Creating and Evaluating Effective Educational Assessments

Chapter 4

Fairness and Accessibility

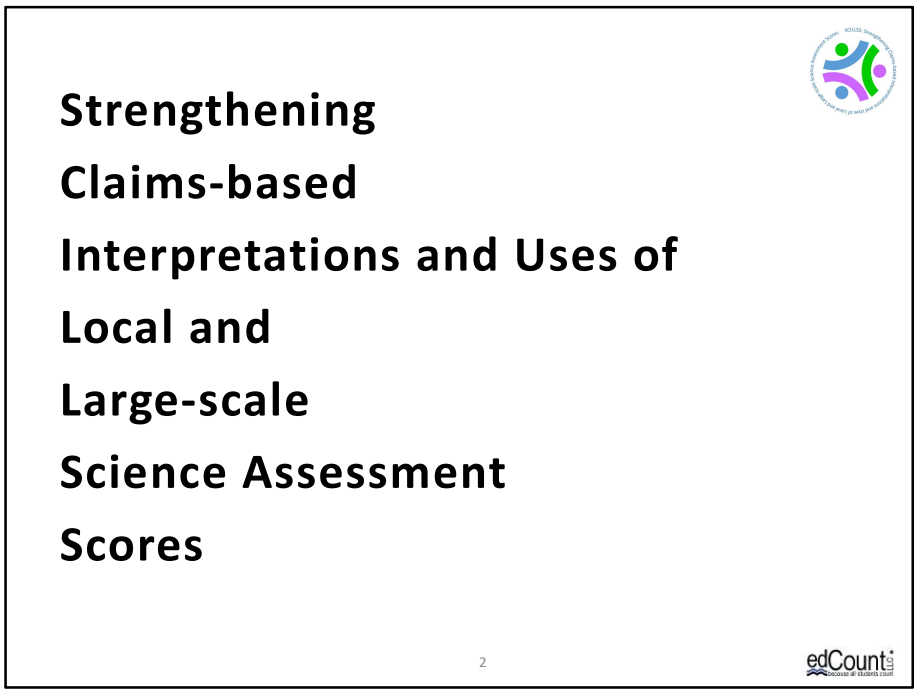
This digital workbook on educational assessment design and evaluation was developed by edCount, LLC, under
Enhanced Assessment Grants Program, CFDA 84.368A.

1

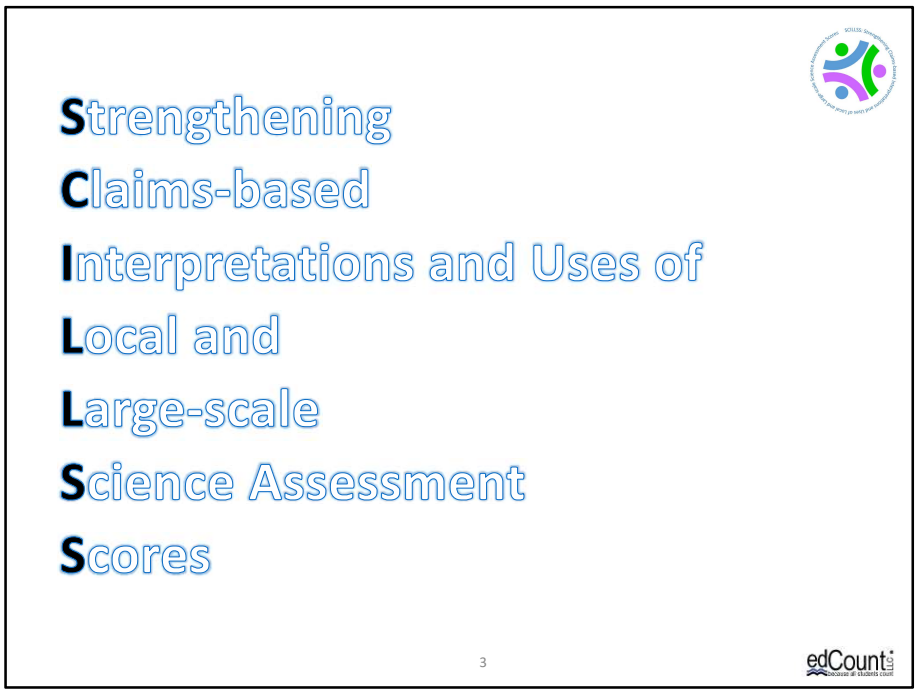
Chapter 4:

Welcome to the fourth 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.

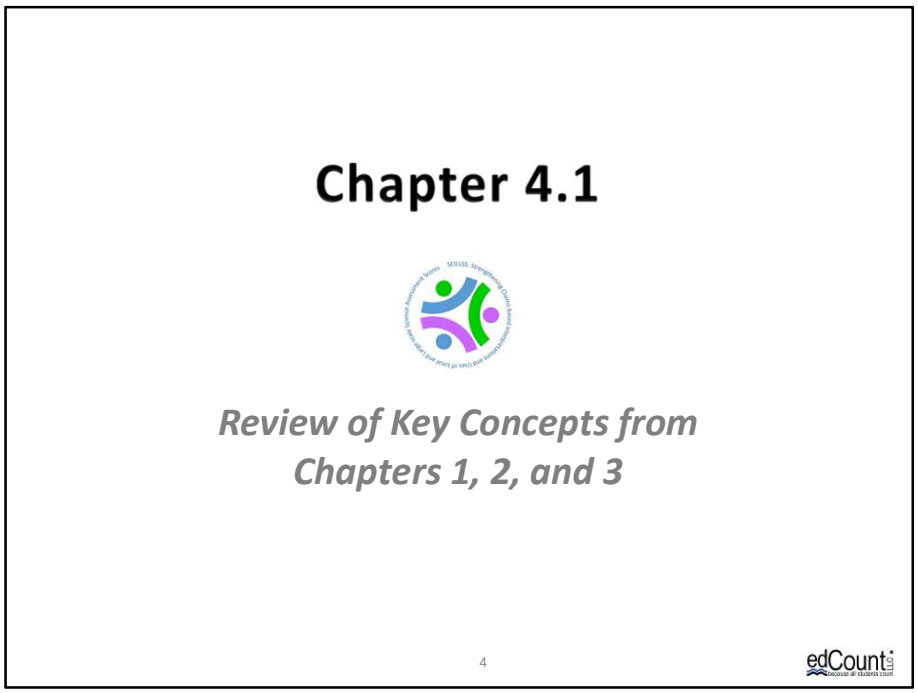
This digital workbook was developed by edCount, LLC, under the US Department of Education's Enhanced Assessment Grants Program, CFDA 84.368A.



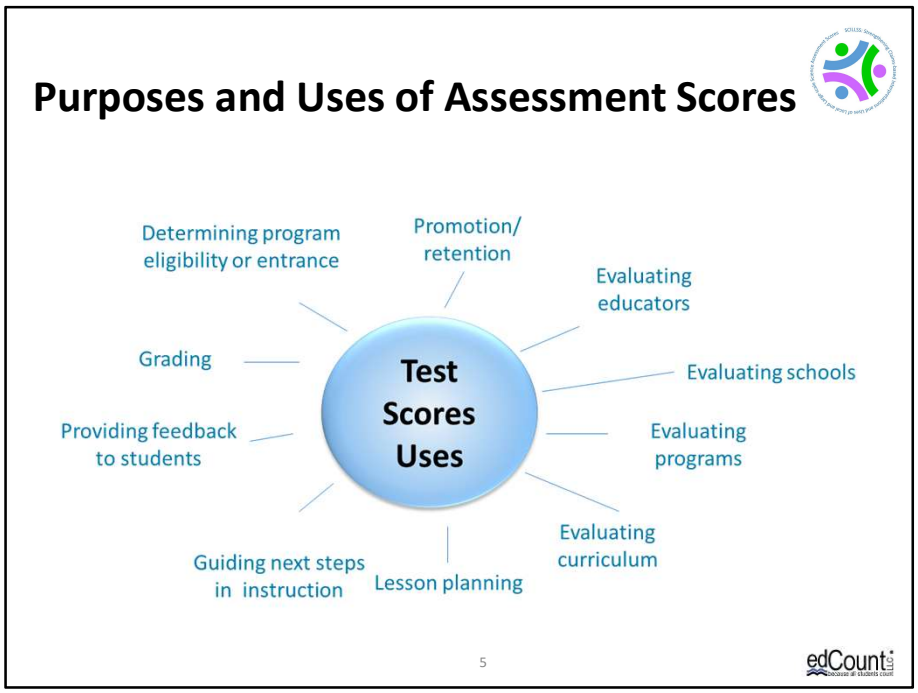
The grant project is titled the Strengthening Claims-based Interpretations and Uses of Local and Large-scale Science Assessment Scores…



or its acronym, “SCILLSS.”

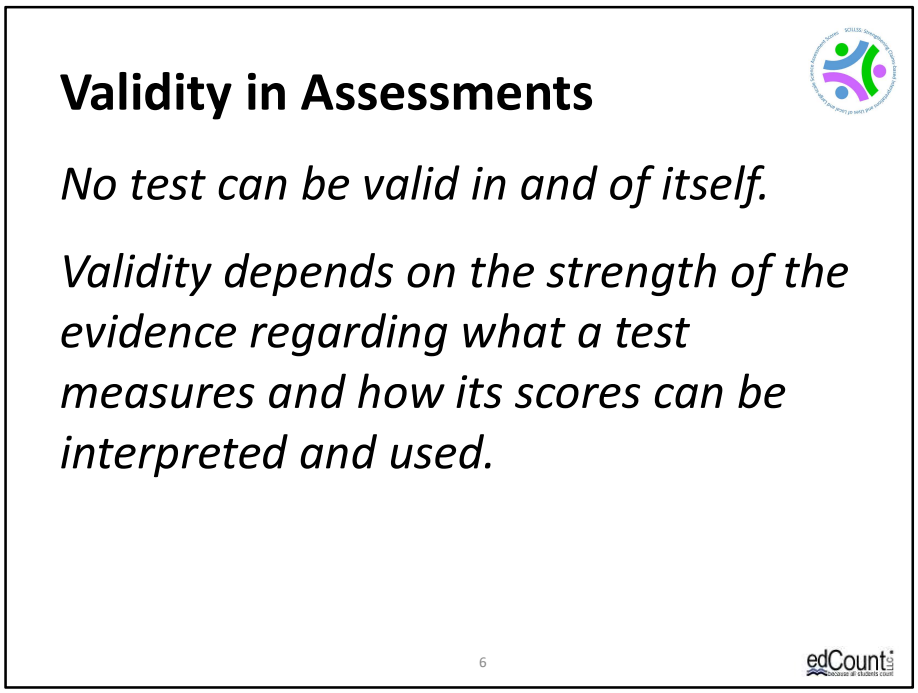


Chapter 4.1. Review of Key Concepts from Chapters 1, 2, and 3.

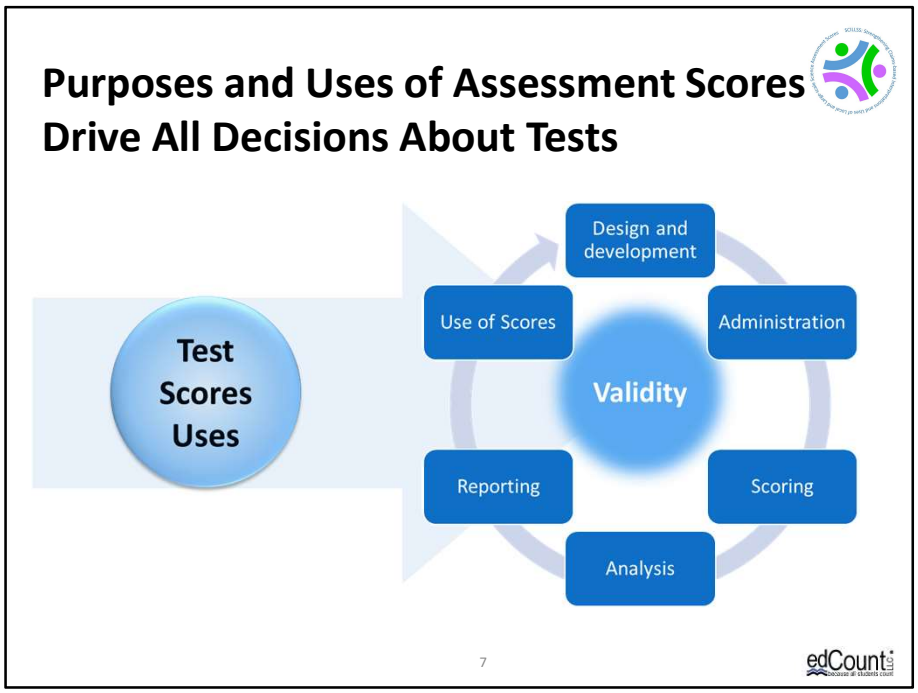


Let's begin with a brief recap of the key concepts covered in the first three chapters of this series.

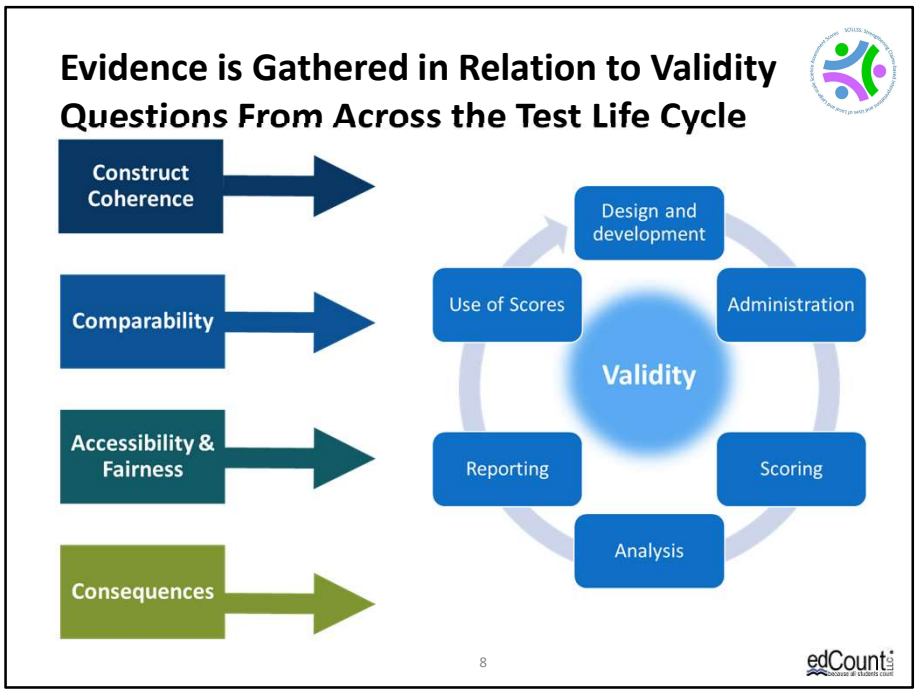
Chapter 1 focused on common reasons why we administer assessments of students' academic knowledge and skills and how we use those assessment scores. We learned that these purposes for administering assessments and the intended uses of assessment scores should drive all decisions about how assessments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



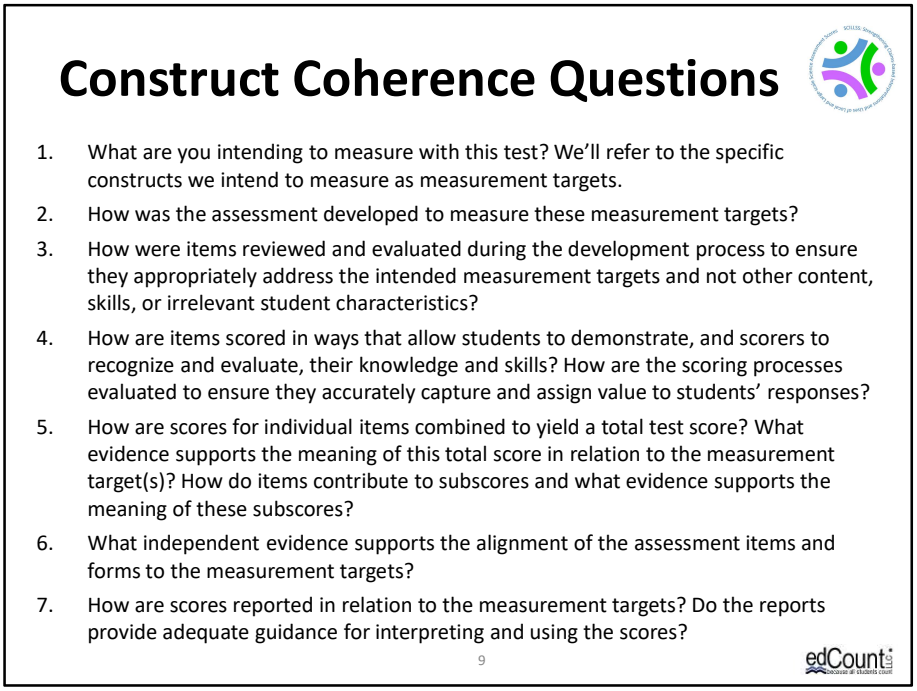
We evaluate validity by gathering and judging evidence. This validity evidence is gathered from across the entire life cycle of a test from design and development through score use. Judgments about validity are based upon the quality and adequacy of this evidence in relation to assessment score interpretations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

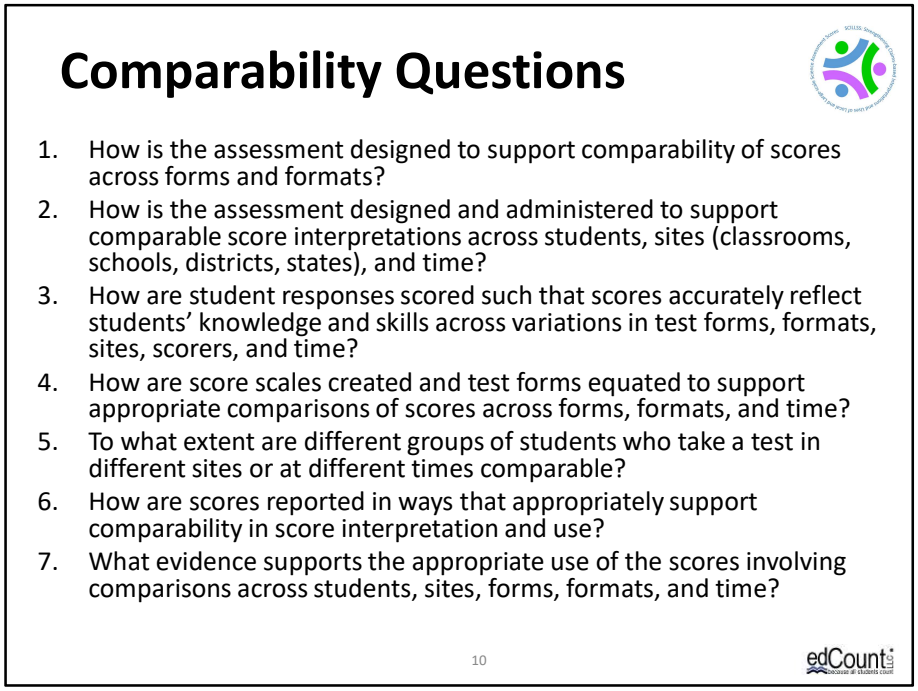
The four validity question categories are:

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



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.

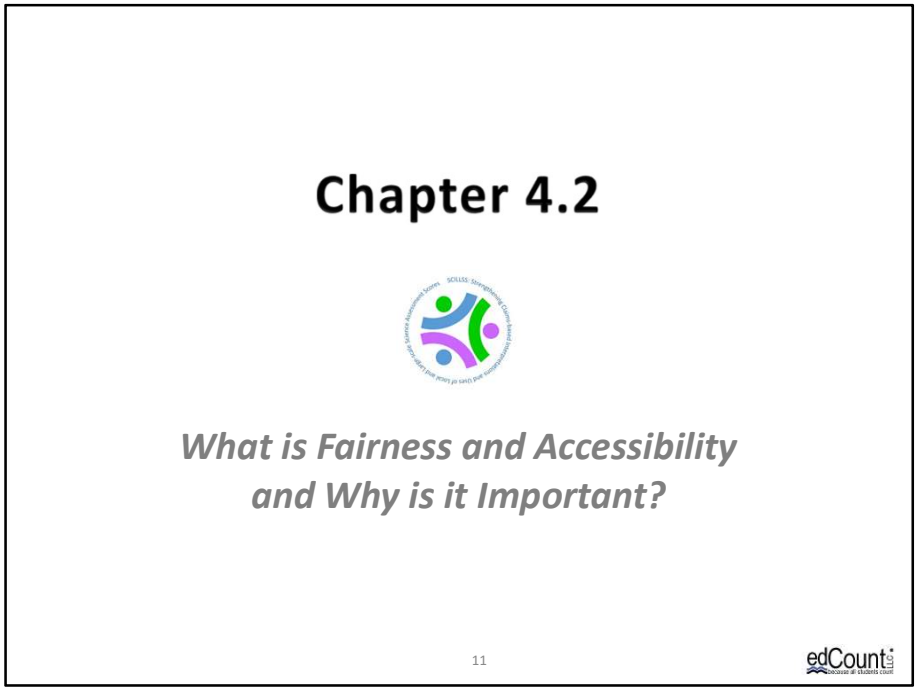
1. What are the measurement targets for this test? That is, what are you intending to measure with this test?
2. How was the assessment developed to measure these measurement targets?
3. How were items reviewed and evaluated during the development process to ensure they appropriately address the intended 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:

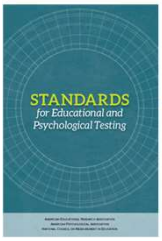
1. How is the assessment designed to support comparability of scores across forms and formats?
2. How is the assessment designed and administered to support comparable score interpretations across students, sites (classrooms, schools, districts, states), and time?
3. How are student responses scored 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 the scores involving comparisons across students, sites, forms, formats, and time?

Now, in this chapter, we turn our attention to the third set of validity questions, which relate to the notion of fairness and accessibility.



Chapter 4.2: What is Fairness and Accessibility and Why is it Important?

Our Professional Standards: Fairness and Accessibility



- A test is fair when it "reflects the same construct(s) for all test takers and scores from it have the same meaning for all individuals in the intended population"

  (AERA, APA, & NCME, 2014, p. 49)

- "…a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct."

  (AERA, APA, & NCME, 2014, p. 50)

- Accessibility is related to fairness and means that all test takers should "have an unobstructed opportunity to demonstrate their standing on the construct(s) being measured

  (AERA, APA, & NCME, 2014, p. 111)

12

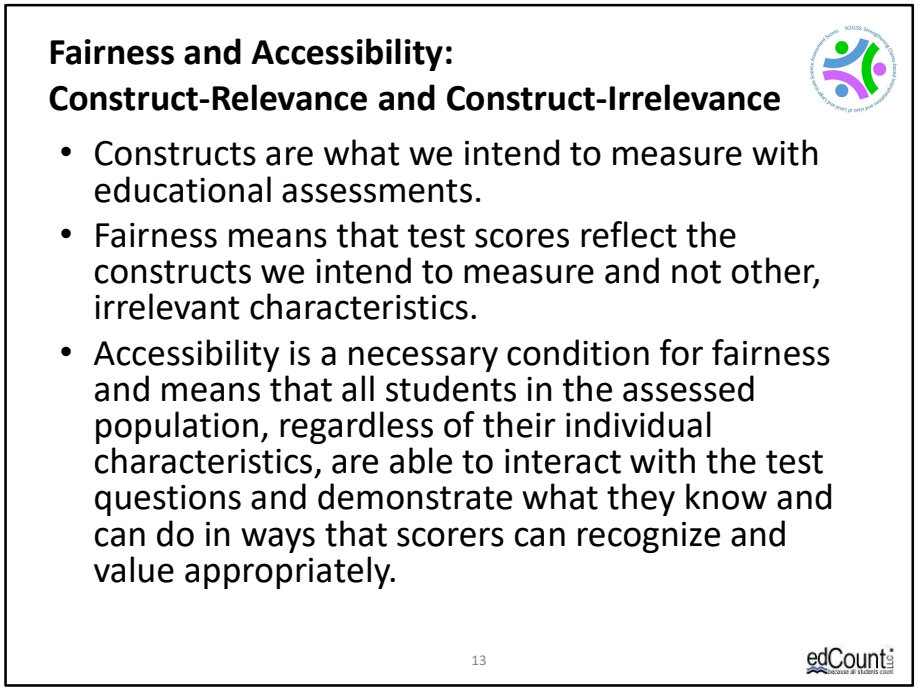
edCount

Fairness is such a critical issue for educational measurement that our professional standards, called *The Standards for Educational and Psychological Testing*, devote one of the three foundational chapters to fairness in testing. The other two foundational chapters focus on validity and reliability.

For educational testing, fairness means responsiveness to individual characteristics and testing contexts in ways that will support valid interpretations of scores for their intended uses. According to the Standards, a test is fair when it "reflects the same construct(s) for all test takers and scores from it have the same meaning for all individuals in the intended population" (p. 49).

Further, the Standards tell us that "a fair test does not advantage or disadvantage some individuals because of characteristics irrelevant to the intended construct" (p. 50).

Accessibility is related to fairness and means that all test takers should "have an unobstructed opportunity to demonstrate their standing on the construct(s) being measured" (p. 111).

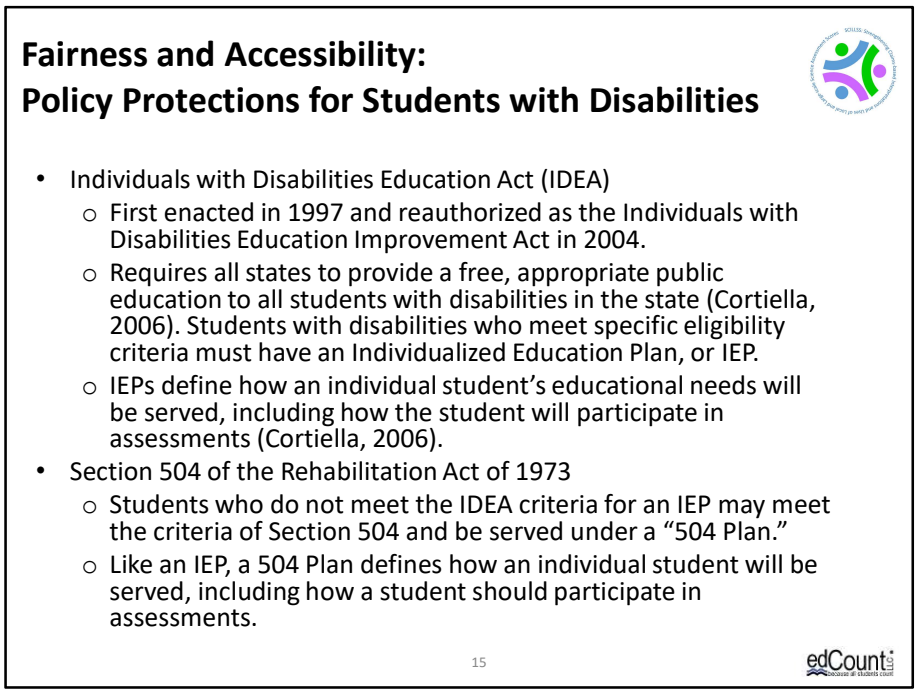


Recall from the first chapter in this workbook series that constructs are what we intend to measure with educational assessments. Reading comprehension and problem-solving skills in a particular content area and context are examples of constructs. Fairness means that test scores reflect the constructs we intend to measure and not other, irrelevant characteristics. Accessibility is a necessary condition for fairness and means that all students in the assessed population, regardless of their individual characteristics, are able to interact with the test questions and demonstrate what they know and can do in ways that scorers can recognize and value appropriately.

As an educator, you already understand the importance of fairness and accessibility in your own classrooms, schools, and communities. You understand that we must be attentive to the distinct needs of each student throughout their school experiences. Ensuring fair and accessible assessment opportunities is just one part of our ethical and professional duty as educators.

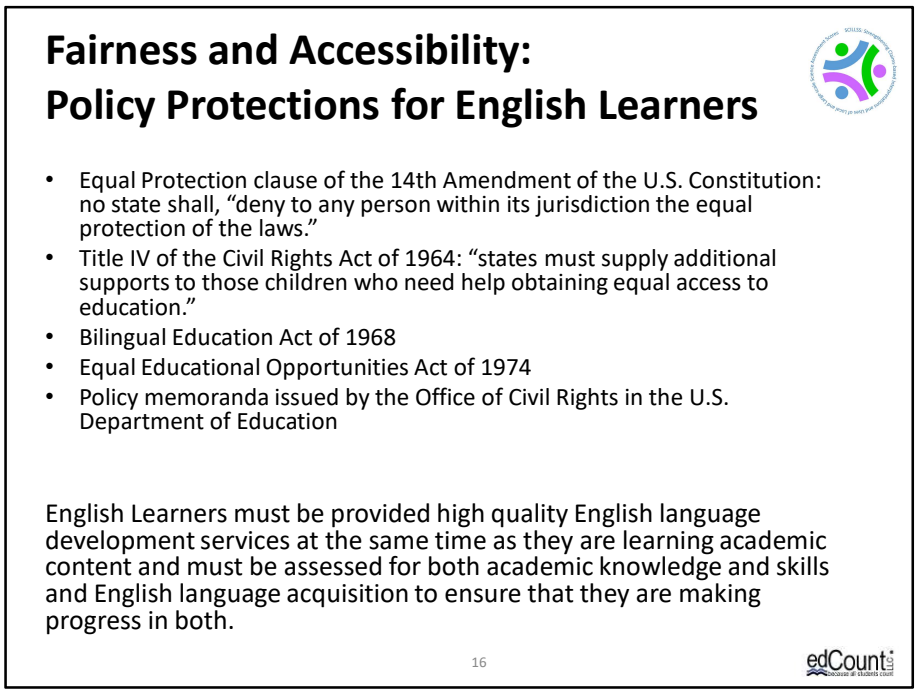
Our duty as educators align with the obligations articulated in our professional standards for educational measurement. We must provide each and every student with a fair and legitimate opportunity to demonstrate their knowledge and skills.

Unfortunately, not all assessments have been designed to support fairness and accessibility for all students. In particular, the accessibility needs of students from some cultural and racial groups, English learners, and students with disabilities require focused attention when designing assessments. Beyond our professional and ethical obligations to ensure these students have opportunities to learn and to demonstrate their knowledge and skills, federal policies specifically mandate fair and accessible assessments for these students. We'll briefly describe some of these protections next.



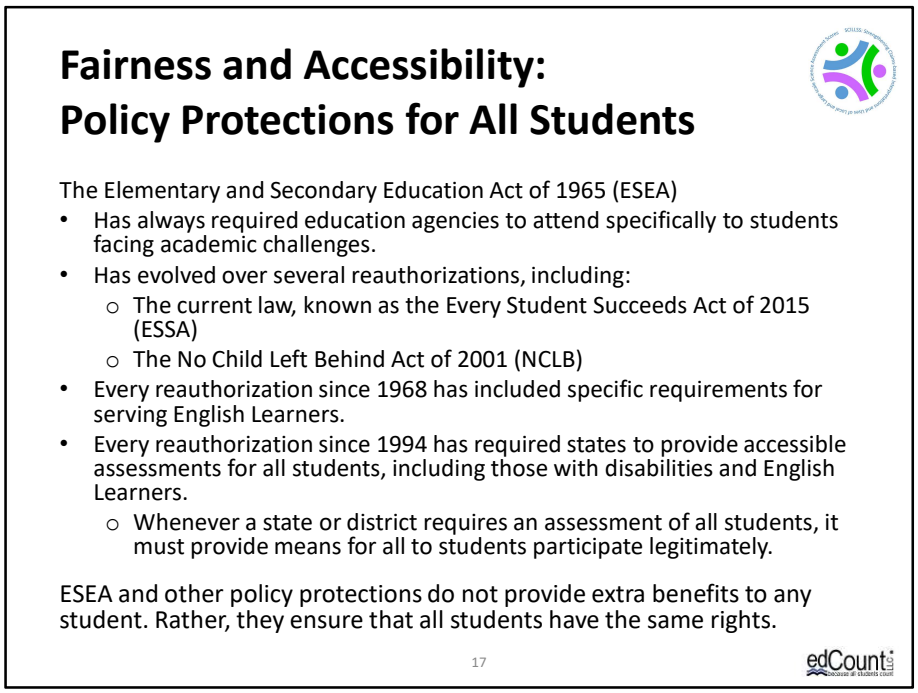
The Individuals with Disabilities Education Act (IDEA), first enacted in 1997 and reauthorized as the Individuals with Disabilities Education Improvement Act in 2004, requires all states that accept IDEA funds – which all states do – provide a free, appropriate public education to all students with disabilities in the state (Cortiella, 2006). Students with disabilities who meet specific eligibility criteria must have an Individualized Education Plan, or IEP, under the IDEA rules. IEPs define how an individual student's educational needs will be served, including how the student will participate in assessments (Cortiella, 2006).

Section 504 of the Rehabilitation Act of 1973 includes similar protections and prohibits discrimination of students based on disability. Students who do not meet the IDEA criteria for an IEP may meet the criteria of Section 504 and be served under a "504 Plan." Like an IEP, a 504 Plan defines how an individual student will be served, including how a student should participate in assessments.



English learners' rights to fair and accessible assessment opportunities are also protected in federal policy. These protections stem from the Equal Protection clause of the 14th Amendment of the U.S. Constitution, which states that no state shall, "deny to any person within its jurisdiction the equal protection of the laws." Title IV of the Civil Rights Act of 1964 stipulates that "states must supply additional supports to those children who need help obtaining equal access to education" and this concept was echoed in the Bilingual Education Act of 1968, the Equal Educational Opportunities Act of 1974, and through several policy memoranda issued by the Office of Civil Rights in the U.S. Department of Education.

The key ideas for English Learners are that they must be provided high quality English language development services at the same time as they are learning academic content and must be assessed for both academic knowledge and skills and English language acquisition to ensure that they are making progress in both.



The legal protections for students with disabilities and English learners include decades of case law in addition to these policies. Further, the Elementary and Secondary Education Act of 1965, known as ESEA, has from its original enactment required education agencies to attend specifically to students facing academic challenges. ESEA has evolved over several reauthorizations, including the current Every Student Succeeds Act of 2015, or ESSA, and its predecessor, the No Child Left Behind Act of 2001, or NCLB. Every reauthorization since 1968, when the Bilingual Education Act was introduced as Title VII of ESEA, has included specific requirements for serving English Learners.

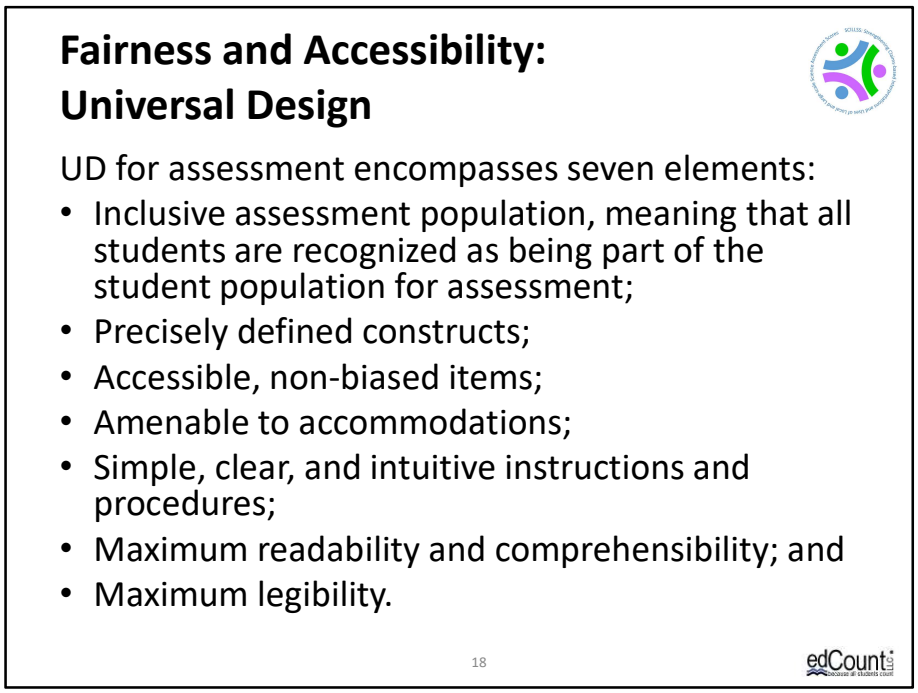
The protections for these two groups of students also intersect. English Learners with a disability must be identified and evaluated for special education services in a timely manner according to IDEA and Section 504.

In each of its reauthorizations since 1994, ESEA has required states to provide accessible assessments for all students, including those with disabilities and English Learners. Whenever a state requires an assessment of all students, it must provide means for all students to participate legitimately. We note that, under IDEA, this requirement extends to all assessments that a district requires students to take, as well.

In terms of legitimate participation, these requirements mean that appropriate

accommodations and alternate assessments must be available for the students who need them. We will return to these ideas in a later section of this chapter.

These protections do not provide extra benefits to any student. Rather, they ensure that all students have the same rights.



While accommodations and alternate assessments are necessary for many students, the approach that addresses the widest range of students' needs is Universal Design. Universal Design, or UD, for assessments emerged from the same philosophy that gave us curb cuts, the ramps built into street curbs that allow everyone, including people who use wheelchairs or push strollers, to access the sidewalk easily.

According to researchers at the National Center for Educational Outcomes at the University of Minnesota, UD for assessment encompasses seven elements:

- Inclusive assessment population, meaning that all students are recognized as being part of the student population for assessment
- Precisely defined constructs
- Accessible, non-biased items
- Amenable to accommodations
- Simple, clear, and intuitive instructions and procedures
- Maximum readability and comprehensibility
- Maximum legibility

It's easy to see that a test developed to reflect all of these elements would not only be accessible for students with disabilities and English Learners, but to all students. And that,

like the curb cuts, is the point of Universal Design. Some adjustments improve access for everyone and reduce the need for some accommodations. If test developers ask themselves whether an assessment and each assessment item address all of these elements, all students will benefit and fewer students may require accommodations. Fairness and accessibility are for each and every student, regardless of whether they have been identified as having a disability or as an English Learner. Every student counts.

Our Professional Standards: Universal Design

If test developers use UD:

"Test items and tasks can then be purposively designed and developed from the outset to reflect the intended construct, to minimize construct irrelevant features that might otherwise impede the performance of intended examinee groups, and to maximize, to the extent possible, access for as many examinees as possible in the intended population regardless of race, ethnicity, age, gender, socioeconomic status, disability, or language or cultural background."

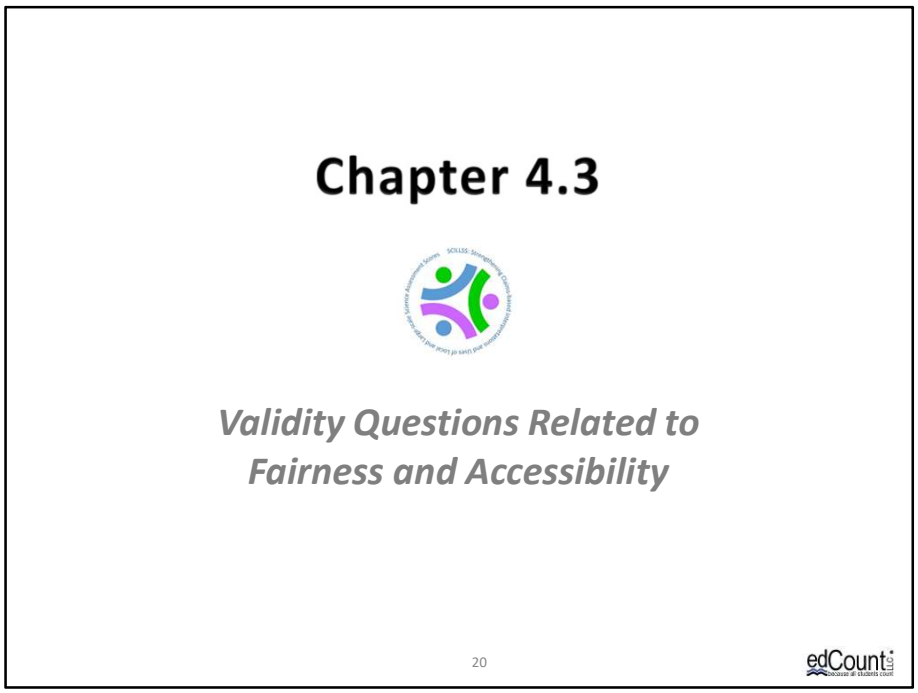
(AERA, APA, & NCME, 2014, p. 50)

19

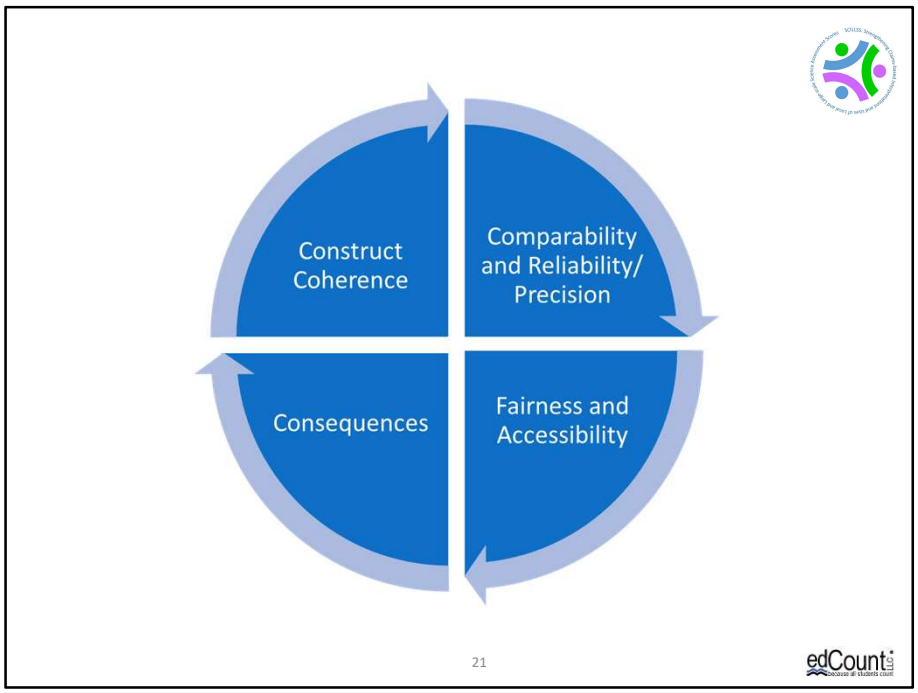
edCount

Our professional standards highlight the benefits of UD in assessment design. If test developers use UD, "Test items and tasks can then be purposively designed and developed from the outset to reflect the intended construct, to minimize construct irrelevant features that might otherwise impede the performance of intended examinee groups, and to maximize, to the extent possible, access for as many examinees as possible in the intended population regardless of race, ethnicity, age, gender, socioeconomic status, disability, or language or cultural background" (p. 50).

Next, we'll consider our validity questions related to fairness and accessibility.

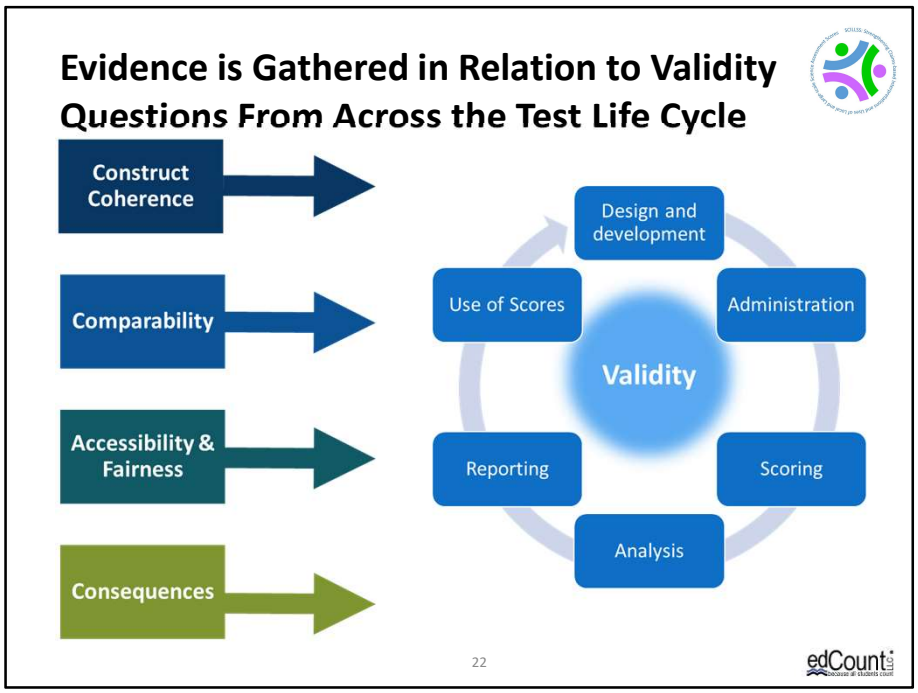


Chapter 4.3: Validity Questions Related to Fairness and Accessibility.

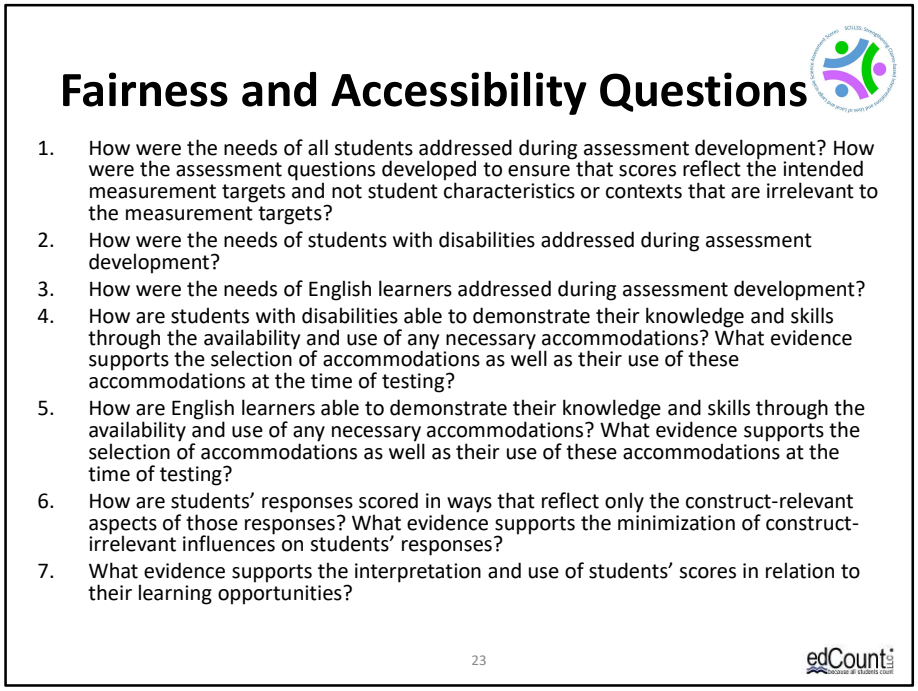


Recall that validity relates to the interpretations and uses of test scores rather than to a test itself. In addition, validity is evaluated based upon a collection of evidence. Key pieces of this evidence relate to how well the test measures the constructs that it is intended to measure, which we addressed in the second chapter of this series, and the degree to which scores can be compared, which we addressed in the third chapter of this series.

Fairness and accessibility evidence is highly related to both of these sets of validity questions. A test might be considered fair if the scores reflect the intended constructs with minimal influence of variations in student characteristics or contexts. A test might be accessible if all students can demonstrate their relevant knowledge and skills without unnecessary obstruction. For scores to be combined or compared across students, schools, forms, administrations, or time, requires evidence of fairness and accessibility in addition to other types of comparability evidence.



Like other areas of validity evidence, evidence related to fairness and accessibility should come from across the testing life cycle. Fairness and accessibility considerations must be integral to the design and development of a test and in each of the subsequent life cycle phases.

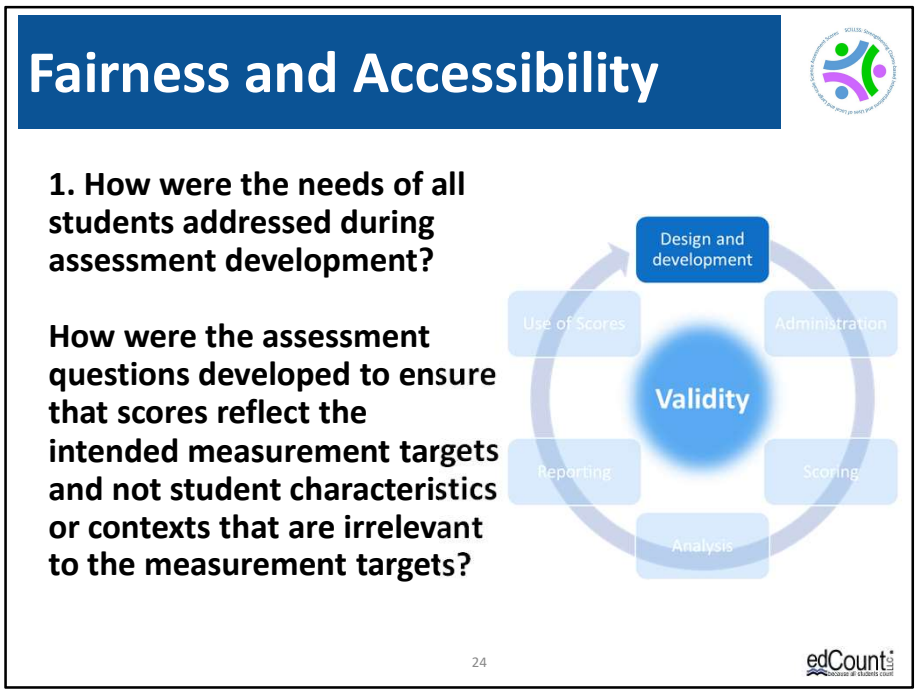


Our validity questions related to fairness and accessibility are:

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?

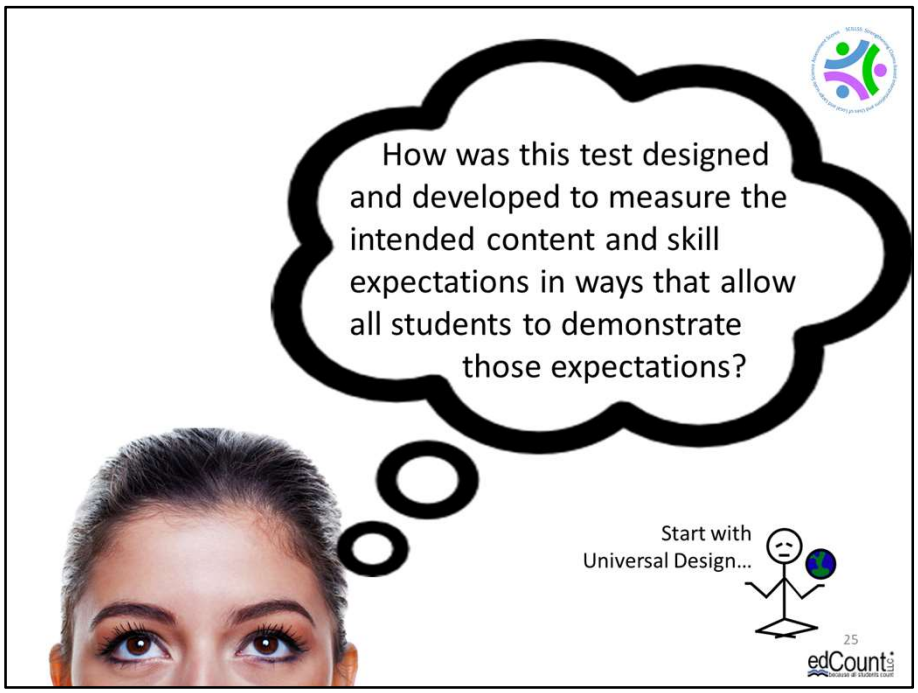
We will consider evidence related to each of these questions in the sections that follow.



Our first validity question related to fairness and accessibility is:

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?

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



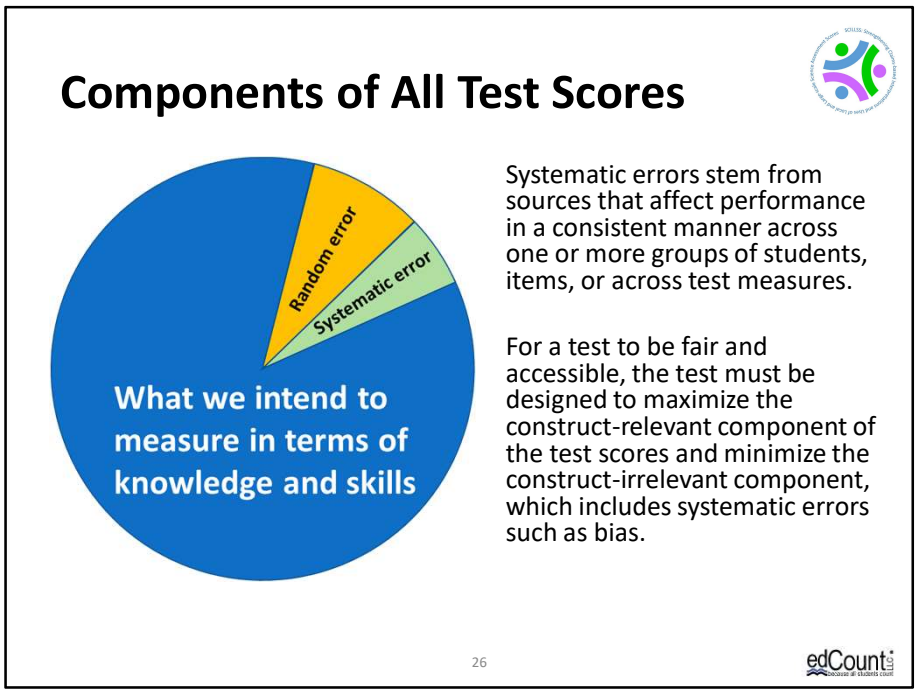
In the second chapter of this series, we focused on validity evidence related to construct coherence. In that case, we described evidence of how a test would be designed to reflect the constructs it was meant to measure, such as academic standards or objectives for a unit.

Here, when we are thinking about fairness and accessibility, we turn our attention to how students interact with the items on a test. This is somewhat like finishing the sentence, "how was this test designed and developed to measure the intended content and skill expectations…" with "…in ways that allow all students to demonstrate those expectations?"

Test designers and developers must consider the characteristics of all students in the population who will be taking the test. Not the characteristics of the average student or of most students. All students. As we've learned, large-scale and local assessments must be accessible to all students, whether they have disabilities or not, are English learners, or come from particular cultural, racial, geographic, or socioeconomic backgrounds.

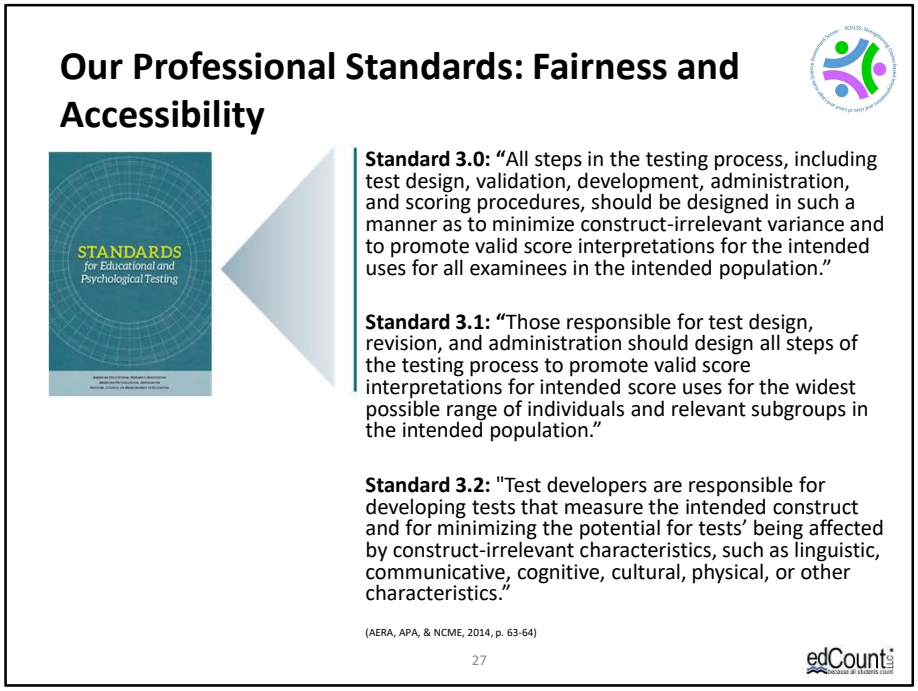
Waiting to address student needs through the use of accommodations during test administration is an unacceptable approach to designing and developing tests.

Instead, thoughtful design that incorporates the principles of Universal Design can generate items that are accessible to the widest range of students and are also amenable to accommodation where necessary.



Some key concepts from chapter 3 of this workbook series, which focused on comparability, are also related to fairness and accessibility. Recall that all test scores have two components: one that reflects what we're intending to measure and one that reflects error. For academic achievement tests, that means one part of the score tells us what students know and can do while the other part, the error, diminishes that meaning of the score. Further, the error component of the score includes random error, which reduces reliability, and systematic error, which reduces validity. Bias is a type of systematic error where the content or features of the test questions give some students an advantage or impose a disadvantage on some students that has nothing to do with the constructs the test is meant to measure.

For a test to be fair and accessible, the test must be designed to maximize the construct-relevant component of the test scores and minimize the construct-irrelevant component, which includes systematic errors such as bias.



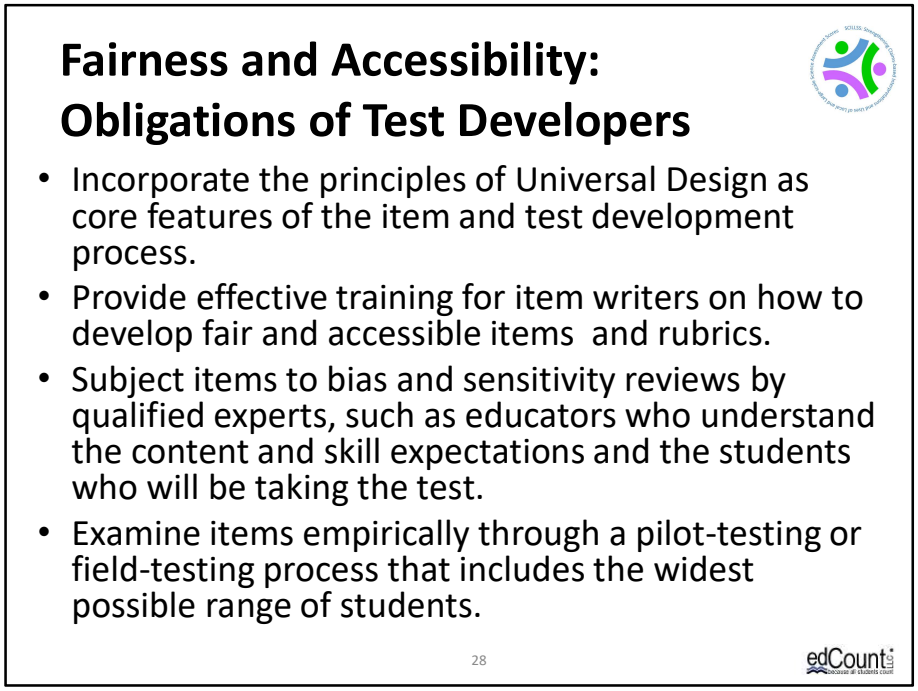
Our professional standards include several expectations related to fairness and accessibility. For example, the first three standards in the chapter on fairness and accessibility address fundamental obligations for supporting valid score interpretations across the full range of the population taking the test. The full population includes all students: that is, students from low income households; male, female, and nonbinary students, students with disabilities, English learners, and students from all racial, ethnic, and religious backgrounds. All students.

Standard 3.0: All steps in the testing process, including test design, validation, development, administration, and scoring procedures, should be designed in such a manner as to minimize construct-irrelevant variance and to promote valid score interpretations for the intended uses for all examinees in the intended population. (p.63)

Standard 3.1: Those responsible for test design, revision, and administration should design all steps of the testing process to promote valid score interpretations for intended score uses for the widest possible range of individuals and relevant subgroups in the intended population. (p.63)

Standard 3.2: Test developers are responsible for developing tests that measure the intended construct and for minimizing the potential for tests' being affected by construct-

irrelevant characteristics, such as linguistic, communicative, cognitive, cultural, physical, or other characteristics. (p. 64)



In practical terms, these obligations require test developers to engage in the following kinds of activities and provide documentation to test users about these aspects of the development process:

- Incorporate the principles of Universal Design as core features of the item and test development process

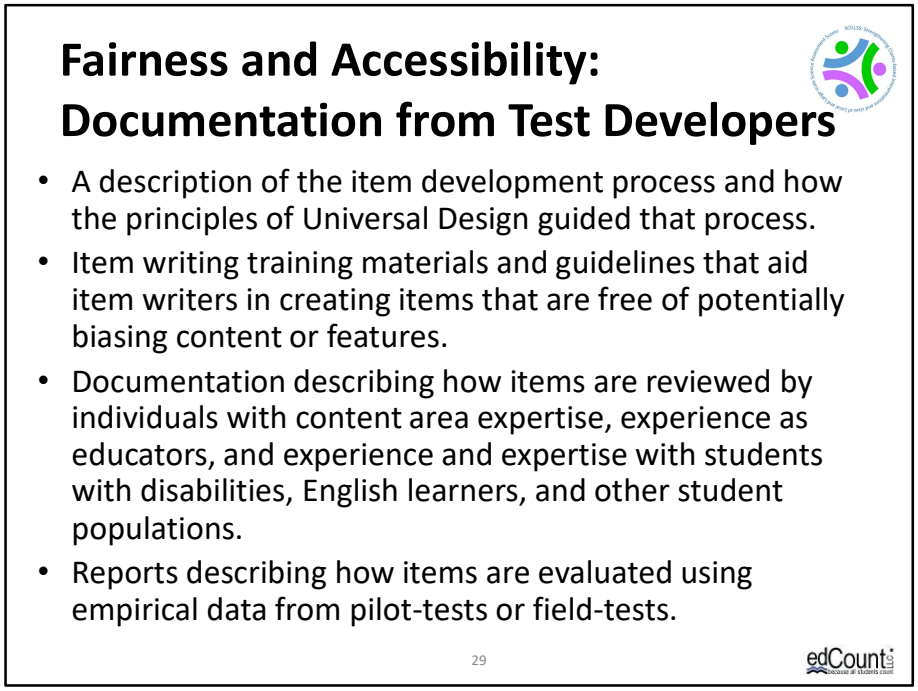
- Provide effective training for item writers on how to develop fair and accessible items and rubrics

- Subject items to bias and sensitivity reviews by qualified experts, such as educators who understand the content and skill expectations and the students who will be taking the test

- Examine items empirically through a pilot-testing or field-testing process that includes the widest possible range of students

These activities apply to local assessments as well as to large-scale commercial ones. Any teacher or administrator who is designing and developing a test should apply the principles of Universal Design and provide guidance to those writing the items about how to avoid

potentially biasing content and features. All items that appear on any test should have some degree of review for accuracy and potential bias, even if pilot-testing isn't a viable option.



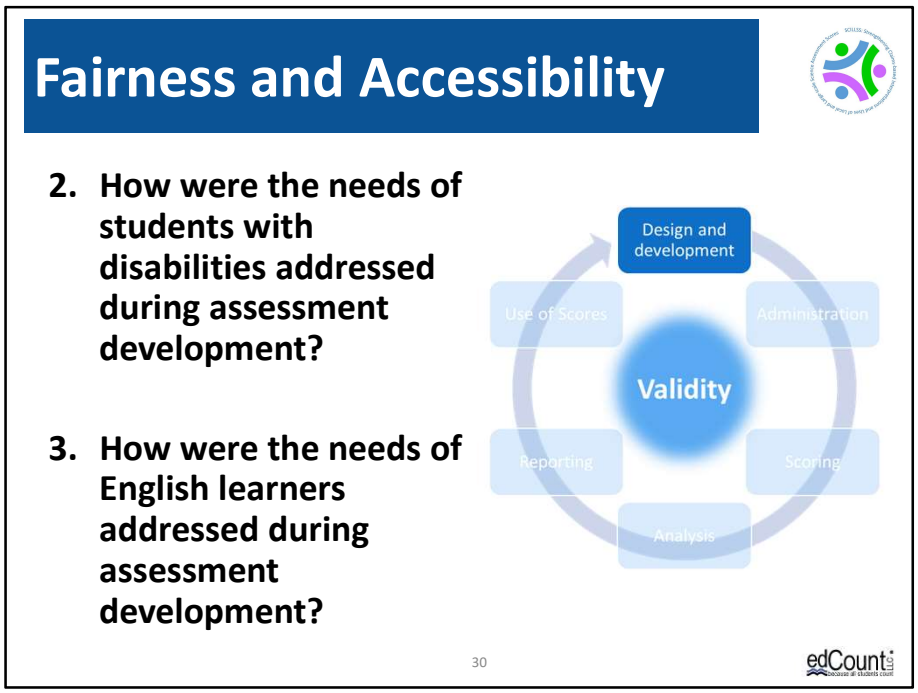
Those who use tests have a right to documentation related to these activities. This documentation should include:

- A description of the item development process and how the principles of Universal Design guided that process
- Item writing training materials and guidelines that aid item writers in creating items that are free of potentially biasing content or features
- Documentation describing how items are reviewed by individuals with content area expertise, experience as educators, and experience and expertise with students with disabilities, English learners, and other student populations
- Reports describing how items are evaluated using empirical data from pilot-tests or field-tests

Some of this evidence may be found in the technical manual for an assessment and some may be found in separate reports.

Next, we consider how these obligations apply specifically to address the needs of students

with disabilities under validity question two and to the needs of English learners under validity question three.



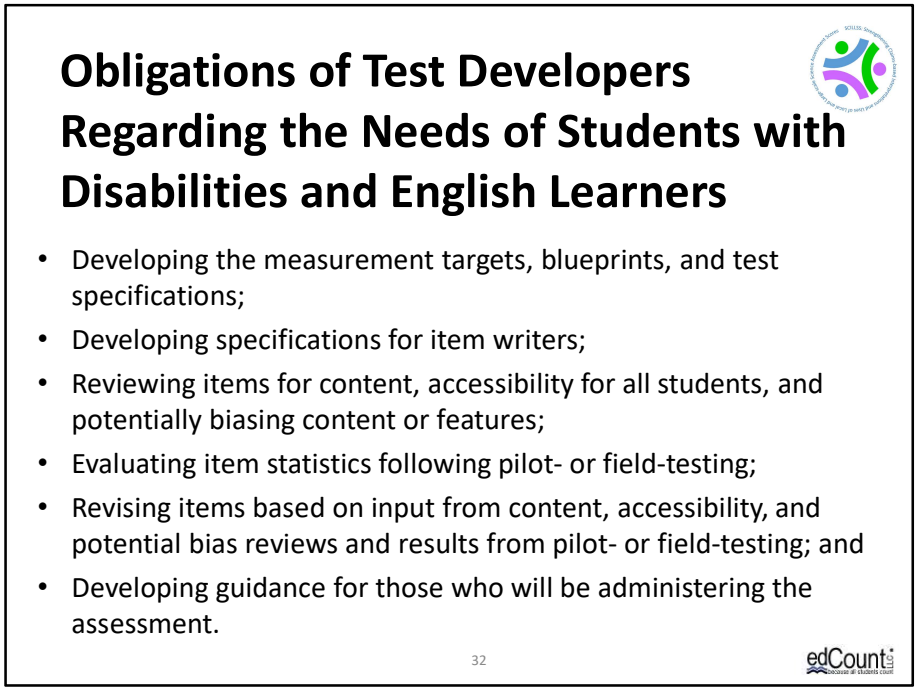
Our second and third validity questions in this chapter focus on the needs of students with disabilities and English learners.

How were the needs of students with disabilities addressed during assessment development?

How were the needs of English learners addressed during assessment development?

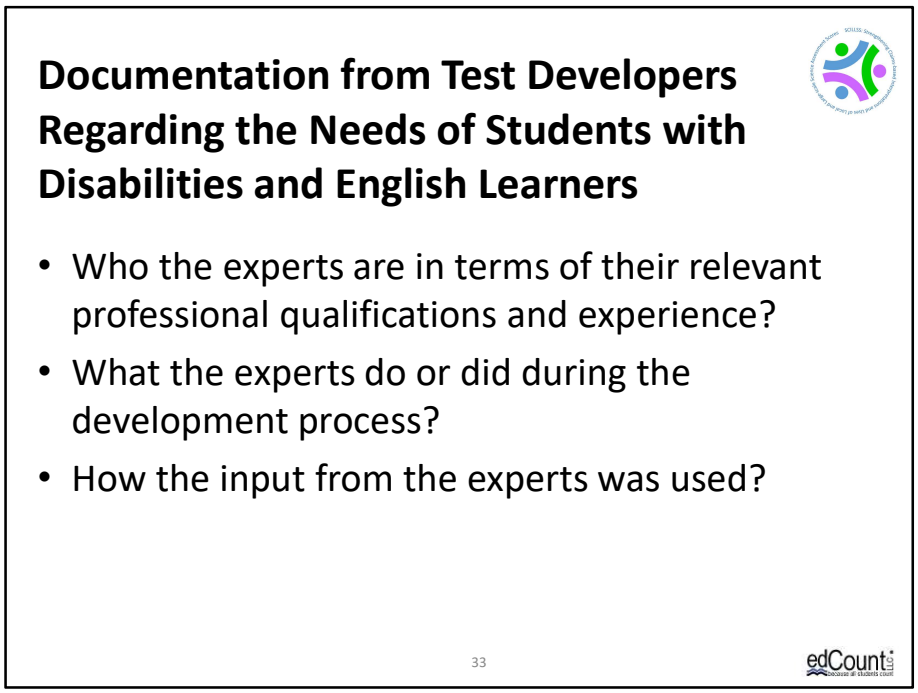
Evidence related to these questions comes from the design and development phase of the testing life cycle.

Hopefully, it's clear by now that fairness and accessibility obligations apply for all students. Every student should have unobstructed opportunities to demonstrate what they know and can do when they take tests. Some students need specific supports to allow them access to tests even when the principles of Universal Design have been applied effectively. It's important for the teams of people who design and develop assessments to include individuals with specific expertise about how students with disabilities and English learners learn and participate in assessment situations to help ensure that the principles of Universal Design are applied well and that items are amenable to accommodation when necessary.



A testing vendor should be able to demonstrate that individuals with expertise related to students with disabilities and English learners were involved in:

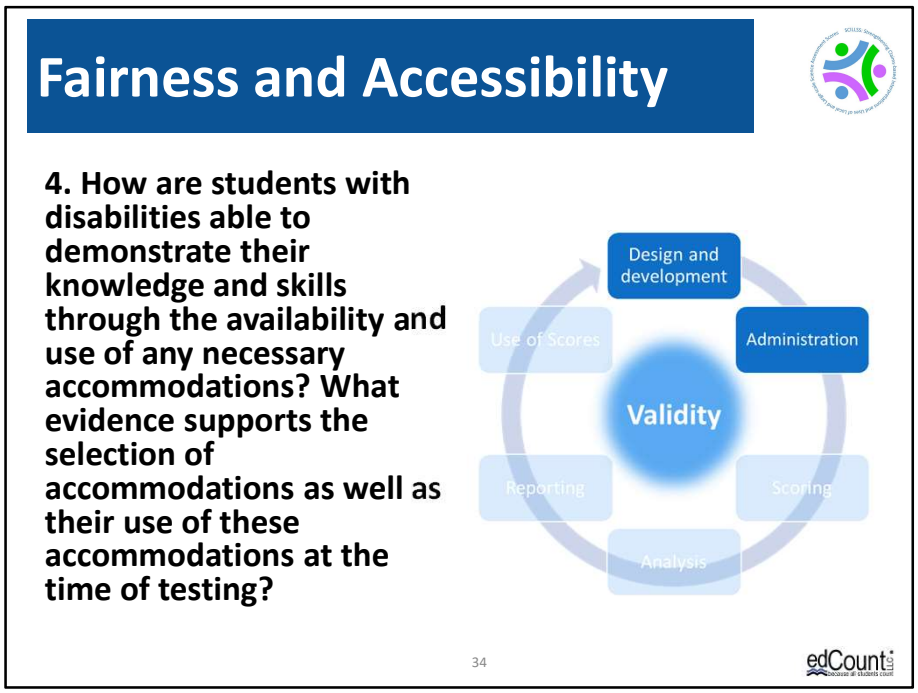
- Developing the measurement targets, blueprints, and test specifications;
- Developing specifications for item writers;
- Reviewing items for content, accessibility for all students, and potentially biasing content or features;
- Evaluating item statistics following pilot- or field-testing;
- Revising items based on input from content, accessibility, and potential bias reviews and results from pilot- or field-testing; and
- Developing guidance for those who will be administering the assessment.



Vendors should provide documentation regarding the involvement of these experts. This documentation may be found in technical manuals or in other reports about the development process and should describe:

- Who the experts are in terms of their relevant professional qualifications and experience;
- What the experts do or did during the development process; and
- How the input from the experts was used.

Merely inviting individuals with expertise in working with students with disabilities or English learners is not enough. Input from these experts must be heard, understood, and used. In all cases, the role of these experts is to help ensure that the test and the test items are developed to allow students to demonstrate what they know and can do so that their scores reflect their knowledge and skills rather than characteristics of their disability or English language proficiency.



Our fourth validity question for fairness and accessibility is:

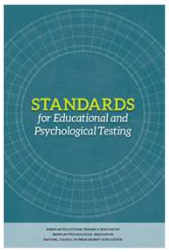
How are students with disabilities able to demonstrate their knowledge and skills through the availability and use of any necessary accommodations? What evidence supports the selection of accommodations as well as their use of these accommodations at the time of testing?

Evidence related to this question comes from the design and development and the administration phases of the testing life cycle.

As we noted for the second validity question in this chapter, individuals with expertise in how students with disabilities learn and participate in assessments must be engaged in the design process. Their input is critical to ensuring that the test items are designed to support legitimate participation opportunities for students with disabilities. This includes the effective use of Universal Design principles and the consideration of how the items support the use of testing accommodations for students with disabilities who need such accommodations.

In addition to these design considerations, teams of individuals who know the specific students who will be taking the test also define the conditions for testing. For students with disabilities who have IEPs or 504 plans, a team that includes educators who know the student, parents, special services personnel, and the student him- or herself must make decisions about how the student participates in assessments. The student may require testing accommodations or, when the student has significant cognitive disabilities and cannot access the general assessment even with accommodations, an alternate assessment that measures the same knowledge and skills. Simply excluding a student from testing is not appropriate because that suggests that that student's learning is unimportant and not valued by educators, parents, and the student.

Our Professional Standards: Fairness and Accessibility in Test Administration



Standard 3.4: "Test takers should receive comparable treatment during the test administration and scoring process."

(AERA, APA, & NCME, 2014, p. 65)

Standard 3.9: "Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs."

(AERA, APA, & NCME, 2014, p. 67)

36

edCount

The notion that some students need accommodations to participate in a test while other students do not may be a source of confusion for some. For example, our professional standards state that:

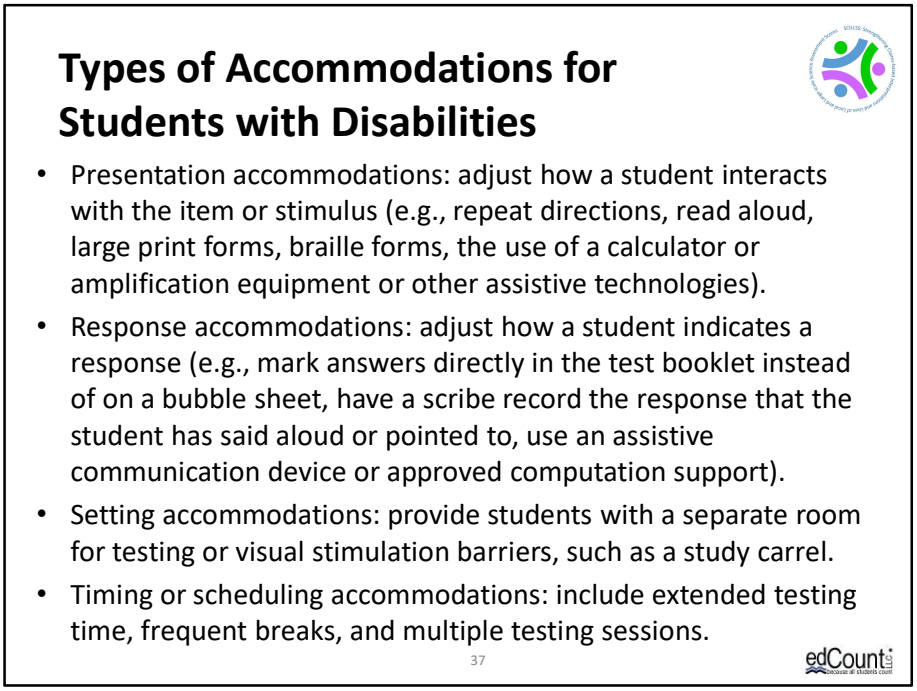
Standard 3.4: Test takers should receive comparable treatment during the test administration and scoring process. (p. 65)

Standard 3.9: Test developers and/or test users are responsible for developing and providing test accommodations, when appropriate and feasible, to remove construct-irrelevant barriers that otherwise would interfere with examinees' ability to demonstrate their standing on the target constructs. (p. 67)

Use of the word "comparable" in Standard 3.4 is important. The authors of the Standards chose it rather than the term "the same" because two students may need different conditions for testing to yield the same information about their knowledge and skills; this is the point of accommodations.

As a concrete example, imagine we are intending to measure student's skills in evaluating a model to explain interactions among two different chemicals. If, as part of the task the student was asked to pour a specific volume of a chemical into a beaker, that would require

the motor skills necessary to lift and pour as well as the visual acuity to read the volume levels on the beaker. If this were a technology-enhanced item, the student would still need to be able to take the action necessary to pour the virtual liquid into the virtual beaker and read the measure. Most students could do all of these things. Some students could not. Those students may need one or more accommodations to assist with the incidental activities such as lifting, pouring, moving virtual objects, and reading quantities so that they can participate in the cognitive aspects of the model evaluation the test is meant to measure. Appropriate accommodations would support these students in ways that even the playing field by removing barriers without providing undue advantages.



Designed and used appropriately, accommodations do not alter the construct that the test is measuring. When an adjustment changes the construct that is being measured, it is no longer an appropriate accommodation. It is a modification that renders the student's score uninterpretable. It may be helpful to use a framework to think about how and when accommodations are applied:

Presentation accommodations are those that adjust how a student interacts with the item or stimulus. These include such options as repeat directions, read aloud, large print forms, braille forms, the use of a calculator or amplification equipment or other assistive technologies.

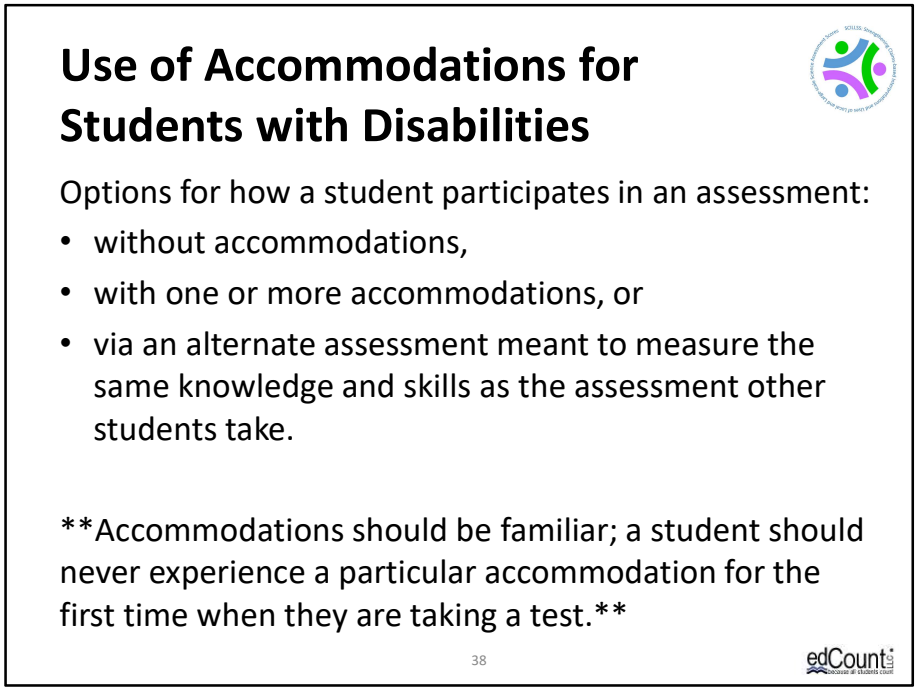
Response accommodations adjust how a student indicates a response. For example, the student may mark answers directly in the test booklet instead of on a bubble sheet, or have a scribe record the response that the student has said aloud or pointed to or use an assistive communication device or approved computation support.

Setting accommodations can provide students with a separate room for testing or visual stimulation barriers such as a study carrel.

Timing or scheduling accommodations include extended testing time, frequent breaks, and

multiple testing sessions.

Appropriate IEP processes should be in place and educators must understand which accommodations are available and most suitable for each student. Once an accommodation is assigned to a student, they should have opportunities to use the accommodation in the instructional setting prior to use on the assessment.



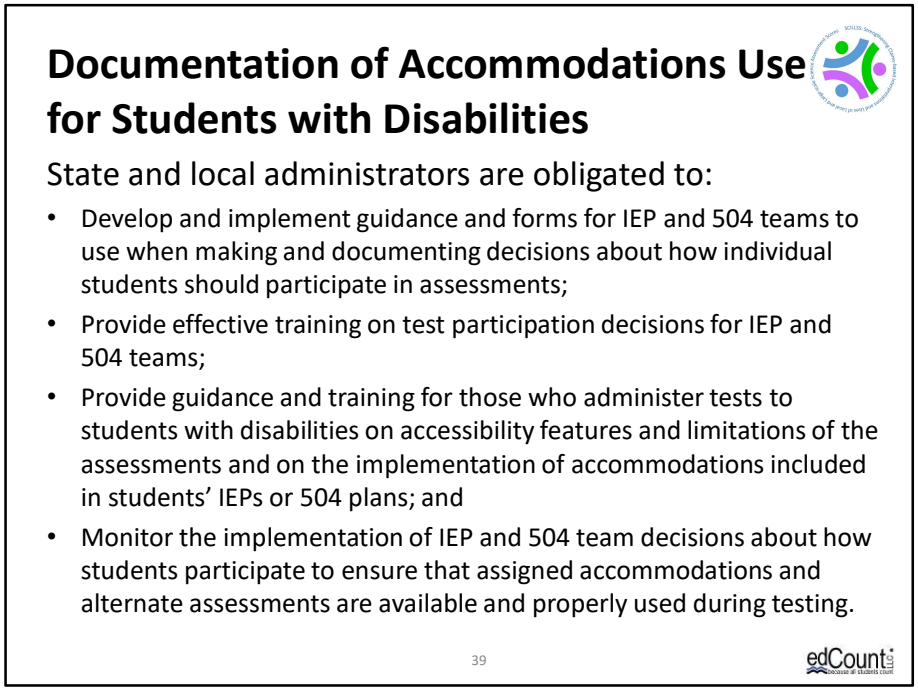
IEP and 504 teams must make test participation decisions for each individual student with an IEP or 504 plan. Every decision about how a student with disabilities participates in an assessment must be made on a case-by-case basis. No student should be assigned to an assessment or an accommodation based solely on his or her disability status or type.

The decision these teams make is how, not whether, a student participates in assessments. Students can participate in assessments:

- without accommodations,
- with one or more accommodations, or
- via an alternate assessment meant to measure the same knowledge and skills as the assessment other students take.

How a student participates in large-scale assessments should reflect how a student participates in instruction and classroom assessments. Accommodations should be familiar; a student should never experience a particular accommodation for the first time when they are taking a test.

Alternate assessments are subject to all of the same validity evidence expectations as assessments meant for students in the general student population. They must measure the same knowledge and skills even if expectations for how students demonstrate their knowledge and skills is very different. We encourage educators to go over all validity questions in the other chapters in this series again with a specific focus on alternate assessments.



State and local administrators are obligated to support local IEP and 504 teams in making appropriate decisions by about how students with disabilities participate in assessments. These obligations, which should be well-documented, include:

- Developing and implementing guidance and forms for IEP and 504 teams to use when making and documenting decisions about how individual students should participate in assessments;

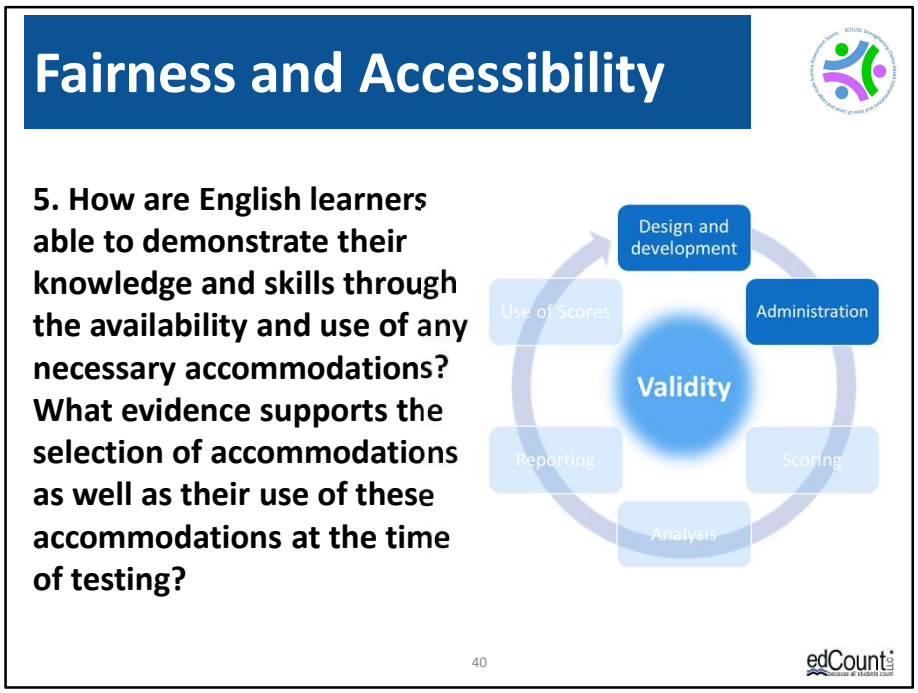
- Providing effective training on test participation decisions for IEP and 504 teams;

- Providing guidance and training for those who administer tests to students with disabilities on accessibility features and limitations of the assessments and on the implementation of accommodations included in students' IEPs or 504 plans; and

- Monitoring the implementation of IEP and 504 team decisions about how students participate to ensure that assigned accommodations and alternate assessments are available and properly used during testing.

The monitoring obligation is cited in Standard 3.10 of our professional standards and means that state and local administrators must have protocols and a process in place for

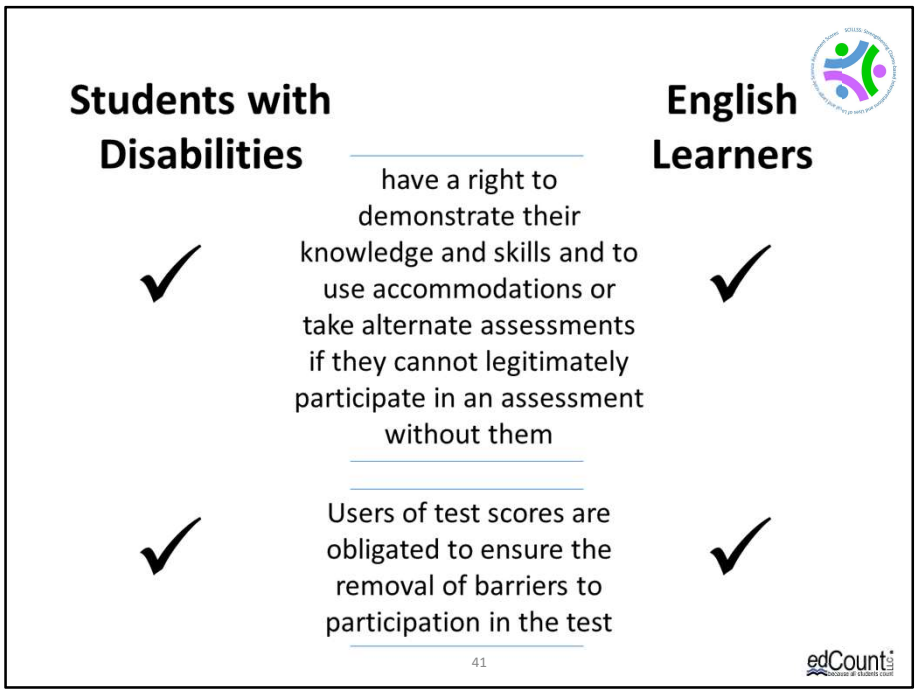
checking on the appropriate implementation of IEP and 504 team decisions about how students participate in the assessments. Not providing the accommodations or the opportunity to take an alternate assessment that an IEP or 504 team has determined is necessary for a student undermines the meaning of that student's score.



Our fifth validity question in this chapter is:

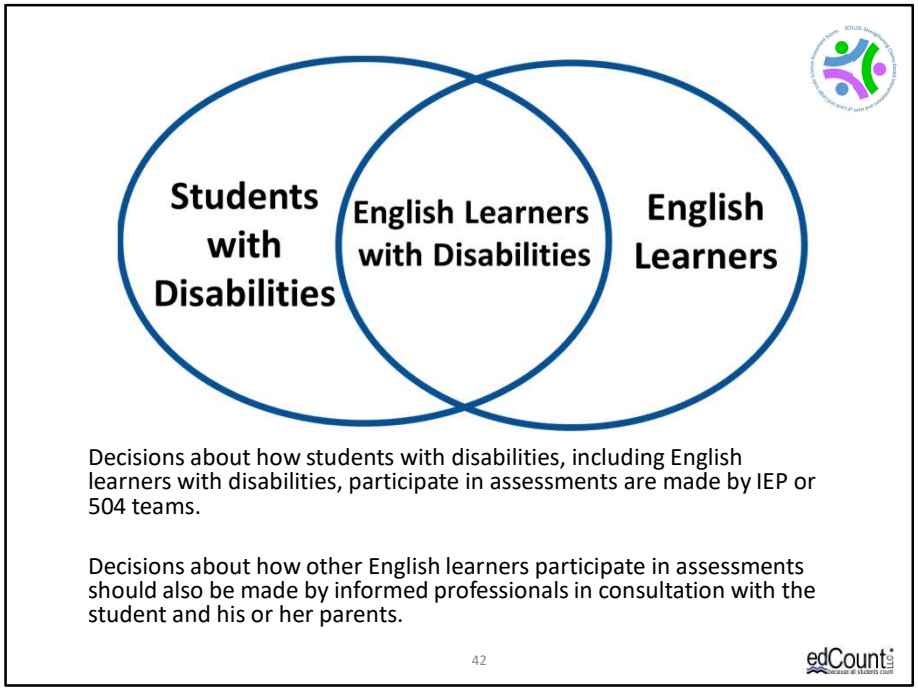
How are English learners able to demonstrate their knowledge and skills through the availability and use of any necessary accommodations? What evidence supports the selection of accommodations as well as their use of these accommodations at the time of testing?

Evidence related to this question comes from the design and development and the administration phases of the testing life cycle.



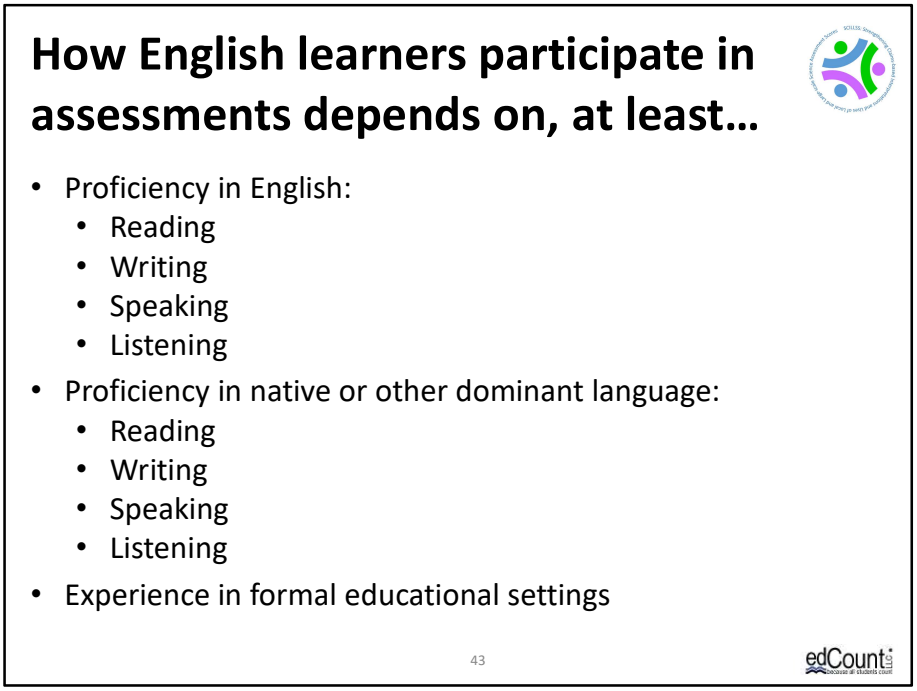
In many ways, the validity evidence for this question is similar to the validity evidence for our fourth question. The two questions are parallel, with the fourth question addressing the needs of students with disabilities and the fifth question addressing the needs of English learners.

Many of the basic concepts apply to students in both groups. Like students with disabilities, English learners have a right to demonstrate their knowledge and skills and to use accommodations or take alternate assessments if they cannot legitimately participate in an assessment without them. Those who use the scores from a test are obligated to ensure the removal of barriers to English learners' participation in that test.



Unlike students with disabilities, English learners may not have a team of educators, parents, and other informed adults to help make participation decisions like IEP and 504 teams make. In addition, the array of accommodations options available to students with disabilities are not the same as the accommodations options that are appropriate for English learners. The exception is for English learners who are also students with disabilities. For these students, accommodations and alternate assessments that address their disabilities and the accommodations and alternate assessments that address their linguistic needs are all available options for the IEP or 504 team to consider.

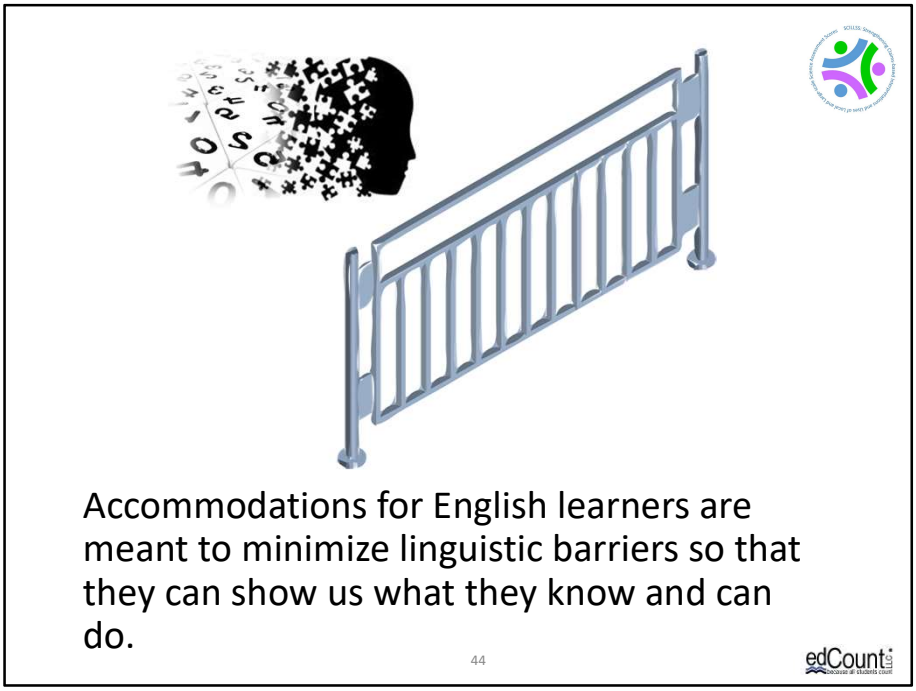
In all other cases, professionals with expertise in the content area being assessed and professionals with expertise in supporting students' English language acquisition should join with the student and the student's parents to make coordinated decisions for each individual student on how that student should best participate in assessments.



English learners are, by definition, students who are not yet proficient enough in English to participate fully in the discourse in school settings where English is the language of instruction, learning, and assessment. While they are learning academic content and skills, they must be supported in acquiring English proficiency in reading, writing, speaking, and listening. Some English learners may be proficient in another language in these areas and some are not. It is inappropriate to assume that an English learner who is a native speaker of, for example, Spanish, is proficient in reading and writing in Spanish unless that has been tested and confirmed. Further, if an English learner is learning academic content in English, it is inappropriate to assume that simply translating a test into his or her native language will remove barriers to that student's participation in the test. The student may not know the language of mathematics or science in his or her native language, either.

In addition, an English learner may be very experienced in formal educational settings or, if the student is a newcomer, may not be experienced in such settings. This has implications for how he or she understands and can best participate in assessments.

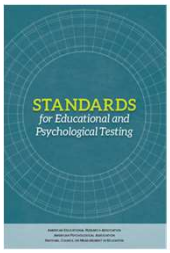
Thus, those who make decisions about how an English learner participates in an assessment must understand, among other things, the student's levels of English language proficiency, proficiency in his or her native language, and educational experiences when making participation decisions.



Appropriate accommodations options for English learners are those that minimize linguistic barriers. What poses a linguistic barrier varies across students and the decision about whether to use accommodations and which accommodations to use must be for each individual student, not for English learners as a group or based solely on English language proficiency assessment scores.

Accommodations for English learners may include, for example, glossaries that define selected words on the test in English or in another language if the student reads proficiently in that language. Presenting an English learner with a dual-language dictionary for use during testing may not be helpful unless that student is experienced with such dictionaries outside of testing and is proficient enough in the other language to read and understand the entries. Extra time to take the test may be helpful but may also frustrate the student and hinder test performance. Those who know the student and have the necessary professional expertise can help make the right decisions.

Our Professional Standards: Testing in Other Languages



Standard 3.12: "When a test is translated and adapted from one language to another, test developers and/or test users are responsible for describing the methods used in establishing the adequacy of the adaptation and documenting empirical or logical evidence for the validity of test score interpretations for the intended uses."

(AERA, APA, & NCME, 2014, p. 68)

Standard 3.13: "A test should be administered in the language that is most relevant and appropriate to the test purpose."

(AERA, APA, & NCME, 2014, p. 69)

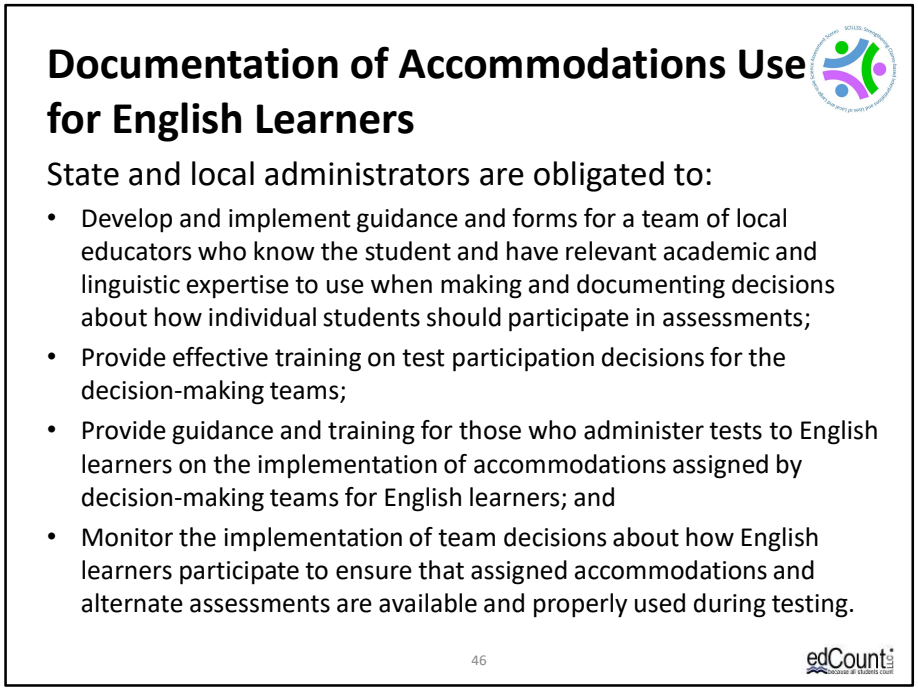
45

If a testing publisher offers tests or portions of tests in languages other than English that are intended to yield scores that are comparable to those from the English versions, our professional standards include expectations regarding test translation or adaptation. For example:

Standard 3.12: When a test is translated and adapted from one language to another, test developers and/or test users are responsible for describing the methods used in establishing the adequacy of the adaptation and documenting empirical or logical evidence for the validity of test score interpretations for the intended uses.

Standard 3.13: A test should be administered in the language that is most relevant and appropriate to the test purpose.

Standard 3.13 underscores that decisions about the language in which a student takes a test depends on what the test is meant to measure as well as the student's proficiency in both English and the other language that may be relevant to him or her.



In addition to the design and development evidence from a test publisher, state and local administrators are obligated to support local educators in making appropriate decisions about how English learners participate in assessments. As is the case for the decisions for students with disabilities, these obligations, which should be well-documented, include:

Developing and implementing guidance and forms for a team of local educators who know the student and have relevant academic and linguistic expertise to use when making and documenting decisions about how individual students should participate in assessments;

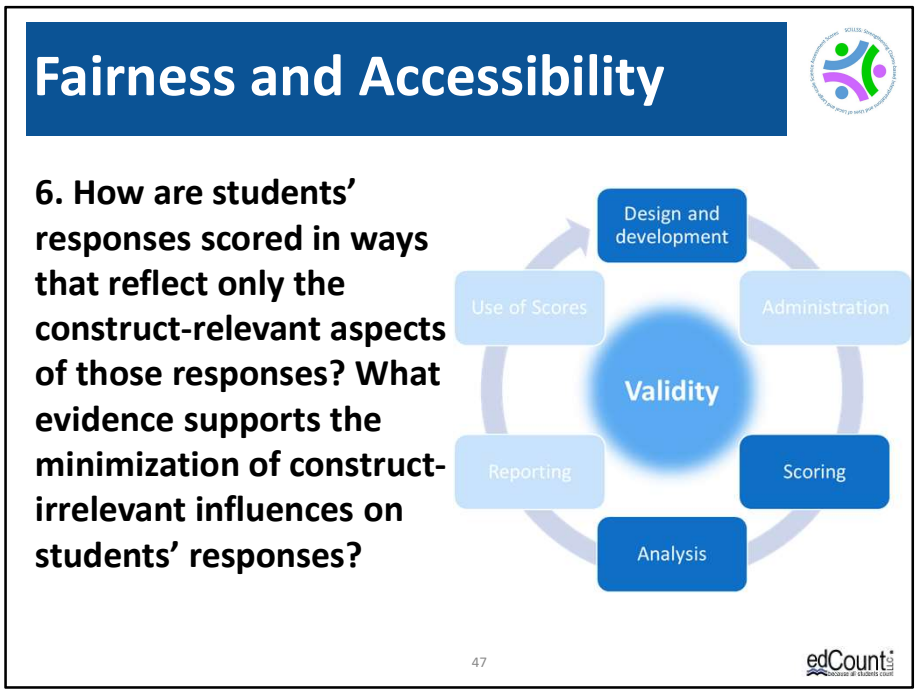
Providing effective training on test participation decisions for the decision-making teams;

Providing guidance and training for those who administer tests to English learners on the implementation of accommodations assigned by decision-making teams for English learners; and

Monitoring the implementation of team decisions about how English learners participate to ensure that assigned accommodations and alternate assessments are available and properly used during testing.

As is the case for monitoring assessment participation for students with disabilities, state

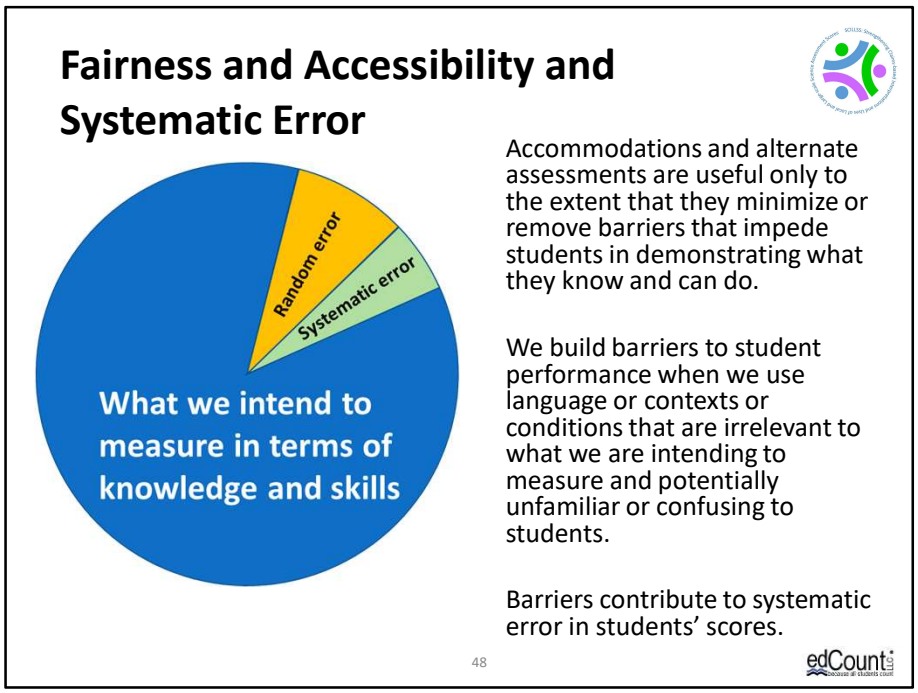
and local administrators must have protocols and a process for checking on the appropriate implementation of decisions about how English learners participate in assessments.



Our sixth question in this chapter is:

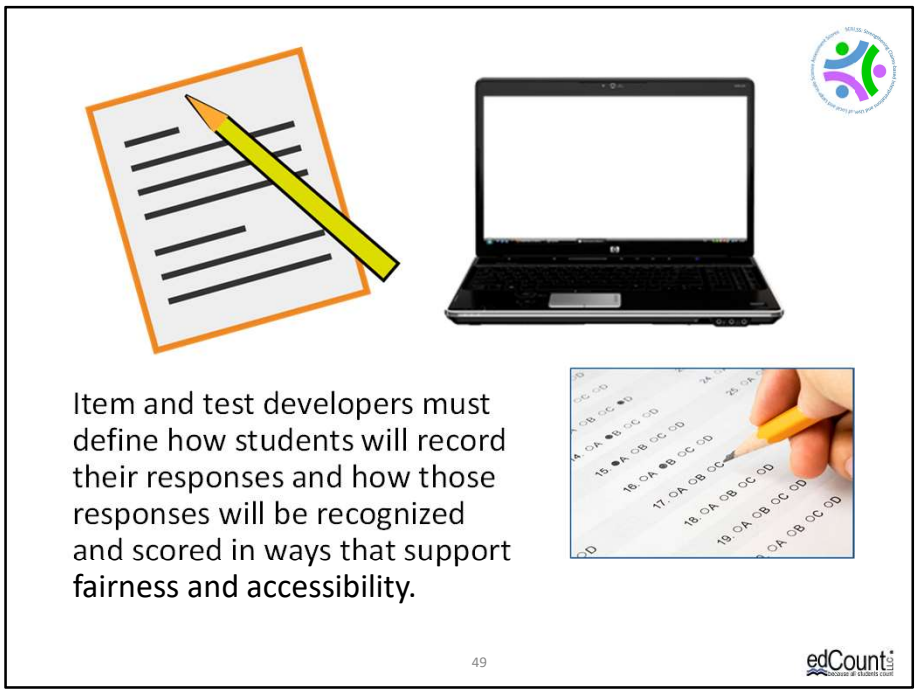
How are students' responses scored in ways that reflect only the construct-relevant aspects of those responses? What evidence supports the minimization of construct-irrelevant influences on students' responses?

Evidence for this question comes from the design and development, scoring, and analysis phases of the test life cycle.



In our third chapter in this series, we focused on issues related to comparability and reliability/precision. We learned that every test score reflects something about what we are intending to measure, which is the construct-relevant portion of the score, and also some error, which is construct-irrelevant. As we reminded ourselves earlier in this chapter, error in test scores encompasses random error, which reduces score reliability, and systematic error, which reduces validity. Systematic error can stem from characteristics of a test, how it is administered and scored, and how the scores are reported and used.

All concerns about fairness and accessibility, like all other concerns about score validity, relate to our obligations to maximize the construct-relevant portion of scores and minimize the construct-irrelevant portion. Accommodations and alternate assessments are useful only to the extent that they minimize or remove barriers that impede students in demonstrating what they know and can do. We build barriers to student performance when we use language or contexts or conditions that are irrelevant to what we are intending to measure and potentially unfamiliar or confusing to students. Barriers contribute to systematic error in students' scores.

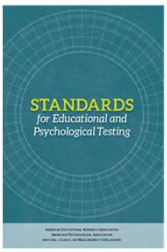


During the design and development phase, test developers must consider how students will record responses and how those responses will be recognized and assigned value during the scoring process. This is a fundamental application of Universal Design. Most students will be able to fill in a bubble or click on an option or use a pencil or a keyboard to write a response. Most can read and write proficiently in English and understand most of the words and contexts in test questions. Some students cannot do these things and, therefore, face barriers in showing us what they know. Item writers and those creating the test booklets, answer documents, and computer-based testing platforms must allow for a range of ways for students to respond and to capture those responses for scoring. In addition, they must describe how scorers – even if those scorers are machines – are to recognize and value students' responses. Those doing the scoring, must demonstrate that they are adhering to those scoring rules and not introducing bias or other forms of error by, for example, assigning lower scores when a student's response includes some misspellings or is captured using assistive technology when spelling and typing are not what the test is intended to measure.

For large-scale assessments such as those a state or large school district requires to be administered to all students each year, psychometricians – the statisticians who specialize in measuring constructs like those that educational assessments target – usually conduct analysis of students' responses to individual test questions to determine whether there are differences in how students in different groups perform. These analyses compare the performance on individual test items of students in different groups, such as students with disabilities and students who have not been identified as having disabilities, who have the same total test scores. The assumption is that students who score the same on the test overall are likely to perform similarly on individual items.

When psychometricians find differences in item performance, they identify those items for further review. Test vendors are obligated to analyze items for potential sources of bias and evaluate what may underlie group differences. Such analyses can include reviews by experts who understand the content and skills the assessment is targeting and studies that directly examine how students interact with and respond to test items.

Our Professional Standards: Evaluating Group Differences in Item and Test Performance



Standard 3.6: "Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretations for intended uses for individuals from those subgroups. What constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws."

(AERA, APA, & NCME, 2014, p. 65)

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."

(AERA, APA, & NCME, 2014, p. 70)

51

edCount

Our professional standards cite several obligations for vendors to evaluate differences in test scores across student populations and provide information about score interpretations for individuals in specific groups, including:

Standard 3.6: Where credible evidence indicates that test scores may differ in meaning for relevant subgroups in the intended examinee population, test developers and/or users are responsible for examining the evidence for validity of score interpretations for intended uses for individuals from those subgroups. What constitutes a significant difference in subgroup scores and what actions are taken in response to such differences may be defined by applicable laws.

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.

Test Publishers' Obligations for Documenting Fair and Accessible Scoring Processes

- Descriptions of how the scoring processes, including rubrics for scoring constructed-response items, have been designed to recognize and appropriately value construct-relevant aspects of students' responses and minimize the influence of construct-irrelevant aspects;
- Descriptions of how the scoring process maintains fidelity with its design and minimized the influence of construct-irrelevant aspects of student responses as well as other sources of error. This evidence should include details about who the scorers are, how they are trained, and how they and other components of the scoring process are monitored and evaluated;
- Results of DIF and other analyses designed to identify and explain group differences in test or item performance.

52

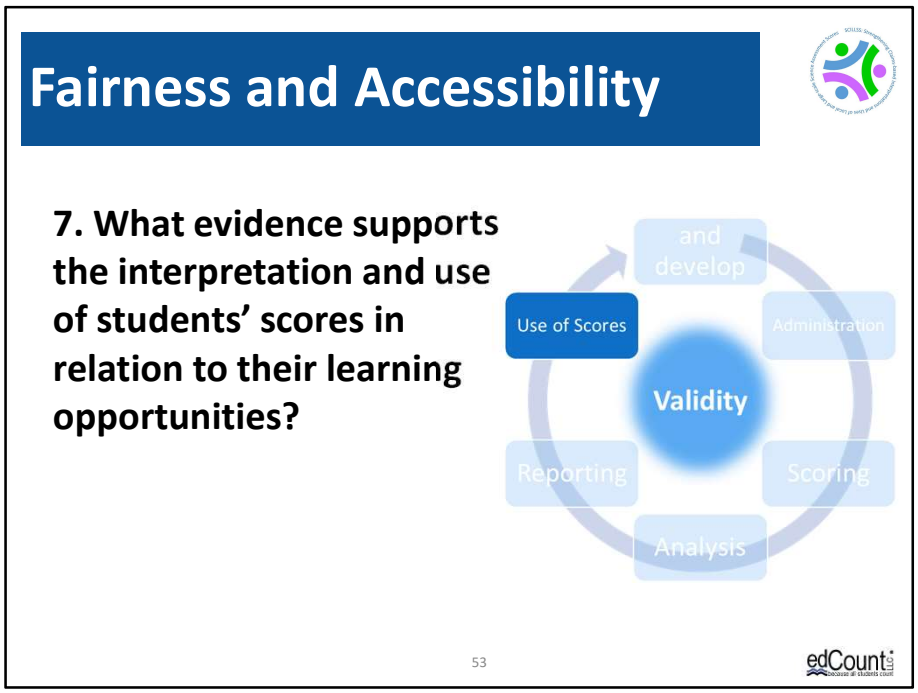
edCount

Evidence related to fair and accessible scoring processes can take several forms. Test publishers must include the following types of information and typically do so in technical manuals that describe test design and development processes, scoring processes, and analyses of differential performance across specific student groups.

- Descriptions of how the scoring processes, including rubrics for scoring constructed-response items, have been designed to recognize and appropriately value construct-relevant aspects of students' responses and minimize the influence of construct-irrelevant aspects;

- Descriptions of how the scoring process maintains fidelity with its design and minimizes the influence of construct-irrelevant aspects of student responses as well as other sources of error. This evidence should include details about who the scorers are, how they are trained, and how they and other components of the scoring process are monitored and evaluated;

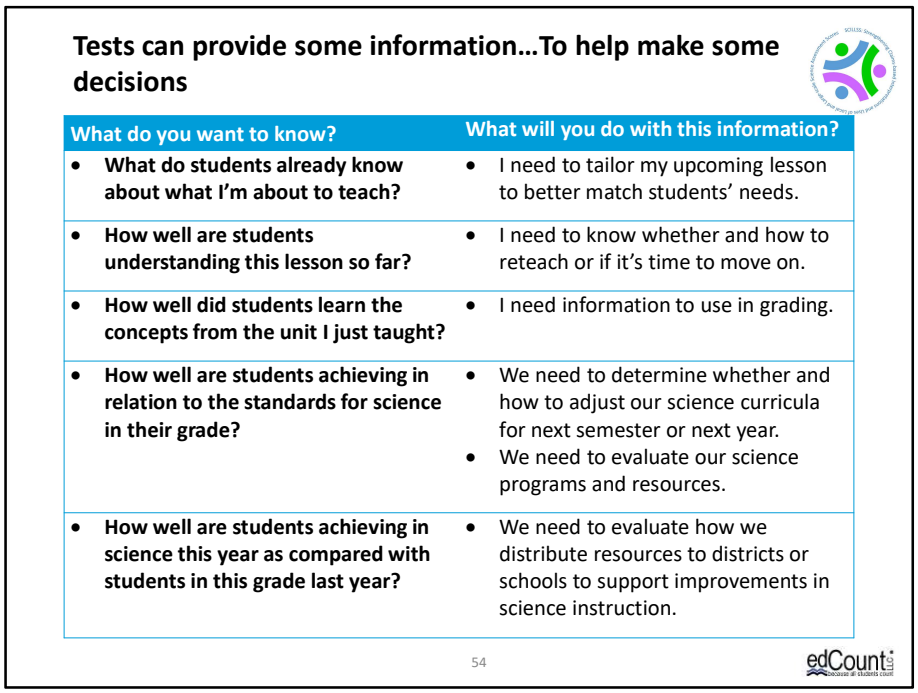
- Results of r analyses designed to identify and explain group differences in test or item performance.



Our final validity question in this chapter is:

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

Evidence related to this question comes from the score use phase of the test life cycle.

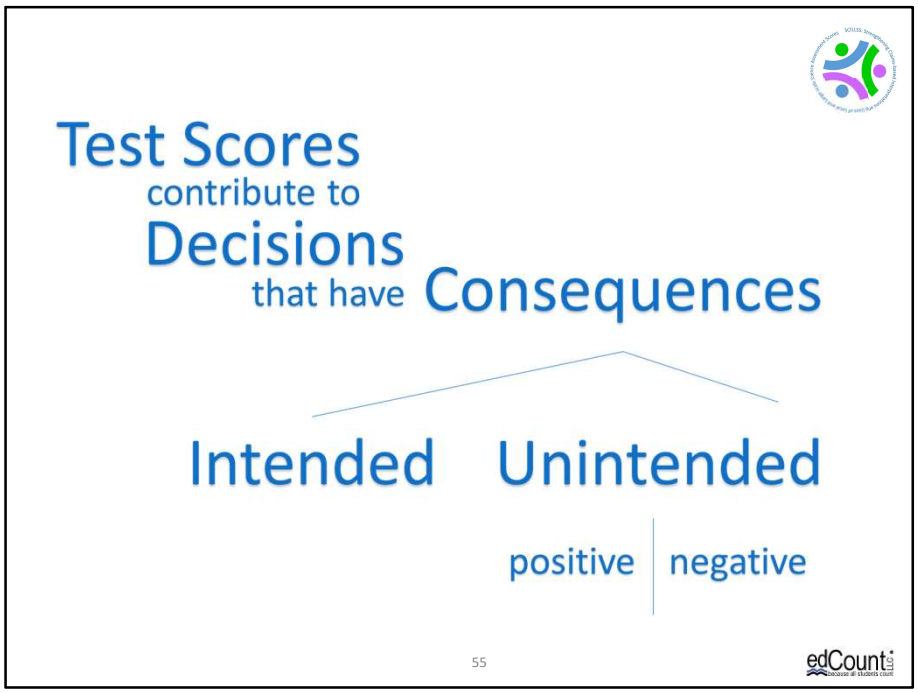


Our fifth and final chapter in this workbook series will focus entirely on validity questions related to the consequences associated with how assessment scores are interpreted and used for particular purposes. Here, we specifically target the fairness and accessibility issues in how test scores are interpreted and used.

In our first chapter, we identified some common purposes for using test scores. We framed those purposes around the questions, "what do you want to know" and "what will you do with this information?" This pair of questions helps us to see that every test is meant to elicit some information that we intend to use for a particular purpose. If we cannot clearly identify what we want to know and why we need that information in making a decision, we should not give a test.

What do you want to know?
What will you do with this information?
What do students already know about what I'm about to teach?
I need to tailor my upcoming lesson to better match students' needs.
How well are students understanding this lesson so far?
I need to know whether and how to reteach or if it's time to move on.
How well did students learn the concepts from the unit I just taught?
I need information to use in grading.

How well are students achieving in relation to the standards for science in their grade?
We need to determine whether and how to adjust our science curricula for next semester or next year.
We need to evaluate our science programs and resources.
How well are students achieving in science this year as compared with students in this grade last year?
We need to evaluate how we distribute resources to districts or schools to support improvements in science instruction.



We use test scores to inform decisions for individual students and for groups of students and every decision is associated with a consequence. What instruction will a student experience next? What grade will a student get? Will a student get into a program or receive a service? Will an instructional program be seen as effective? Will a school or its teachers get rewarded for improving student achievement or lose some degree of autonomy if scores go down?

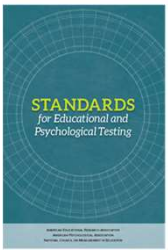
Some consequences are intended, such as when a test is designed to measure a student's readiness for a particular instructional sequence and the scores are used to inform that instructional decision. Of course, the interpretation and use of scores for this purpose still requires evidence of the types we've discussed in all of the chapters of this workbook.

Some consequences are unintended and some of those are negative. For example, if the results of a test are used for accountability purposes and individuals within a school decide to adjust their students' responses on the test, that would be the negative, unintended consequence of testing also known as a form of cheating.

However, other negative, unintended consequences may not be so obvious. If tests are not fair and accessible for all students, then the scores for any particular student, especially if that student has disabilities or comes from a linguistic, cultural, or socioeconomic

background that was not appropriately considered during test design and development, administration, or scoring, may not be interpretable. Likewise, the separate or disaggregated scores for different groups of students may lack meaning.

Our Professional Standards: Fairness and Accessibility Issues for Testing Consequences



Standard 12.10: "In educational settings, a decision or characterization that will have a major impact on a student should take into account not just scores from a single test but other relevant information."
(AERA, APA, & NCME, 2014, p. 198)

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."

(AERA, APA, & NCME, 2014, p. 71)

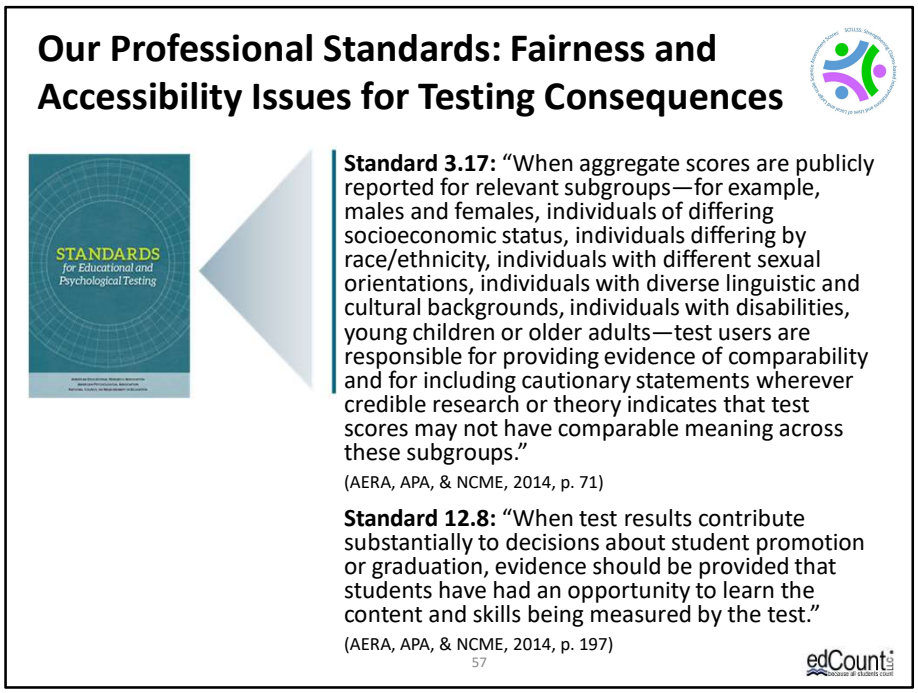
56

edCount

For these reasons, and because all scores reflect error, our professional standards state very clearly that no decision should be made solely on the basis of test scores. For example:

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

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)



Further,

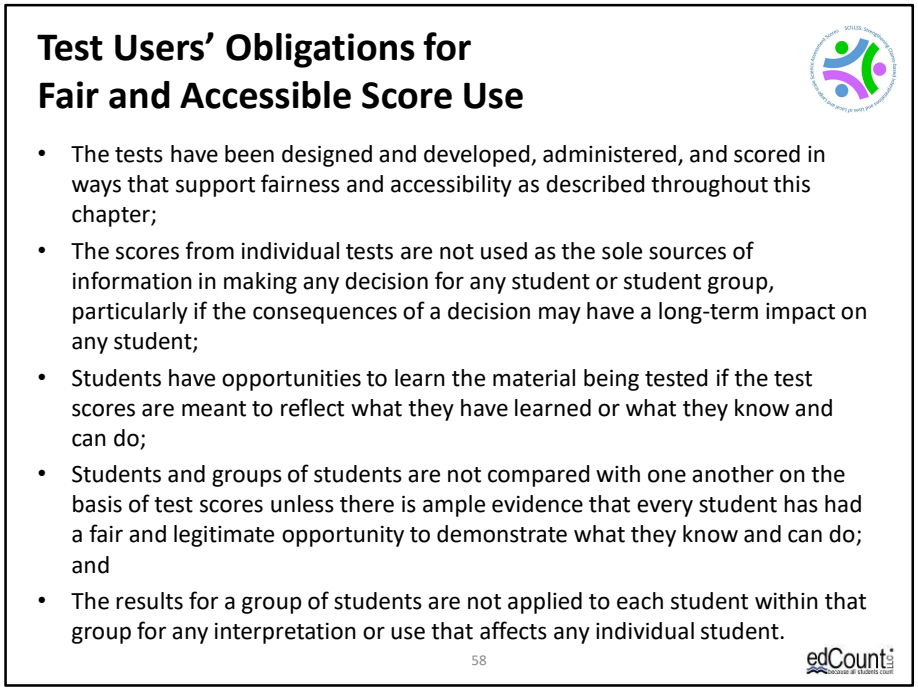
Standard 3.17: When aggregate scores are publicly reported for relevant subgroups—for example, males and females, individuals of differing socioeconomic status, individuals differing by race/ethnicity, individuals with different sexual orientations, individuals with diverse linguistic and cultural backgrounds, individuals with disabilities, young children or older adults—test users are responsible for providing evidence of comparability and for including cautionary statements wherever credible research or theory indicates that test scores may not have comparable meaning across these subgroups. (p. 71)

These prohibitions against using test scores as sole indicators in making decisions apply to scores for all students and all student groups, but require additional attention when the decisions affect students with disabilities and students from non-white, non-English linguistic, cultural, or socioeconomic backgrounds.

In addition, test users must have reasonable evidence that students taking a test meant to reflect what they have learned have had a legitimate opportunity to learn the material.

Standard 12.8: When test results contribute substantially to decisions about student promotion or graduation, evidence should be provided that students have had an

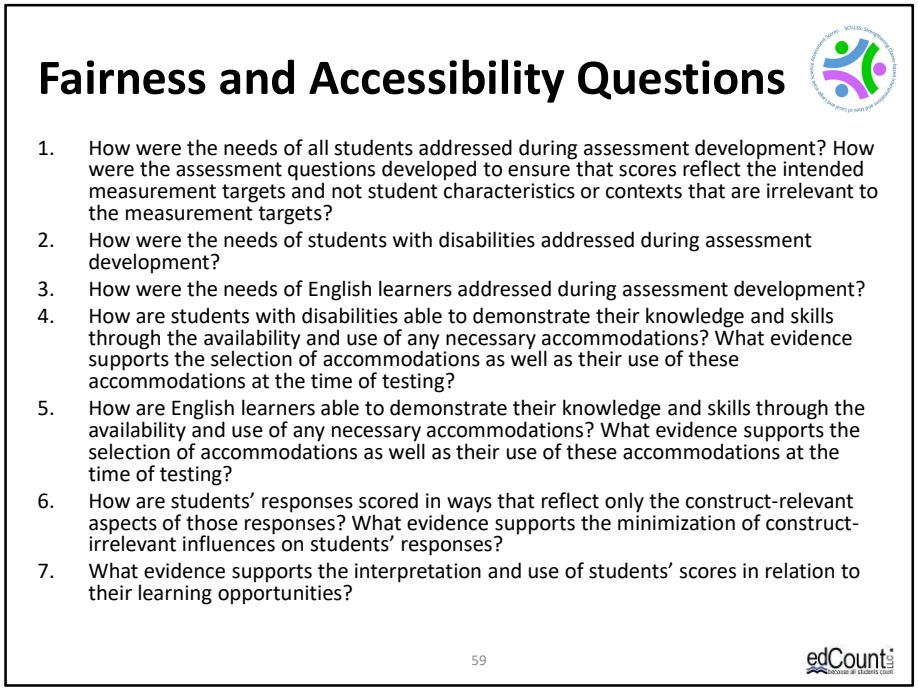
opportunity to learn the content and skills being measured by the test. (p. 197)



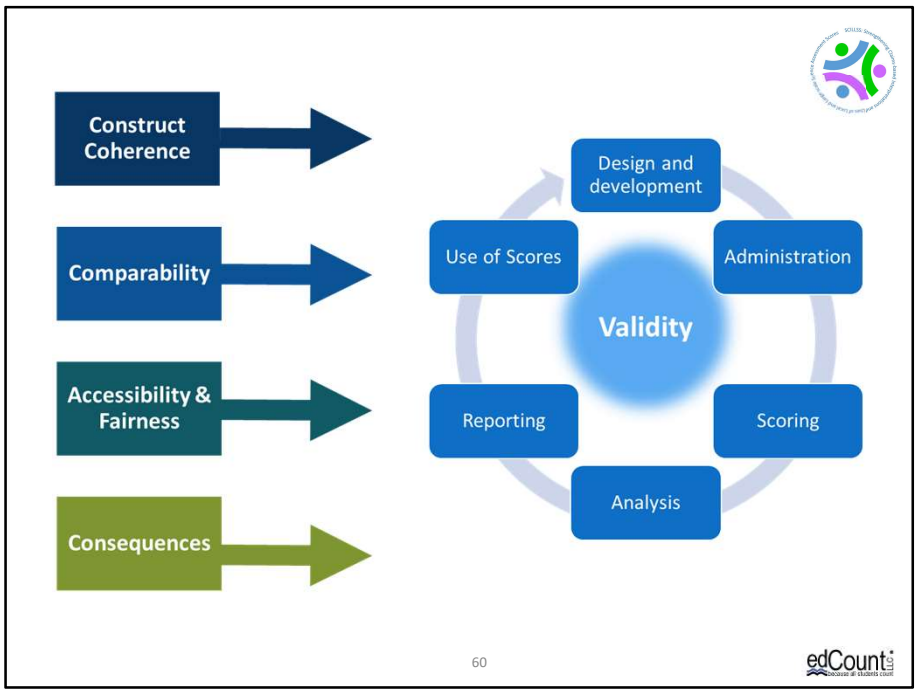
Individuals or agencies that intend to use test scores to make decisions about individual students or groups of students must gather and consider evidence that:

- The tests have been designed and developed, administered, and scored in ways that support fairness and accessibility as described throughout this chapter;
- The scores from individual tests are not used as the sole sources of information in making any decision for any student or student group, particularly if the consequences of a decision may have a long-term impact on any student;
- Students have opportunities to learn the material being tested if the test scores are meant to reflect what they have learned or what they know and can do;
- Students and groups of students are not compared with one another on the basis of test scores unless there is ample evidence that every student has had a fair and legitimate opportunity to demonstrate what they know and can do; and
- The results for a group of students are not applied to each student within that group for any interpretation or use that affects any individual student.

Although some relevant evidence may come from test vendors, these obligations lie with those using test scores.



We have reached the end of our seven questions in this chapter. The key takeaway from this chapter is that test scores should never be used in isolation for any decision. Every student has a right to learn and to demonstrate what they have learned so that they benefit from appropriate decisions about the opportunities available to them next instructionally and over the course of their educational careers and beyond. Those using test scores have the primary obligation to gather and weigh evidence regarding fairness and accessibility and to ensure that the evidence that they obtain from vendors meets professional standards for fairness and accessibility.



Thus far in this series, we have addressed questions related to construct coherence, comparability and reliability/precision, and fairness and accessibility. In the final chapter, we will address questions related to consequences of test use.

Resources and Additional Information

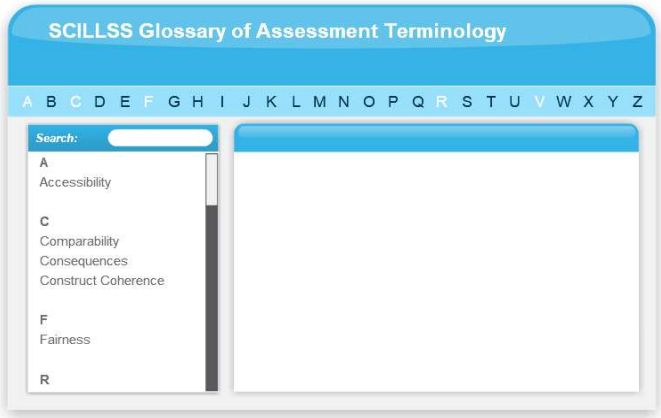
61

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 third 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

Web links

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

- National Center on Educational Outcomes
- Council of Chief State School Officers
- Standards and Assessment Approaches for Students with Disabilities Using a Validity Argument
- The Impact of Construct-Irrelevant Variance and Construct Under-Representation in Assessing Teachers' Coaching Competence
- Detecting Construct-Irrelevant Variance in an Open-Ended, Computerized Mathematics Task
- Smarter Balanced Assessment Consortium: Bias and Sensitivity Guidelines

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.

Cortiella, C. (2006). NCLB and IDEA: *What parents of students with disabilities need to know and do* [PDF]. Minneapolis, MN: University of Minnesota, National Center on Educational Outcomes. Retrieved December 31, 2018. Retrieved from https://nceo.umn.edu/docs/OnlinePubs/Parents.pdf