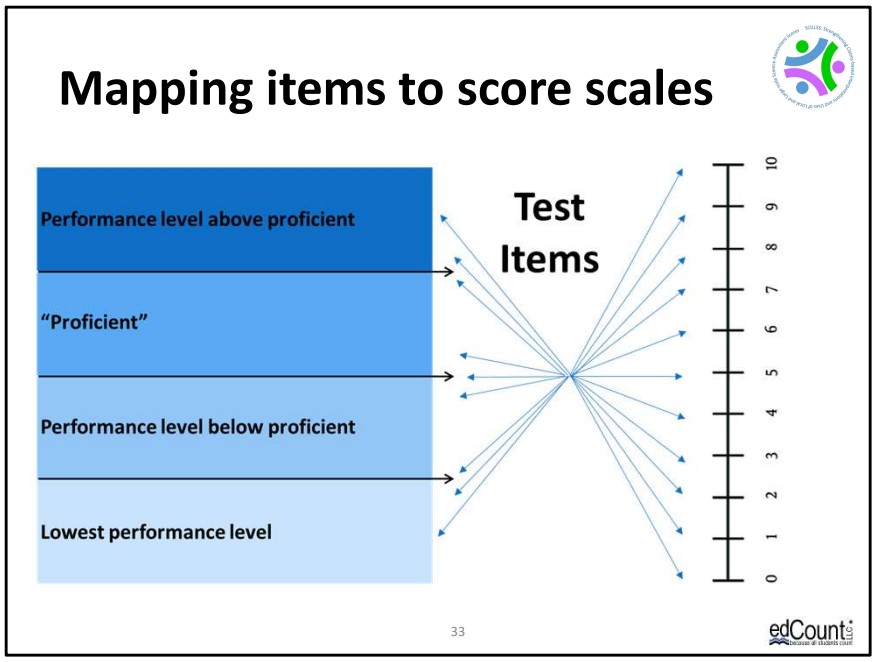


When designing a test, as part of defining what the test is meant to measure, one must determine the range of information the test is meant to provide. Consider a test that is meant to yield information about where students in a school are on a particular learning progression so that scores can be used to guide instructional next steps. That test would have to include a sufficient number of items for all parts of that learning progression so that students' scores can be trustworthy whether they are at the lower end of the scale, in the middle, or at the high end of the scale.



Another test might be intended for accountability uses where the key decisions relate to the part of the scale associated with what is considered "proficient." That test would still need to have items along the entire score scale, but would need a particular focus on the part of the scale that is just below, right at, and just above that proficient cut point, which we call a cut score.

Recall that some tests, such as this accountability test, report scores using performance or achievement levels. This means that the score scale has been divided into two or more ranges. Using a process called standard setting, a test developer or test user identifies the cut scores, which are the scores that separate the score ranges.

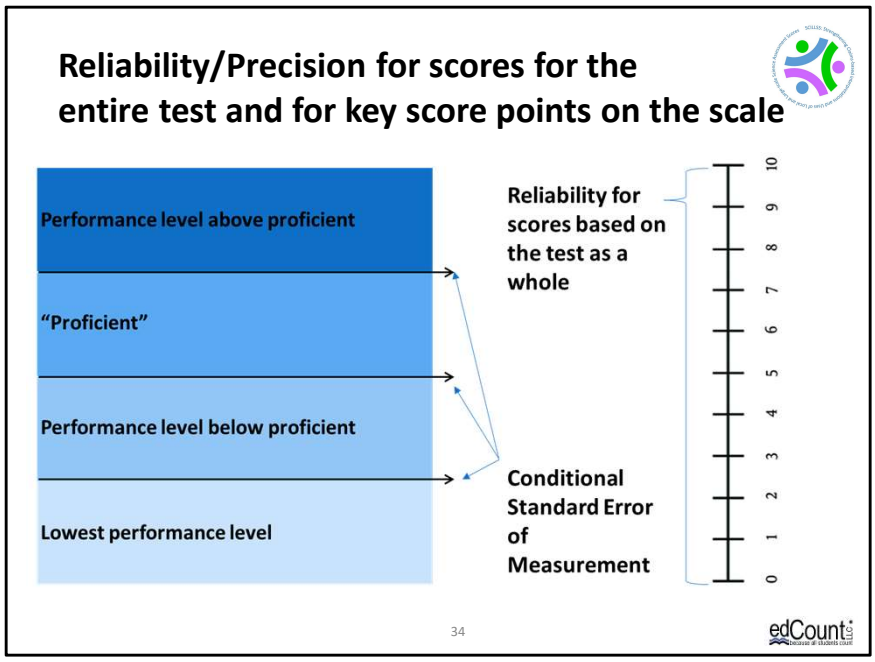


We can see that different intended uses of test scores, as in these two examples, translate into different test and item development strategies. A test should be developed to maximize the quality of information at the points on the score scale that are associated with intended interpretations and uses. Items should be mapped to the score scale as they are developed, meaning that the item writer is not simply writing an item to reflect an objective, but also to reflect performance at a point or in a range on the score scale. Is the item meant to reflect a high degree of sophistication or a relatively low level of sophistication? Similarly, if it is a performance-based item that is scored using a rubric, how does a response that earns a single point differ from one that earns all five possible points?

If scores are to be reported using performance or achievement levels, test developers must establish performance level descriptors (PLDs) or achievement level descriptors (ALDs). Note that the use of "performance" or "achievement" is a local decision and the terms are interchangeable.

These descriptors define the characteristics of student performance associated with each range on the score scale. Performance in the lowest level on the scale is the least sophisticated that the test is intended to measure and performance in the

highest level of the score scale is the most sophisticated that the test is intended to measure.



In addition to considerations during development, one must evaluate the quality of measurement for the score scale using student responses to the items. Test publishers or others who score student responses and report scores must evaluate not only overall test reliability, but also reliability at the points on the score scale that are associated with specific interpretations and uses. Overall reliability, reported in the form of precision, a reliability coefficient, or standard error of measurement, is always mandatory.

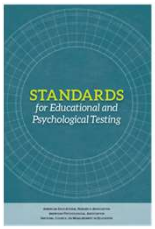
If a test yields scores that teachers are supposed to use in making instructional decisions, then each score on the scale, from the lowest to the highest, must have sufficient reliability evidence as well. For accountability tests, reliability evidence must be reported for each cut score. Note that if a teacher or a school or a district decides on their own to designate a particular score as a cut score for any type of decision, they are responsible for establishing evidence regarding the reliability of that score as well as validity evidence to support that score interpretation and use.

Psychometricians calculate the standard error of measurement for a test as a whole and calculate the conditional standard error of measurement for specific score points on a test. In addition, when scores are used to classify scores into ranges as is done

for performance levels, psychometricians must calculate indicators of decision consistency and accuracy.

As we learned in our third chapter in this series, reliability information is population specific. As is the case for validity, a test is neither reliable or unreliable. Scores are reliable or less so. The scores for a test as administered in a specific instance for a specific population may be reliable, and evidence of reliability across administrations supports a claim that the test can yield reliable information. But a test cannot be called reliable in the absence of evidence based on a set of scores.

Our Professional Standards: Consequences



Standard error of measurement: the standard deviation of an individual's observed scores from repeated administrations or parallel forms of a test under identical conditions.

Decision consistency: The extent to which decisions based on the scores from one measure are the same as those made with scores from a comparable measure, or the same measure on another occasion.

Standard 2.13: "The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score". (p. 45)

Standard 2.14: "When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurements should be reported in the vicinity of each cut score." (p. 46)

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

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)

35

edCount

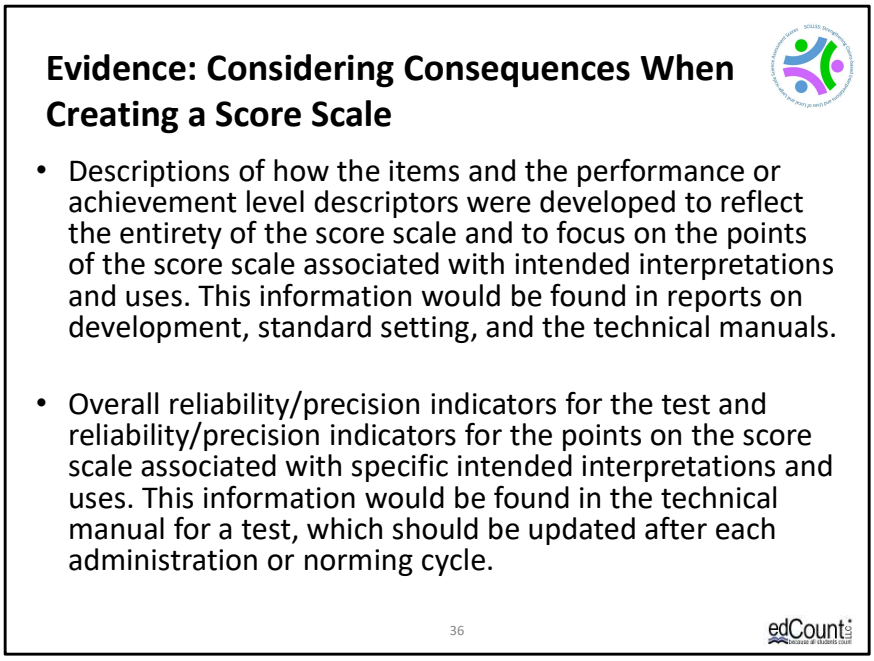
Our professional standards include several statements about these obligations to build and evaluate the quality of measurement across the score scale. For example:

Standard 2.13: The standard error of measurement, both overall and conditional (if reported), should be provided in units of each reported score. (p. 45)

Standard 2.14: When possible and appropriate, conditional standard errors of measurement should be reported at several score levels unless there is evidence that the standard error is constant across score levels. Where cut scores are specified for selection or classification, the standard errors of measurements should be reported in the vicinity of each cut score. (p. 46)

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

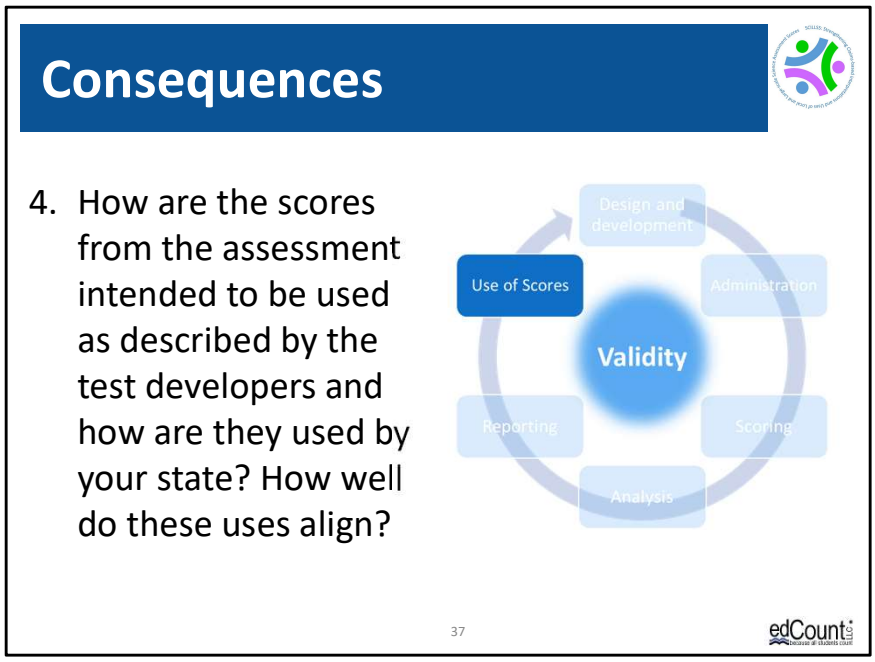
Standard 5.21: When proposed score interpretations involve one or more cut scores, the rationale and procedures used for establishing cut scores should be documented clearly. (p. 107)



Test users must have evidence regarding the quality of measurement at the points of the score scale associated with intended interpretations and uses of the test scores. This evidence should include:

Descriptions of how the items and the performance or achievement level descriptors were developed to reflect the entirety of the score scale and to focus on the points of the score scale associated with intended interpretations and uses. This information would be found in reports on development or standard setting, but are usually found in the technical manuals.

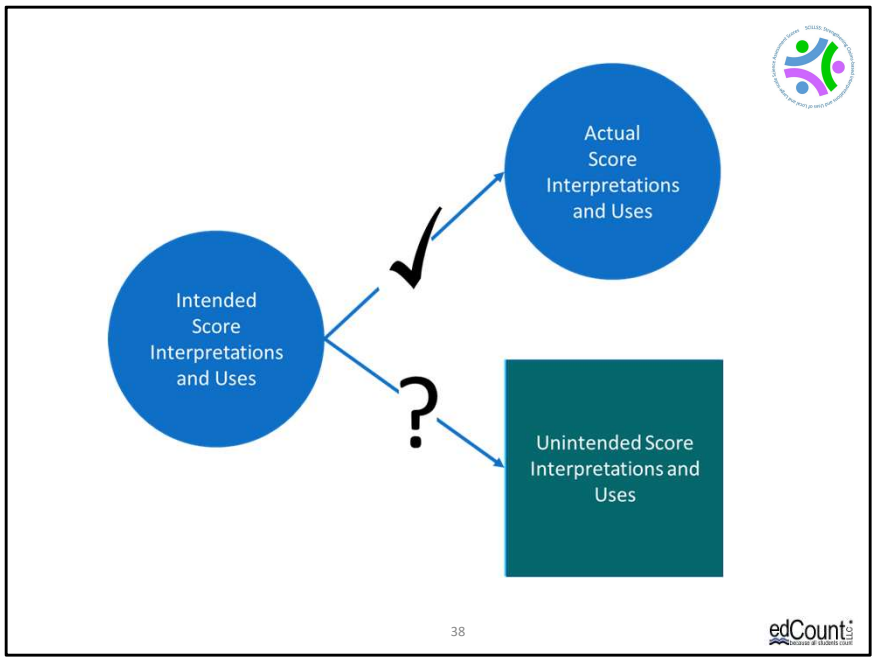
Overall reliability/precision indicators for the test and reliability/precision indicators for the points on the score scale associated with specific intended interpretations and uses. This information would be found in the technical manual for a test, which should be updated after each administration or norming cycle.



Our fourth question in this chapter is:

4. How are the scores from the assessment intended to be used as described by the test developers and how are they used by your state? How well do these uses align?

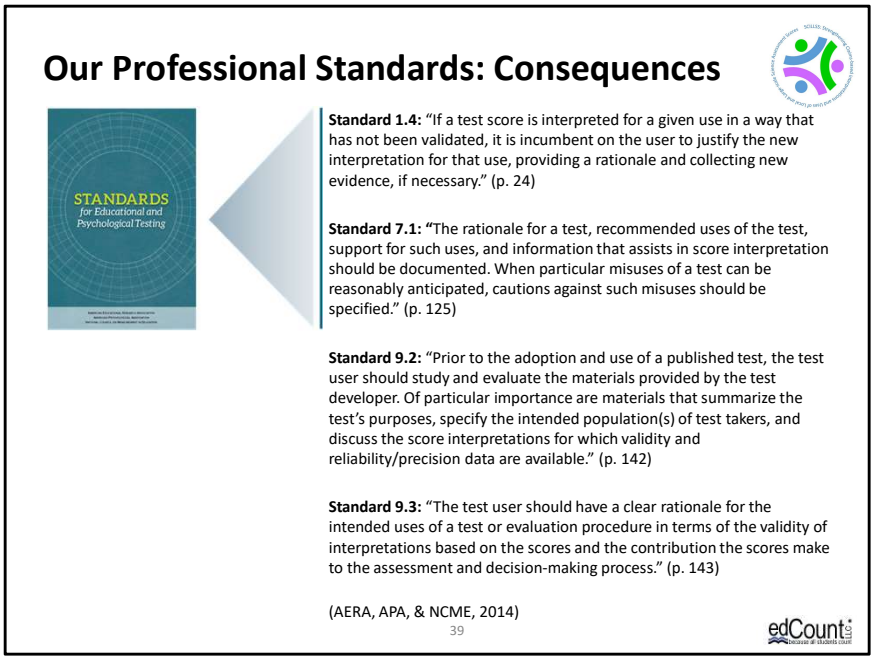
Evidence related to this question can be found in the score use phase of the testing life cycle.



This question addresses the alignment of how the scores from a test are intended to be used and how they are actually used by a variety of individuals and groups, such as administrators, teachers, students, parents, politicians, and other members of the community.

As we've learned, using scores for purposes other than those for which a test was designed can be problematic. Validity depends upon score interpretations and uses.

Those who develop tests, and particularly those who publish or otherwise sell them to users, must be clear about how the test scores are intended to be interpreted and used. While test publishers may not be able to envision all possible interpretations and uses beyond the intended ones, they are obligated to identify likely misinterpretations and misuses, with special attention to ones associated with negative consequences, and caution against those in the materials that accompany the tests and the test score reports.



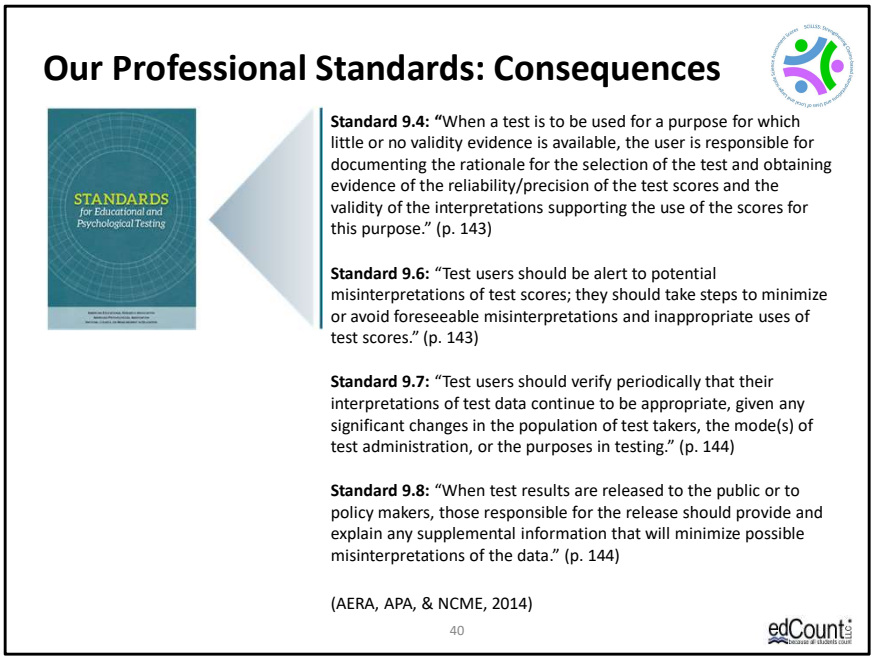
Our professional standards speak to these obligations repeatedly. Some standards relate to the obligations of test publishers, but most target those who use test scores. For example:

Standard 1.4: If a test score is interpreted for a given use in a way that has not been validated, it is incumbent on the user to justify the new interpretation for that use, providing a rationale and collecting new evidence, if necessary (p. 24).

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

Standard 9.2: Prior to the adoption and use of a published test, the test user should study and evaluate the materials provided by the test developer. Of particular importance are materials that summarize the test's purposes, specify the intended population(s) of test takers, and discuss the score interpretations for which validity and reliability/precision data are available. (p. 142)

Standard 9.3: The test user should have a clear rationale for the intended uses of a test or evaluation procedure in terms of the validity of interpretations based on the scores and the contribution the scores make to the assessment and decision-making process. (p. 143)



Standard 9.4: When a test is to be used for a purpose for which little or no validity evidence is available, the user is responsible for documenting the rationale for the selection of the test and obtaining evidence of the reliability/precision of the test scores and the validity of the interpretations supporting the use of the scores for this purpose. (p. 143)

Standard 9.6: Test users should be alert to potential misinterpretations of test scores; they should take steps to minimize or avoid foreseeable misinterpretations and inappropriate uses of test scores. (p. 143)

Standard 9.7: Test users should verify periodically that their interpretations of test data continue to be appropriate, given any significant changes in the population of test takers, the mode(s) of test administration, or the purposes in testing. (p. 144)

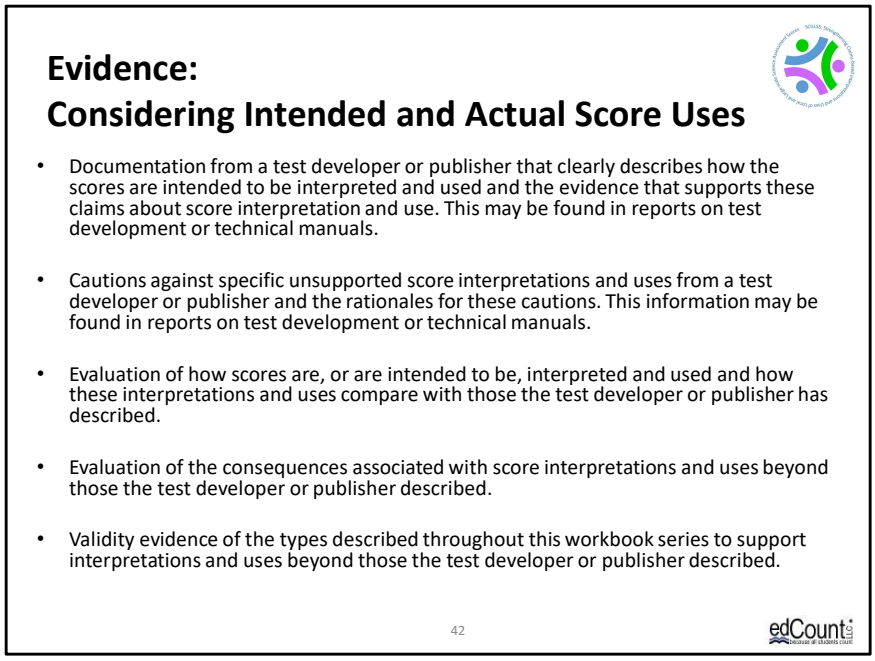
Standard 9.8: When test results are released to the public or to policy makers, those responsible for the release should provide and explain any supplemental information that will minimize possible misinterpretations of the data. (p. 144)

Test scores cannot alone reflect the quality of teaching or learning or the quality of other important components of school quality.

41

edCount

The professional standards are so prolific with regard to the obligations of test score users because misinterpretations and misuses can be so harmful to students, teachers, and other educators. It is not uncommon for individuals or agencies to misinterpret scores and base decisions on those misinterpretations. For example, annual academic test scores for a particular school or district may be used by real estate agents as indicators of school quality in one neighborhood as compared with another for the purpose of guiding home buyers to more expensive areas. This is nearly always an inappropriate interpretation and use of the scores because, among other reasons, they cannot alone reflect the quality of teaching or learning or the quality of other important components of school quality.



To ensure that they are meeting these obligations, those who intend to use test scores should gather and evaluate many forms of evidence. These include:

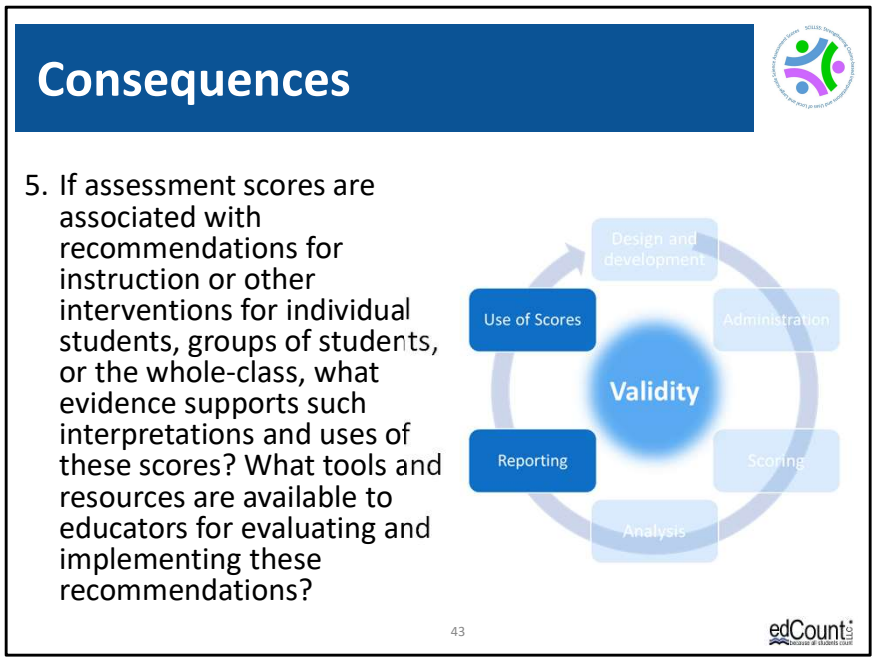
Documentation from a test developer or publisher that clearly describes how the scores are intended to be interpreted and used and the evidence that supports these claims about score interpretation and use. This may be found in reports on test development or technical manuals.

Cautions against specific unsupported score interpretations and uses from a test developer or publisher and the rationales for these cautions. This information may be found in reports on test development or technical manuals.

Evaluation of how scores are, or are intended to be, interpreted and used and how these interpretations and uses compare with those the test developer or publisher has described.

Evaluation of the consequences associated with score interpretations and uses beyond those the test developer or publisher describes.

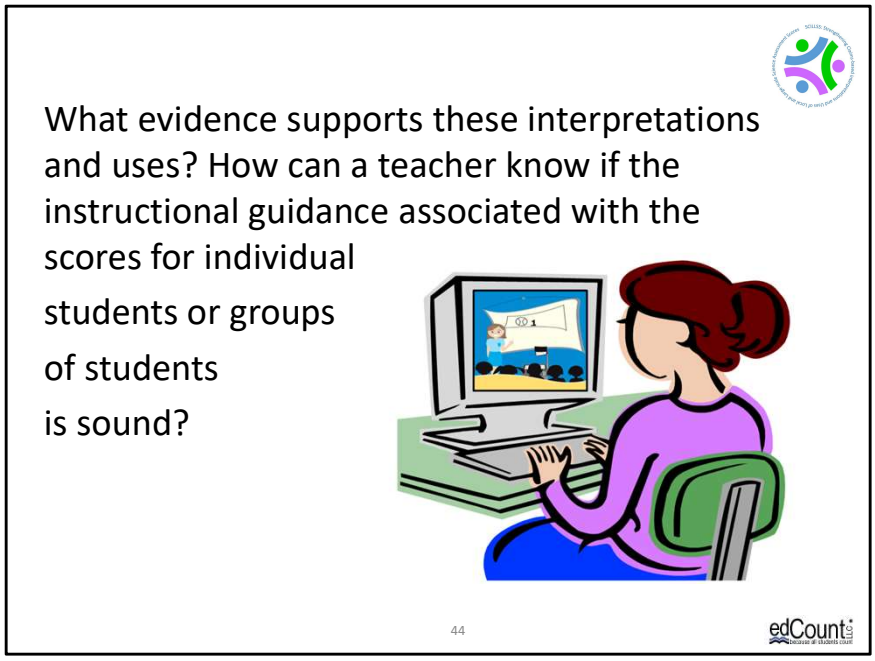
Validity evidence of the types described throughout this workbook series to support interpretations and uses beyond those the test developer or publisher describes.



Our fifth question in this chapter is:

5. If assessment scores are associated with recommendations for instruction or other interventions for individual students, groups of students, or the whole-class, what evidence supports such interpretations and uses of these scores? What tools and resources are available to educators for evaluating and implementing these recommendations?

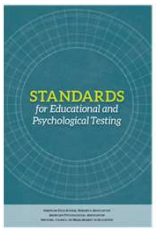
Evidence related to this question comes from the reporting and score use phases of the test life cycle.



Most school districts and even some schools purchase commercial tests to use in classrooms and most students experience many of these tests each school year. Districts and school generally buy these tests because they want information to guide teachers in making instructional decisions and to evaluate students' progress across an academic year. To support these uses, vendors may provide guidance for interpreting their test scores that includes direction for next instructional steps or interventions.

But, what evidence supports these interpretations and uses? How can a teacher know if the instructional guidance associated with the scores for individual students or groups of students is sound?

Our Professional Standards: Consequences



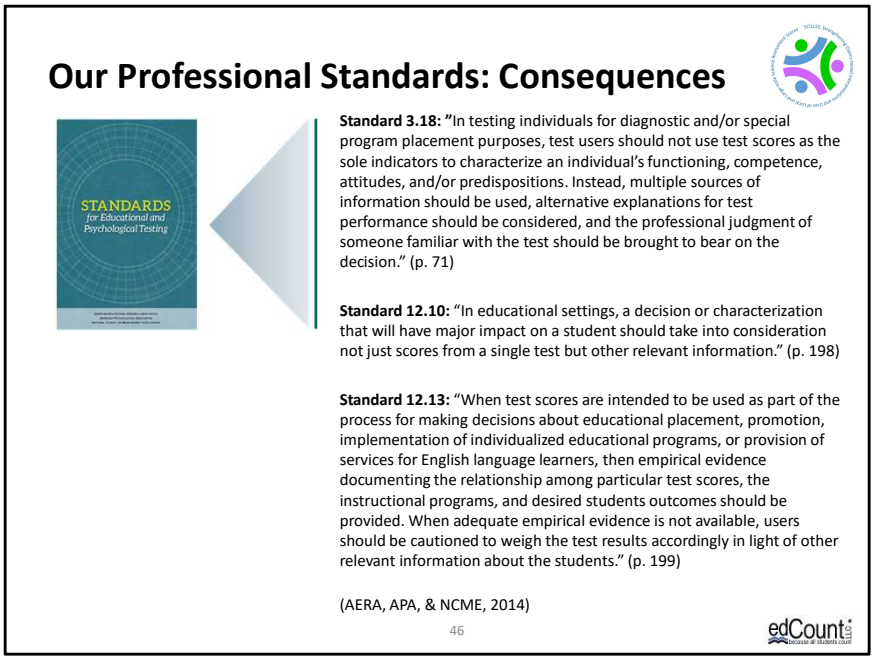
Standard 12.19: "In educational settings, when score reports include recommendations for instructional intervention or are linked to recommended plans or materials for instruction, a rationale for and evidence to support these recommendations should be provided." (p. 201)

(AERA, APA, & NCME, 2014)

45

Recommendations for instruction based on assessment results must be accompanied by logical, theoretical, and empirical evidence to support those recommendations. Obligations of test developers and publishers who provide instructional guidance in relation to the scores on their tests clearly demand evidence to support such guidance. For example:

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)



Our professional standards also speak to the obligations of those using test scores to make instructional decisions. Simply following a test publisher's guidance is insufficient, even when that guidance is accompanied by adequate validity evidence.

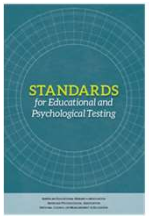
Standard 3.18: In testing individuals for diagnostic and/or special program placement purposes, test users should not use test scores as the sole indicators to characterize an individual's functioning, competence, attitudes, and/or predispositions. Instead, multiple sources of information should be used, alternative explanations for test performance should be considered, and the professional judgment of someone familiar with the test should be brought to bear on the decision. (p. 71)

Standard 12.10: In educational settings, a decision or characterization that will have major impact on a student should take into consideration not just scores from a single test but other relevant information. (p. 198)

Standard 12.13: When test scores are intended to be used as part of the process for making decisions about educational placement, promotion, implementation of individualized educational programs, or provision of services for English language learners, then empirical evidence documenting the relationship among particular test

scores, the instructional programs, and desired students outcomes should be provided. When adequate empirical evidence is not available, users should be cautioned to weigh the test results accordingly in light of other relevant information about the students. (p. 199)

Our Professional Standards: Consequences



Standard 12.14: "In educational settings, those who supervise others in test selection, administration, and score interpretation should be familiar with the evidence for the reliability/precision, the validity of the intended interpretations, and the fairness of the scores. They should be able to articulate and effectively train others to articulate a logical explanation of the relationships among the tests used, the purposes served by the tests, and the interpretations of the test scores for the intended uses." (p. 199)

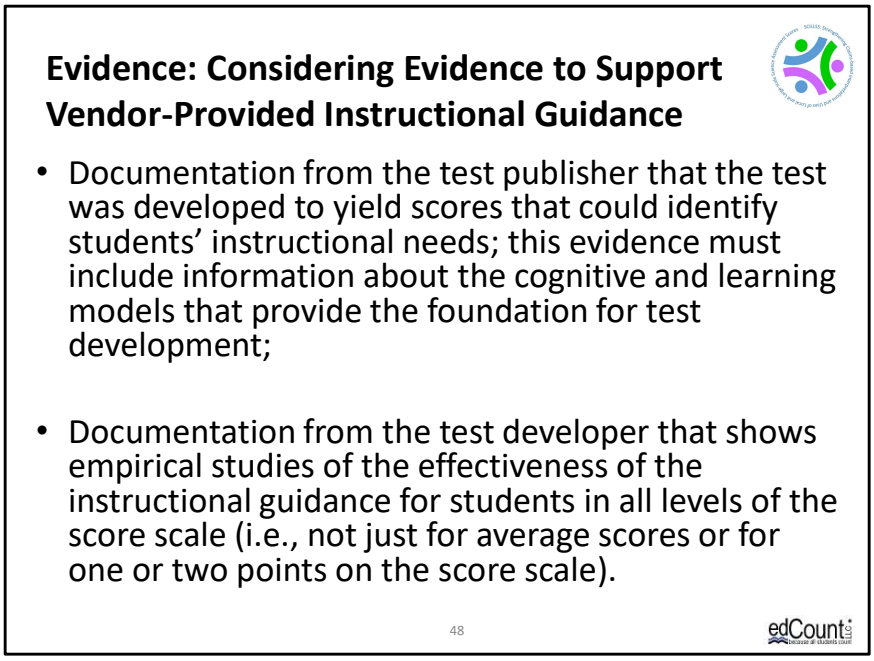
Standard 12.15: "Those responsible for educational testing programs should take appropriate steps to verify that the individuals who interpret the test results to make decisions within the school context are qualified to do so or are assisted by and consult with persons who are so qualified." (p. 199)

(AERA, APA, & NCME, 2014)

47

Standard 12.14: In educational settings, those who supervise others in test selection, administration, and score interpretation should be familiar with the evidence for the reliability/precision, the validity of the intended interpretations, and the fairness of the scores. They should be able to articulate and effectively train others to articulate a logical explanation of the relationships among the tests used, the purposes served by the tests, and the interpretations of the test scores for the intended uses. (p. 199)

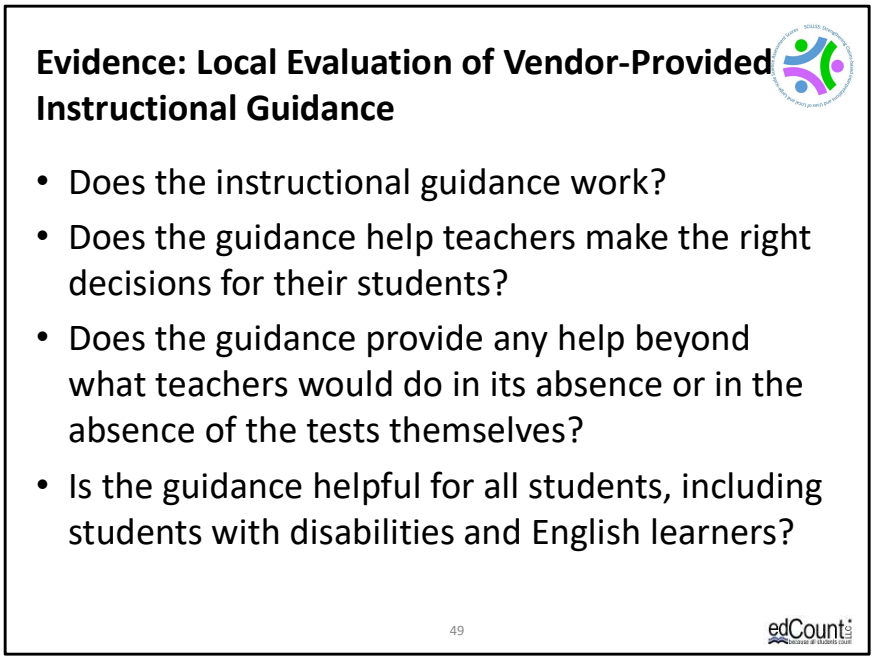
Standard 12.15: Those responsible for educational testing programs should take appropriate steps to verify that the individuals who interpret the test results to make decisions within the school context are qualified to do so or are assisted by and consult with persons who are so qualified. (p. 199)



When evaluating a test as part of making a decision to purchase it or to determine whether it is an effective part of an assessment system, a test user must consider the following kinds of evidence:

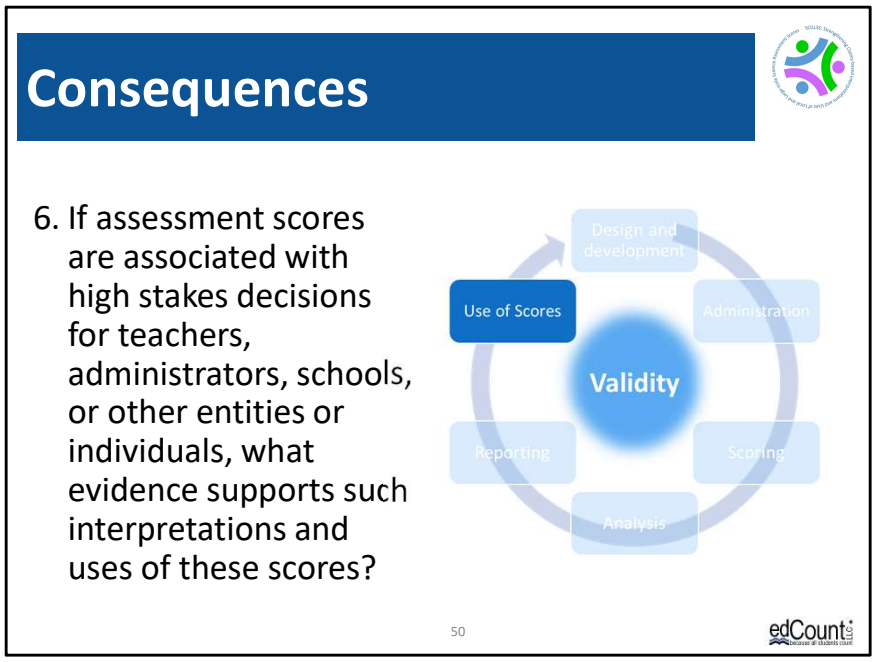
Documentation from the test publisher that the test was developed to yield scores that could identify students' instructional needs; this evidence must include information about the cognitive and learning models that provide the foundation for test development;

Documentation from the test developer that documents empirical studies of the effectiveness of the instructional guidance for students in all levels of the score scale (for example, not just for average scores or for one or two points on the score scale).



In addition, test users must evaluate the efficacy of the instructional guidance as it is implemented in their classrooms. That is, does it work? Does the guidance help teachers make the right decisions for their students? Does the guidance provide any help beyond what teachers would do in its absence or in the absence of the tests themselves? Is the guidance helpful for all students, including students with disabilities and English learners?

At the end of the day, if scores from a test are not useful, then it is not worth the time and money to take the test.



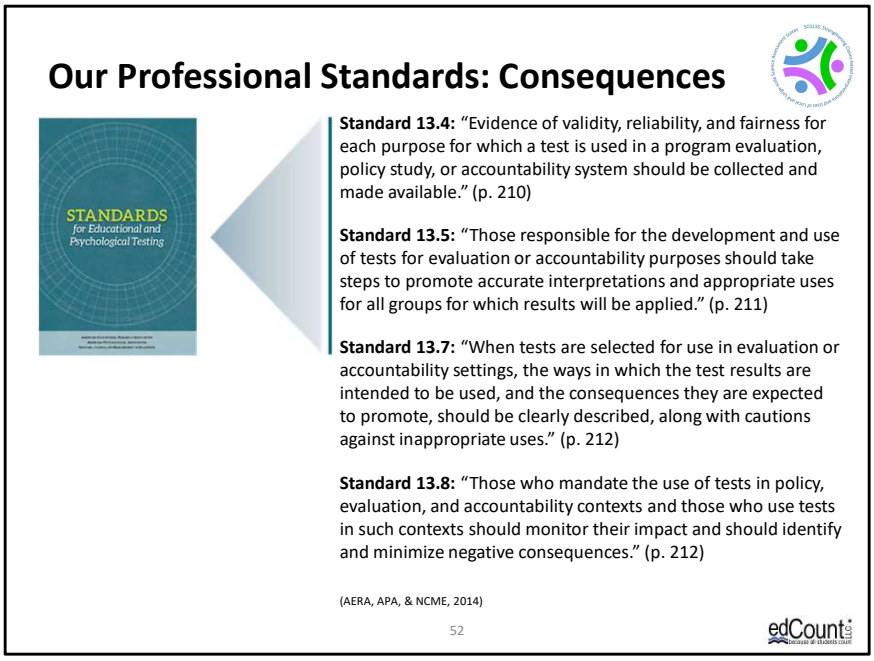
Our sixth question in this chapter is:

6. If assessment scores are associated with high stakes decisions for teachers, administrators, schools, or other entities or individuals, what evidence supports such interpretations and uses of these scores?

Evidence related to this question would be found in the score use phase of the testing life cycle.

As we described earlier in this chapter, as well as in Chapter 1 of this series, every test score use is associated with some degree of stakes. Some stakes are low, such as when the decisions do not have a significant impact on an individual or group or are easily reversible. Some stakes are high, such as when decisions lead to a diagnosis, grade promotion or retention, program enrollment, or financial award or loss. As we've cautioned several times, no decision should ever be made on the basis of a sole test score and no high stakes decision should depend solely on scores from tests.

The truth is, test scores are used everyday to make decisions that affect people's lives. The effects may be small in the moment, but have lasting impact as a student or teacher or school is set on a path that, if wrong, cannot be corrected quickly and without negative consequences. Thus, it's the responsibility of the test user to ensure that there is adequate evidence to support test scores uses, particularly when intended uses may be associated with high stakes.



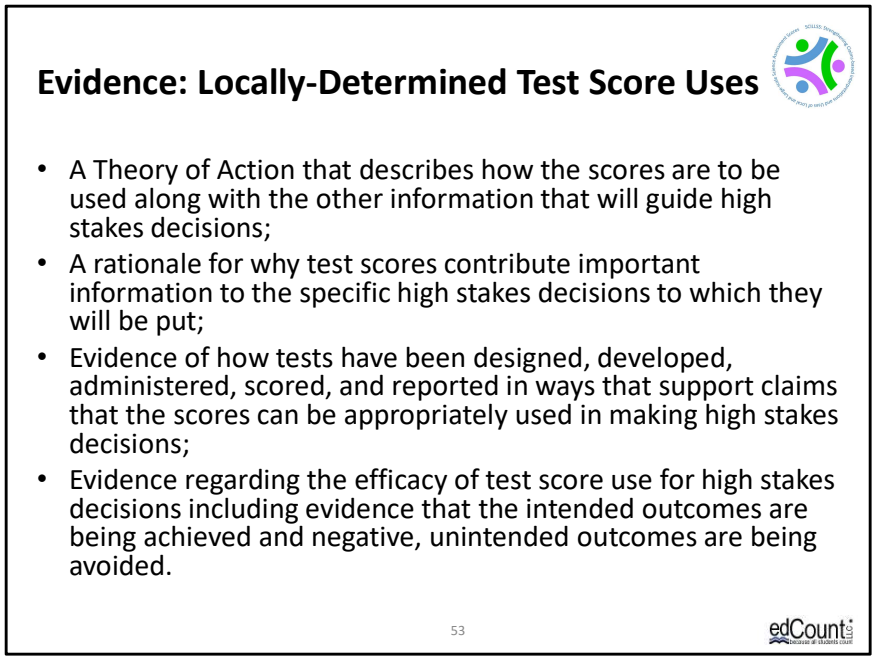
As always, we turn to our professional standards for guidance.

Standard 13.4: Evidence of validity, reliability, and fairness for each purpose for which a test is used in a program evaluation, policy study, or accountability system should be collected and made available. (p. 210)

Standard 13.5: Those responsible for the development and use of tests for evaluation or accountability purposes should take steps to promote accurate interpretations and appropriate uses for all groups for which results will be applied. (p. 211)

Standard 13.7: When tests are selected for use in evaluation or accountability settings, the ways in which the test results are intended to be used, and the consequences they are expected to promote, should be clearly described, along with cautions against inappropriate uses (p. 212).

Standard 13.8: Those who mandate the use of tests in policy, evaluation, and accountability contexts and those who use tests in such contexts should monitor their impact and should identify and minimize negative consequences (p. 212).



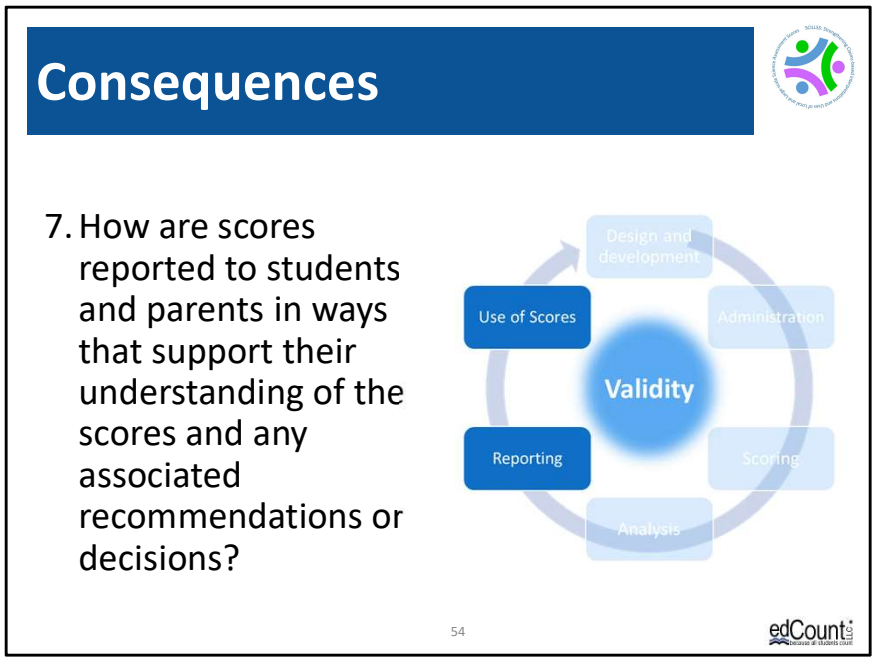
It is beyond the scope of this series to articulate a full scope for the research and evaluation studies necessary to support the uses of scores for high stakes decisions that include those for accountability. However, the following are key aspects of such research and evaluation scopes:

A Theory of Action that describes how the scores are to be used along with the other information that will guide high stakes decisions;

A rationale for why test scores contribute important information to the specific high stakes decisions to which they will be put;

Evidence of how tests have been designed, developed, administered, scored, and reported in ways that support claims that the scores can be appropriately used in making high stakes decisions;

Evidence regarding the efficacy of test score use for high stakes decisions including evidence that the intended outcomes are being achieved and negative, unintended outcomes are being avoided.



Our final question in this chapter is:

7. How are scores reported to students and parents in ways that support their understanding of the scores and any associated recommendations or decisions?

Evidence for this question comes from the reporting and use phases of the testing life cycle.

Those who use test scores must take care to report them in ways that students and parents can understand what they do and do not mean. Students and parents have a right to understand the scores and how they can and should be used as well as how they should not be used.

55

edCount

Tests and assessment systems can seem so complicated and unwieldy that we can forget that the most important result from testing is that students better understand what they know and can do and that their teachers and parents understand how to help them learn. Students and parents are often the forgotten elements of assessment systems even as they, along with their educator partners, have the most at stake in testing.

Those who use test scores must take care to report them in ways that students and parents can understand what they do and do not mean. Students and parents have a right to understand the scores and how they can and should be used as well as how they should not be used.

Our Professional Standards: Consequences

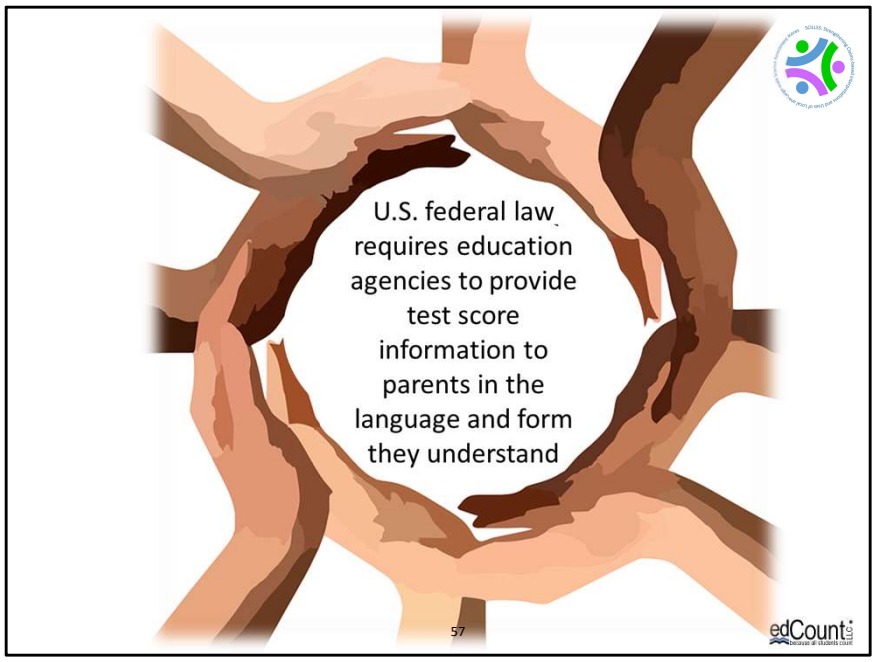
Standard 6.10: "When test score information is released, those responsible for testing programs should provide interpretations appropriate to the audience. The interpretations should describe in simple language what the test covers, what scores represent, the precision/reliability of the scores, and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

(AERA, APA, & NCME, 2014)

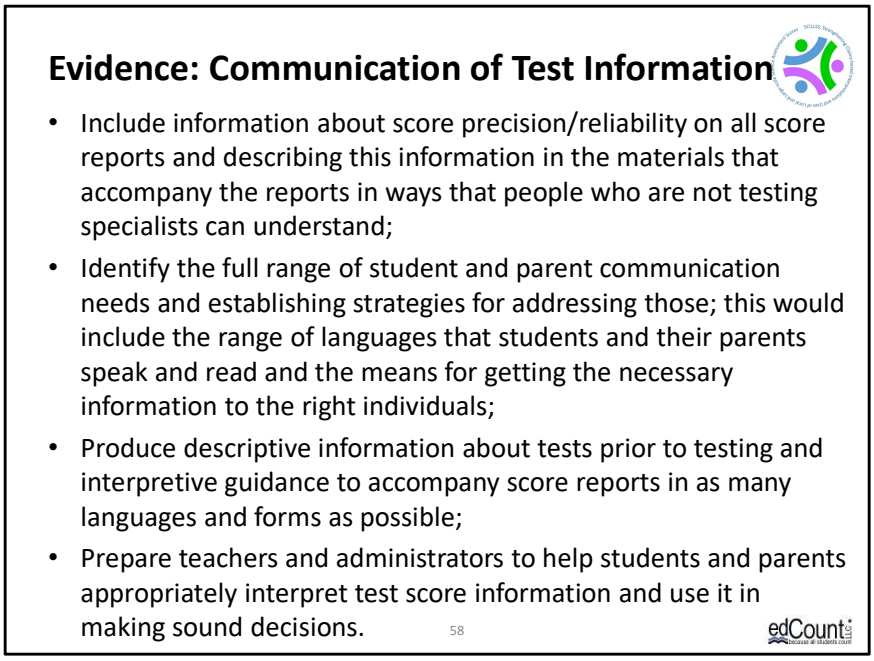
56

Our trusty professional standards underscore the obligations of those who publish tests as well as those who use the scores in terms of how test score information is conveyed.

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 scores represent, the precision/reliability of the scores, and how scores are intended to be used. (p. 119)



In addition to our professional standards, federal law requires education agencies to provide test score information to parents in the language and form they understand. It may not be enough for a school or district to distribute score reports written in English; parents who do not speak or read English have the same right to information about their children as any other parent does. Further, all students and all parents have a right to information about tests and how the scores from those tests will be used before the tests are administered.



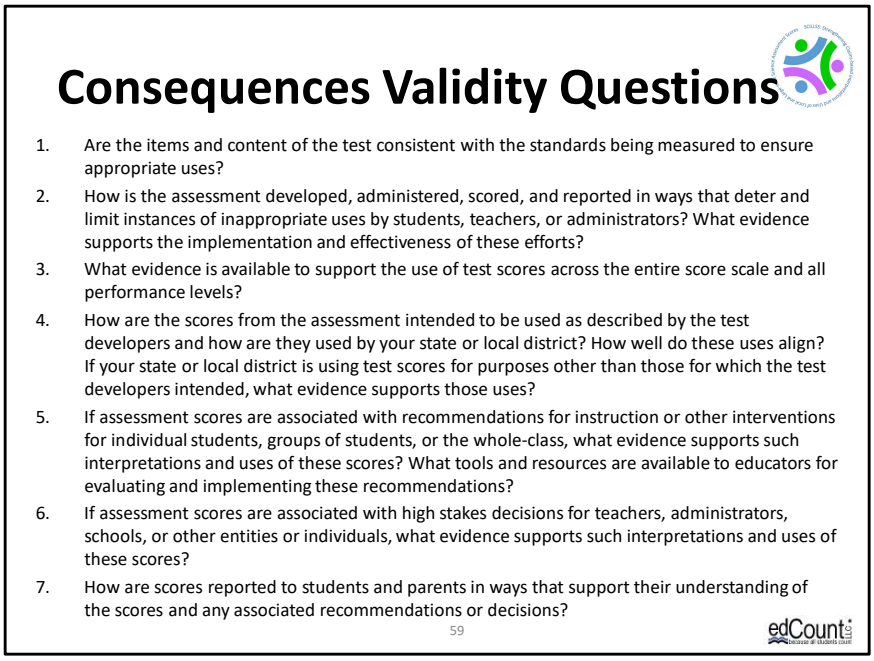
To support students' and parents' rights to testing information, test users should work with publishers to identify and implement the best means for communicating test score information to these stakeholders. This includes:

Including information about score precision/reliability on all score reports and describing this information in the materials that accompany the reports in ways that people who are not testing specialists can understand;

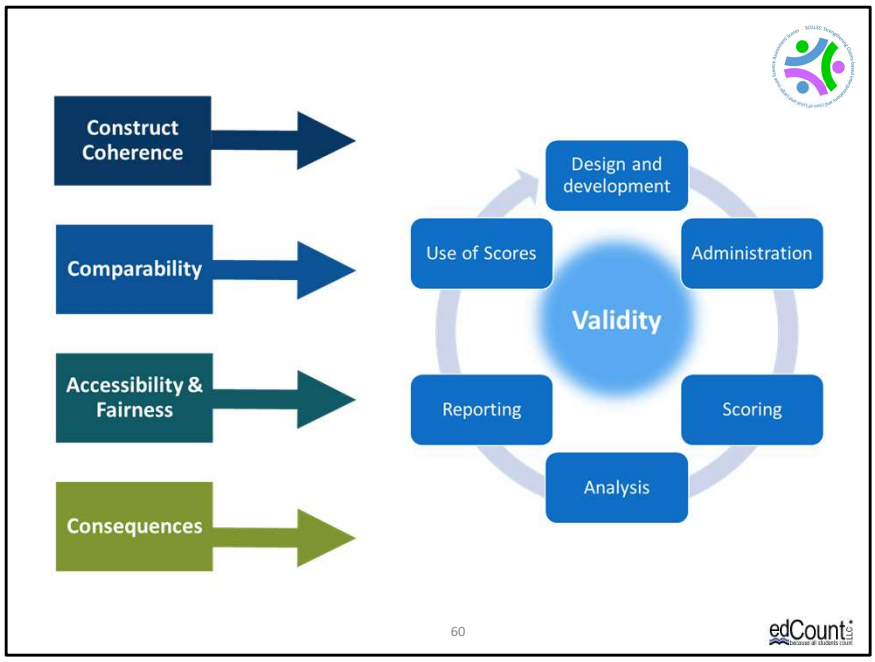
Identifying the full range of student and parent communication needs and establishing strategies for addressing those; this would include the range of languages that students and their parents speak and read and the means for getting the necessary information to the right individuals;

Producing descriptive information about tests prior to testing and interpretive guidance to accompany score reports in as many languages and forms as possible;

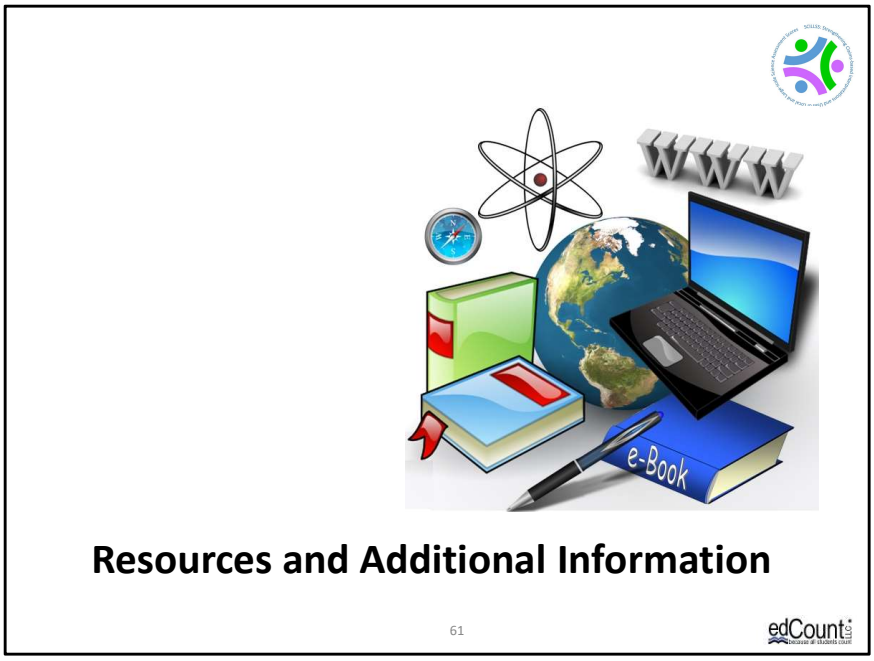
Preparing teachers and administrators to help students and parents appropriately interpret test score information and use it in making sound decisions.



We have reached the end of our seven questions in this chapter. The key takeaway from this chapter is that test scores are associated with consequences. Those publishing and using scores are obligated to consider these consequences and to take steps both to support intended test score interpretations and uses and avoid untended ones, particularly those that are negative for any individual or group.



In this series, we have addressed questions related to construct coherence, comparability and reliability/precision, fairness and accessibility, and consequences. We hope that this series has been helpful in understanding how to evaluate the validity and reliability of assessment scores.

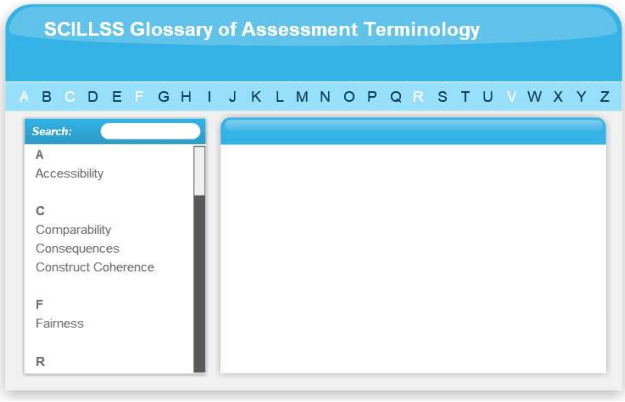


Finally, we offer additional resources that may be helpful to anyone interested in learning more about the concepts presented in this chapter. A glossary of terms and our reference list follow.

Thank you for your engagement in this fifth 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



62

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

63

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.

Booher-Jennings, J. (2005). *Below the Bubble: "Educational Triage" and the Texas Accountability System.* American Educational Research Journal, 42(2), 231-268.

Stecher, B. (2002). *Consequences of Large-Scale, High-Stakes Testing on School and Classroom Practices.* In L.S. Hamilton, B.M.