Creating and Evaluating Effective Educational Assessments

Chapter 1

Purposes and Uses of Assessment Scores

Validity

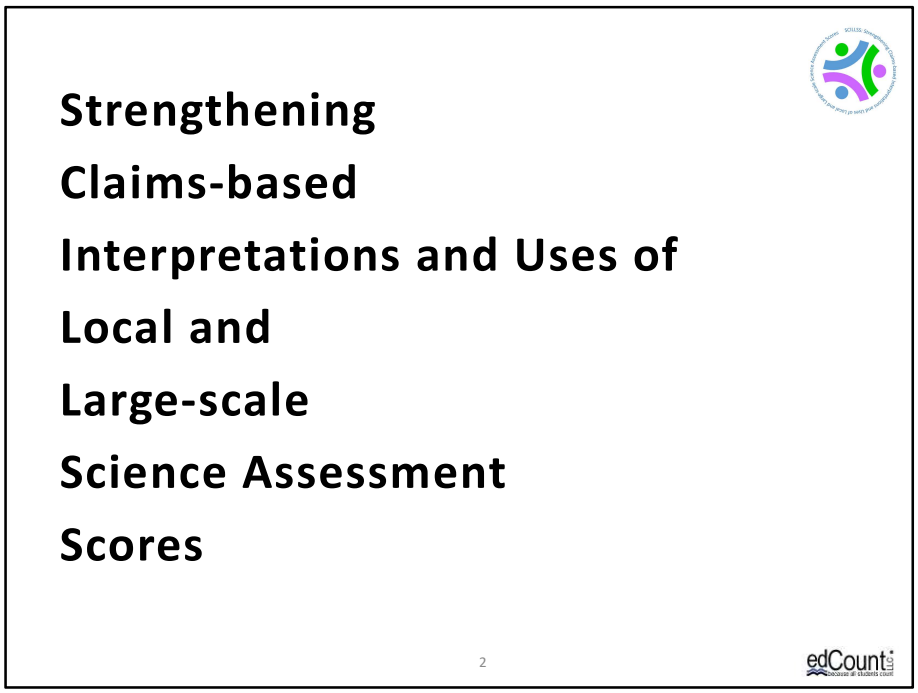
Validity Questions

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

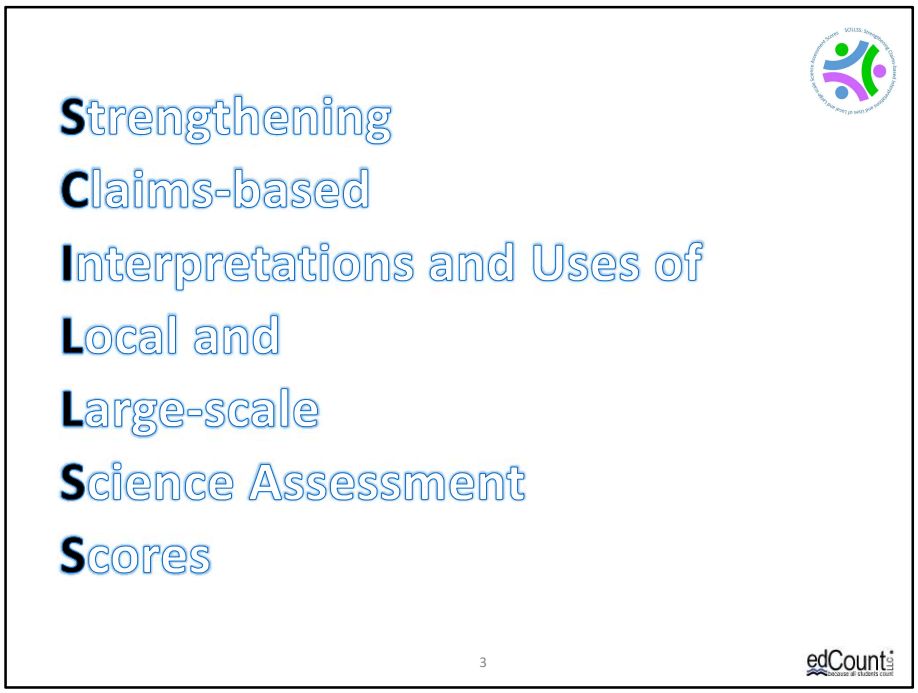
1

Welcome to the first 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.

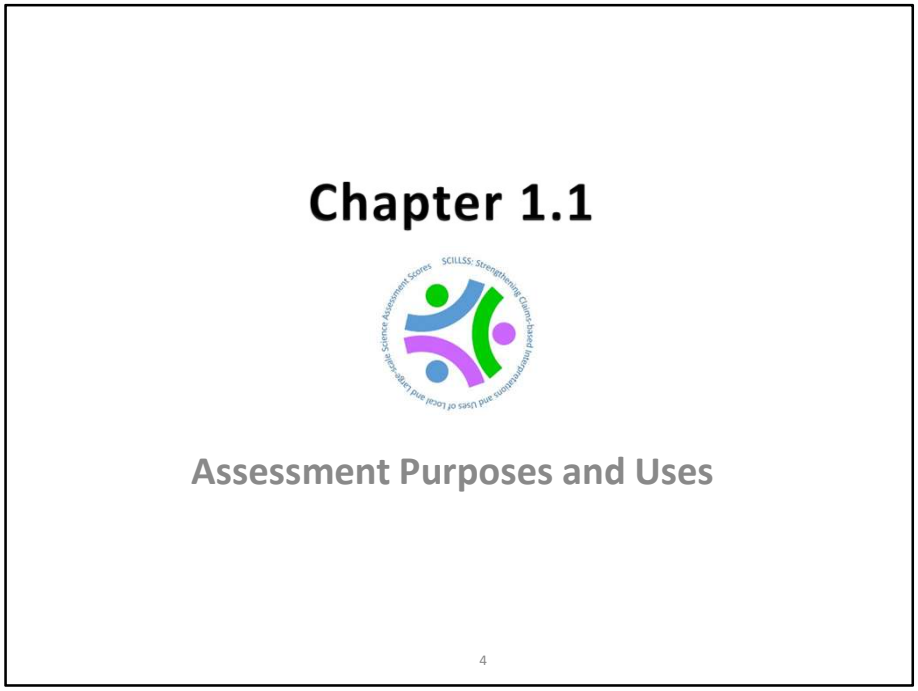
This digital workbook was developed by edCount, LLC, under the US Department of Education's Enhanced Assessment Grants Program, CFDA 84.368A.



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

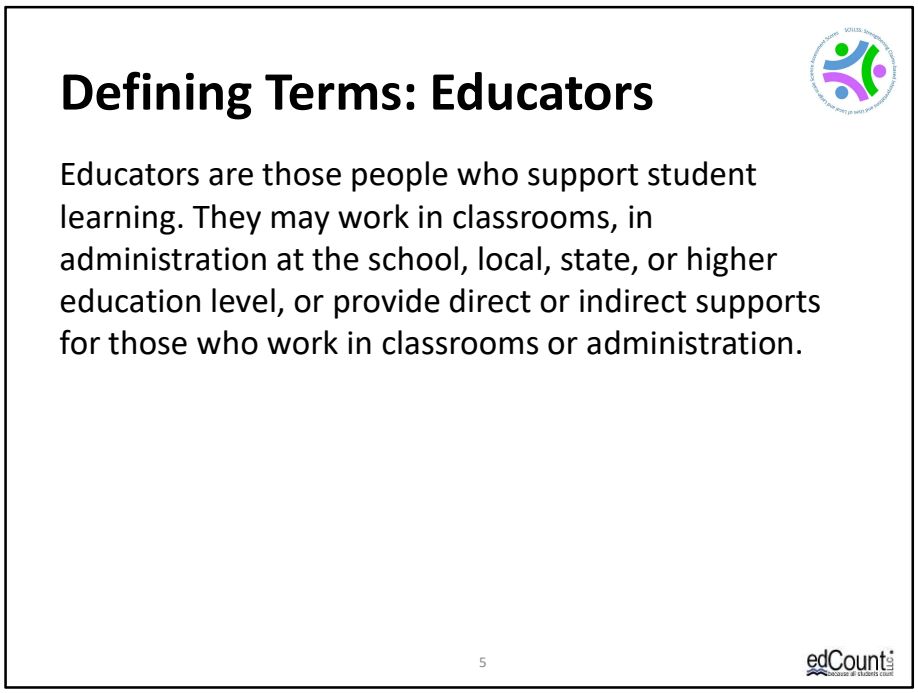


or its acronym, “SCILLSS.”

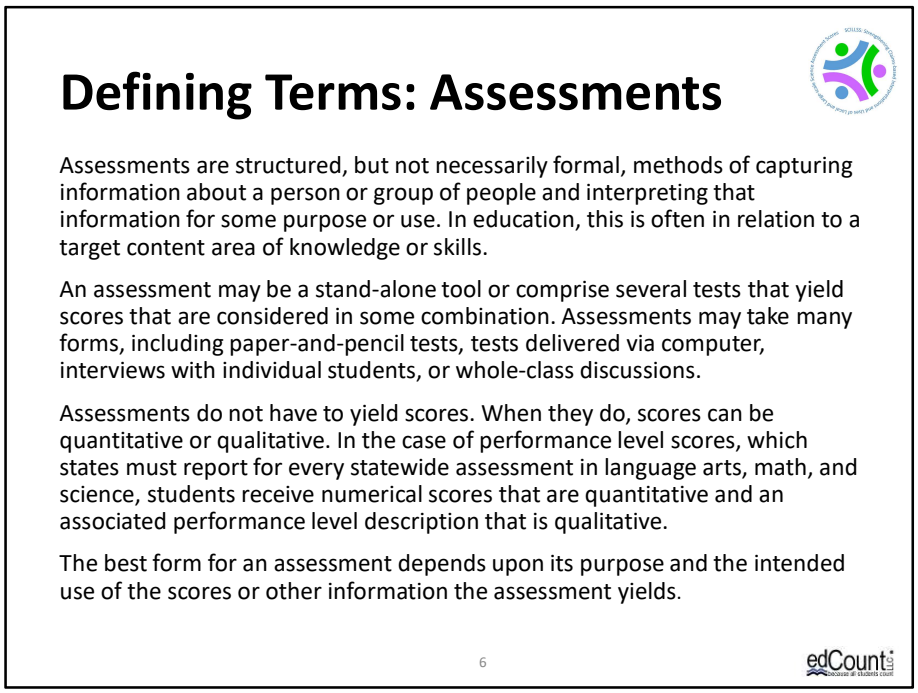


Chapter 1 of this series focuses on why we administer assessments of students' knowledge and skills and how we use assessment scores. We break down the phases of an assessment life cycle and lay the groundwork for examining what assessment scores mean and how they can be used appropriately.

In subsequent chapters, we will explore how to gather and evaluate evidence about the meaning and use of assessment scores.



Let's start by clarifying a couple of the terms we'll be using. First, we will often refer to "educators." By educators we mean people who support student learning. They may work in classrooms, in administration at the school, local, state, or higher education level, or provide direct or indirect supports for those who work in classrooms or administration.



Second, we will use the terms “assessment” and “test”, often interchangeably. By assessments we mean methods of capturing information about what a person or a group of people knows and can do in a particular content or skill area. An assessment may be a stand-alone tool or an assessment may comprise several tests with scores that are considered in some combination.

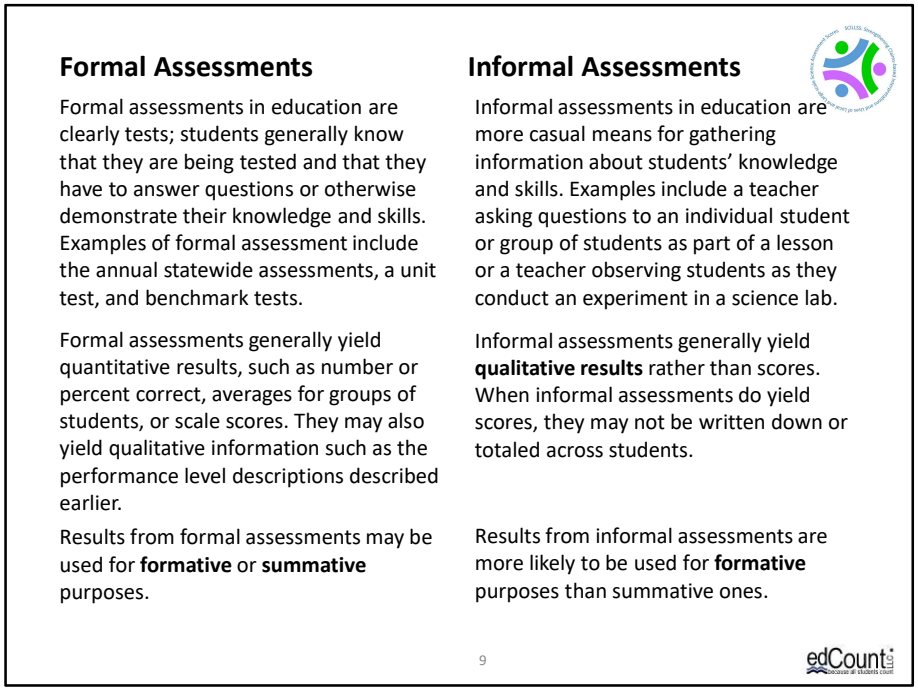
An assessment is not necessarily a formal or even a physical test and doesn’t always yield scores. Rather, an assessment can involve a teacher or even another student posing questions and considering the responses.

Some may prefer to think of a test as a means for gathering information and an assessment as the judgment based on the results of one or more tests. We consider any of these uses appropriate and will do our best to avoid confusion.

Much of what teachers do in their classrooms every day is assess what their students know and can do. Most of these assessments are informal and intended to be formative in purpose. That is, the results of the assessments, which are generally qualitative rather than numerical or letter grades, are used to guide next steps in instruction. For example, teachers may use exit tickets to take stock of students' knowledge and skills as part of their instructional planning for the following day. Or, they may ask questions to help formulate very next steps in the flow of a lesson.

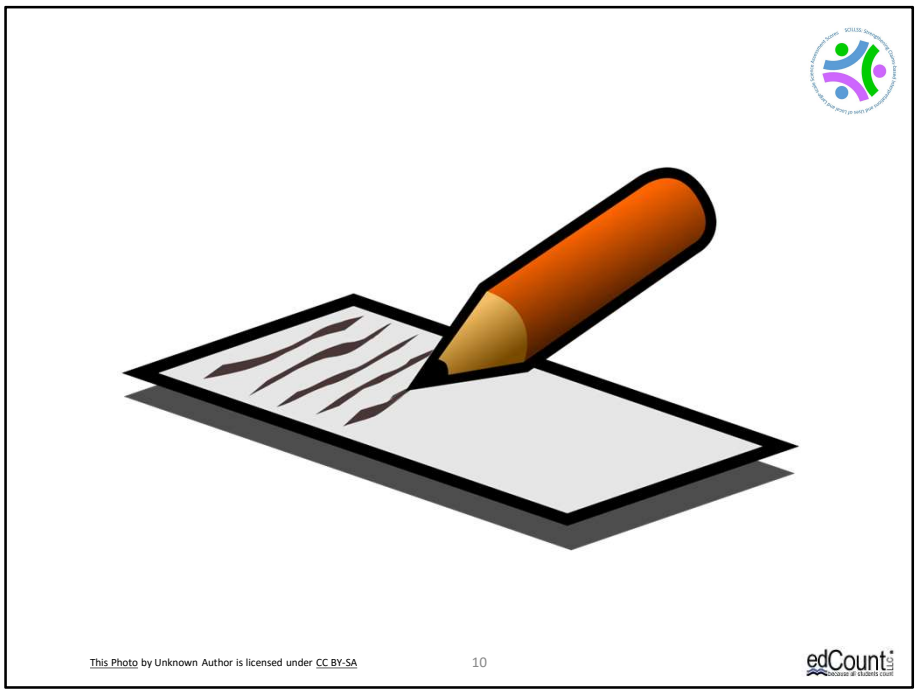
Although many of these teacher actions do not involve a formal test, per se, they all involve gathering and evaluating information and then using that information to make decisions. They all require teachers to understand what they want to know, identify an effective approach for gathering information related to this goal, apply that approach, and interpret the information they get in response.

Sometimes, assessments are formal, such as when a teacher distributes a quiz or a unit test or administers an assessment required by the school, the school district, or the state. In the cases of these required assessments, students' answers are sometimes aggregated – that is, totaled or averaged – at the classroom, grade, school, district, or state levels. Scores may be reported in numerical form and also in the form of performance levels, such as basic, proficient, or advanced. Aggregations of performance level information is typically reported as the percent of students scoring in a given level.

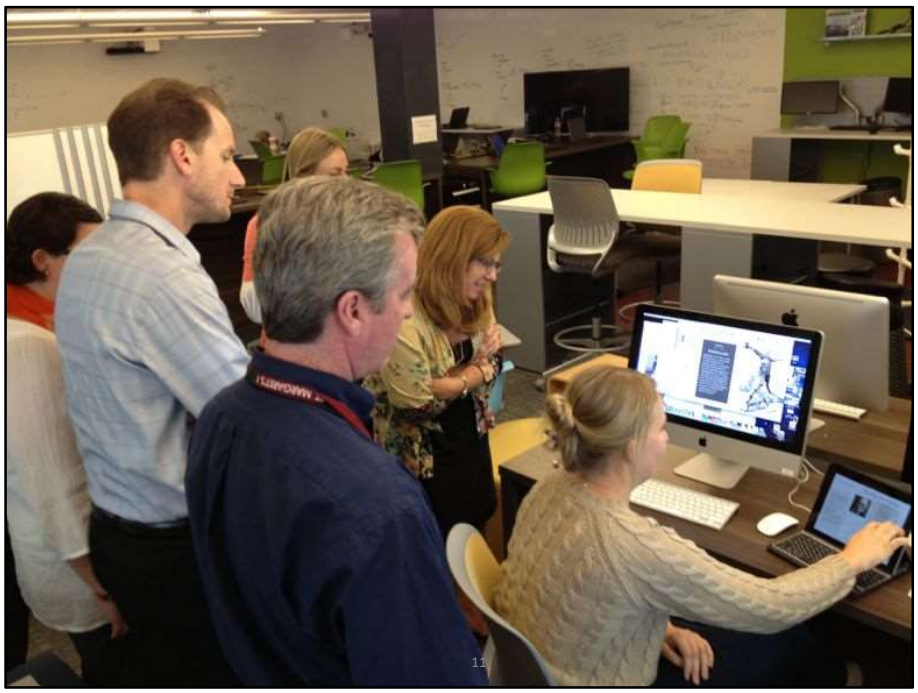


All assessments, formal or informal, developed and used within classrooms or required by those external to the classrooms, whether they yield results that are quantitative or qualitative, provide some type of information. In all cases, **the value of the assessment depends on the quality and usefulness of the information it provides**.

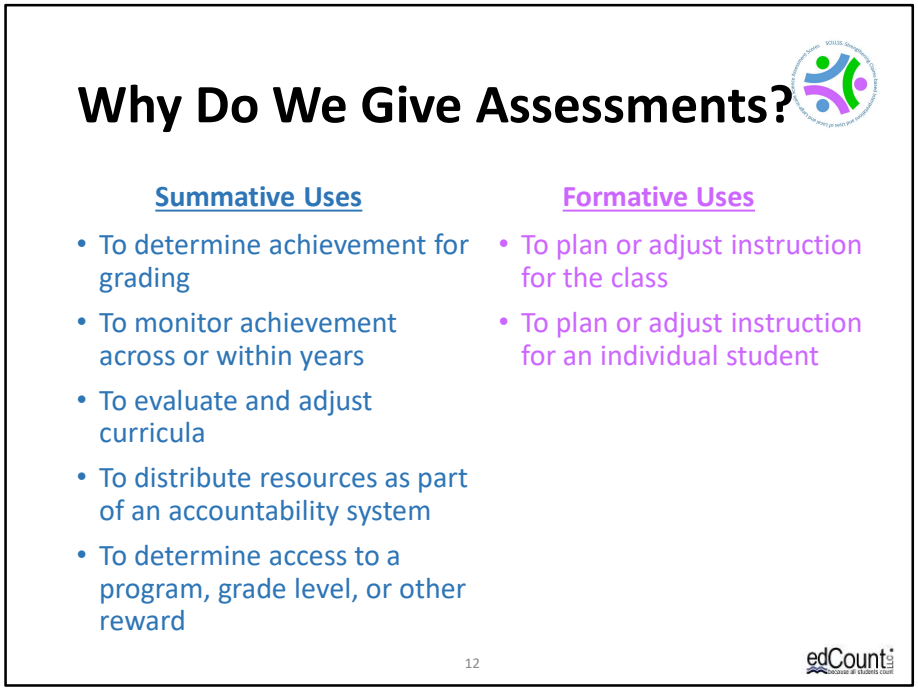
The purpose of this workbook is to help educators understand how to determine the quality and usefulness of the information each of their assessments provides. In particular, we target formal assessments that are meant to provide either formative or summative information: these assessments make up what we can call an "assessment system" at a school, district, or state level. Scores from these assessments are considered when making decisions that affect individual students or groups of students.



As noted previously, teachers assess their students frequently. Whether a teacher or administrator creates a formal test or chooses one from a vendor, he or she should consider what constitutes "good enough" evidence that the test is measuring what it is meant to measure, how it functions to gather information from students about what they know and can do, and how useful the scores are for supporting solid decisions. While we are not directly addressing strategies for developing classroom-based tests in this workbook, the same fundamental concepts about quality apply to these tests as well.



The very first step in building or adopting an assessment is to establish a clear purpose for doing so. Teachers may pose questions to take stock of what students already know at the beginning of a lesson or to determine whether students fully understand a concept or need additional instruction. A district or a state may require students to take specific tests for the purpose of monitoring achievement as part of its accountability system or to provide teachers with information meant to inform classroom instruction. A school counselor or psychologist may administer a test to an individual student to help diagnose specific learning challenges that may require tailored instruction or other accommodations in classrooms.



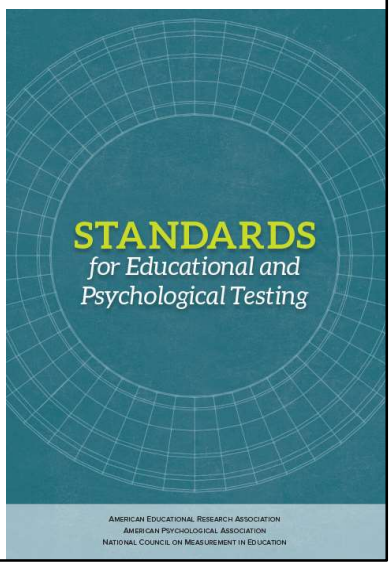
So, why do we give assessments? Here are some common reasons for giving tests. The column on the left includes purposes associated with tests teachers often given in their classrooms. The column on the right lists purposes that are more often associated with tests that are required by administrators outside of the classroom.

Tests must be designed for a given purpose. Although some tests may yield information that could be used for more than one purpose, a single test could never serve all of these purposes. What makes a test effective for informing instruction is entirely different from what makes a test effective for making high stakes decisions for accountability purposes.

Purposes and Uses of Assessment Scores

"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."

(AERA, APA, & NCME, 2014, p. 23)

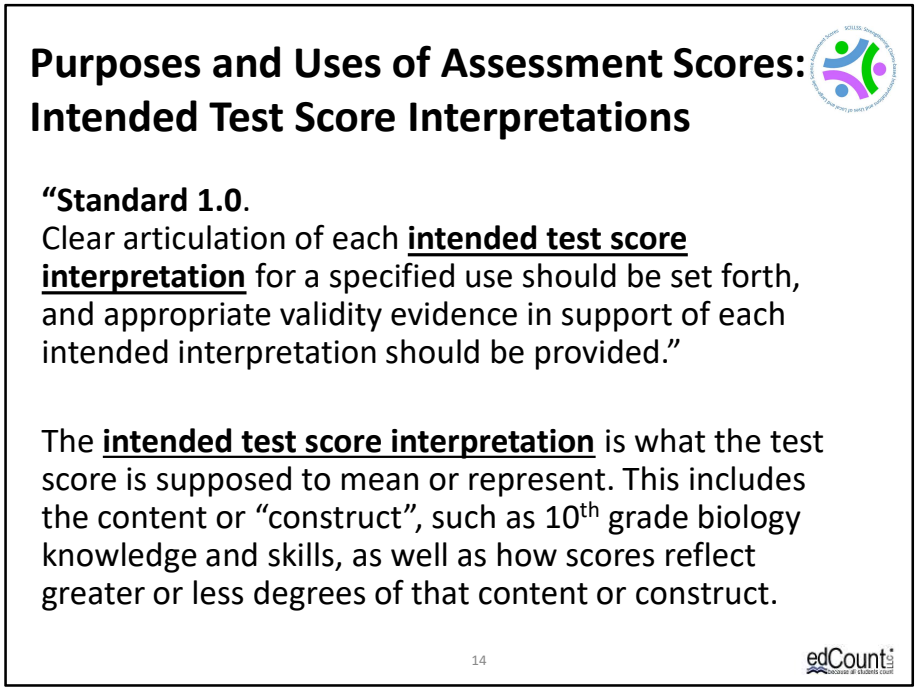


13

This notion of purpose is fundamental to all questions about test quality. Without a clear purpose for giving an assessment, it is not possible to determine whether the assessment yields meaningful information or if it is appropriate to use the scores in making specific decisions. 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. This first standard for testing practice states the following:

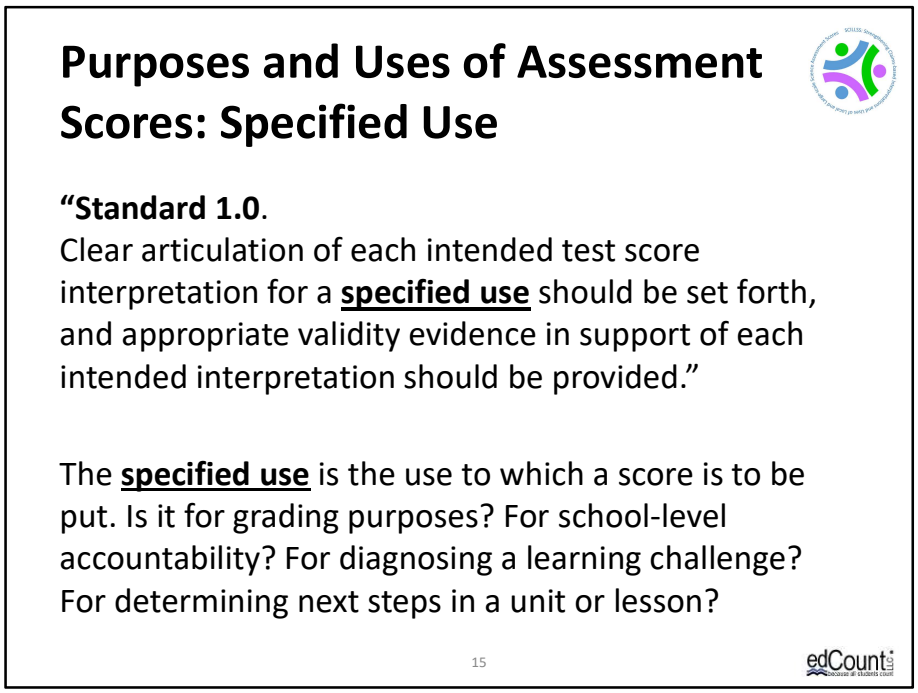
Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided.

AERA, APA, NCME, 2014



We can unpack this standard to help capture its full meaning. Consider these three key pieces of the standard:

Intended score interpretations are what the test score is supposed to mean or represent. This includes the content or "construct", such as 10th grade biology knowledge and skills, as well as how scores reflect greater or lesser degrees of that content or construct.



The specified purpose is the use to which a score is to be put. Is it for grading purposes? For school-level accountability? For diagnosing a learning challenge? For determining the next steps in a unit or lesson?

Purposes and Uses of Assessment Scores: Validity Evidence

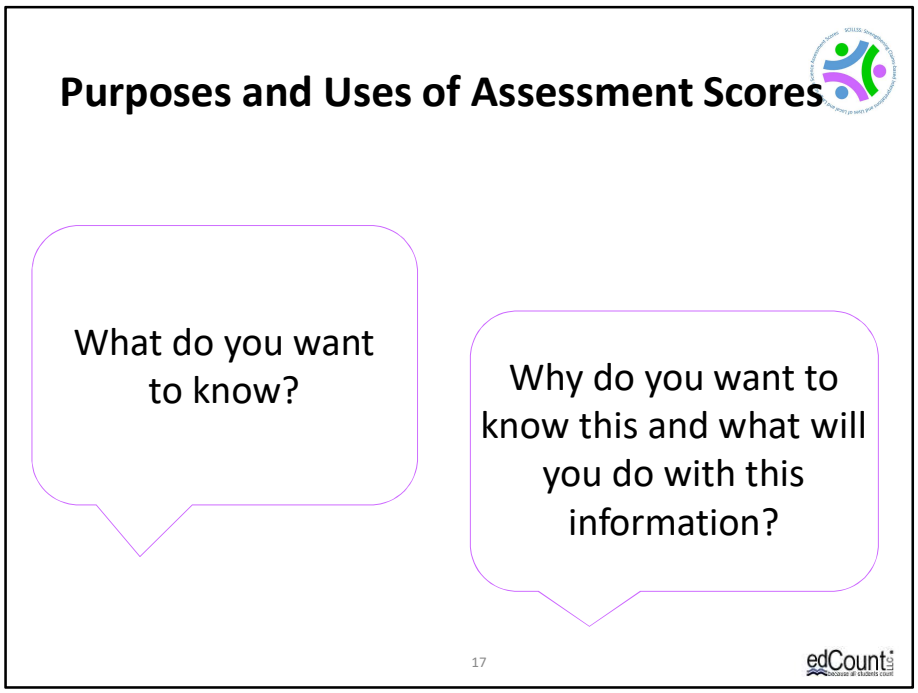
"Standard 1.0.
Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate **validity evidence** in support of each intended interpretation should be provided."

Validity evidence encompasses specific types of evidence regarding what a test measures and how scores can be interpreted and used.

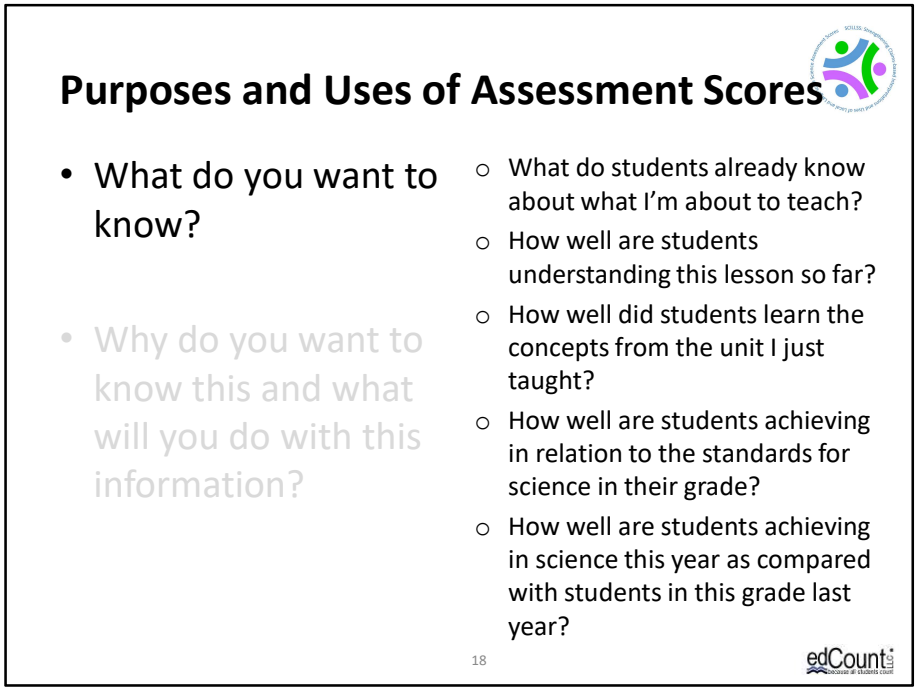
16

edCount

Validity evidence encompasses specific types of evidence regarding what a test measures and how scores can be interpreted and used. How can we know what the scores mean and if we can use them for informing instruction? For grading? For monitoring achievement trends at a school or school district level? For making accurate diagnoses?

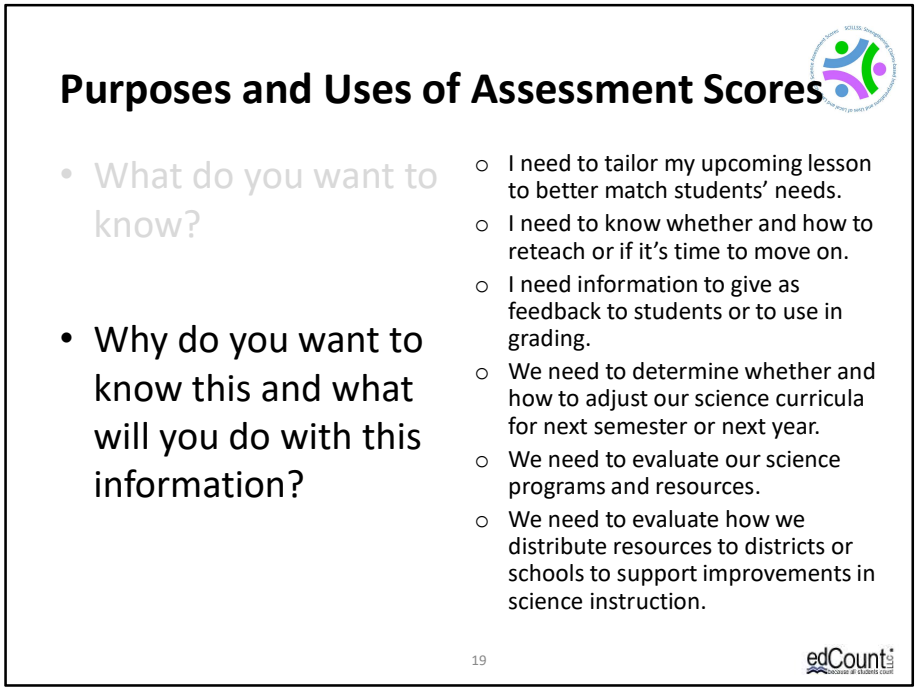


This first standard tells us that validity evidence has to do with the purposes and uses of tests scores. To understand our purposes and uses, we must consider what it is we want to know and what we will do with the information the test provides.



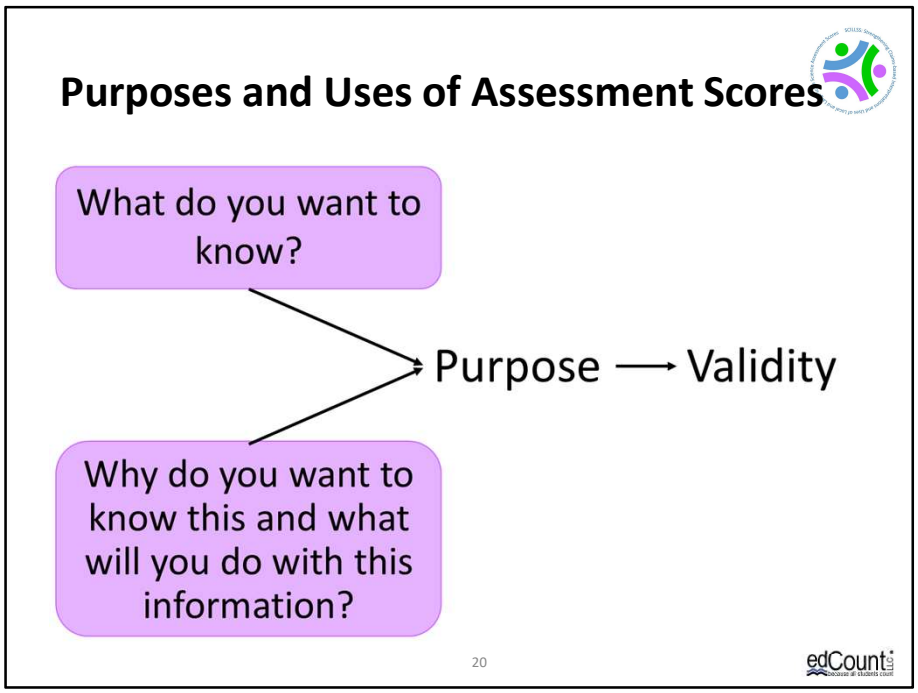
First, what do you want to know? Just as a teacher needs to be clear about what she's trying to find out when formulating questions during a lesson, anyone attempting to build a test must identify what the test is meant to measure with as much specificity as possible. A teacher or administer may want to know any number of things, for example:

- What do students already know about what I'm about to teach?
- How well are students understanding this lesson so far?
- How well did students learn the concepts from the unit I just taught?
- How well are students achieving in relation to the standards for science in their grade?
- How well are students achieving in science this year as compared with students in this grade last year?



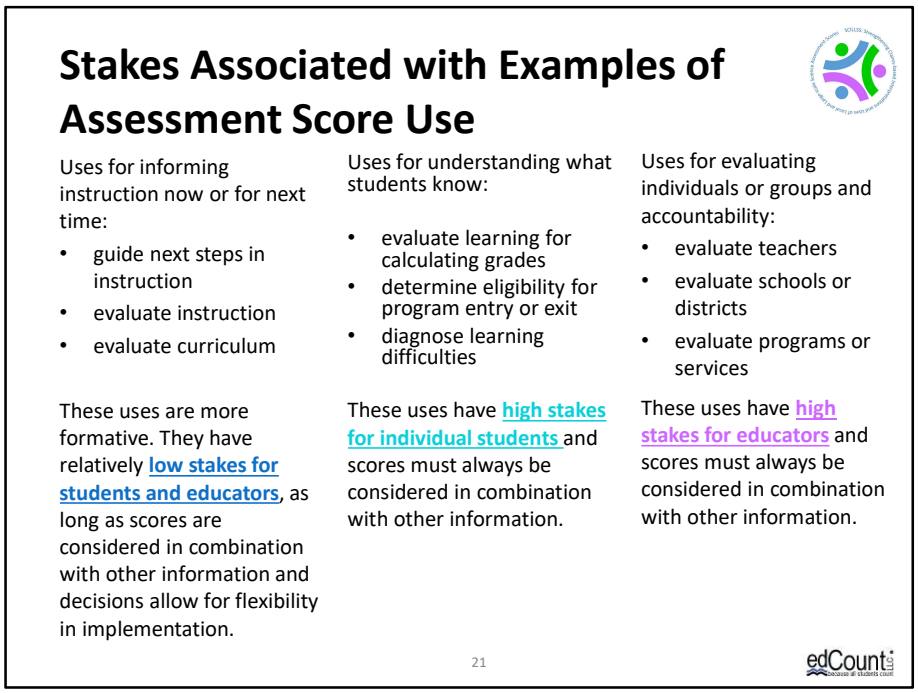
Second, what will you do with this information? What decisions will it inform? Being clear about this is necessary to ensure that the information the test provides is relevant and useful for this purpose. For example:

- I need to tailor my upcoming lesson to better match students' needs.
- I need to know whether and how to reteach or if it's time to move on.
- I need information to use in grading.
- 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.
- We need to evaluate how we distribute resources to districts or schools to support improvements in science instruction.



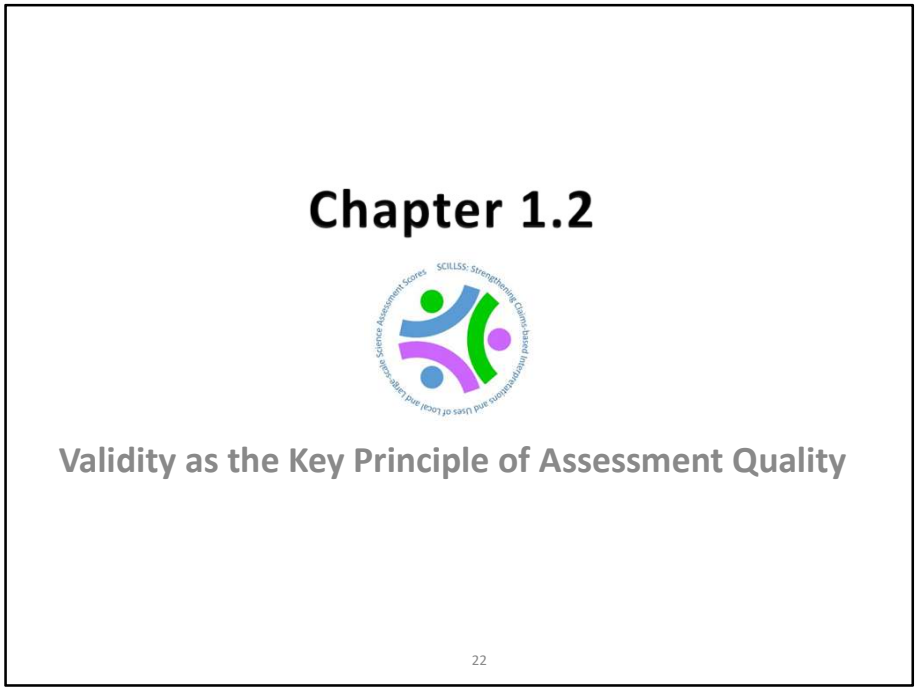
Only after an educator has identified how the scores from a test will be used – that is, the purpose for giving the assessment – can he or she determine whether that test is a good one for that purpose.

Said another way, no test is inherently good or bad. Judgments about test quality are always related to how the scores are interpreted and used. This is the crux of the concept of validity in testing. The purposes for which test scores are used drive judgments about validity.



In addition to understanding the basic purposes for giving a test and how the scores are meant to be used, we need to recognize that different uses of scores are associated with different stakes. Stakes are the implications of the scores for individuals or groups.

Higher stakes are associated with scores used in making decisions with significant implications for individuals or groups, such as placement and even grading. Stakes are particularly high when decisions are based solely or mostly on test scores and when the decisions are permanent or hard to over turn. Note that our professional standards in educational measurement, which we described earlier in this chapter, advise strongly that no one ever makes a decision for a student on the basis of a single test score.



Chapter 1.2. Validity as the Key Principle of Assessment Quality

Defining Terms: Validity

The quality of being logically or factually sound; soundness or cogency

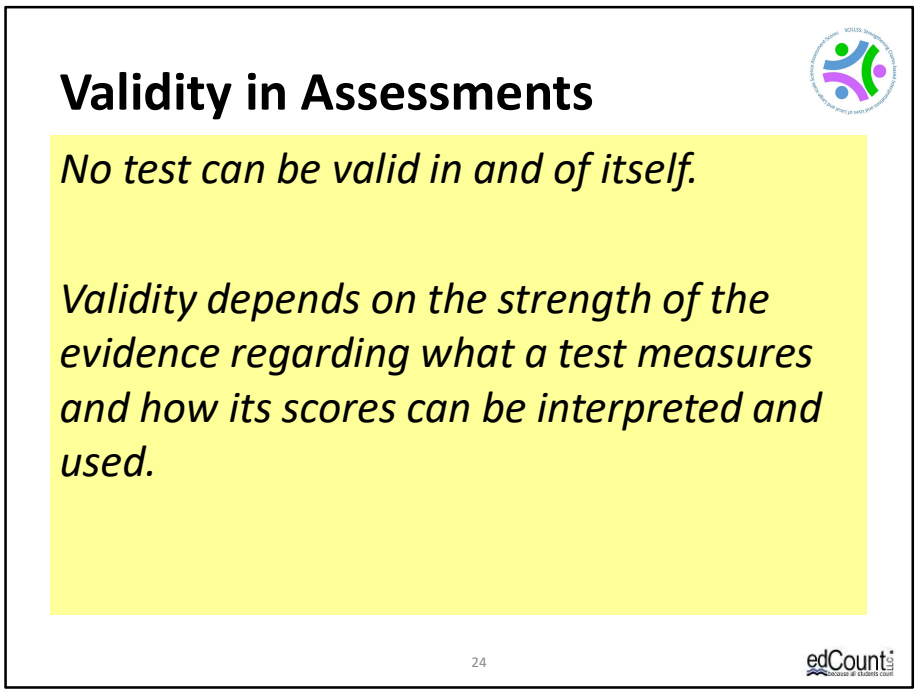
(Oxford English Dictionary, 2018)

23

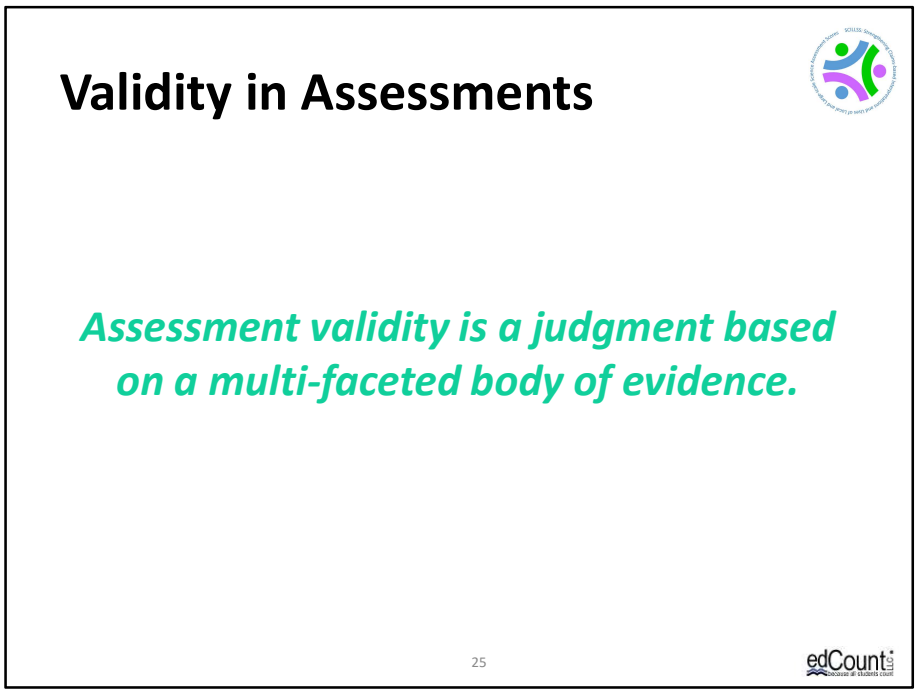
edCount

Validity is the cornerstone concept in testing and its meaning in this area is essentially the same as what is found in the dictionary:

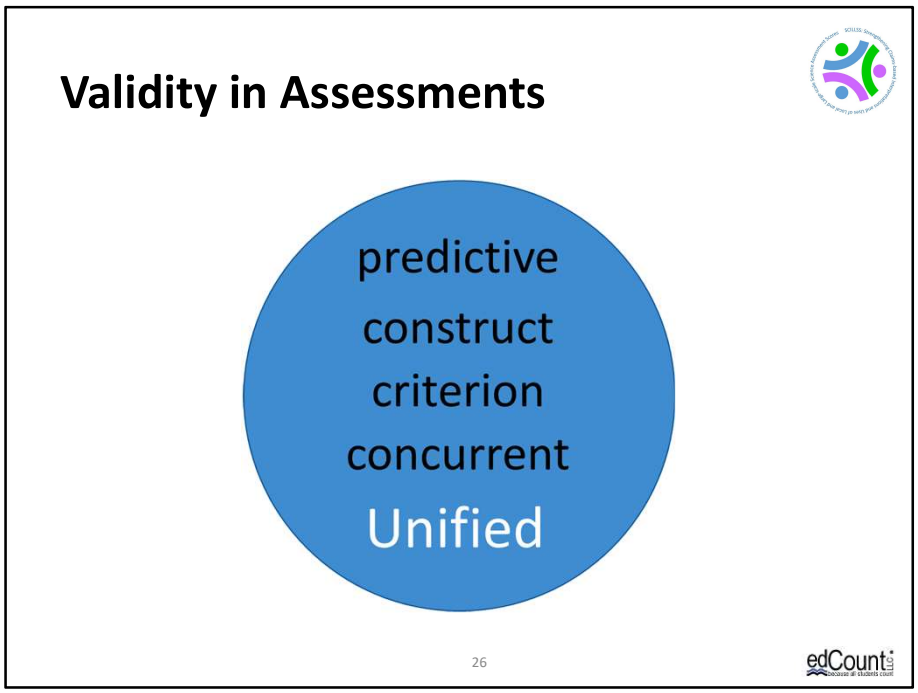
The quality of being logically or factually sound; soundness or cogency. (Oxford English Dictionary)



What may be surprising to some is that assessment validity relates to assessment scores, not to the assessments themselves. Assessment scores may support valid interpretations and uses, but a test itself cannot be inherently valid or invalid.

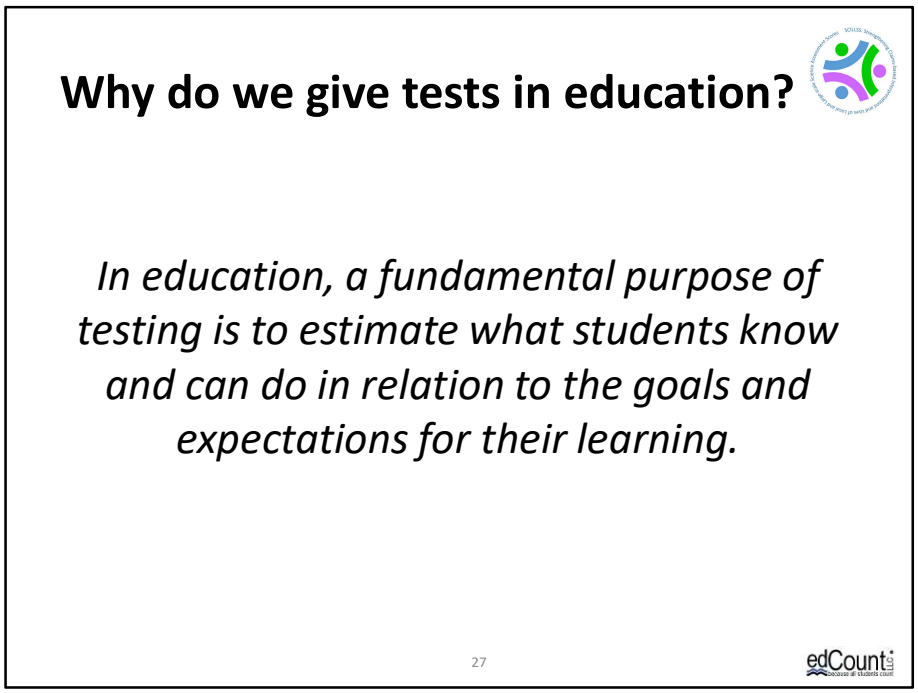


Further, assessment validity is a judgment based on a multi-faceted body of evidence. No single piece of evidence is sufficient and validity cannot be characterized with a statistic.



In the past, some characterized different types of validity such as predictive, criterion, and concurrent. Our more recent views are of a unified theory of validity. The old notions of validity are apparent within the unified theory: if the purpose of a test is to predict successful performance in some setting – say, in a college or law school – then the test maker should provide evidence that the test scores are predictive as claimed. Prediction is a use of test scores. Evidence related to that use is validity evidence. Evidence related to prediction is not important if the scores are not used for predictive purposes.

Validity is a wholistic concept and its evaluation relies on evidence related to assessment purposes and uses.



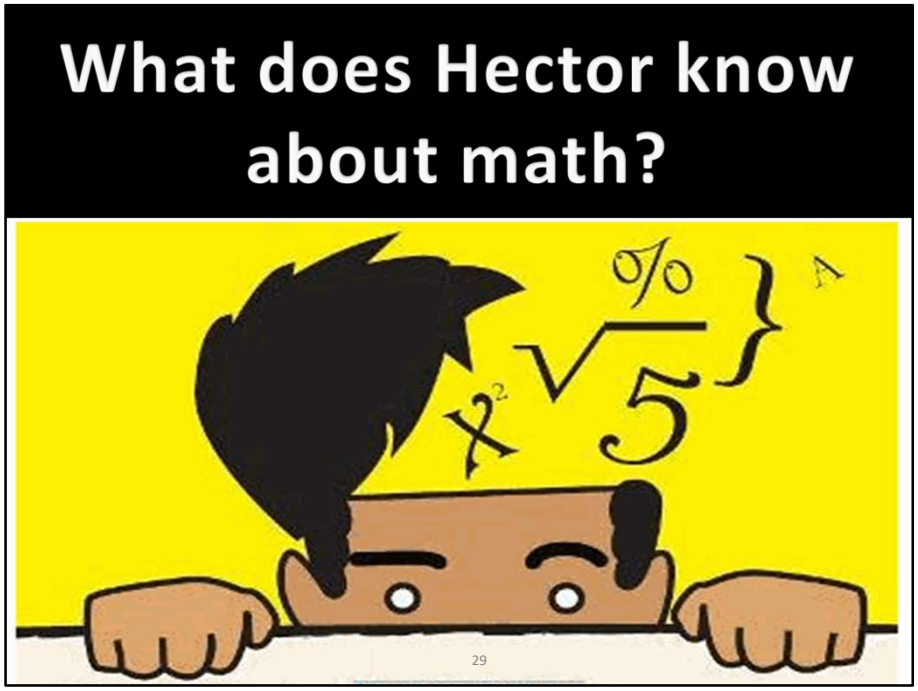
So, why do we give tests in education? What are their purposes? In education, a fundamental purpose of testing is to estimate what students know and can do in relation to the goals and expectations for their learning, often defined through standards.

This Photo by Unknown Author is licensed under CC BY-SA

28

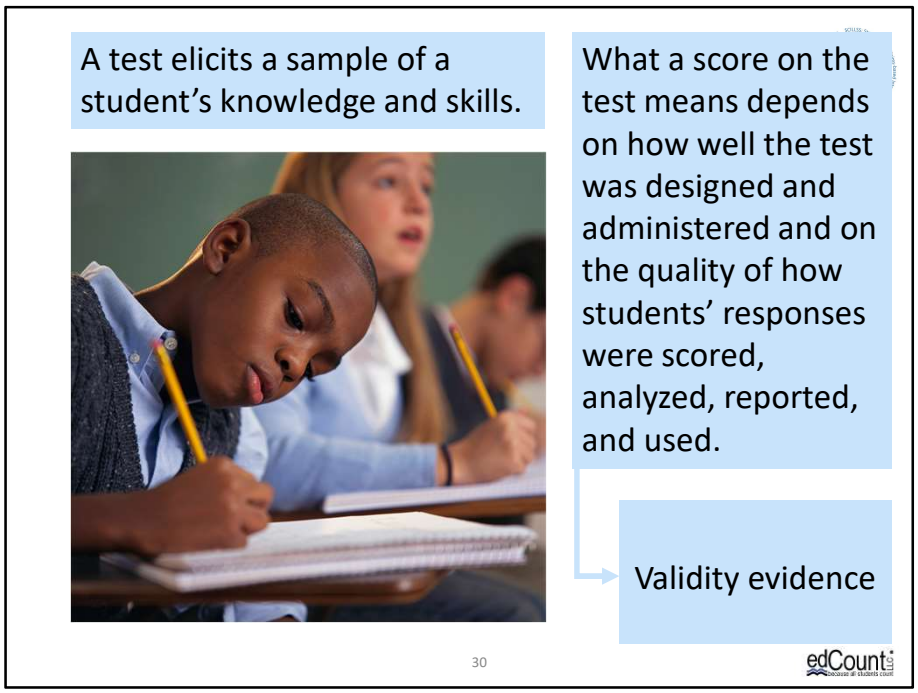
What does Maya know about science?

Consider that tests in education are ways for us to find out what students know and can do. Tests are ways for us to observe students as they perform a task or to gather data about their knowledge based on how they respond to questions.



We cannot peer inside someone's brain or observe students in every possible situation where they might demonstrate what they know and can do.

This is true even when it comes to a very specific concept like multiplying single-digit integers. Certainly, we could build a test where students are asked to multiply every combination of single-digit integers. But, that's not the end of the story. We would eventually want to know whether students can recognize when multiplication is the correct operation to apply in real-world situations. We would need to pose a variety of opportunities where students would have to identify multiplication as the right operation, identify the numbers to multiply, and apply the operation accurately. But, we would never be able to pose all possible opportunities.

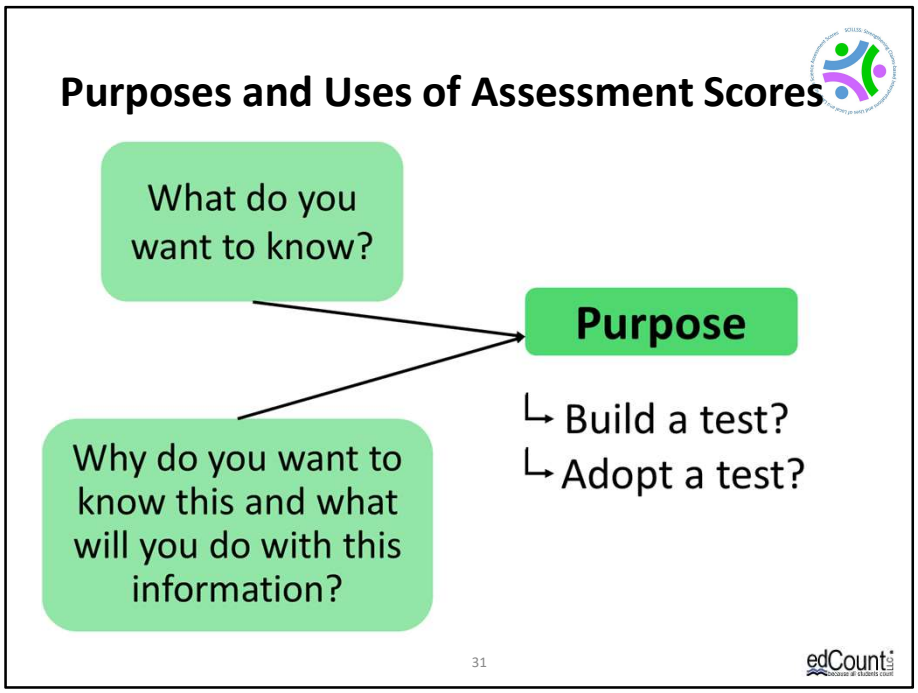


So, we create tests to elicit samples of what students know and can do and then make interpretations based on those samples.

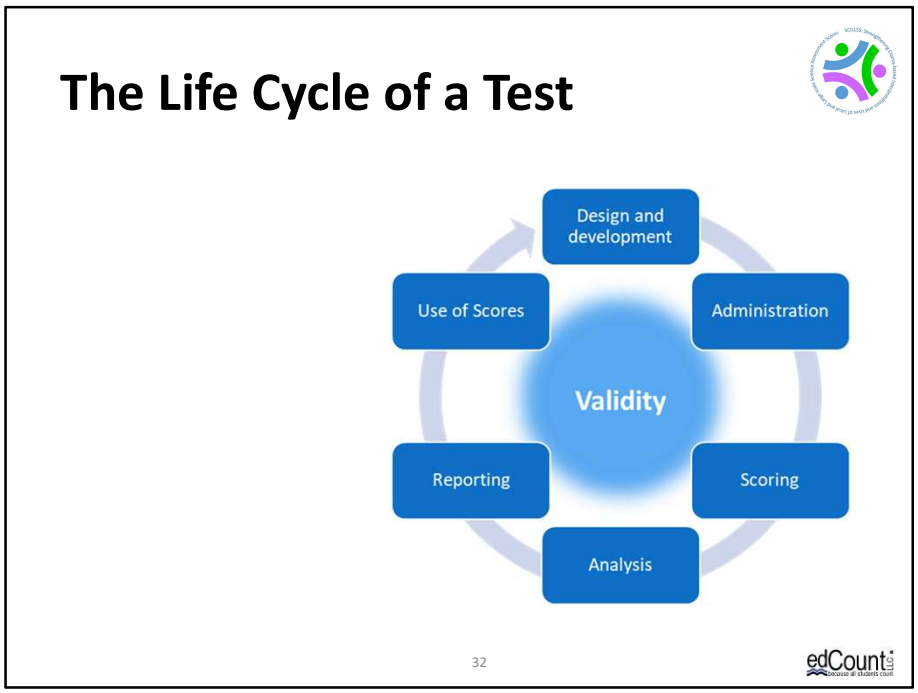
How do we know if those interpretations are sound? We gather evidence of validity.

As we will see in this chapter and those that follow, validity evidence can take many forms. Judgments about validity are based on the body of that evidence and always in relation to what the scores are supposed to mean and how they are supposed to be used.

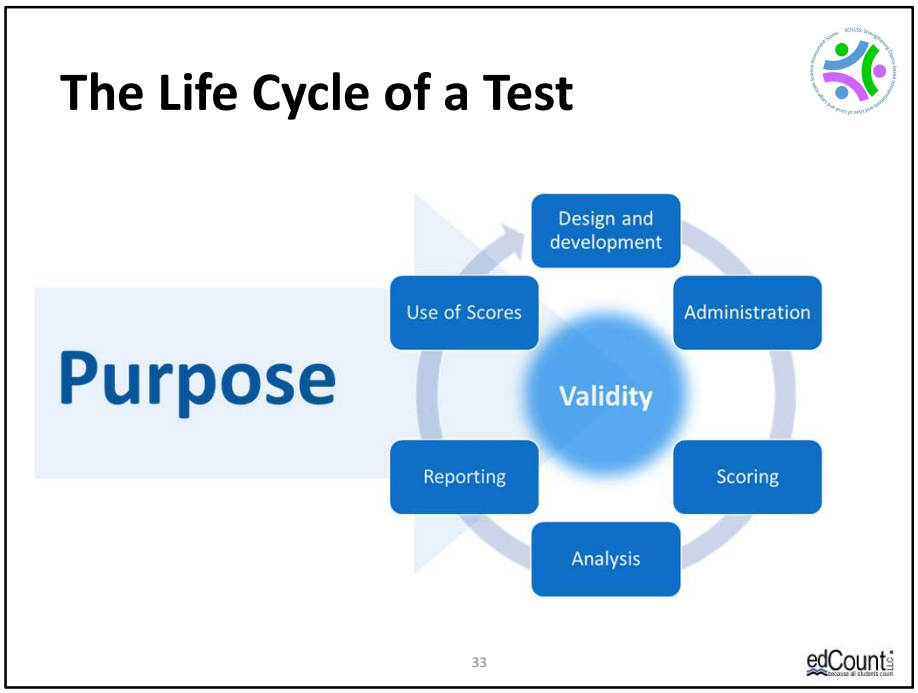
What a score on the test means depends on how well the test was designed and administered and on the quality of how students' responses were scored, analyzed, and reported.



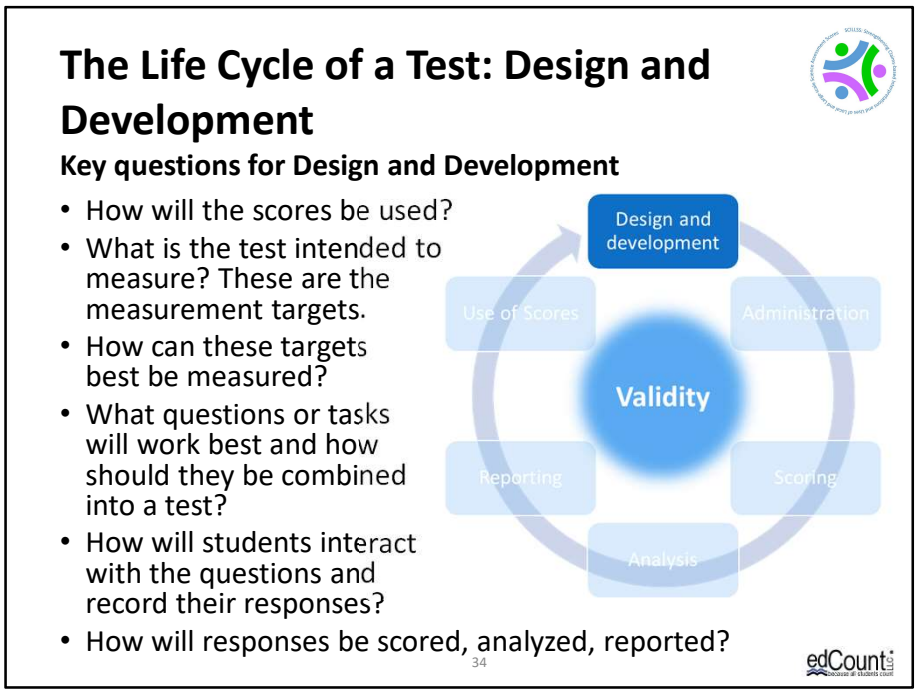
Once you have identified a need and purpose for an assessment, that is, you have a question that assessment scores may inform, it's time to either build a test for that purpose or to adopt one that already exists. As we mentioned earlier, this workbook will not focus on how to create tests; that is a topic that requires far more in-depth coverage than we can offer here. However, how a test is developed has major implications for how the scores can be interpreted and used.



The evidence necessary to make a judgment about assessment validity comes from every part of what we can think of as an assessment life cycle. This life cycle encompasses design and development, administration, scoring, analysis, reporting, and use of scores phases. This cycle is iterative and on-going for as long as the scores from a test are used.



We'll start at the beginning. As we've already seen, every assessment begins with an intent: the scores are intended to mean something in particular and to be used for a particular purpose. "I want to know X so that I can Y." This intent drives every subsequent decision about the design of a test, how it is administered, how students record their responses, how those responses are scored, how the data are analyzed, and how scores are reported. That is, how a test evolves into reality depends on the test's purpose. The nature and quality of how a test evolves relates to the validity of its scores.



When beginning the design and development process, a test developer must address fundamental questions about the nature of what the test is to measure. If I want to know something, how can I find that out? What would a student need to do to demonstrate his or her knowledge and skills in ways that allow me to make a sufficient judgment? What kinds of tasks, activities, or questions – often referred to as "items" – would allow students to demonstrate their knowledge and skills and how many of these, in what combination, would I need for this test?

Many questions present students with a prompt or statement and require the student to pick or produce a response. Some items require students to create or interact with objects or documents. In every case, the test developer must establish a clear set of rules about how students interact with the questions or tasks and how students' responses are scored. From a student's perspective, this could be thought of as answering questions like "what do I have to do" and "how will my answers be rated or counted?"

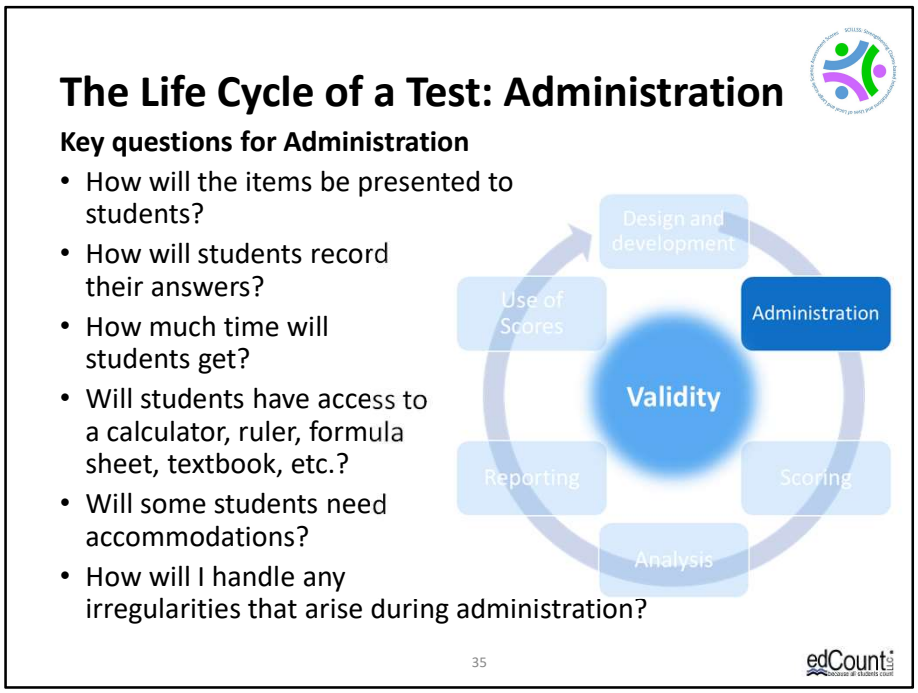
As we suggested with the multiplication example, the questions or tasks on a test are usually just samples of a much larger set of possible questions or tasks that could be used to obtain information about what students know and can do. Given this notion of "sampling", we can see how a student's score on a test is best thought of as an estimate of his or her actual knowledge and skills. This is a particularly important concept for both validity and reliability and one we'll return to in subsequent chapters.

Only after careful planning should a developer begin writing the questions or tasks that will make up an assessment. In addition, and particularly for assessments used to make decisions that have any stakes for students, those who build tests are obligated to provide evidence of how they built them. Further, developers should provide evidence that the processes they used were sound and yielded an appropriate collection of questions. Just saying that a test measures multiplication skills is not enough. The test developer must provide evidence to support that claim. How were items designed? How were they developed?

Chapters 2 through 5 of this workbook will address how to gather and evaluate this evidence.

The development phase for large-scale tests that schools, districts, and states purchase always includes one or more types of "try-outs" of the items they've built. This can be called pilot-testing or field-testing and begins the connection between the design and development phase and the administration phase.

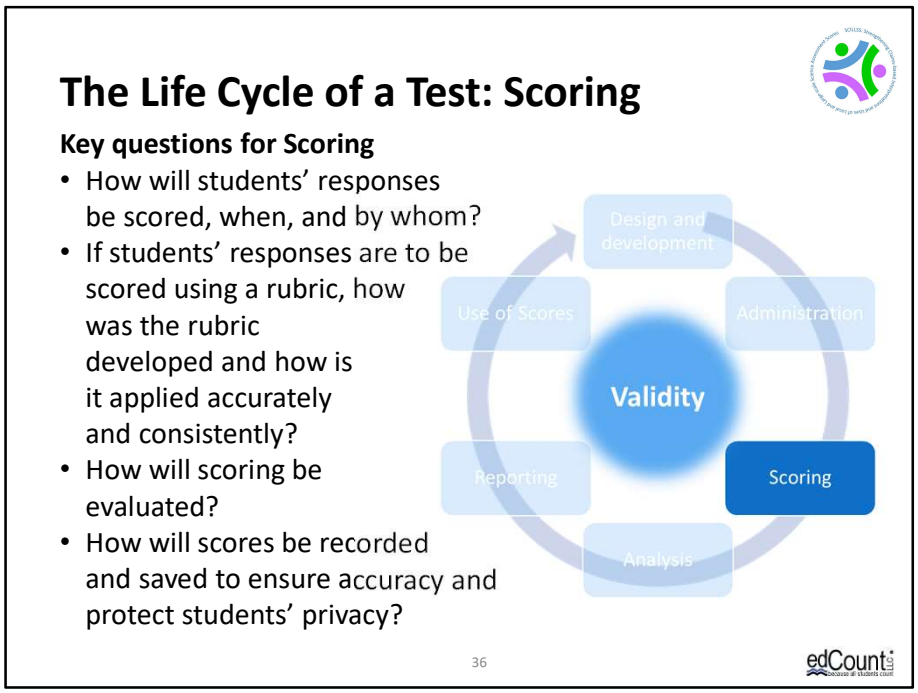
Note that the design and development phase is really on-going; information gathered through the rest of the testing life cycle can and should be used to inform subsequent adjustments in design and development. The cycle is just that: an iterative process. Ideally, each iteration improves on the last.



Test administration involves “giving” the test: having students interact with the items and produce responses. Those who build tests must determine every detail about that interaction.

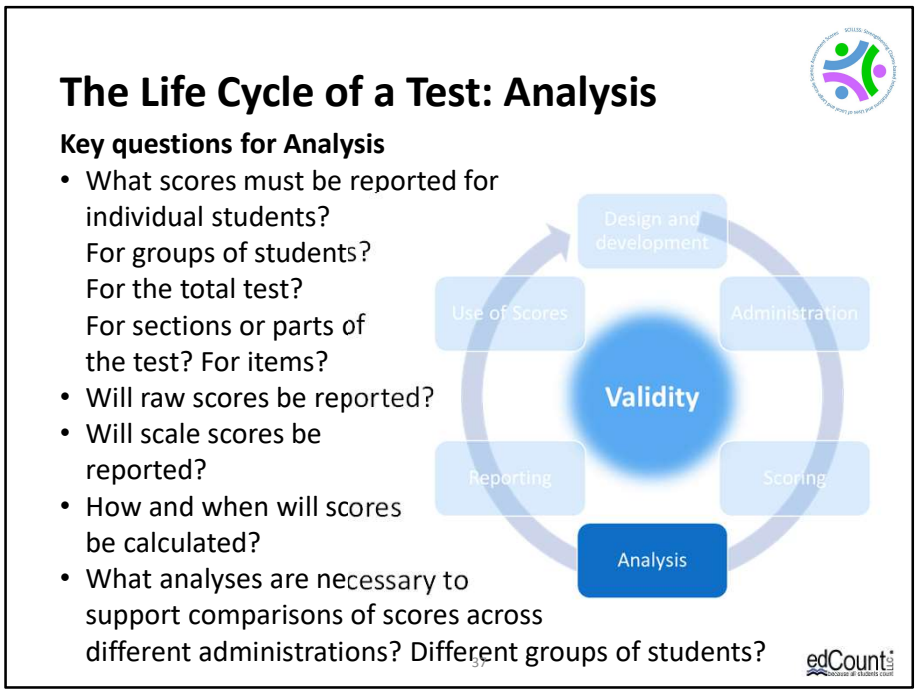
Will the items be presented on paper or on a computer?
How will students record their answers?
How much time will students get to take the test?
Will students have access to other resources, like a calculator, ruler, formula sheet, textbook, dictionary, or the internet, while taking the test?
How will answers be collected?
Will students answer questions independently or work as teams?
Will some students need accommodations that allow us to understand what they know and can do?
How will I handle any irregularities that arise during administration, such as student illness, fire drills, or power outages?

Those who build tests are obligated to provide answers to all of these questions as well as the rationales and evidence that back-up their methods.



Once students have answered the questions on the test, the scoring phase can begin. Scoring has its roots in the design and development phase just as the administration phase does. Those building tests must determine how each item will be scored; these decisions are not to be made "on the fly" after students have taken a test. If the item has a multiple-choice or selected-response design, the item writer must indicate which of the response options is correct. It is also very helpful to consider what each wrong answer might tell you about what a student's particular misconceptions may be.

If the item is performance-based, which requires a student to demonstrate a response through actions like writing, speaking, drawing, performing an experiment, making music, or dancing, the item writer must provide the rubric and rules for those scoring that performance.



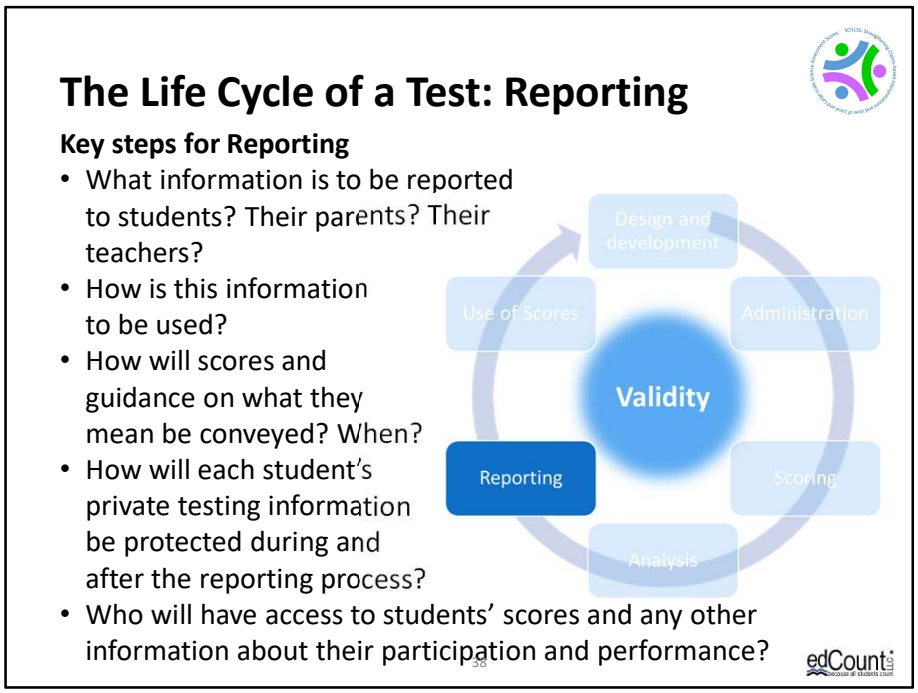
The scoring process results in data. These data must be analyzed before they can be reported back to students or others. Within a classroom, test data analyses can include calculation of class averages and statistics related to the distribution of scores on the range of possible highest and lowest scores. Large-scale assessment data are typically analyzed in sophisticated ways, among other things, to create score scales and to evaluate the comparability of scores across different administrations of the test.

Most of us have encountered score scales for assessments such as the ACT, the SAT, and statewide academic assessments. Score scales involve the use of statistics to transform raw scores – that is, scores like the number of correct answers for a set of test questions – into a scale that allows for more stable comparisons across versions of a test and across test administrations. Nearly all teacher-made tests will use a simple raw score scale, like the number or percent of correct answers. Large-scale assessments rarely report raw scores without also providing the corresponding scale score.

Analysis of test data can involve a range of other statistical calculations. These include simple statistics like frequencies of each total score; averages such as mean, median, and mode; and standard deviations. More complex psychometric analyses can include reliability estimates, analyses of item and test characteristics, and equating analyses necessary for comparing test performance across administrations and forms. Analyses of test data are how we answer questions such as, “How many students chose option A when the correct answer was B?”, “Has Frank’s achievement improved since the last time he took the assessment?“, and “Did students do better this year than they did last year?”

While the details of these “psychometric” analyses may seem inaccessible to many educators, they contribute critical evidence regarding the meaning of assessment scores. That is, they are part of the body of validity evidence. Those scoring and analyzing test data are obligated to describe these analyses, justify the methods they use, and provide and explain the results.

Smart test developers also consider the results of statistical and psychometric analyses to determine whether and how to make adjustments in the test design, test administration, scoring, analysis, reporting, and use phases.

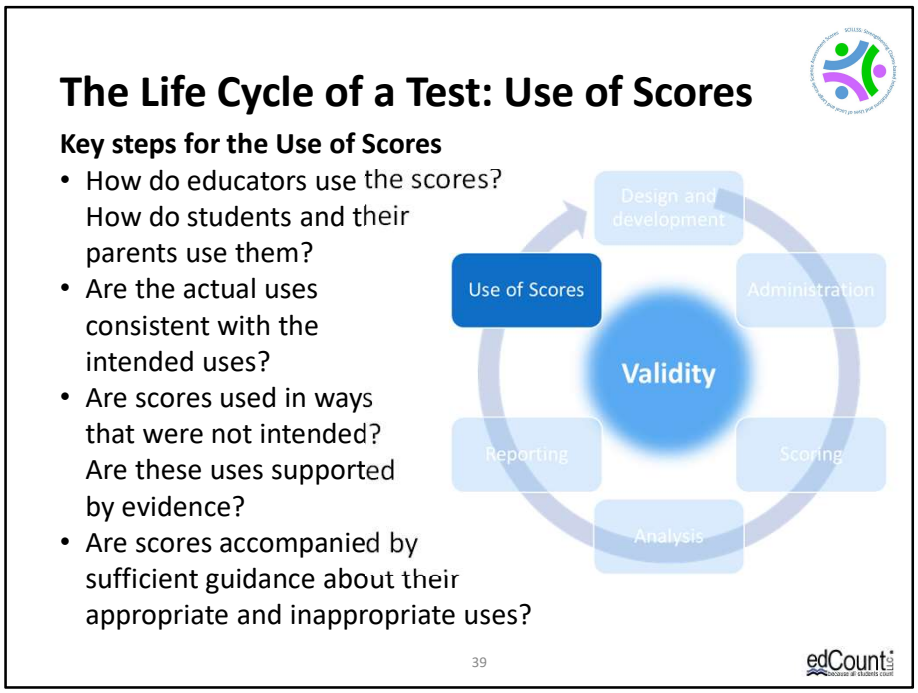


After data analysis comes reporting. If a test does not yield information, it was not worth giving.

Like scoring and analyses, reporting of assessment scores can range from relatively simple communications to sophisticated collections of scores like those provided by commercial test vendors. A test developer must determine how scores are to be reported, to whom, and with what additional information to aid interpretation and use. What information will the students who took the test get? When and how? What information will teachers get? Administrators? Parents? Other stakeholders? How will that information be conveyed in ways that reflect accuracy, protect students' privacy, and support appropriate uses of the scores and inhibit inappropriate uses?

Test developers are wise to consider these questions early in the design and development process even though reporting comes at the end of an assessment

cycle. If you don’t know where you are going, how do you plan your route and know when you’ve arrived?

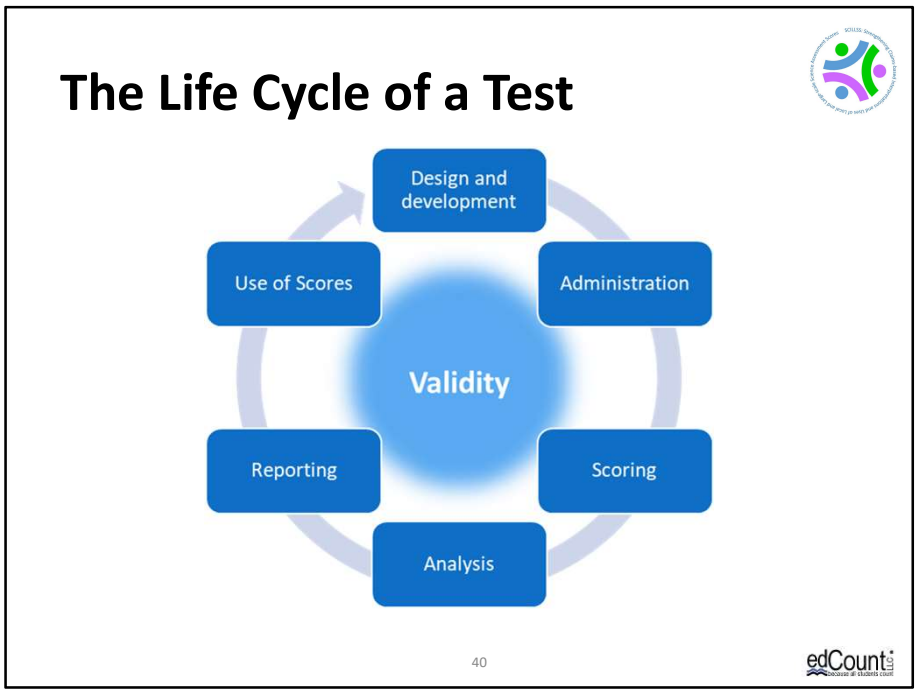


While scores cannot actually be used until after test administration, scoring, analysis, and reporting, decisions about how they are to be used are really made in or even before the test design phase. Score uses are inextricably linked to the purposes for creating or adopting a test in the first place.

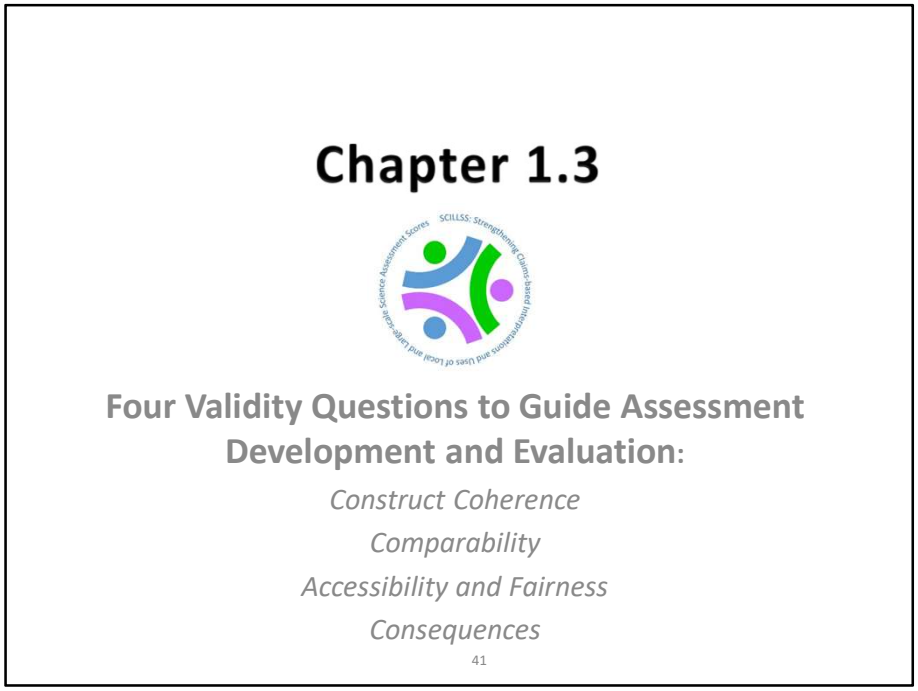
Test developers are obligated to consider how scores are intended to be used and to provide support and guidance for such score uses. In addition, test developers must consider how scores may be used in ways they had not intended and caution test users if such uses are inappropriate. For example, a teachers may use interim tests to monitor students' progress across a school year. Using those scores for promotion and retention decisions at the end of that year would almost certainly be an inappropriate use of those scores.

Questions related to score use include: How do educators use the scores? How do students and their parents use them? Are the actual uses consistent with the

intended uses? Are scores used in ways that were not intended? Are these uses supported by evidence? Are scores accompanied by sufficient guidance about their appropriate and inappropriate uses?



Clearly, the testing process involves many decisions and solid evidence to support those decisions and their outcomes. In the next section of this chapter, we will begin to see how to gather and evaluate validity evidence in each of the phases of the assessment life cycle.



Chapter 1.3. Four Validity Questions to Guide Assessment Development and Evaluation

Validity Questions

Recall from an earlier section:

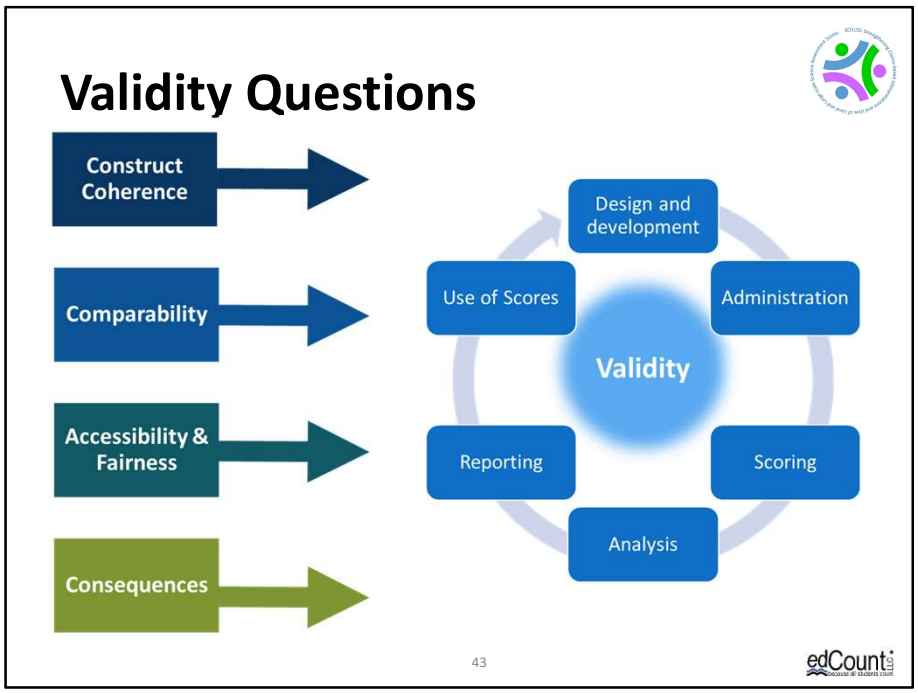
- No test can be valid in and of itself. Validity depends on the strength of the evidence regarding what a test measures and how its scores can be interpreted and used.
- Validity questions relate to the interpretation and use of test scores.
- Validity evidence necessary to answer each question comes from across the range of phases in an assessment life cycle.

42

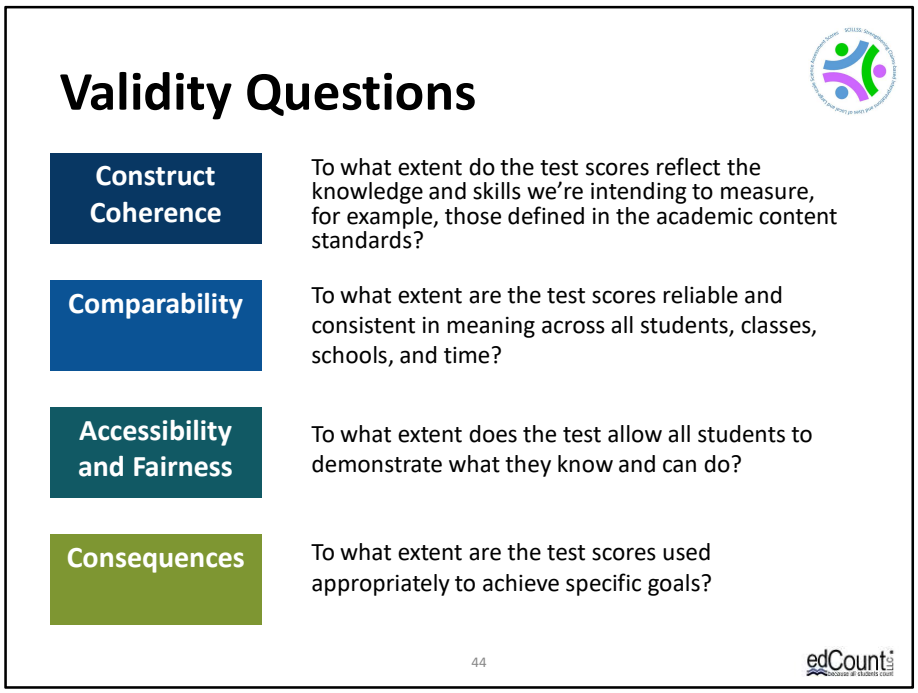
edCount

Now that we have an overview of the testing process, from design and development through reporting and use, we can lay out a framework for how to gather and evaluate validity evidence across all phases of this process.

Recall from an earlier section of this chapter that validity relates the interpretation and use of assessment scores and not directly to an assessment itself. Thus, validity questions relate to the interpretation and use of test scores and validity evidence comes from across the range of phases in an assessment life cycle



The four validity questions represent broad categories and each subsumes many other questions. All require evidence from across the assessment life cycle. The categories are: construct coherence, comparability, accessibility and fairness, and consequences.



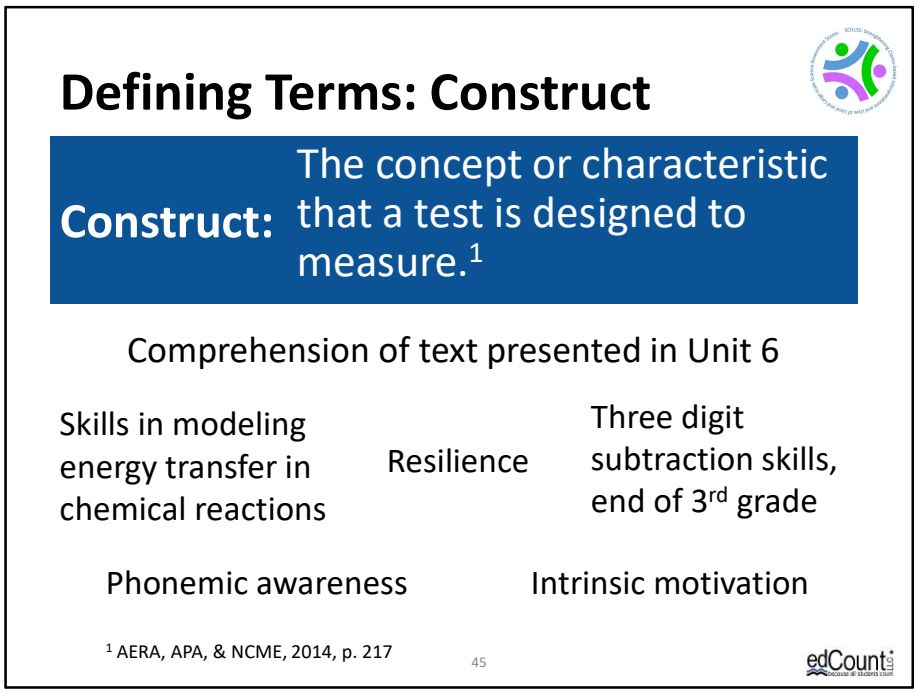
The four validity questions are:

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

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

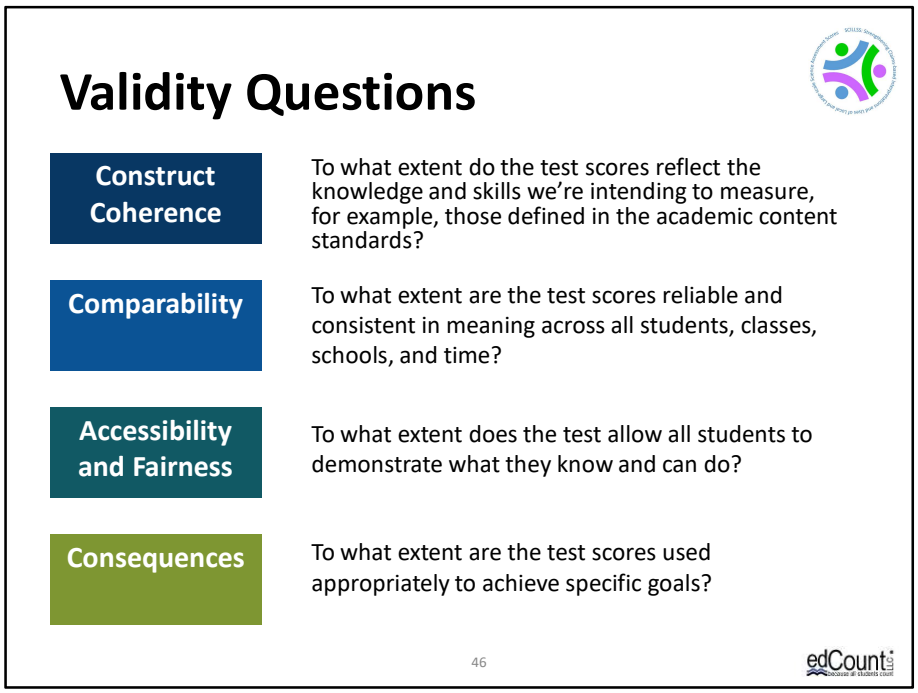
To what extent does the test allow all students to demonstrate what they know and can do? This question addresses the concept of accessibility and fairness. And

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

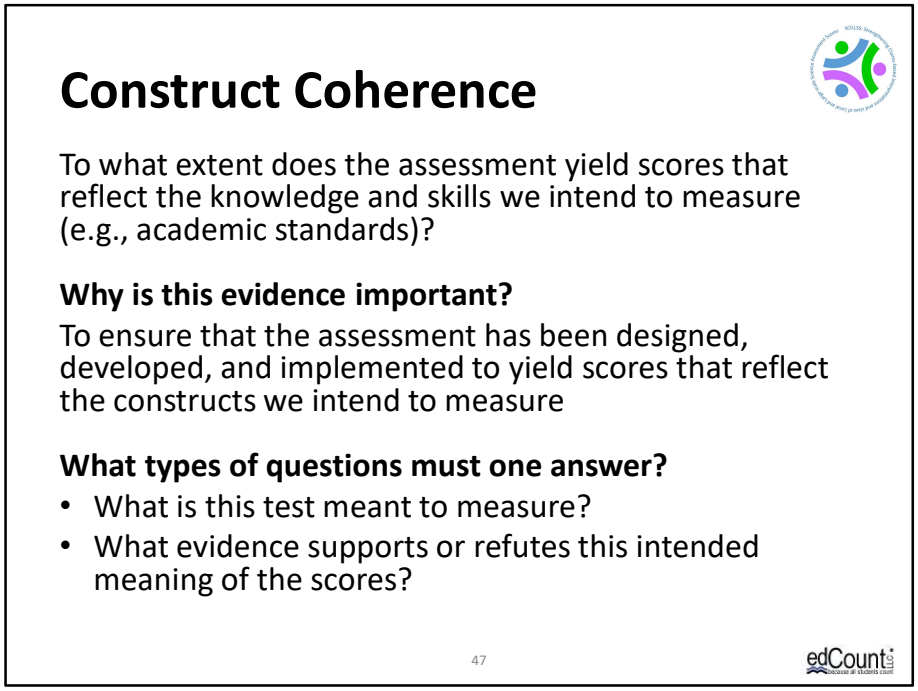


The term "construct" may not be familiar to you or some of your colleagues. A construct is a concept or characteristic that a test is designed to measure. In education settings, the constructs of most interest have to do with content knowledge and skills or personal or social characteristics that often relate to academic performance.

Recall from earlier in this chapter that it's not possible to see directly what a student knows and that we have to present students with opportunities – such as tests – when we can observe them demonstrate their knowledge and skills. Their knowledge and skills, like text comprehension, modeling energy transfer, subtraction, resilience, and motivation are constructs that tests are meant to measure.



Each of the next four chapters of this workbook will focus on one of the four validity questions such that all four are addressed in turn. Here, we'll provide a brief overview of each question.

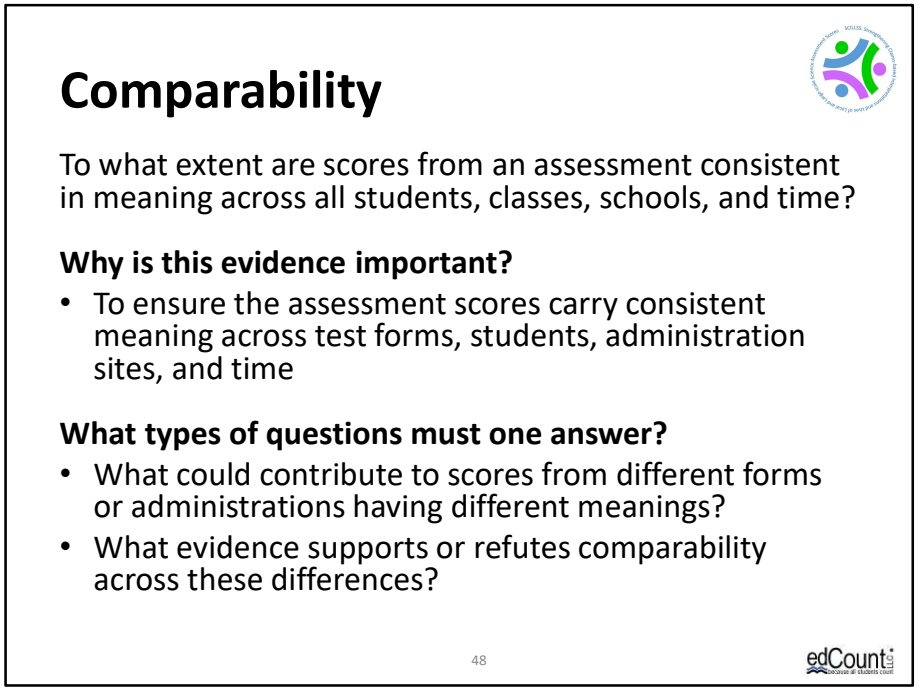


The first validity question involves the notion of construct coherence. What are we intending to measure and how can we know whether and how well we're measuring that? In education, we may be intending to measure students' knowledge and skills as defined in a set of academic standards. In addition, we need consider what we are not intending to measure and take care to avoid tapping into constructs that are not relevant to the ones we're targeting. For example, while "story problems" may offer engaging contexts for applying problem-solving skills, they should not impose an undue reading burden or present vocabulary or settings that are irrelevant to the target skills and may be unfamiliar to some students.

Questions related to construct coherence include:

- What is this test meant to measure?
- What evidence supports or refutes this intended meaning of the scores?

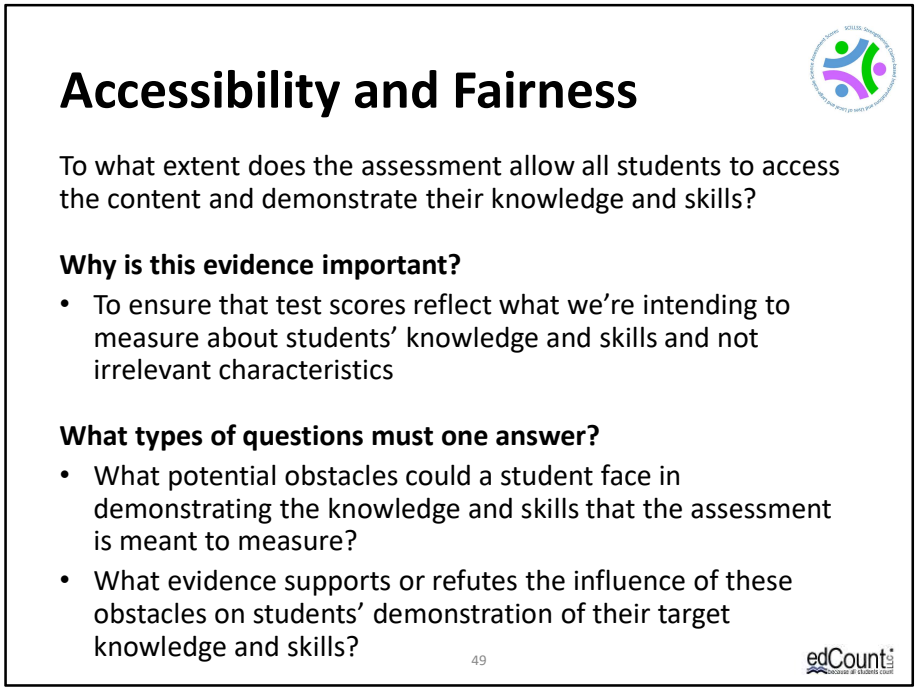
As is the case for the other validity questions, we would gather evidence related to construct coherence from across the assessment life cycle.



Our second question relates to comparability: To what extent are the assessment scores reliable and consistent in meaning across all test forms, students, classes, schools, and time?

Questions related to comparability include:

- What could contribute to scores from different forms or administrations having different meanings?
- What evidence supports or refutes comparability across these differences?



Our third validity question relates to test fairness and accessibility. To what extent does the assessment allow all students to access the content and demonstrate their knowledge and skills?

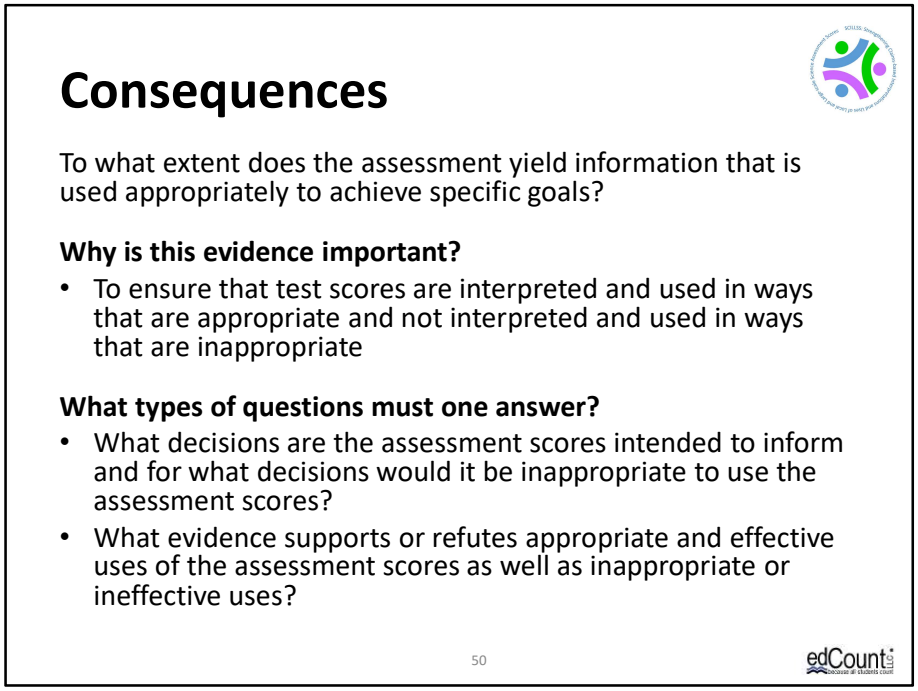
Fairness and accessibility evidence is critical because we must ensure that each student has a legitimate opportunity to demonstrate what he or she knows and can do. No student should be hindered in demonstrating their skills because they have a visual impairment that makes it difficult to read the questions or because they cannot type or write or speak a response or for any other reason unrelated to what we're trying to measure.

Addressing fairness and accessibility involves answering questions such as:

- What potential obstacles could a student face in demonstrating the

knowledge and skills that the assessment is meant to measure?

- What evidence supports or refutes the influence of these obstacles on students' demonstration of their target knowledge and skills?

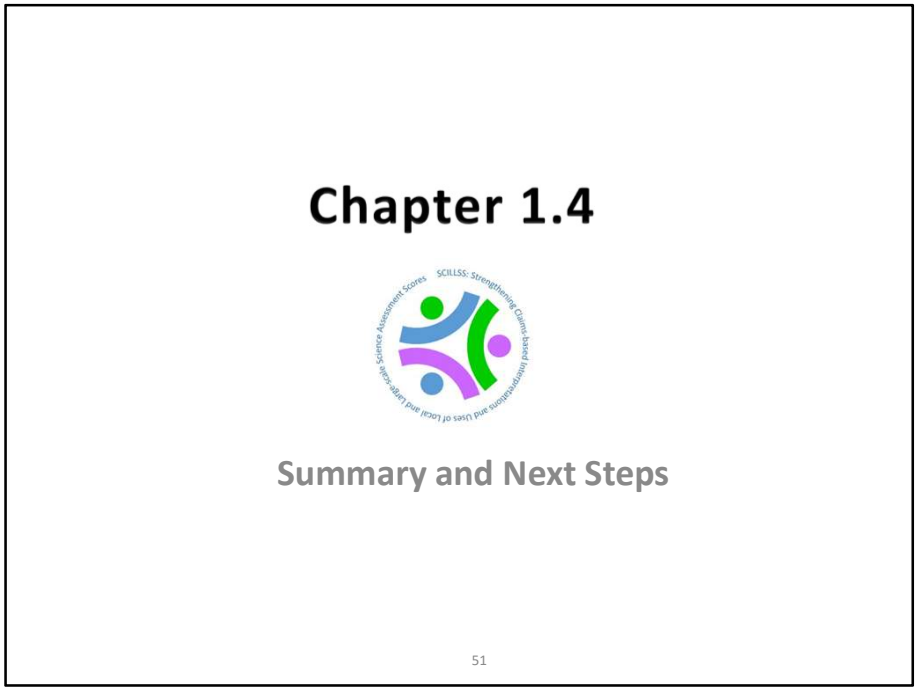


Our fourth validity question relates to consequences. To what extent does the assessment yield information that is used appropriately to achieve specific goals?

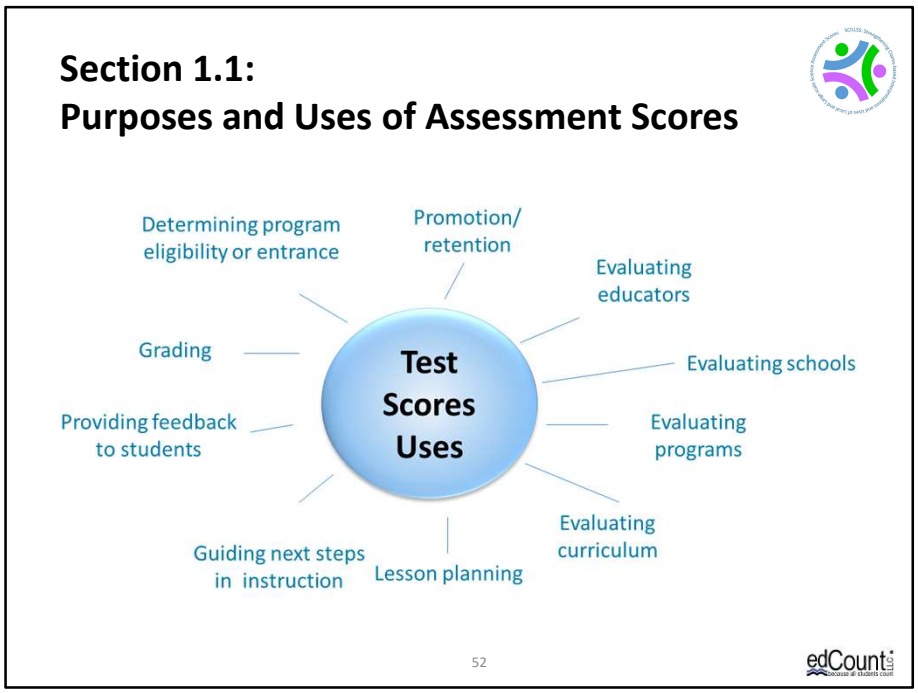
Recall that every test must be associated with a purpose. Tests are given because test scores are intended to be used to inform one or more decisions. Consequences in testing have to do with these intended uses as well as with any possible unintended uses of test scores.

Questions related to consequences include:

- What decisions are the assessment scores intended to inform and for what decisions would it be inappropriate to use the assessment scores?
- What evidence supports or refutes appropriate and effective uses of the assessment scores as well as inappropriate or ineffective uses?

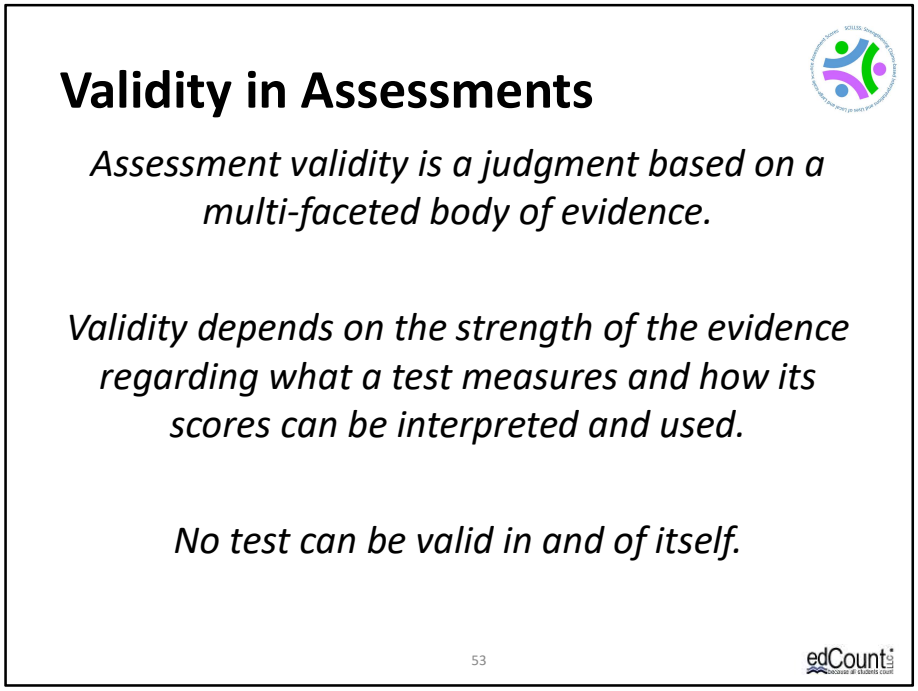


Chapter 1.4: Summary and Next Steps

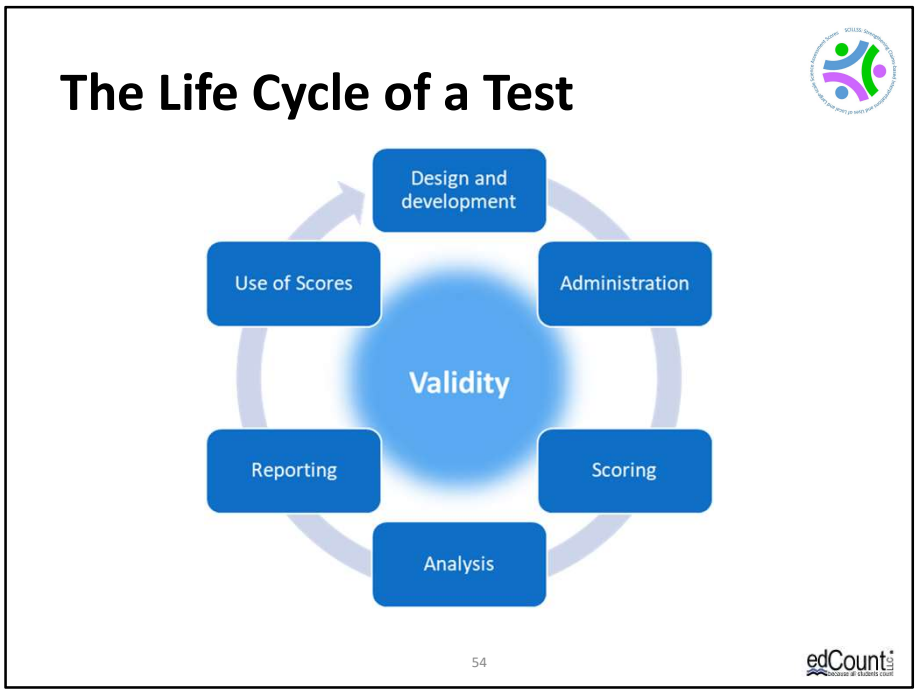


In this first chapter, we have addressed several fundamental concepts about educational assessment.

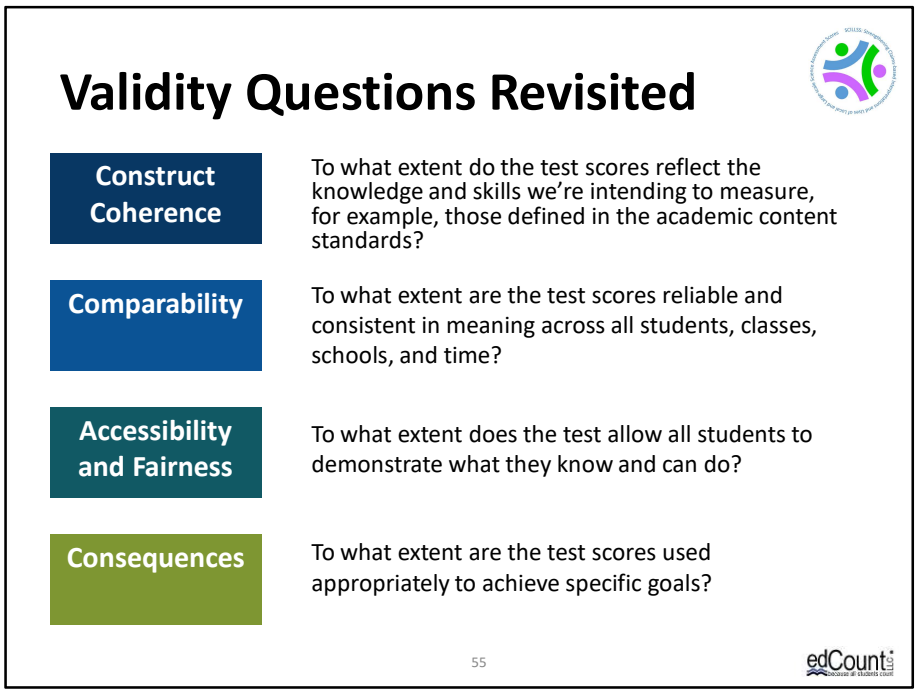
In section 1.1, we learned that all assessments must be associated with specific purposes. A teacher or administrator must have a clear reason for giving a test and understand how the test scores are to be used. A single test cannot serve all purposes.



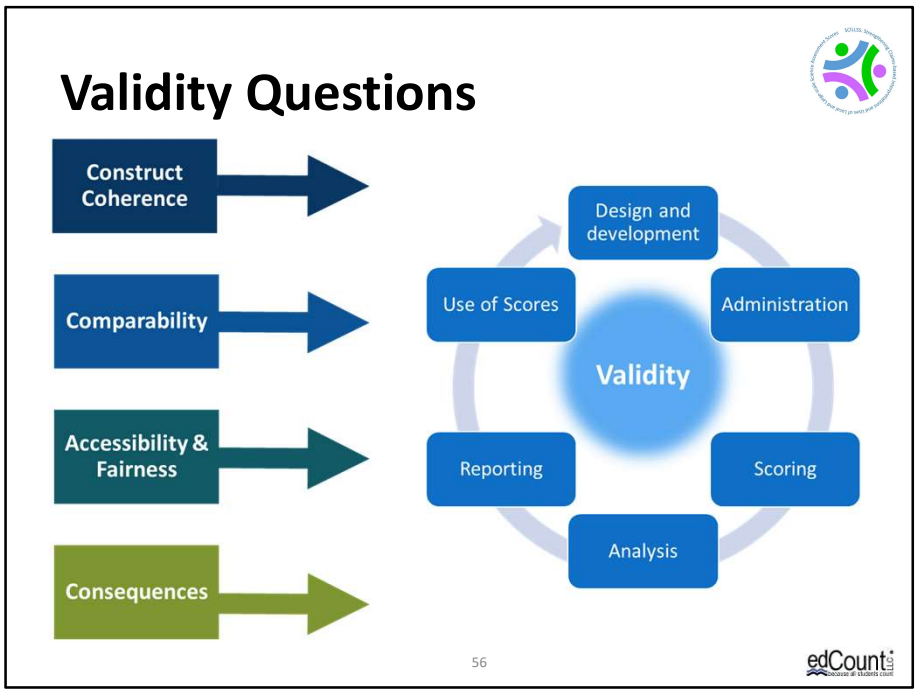
In section 1.2, we learned that validity is the cornerstone of educational testing and that validity relates to the purposes and uses of test scores, not to tests themselves.



We also learned about the assessment life cycle that begins with design and development and proceeds through administration, scoring, analysis, reporting, and score use.



In section 1.3, we briefly explored the four validity questions that the subsequent four chapters of this workbook will address in detail.



These four chapters that follow will connect the validity questions to specific forms of evidence that should be gathered throughout the assessment life cycle.

Resources and Additional Information

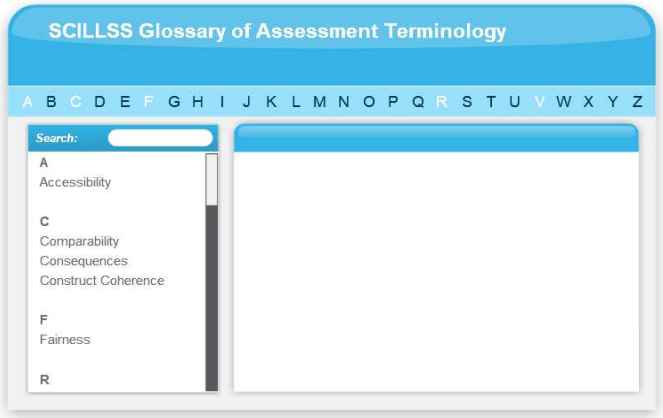
57

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 first 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.



58

Web links

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

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

References

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

Validity. 2018. In *oxforddictionaries.com*. Retrieved January 22, 2018, from https://en./definition/validity