Assessment Literacy

Ellen Forte, Ph.D.
CEO & Chief Scientist
edCount, LLC

Indiana Assessment
Professional Development Day
June 22, 2018

Overview

SCILLSS

Defining assessment literacy

Why use tests?

The life cycle of a test

Validity and how to evaluate it

Asking the right questions

SCILLSS

Strengthening Claims-based Interpretations and Uses of Local and Large-scale Science Assessment Scores

The information presented here comes from a digital workbook on educational assessment design and evaluation developed by edCount, LLC, under a grant to the state of Nebraska from the U.S. Department of Education, Office of Elementary and Secondary Education, Enhanced Assessment Grants Program, CFDA 84.368A.

SCILLSS

Principled-design resources to support development of:

- Large-scale science assessments
- Classroom-based, instructionally-embedded assessments
- Self-evaluation Protocols: Reflecting on and evaluating assessment systems
- Digital Workbook on Educational Assessment Design and Evaluation

If

- Constructs are well-defined and
- Construct definitions are shared across the system and
- The system is well-designed and
- The system is well-implemented

Then scores may reflect…

- what students know and can do

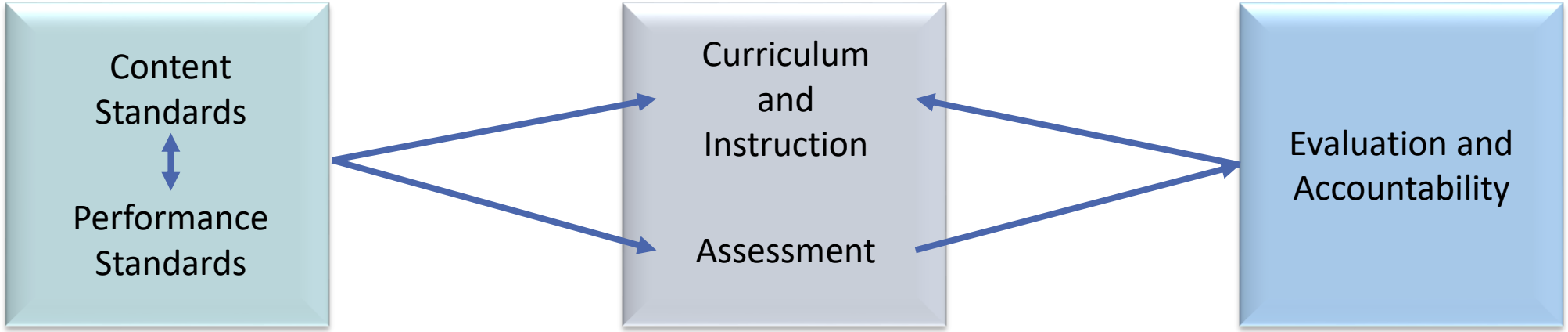
or

- what students have learned this year/ in this course

And may be used…

- to build and deliver instruction aligned with academic expectations
- to monitor or track student progress
- for school accountability decisions and program evaluation

Systemic Reform: Standards-based Assessment and Accountability



- Standards define expectations for student learning
- Curricula and assessments are interpretations of the standards
- Evaluation and accountability rely on the meaning of scores
- Without clear alignment among standards, curricula, and assessment the model falls apart

Overview

SCILLSS

Defining assessment literacy

Why use tests?

The life cycle of a test

Validity and how to evaluate it

Asking the right questions

Assessment Literacy

Being assessment literate means that one understands key principles about how tests are designed, developed, administered, scored, analyzed, and reported upon in ways that yield meaningful and useful scores.

An assessment literate person can accurately interpret assessment scores and use them appropriately for making decisions.

Assessment Literacy – knowledge and skills

- Why, when, and how to use tests and test scores
- When and why tests and test scores may not be the best option
- How to determine which test may be most appropriate for a given purpose
- How to evaluate the quality and usefulness of test scores

Overview

SCILLSS

Defining assessment literacy

Why use tests?

The life cycle of a test

Validity and how to evaluate it

Asking the right questions

Why use tests?

- Every test must have a purpose: what are the scores to be used for?
- Each use is associated with stakes.
- All tests should have some validity evidence to support score meaning.
- The higher the stakes, the more critical validity evidence becomes.

Purposes and Uses of Assessment Scores

- What do you want to know?
- Why do you want to know this and what will you do with this information?

Purposes and Uses of Assessment Scores

- What do you want to know?
- Why do you want to know this and what will you do with this information?
  - 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?

Purposes and Uses of Assessment Scores

- What do you want to know?
- Why do you want to know this and what will you do with this information?
  - 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 give as feedback to students or 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.

Why use tests?

To…

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

Why use tests?

To…

- guide next steps in instruction
- evaluate instruction
- evaluate curriculum

These uses are more formative. They have relatively **low stakes for students and educators**, as long as scores are considered in combination with other information and decisions allow for flexibility in implementation.

- evaluate learning for calculating grades
- determine eligibility for program entry or exit
- diagnose learning difficulties
- predict performance in a later setting

These uses have **high stakes for individual students** and scores must always be considered in combination with other information.

- evaluate teachers
- evaluate schools or districts
- evaluate programs or services

These uses have **high stakes for educators and some administrators** and scores must always be considered in combination with other information.

Overview

SCILLSS

Defining assessment literacy

Why use tests?

The life cycle of a test

Validity and how to evaluate it

Asking the right questions

The Life Cycle of a Test

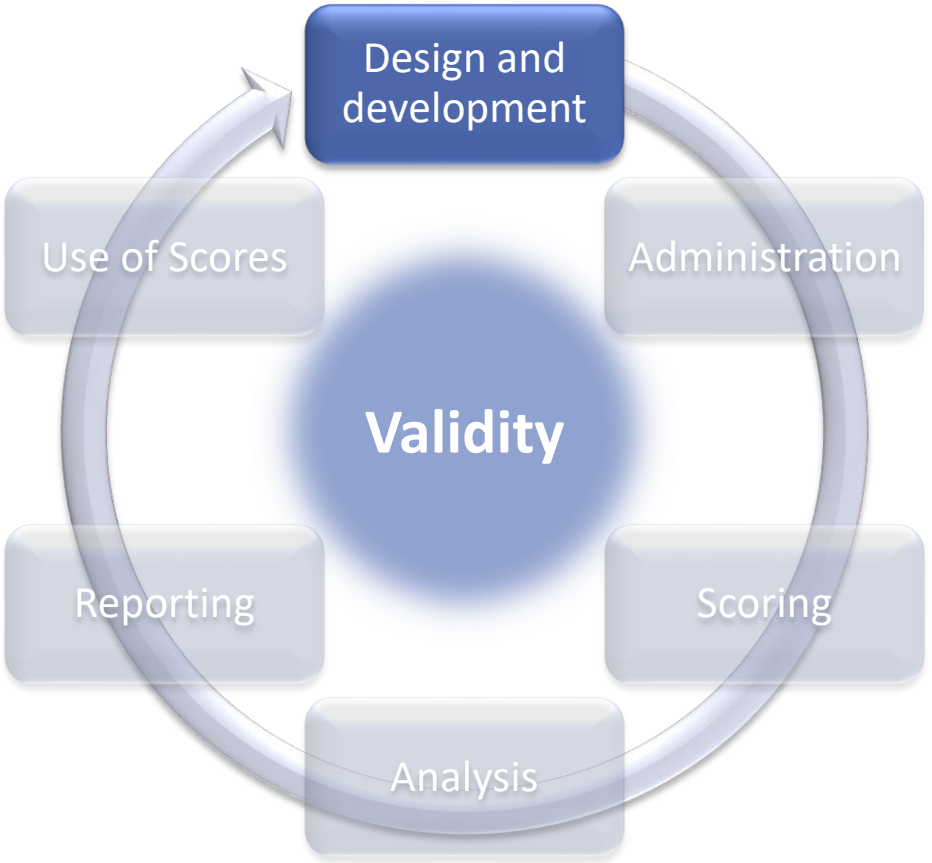
Purpose

18

The Life Cycle of a Test

Key questions for Design and Development

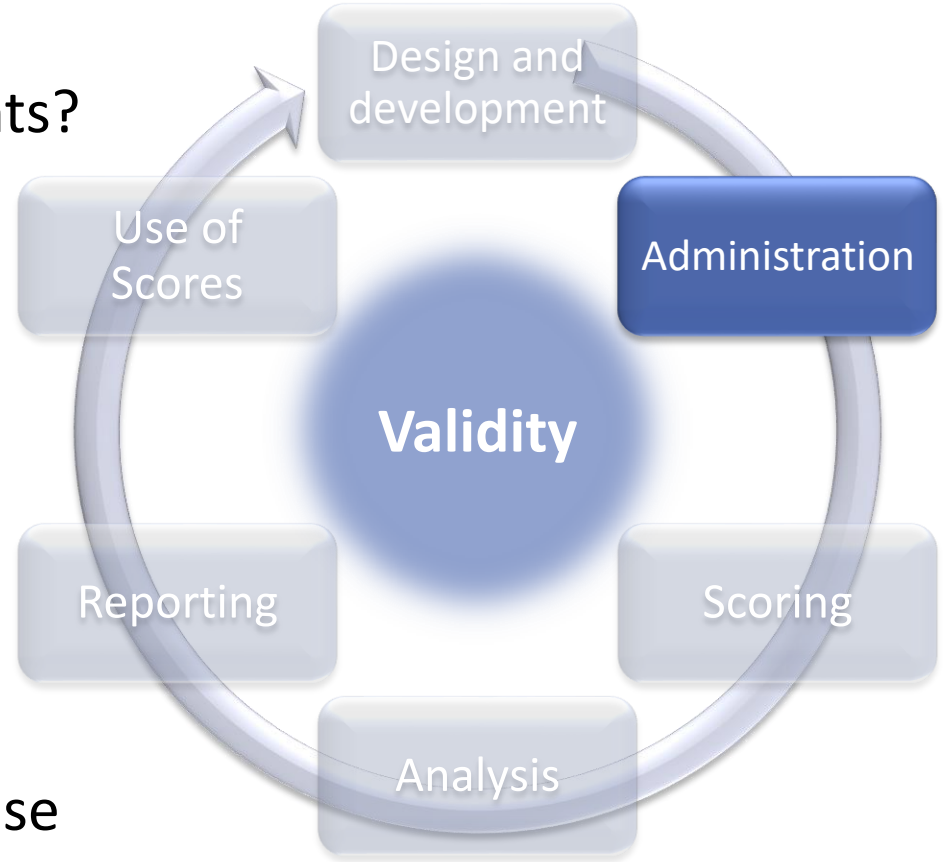
- How will the scores be used?
- What is the test intended to measure? These are the measurement targets.
- What questions or tasks will work best and how should they be combined into a test?
- How will students interact with the questions and record their responses?
- How will responses be scored, analyzed, reported?



The Life Cycle of a Test

Key questions for Administration

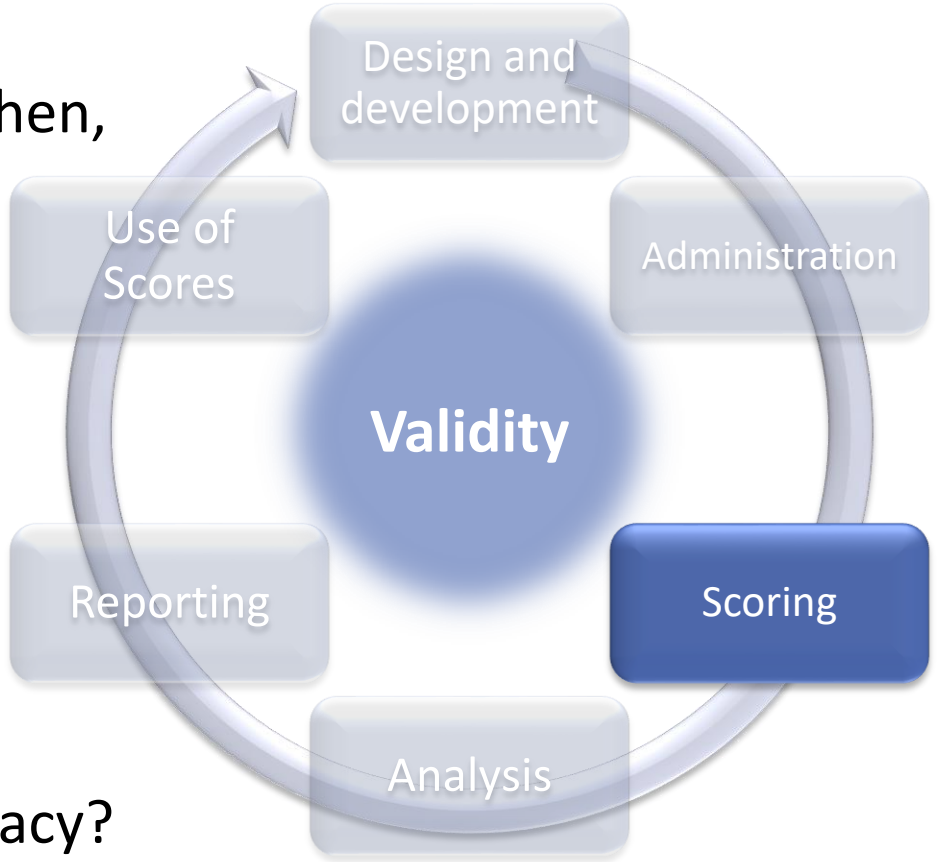
- How will the items be presented to students?
- How will students record their answers?
- How much time will students get?
- Will students have access to a calculator, ruler, formula sheet, textbook, etc.?
- Will some students need accommodations?
- How will I handle any irregularities that arise during administration?



The Life Cycle of a Test

Key questions for Scoring

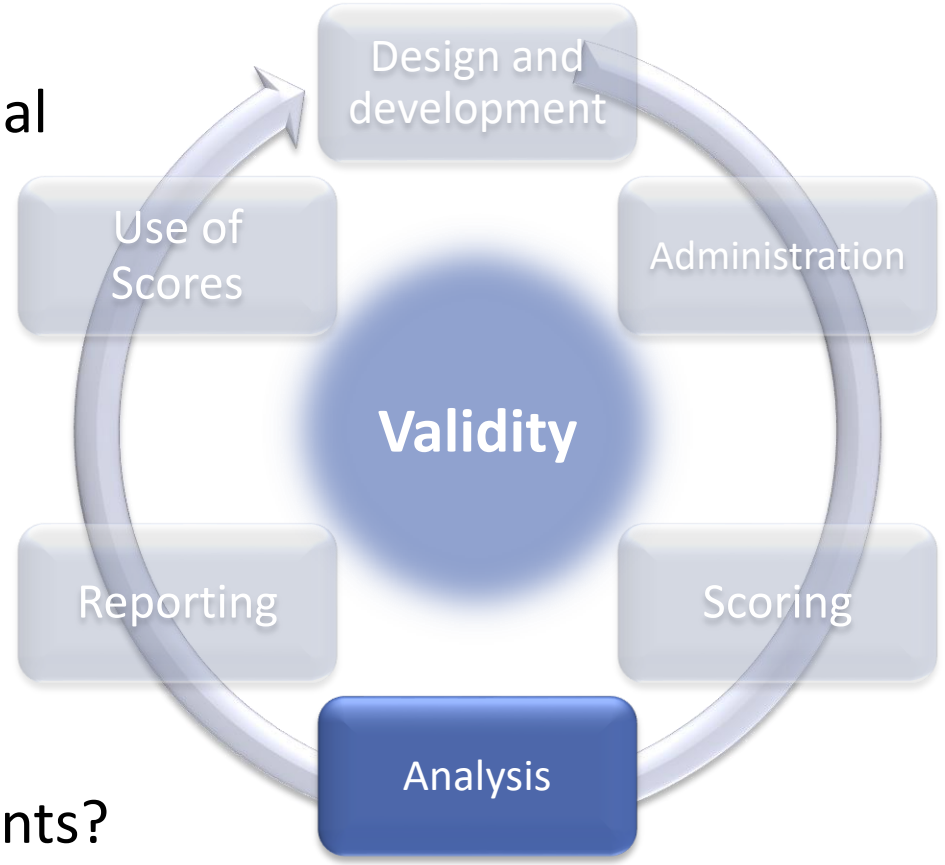
- How will students' responses be scored, when, and by whom?
- If students' responses are to be scored using a rubric, how was the rubric developed and how is it applied accurately and consistently?
- How will scoring be evaluated?
- How will scores be recorded and saved to ensure accuracy and protect students' privacy?



The Life Cycle of a Test

Key questions for Analysis

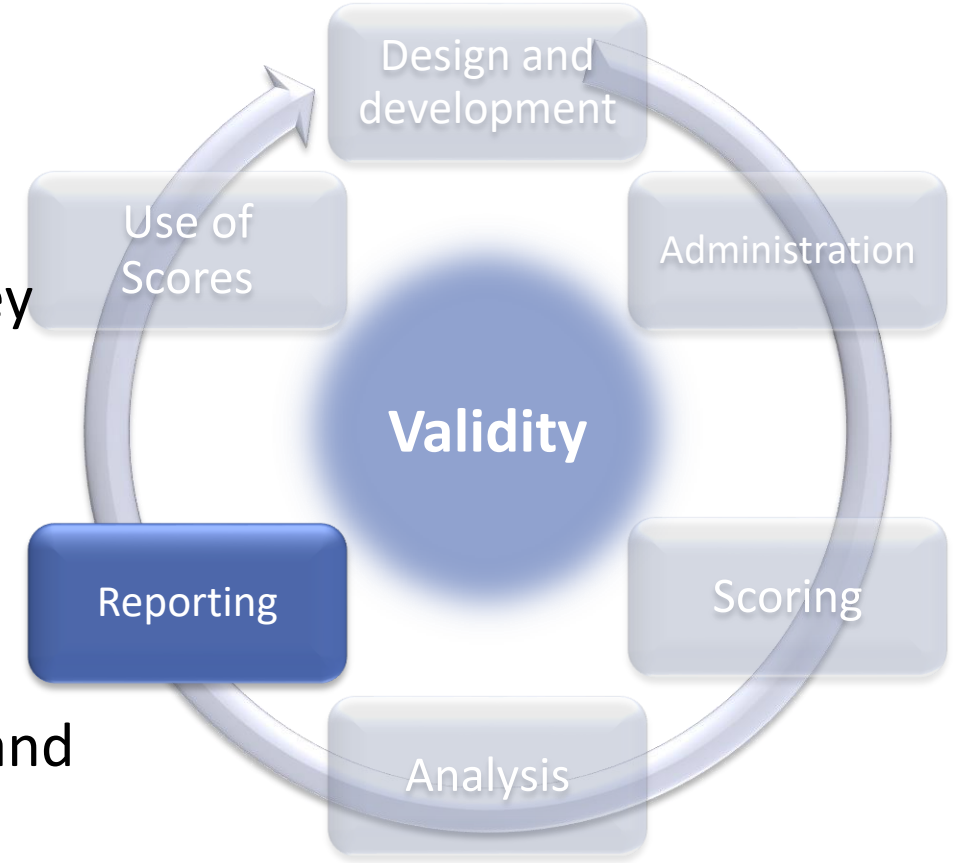
- What scores must be reported for individual students? For groups of students? For the total test? For sections or parts of the test? For items?
- Will raw and/or scores be reported?
- How and when will scores be calculated?
- What analyses are necessary to support comparisons of scores across different administrations? Different groups of students?



The Life Cycle of a Test

Key steps for Reporting

- What information is to be reported to students? Their parents? Their teachers?
- How will scores and guidance on what they mean be conveyed? When?
- How will each student's private testing information be protected during and after the reporting process?
- Who will have access to students' scores and any other information about their participation and performance?



The Life Cycle of a Test

Key steps for the Use of Scores

- 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?

Overview

SCILLSS

Defining assessment literacy

Why use tests?

The life cycle of a test

Validity and how to evaluate it

Asking the right questions

Validity

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

(Oxford English Dictionary)

Validity in Assessments

No test can be valid in and of itself.

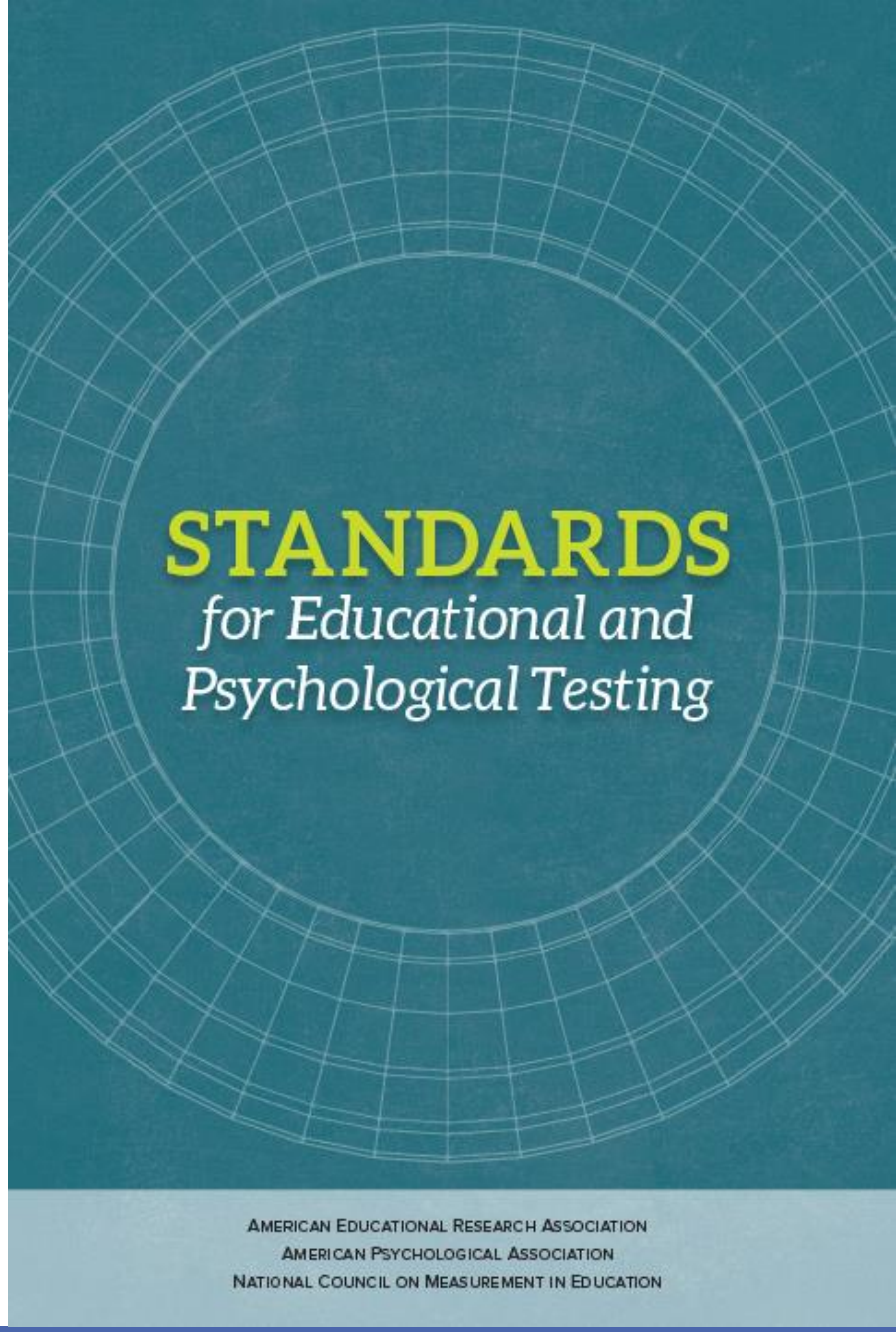
Validity depends on the strength of the evidence regarding what a test measures and how its scores can be interpreted and used.

Assessment validity is a judgment based on a multi-faceted body of evidence.

Validity in Assessments

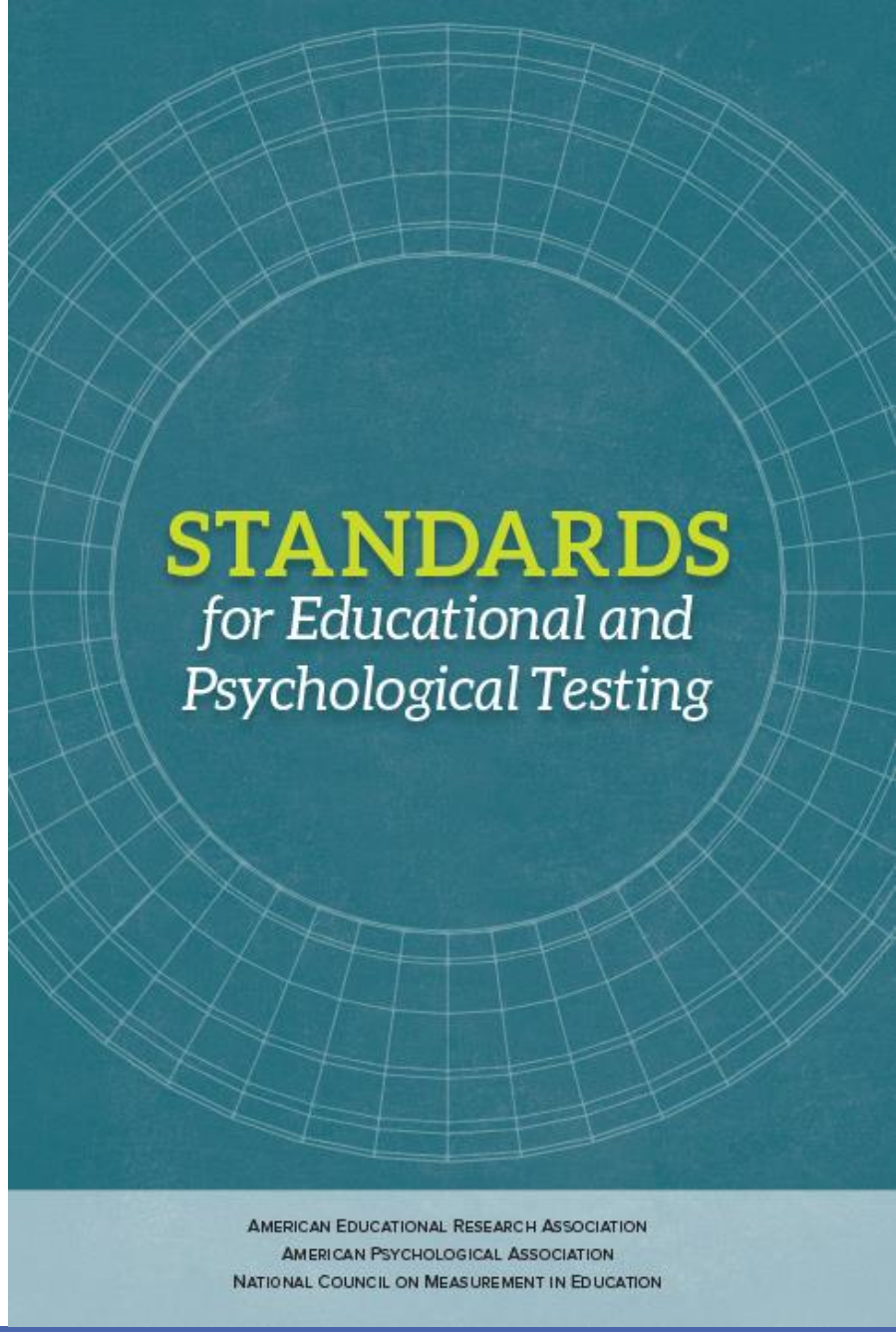
Standard 1.0. Clear articulation of each intended test score interpretation for a specified use should be set forth, and appropriate validity evidence in support of each intended interpretation should be provided.

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



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

(AERA, APA, & NCME, 2014, p. 85)



Overview

SCILLSS

Defining assessment literacy

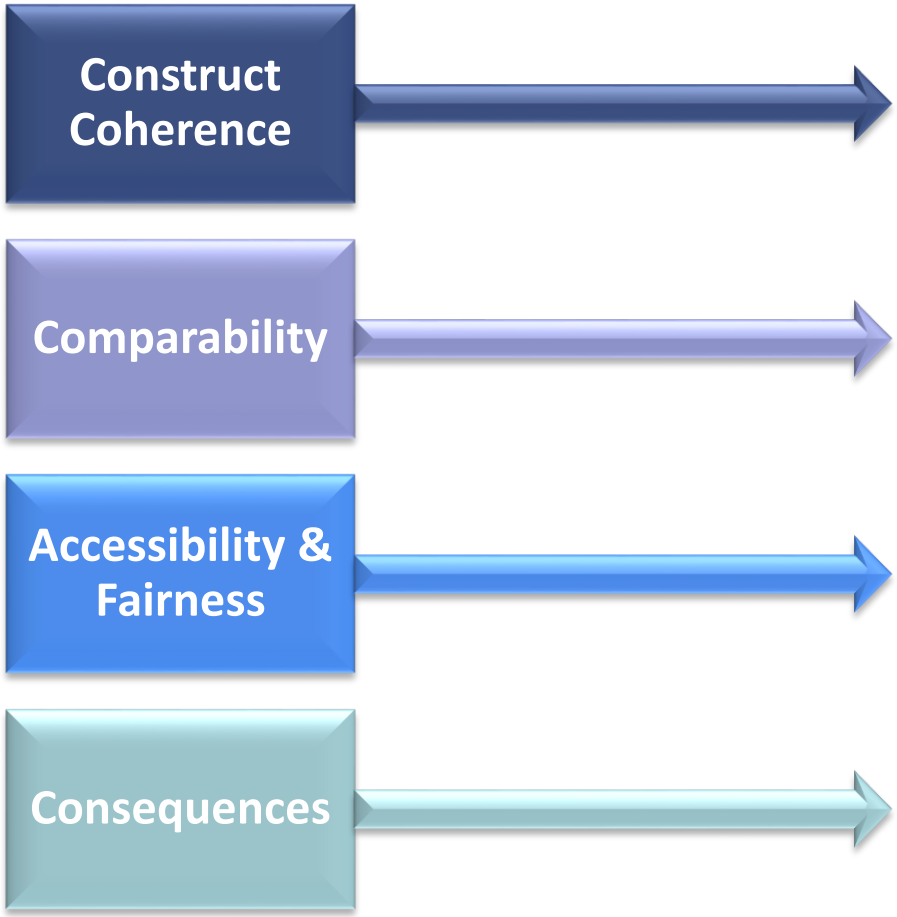
Why use tests?

The life cycle of a test

Validity and how to evaluate it

Asking the right questions

Validity Questions



Validity Questions

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 forms, students, test sites, and time?
Accessibility & Fairness	To what extent does the test allow all students to demonstrate what they know and can do?
Consequences	To what extent are the test scores used appropriately to achieve specific goals?

Construct: The concept or characteristic that a test is designed to measure.*

Comprehension of text presented in Unit 6

Skills in modeling energy transfer in chemical reactions

Resilience

Three digit subtraction skills, end of 3rd grade

Phonemic awareness

Intrinsic motivation

*AERA, APA, & NCME, 2014, p. 217

Construct Coherence

To what extent does the assessment yield scores that reflect the knowledge and skills we intend to measure (e.g., academic standards)?

Why is this evidence important?

To ensure that the assessment has been designed, developed, and implemented to yield scores that reflect the constructs we intend to measure.

Comparability

To what extent are the assessment scores reliable and consistent in meaning across all students, classes, and schools?

Why is this evidence important?

To ensure the assessment scores carry consistent meaning across test forms, students, administration sites, and time

Accessibility and Fairness

To what extent does the assessment allow all students to access the content and demonstrate their knowledge and skills?

Why is this evidence important?

To ensure that test scores reflect what we're intending to measure about students' knowledge and skills and not irrelevant characteristics

Consequences

To what extent does the assessment yield information that is used appropriately to achieve specific goals?

Why is this evidence important?

To ensure that test scores are interpreted and used in ways that are appropriate and not interpreted and used in ways that are inappropriate

Strengthening Claims-based Interpretations and Uses of Local and Large-scale Science Assessment Scores (SCILLSS)

Ensuring Rigor in Local Assessment Systems: A Self-Evaluation Protocol

Self-Evaluation Protocol, Steps One and Two: Identifying Purposes and Assessments Used to Serve those Purposes

Need/purpose	Assessment(s) Used to Serve this Purpose
Evaluate science curricula	ISTEP+: science in grades 4, 6, and 10
Monitor learning and guide instruction in math	MAP Growth in grades K-10
Monitor reading development	Edmentum Reading Eggs

Self-Evaluation Protocol, Step Three: Gather and Evaluate the Evidence for Each Assessment

Name of Assessment:

MAP Growth in grades K-10

Who takes this test?

All* students in grades K-10

Key Validity Area	Score	Low (0-6)	Moderate (7-10)	Strong (11-14)
Construct Coherence:		☐	☐	☐
Comparability & Reliability:		☐	☐	☐
Fairness & Accessibility:		☐	☐	☐
Consequences & Use:		☐	☐	☐

*Except some students with disabilities and English learners?

How are scores used?

Low stakes for educators and students		High stakes for students		High stakes for educators	
To guide next steps in instruction	☑	To evaluate learning for calculating grades	☐	To evaluate teachers	☐
To evaluate instruction	☐	To determine eligibility for program entry or exit	☐	To evaluate schools or districts	☐
To evaluate curriculum	☐	To diagnose learning difficulties	☐	To evaluate programs or services	☐
Other uses:		Other uses:		Other uses:	

Measurement targets (the concepts, knowledge, and skills this test is meant to measure):

When and how often is this test administered?

Four times annually: September, December, February, April

Construct Coherence

Question	Summary of Evidence	Comments on Evidence	Adequacy of Evidence
1. How clear are the definitions of the measurement target(s)? How does/do this/these measurement target(s) align with your intended measurement target(s) for the content area and grade level?			☐ Adequate ☐ Incomplete ☐ Lacking
2. How was the assessment developed to measure the measurement target(s)? What evidence do the developers provide to support the quality of their development processes and their implementation?			☐ Adequate ☐ Incomplete ☐ Lacking
3. How are items reviewed and evaluated during the development process to ensure they appropriately address the intended measurement target(s) and not other content, skills, or irrelevant student characteristics? What evidence supports the quality of these reviews and the use of the feedback they provide to improve item quality?			☐ Adequate ☐ Incomplete ☐ Lacking
4. How are items scored in ways that allow students to demonstrate, and scorers to recognize and evaluate, their knowledge and skills? How are the scoring processes evaluated to ensure they accurately capture and assign value to students' responses?			☐ Adequate ☐ Incomplete ☐ Lacking

Question	Summary of Evidence	Comments on Evidence	Adequacy of Evidence
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)?			☐ Adequate ☐ Incomplete ☐ Lacking
6. What independent evidence supports the alignment of the assessment items and forms to the measurement target(s)?			☐ Adequate ☐ Incomplete ☐ Lacking
7. How are scores reported in relation to the measurement target(s)? Do the reports provide adequate guidance for interpreting and using the scores?			☐ Adequate ☐ Incomplete ☐ Lacking
		Number of Adequate ratings: ____ X 2 =	
		Number of Incomplete ratings: ____ X 1 =	
		Number of Lacking ratings: ____ X 0 =	
		Construct Coherence Total =	

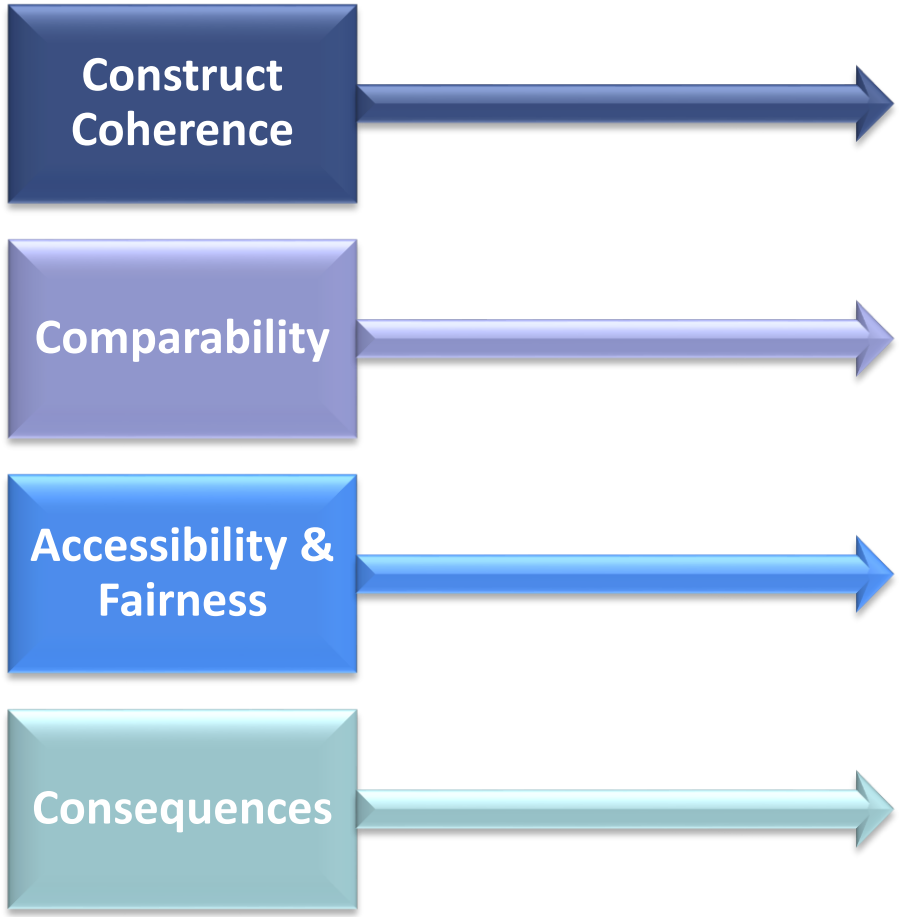
Self-Evaluation Protocol, Step Four: Summary of Individual Assessment Reviews

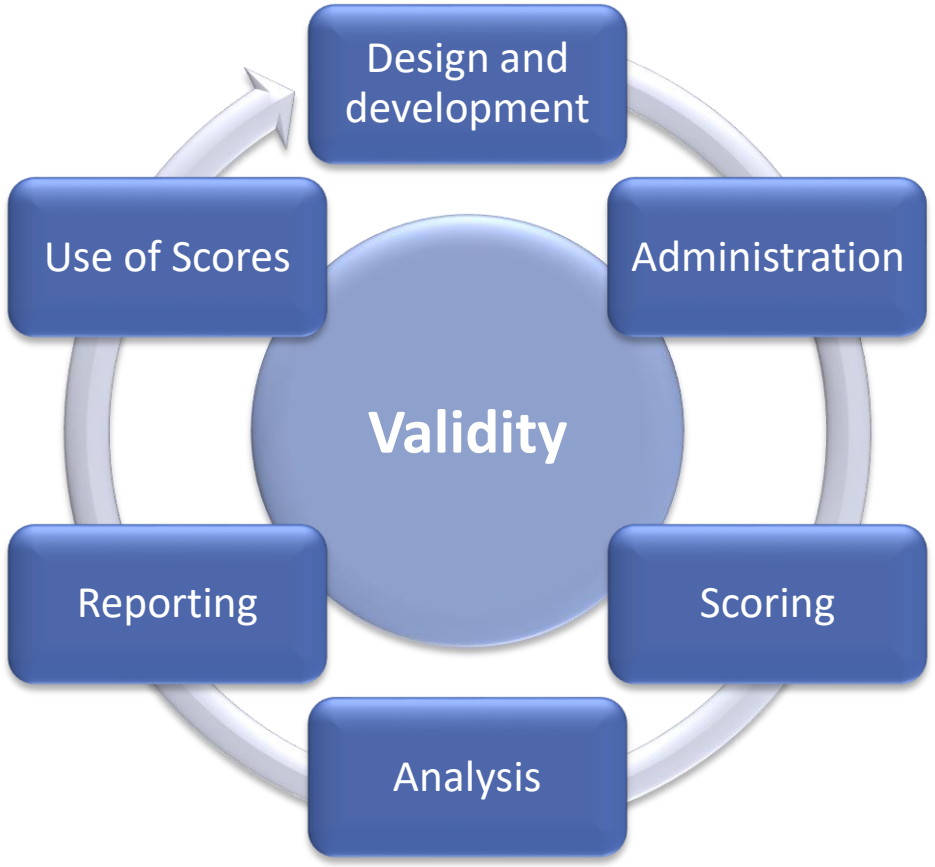
Name of Assessment	Summary of Evidence												Action		
	Construct Coherence			Comparability and Reliability			Fairness & Accessibility			Consequences & Use			Drop	Revisit	Keep as is
	Low 0-6	Moderate 7-10	Strong 11-14	Low 0-6	Moderate 7-10	Strong 11-14	Low 0-6	Moderate 7-10	Strong 11-14	Low 0-6	Moderate 7-10	Strong 11-14			
MAP Growth in grades K-10	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐
	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐

Assessment Literacy

Ellen Forte, Ph.D.
CEO & Chief Scientist
edCount, LLC
eforte@edCount.com

Validity Questions





Construct Coherence

To what extent does the assessment yield scores that reflect the knowledge and skills we intend to measure (e.g., academic standards)?

Why is this evidence important?

To ensure that the assessment has been designed, developed, and implemented to yield scores that reflect the constructs we intend to measure.

Examples of Construct Coherence Questions

1. What are you intending to measure with this test? (What are the measurement targets?)
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 allow students to demonstrate, and scorers to recognize and evaluate, their knowledge and skills? How are the scoring processes evaluated to ensure they accurately capture and assign value to students' responses?
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?
7. How are scores reported in relation to the measurement targets? Do the reports provide adequate guidance for interpreting and using the scores?

Comparability

To what extent are the assessment scores reliable and consistent in meaning across all students, classes, and schools?

Why is this evidence important?

To ensure the assessment scores carry consistent meaning across test forms, students, administration sites, and time

Examples of Comparability Questions

1. How is the assessment designed to yield consistent, reliable scores? What evidence supports score reliability?
2. How is the assessment designed to support comparability of scores across forms? What evidence supports such comparability?
3. How is the assessment designed to support comparability of scores across time? What evidence supports such comparability?
4. How is the assessment designed to support comparability of scores across administration sites? What evidence supports such comparability?
5. How is the assessment administered to protect against various types of cheating so that the scores reflect students' knowledge and skills and not inappropriate access to testing materials? What evidence supports the implementation of these safeguards and security protocols?
6. How is the assessment scored such that scores reflect students' knowledge and skills and not inaccuracies or inconsistencies in the scoring process? What evidence supports implementation of these scoring procedures?
7. How are scores reported in ways that appropriately support or disrupt comparability in score interpretation across time, administration sites, or variations in student characteristics?

Accessibility and Fairness

To what extent does the assessment allow all students to access the content and demonstrate their knowledge and skills?

Why is this evidence important?

To ensure that test scores reflect what we're intending to measure about students' knowledge and skills and not irrelevant characteristics

Examples of Accessibility and Fairness Questions

1. How were the assessment questions developed to ensure that scores do not reflect student characteristics that are irrelevant to the measurement targets?
2. How were the needs of students with disabilities addressed during assessment development? What evidence supports these efforts and their effectiveness?
3. How were the needs of English learners addressed during assessment development? What evidence supports these efforts and their effectiveness?
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 identification and 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 identification and 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. How are assessment scores interpreted in relation to knowledge and skills that test takers have had an **opportunity to learn** or are preparing to learn? What evidence supports the interpretation of students' scores in relation to their learning opportunities?

Consequences

To what extent does the assessment yield information that is used appropriately to achieve specific goals?

Why is this evidence important?

To ensure that test scores are interpreted and used in ways that are appropriate and not interpreted and used in ways that are inappropriate

Examples of Consequences Questions

1. How is the assessment developed, administered, scored, and reported in ways that deter and limit instances of cheating by students or others associated with the assessment or its stakes? What evidence supports the implementation and effectiveness of these efforts?
2. How are the scores from the assessment intended to be used as described by the test developers and how are they used by your state? How well do these uses align?
3. If your state is using test scores for purposes other than those for which the test developers intended, what evidence supports those uses?
4. If assessment scores are associated with recommendations for instruction or other interventions for individual students, what evidence supports such interpretations and uses of these scores? What tools and resources are available to educators for evaluating and implementing these recommendations?
5. If assessment scores are associated with recommendations for whole-class or group instruction, what evidence supports such interpretations and uses of these scores? What tools and resources are available to educators for evaluating and implementing these recommendations?
6. If assessment scores are associated with high stakes decisions for teachers, administrators, schools, or other entities or individuals, what evidence supports such interpretations and uses of these scores?
7. How are scores reported to students and parents in ways that support their understanding of the scores and any associated recommendations or decisions?