



Strengthening the Meaning and Utility of Test Scores for their Intended Uses

2019 OPI Data and Assessment Conference:
Get READY for 2020!

Bozeman, Montana

This presentation was developed with funding from the U.S. Department of Education under Enhanced Assessment Grants Program CFDA 84.368A. The contents do not necessarily represent the policy of the U.S. Department of Education, and no assumption of endorsement by the Federal government should be made.

About SCILLSS

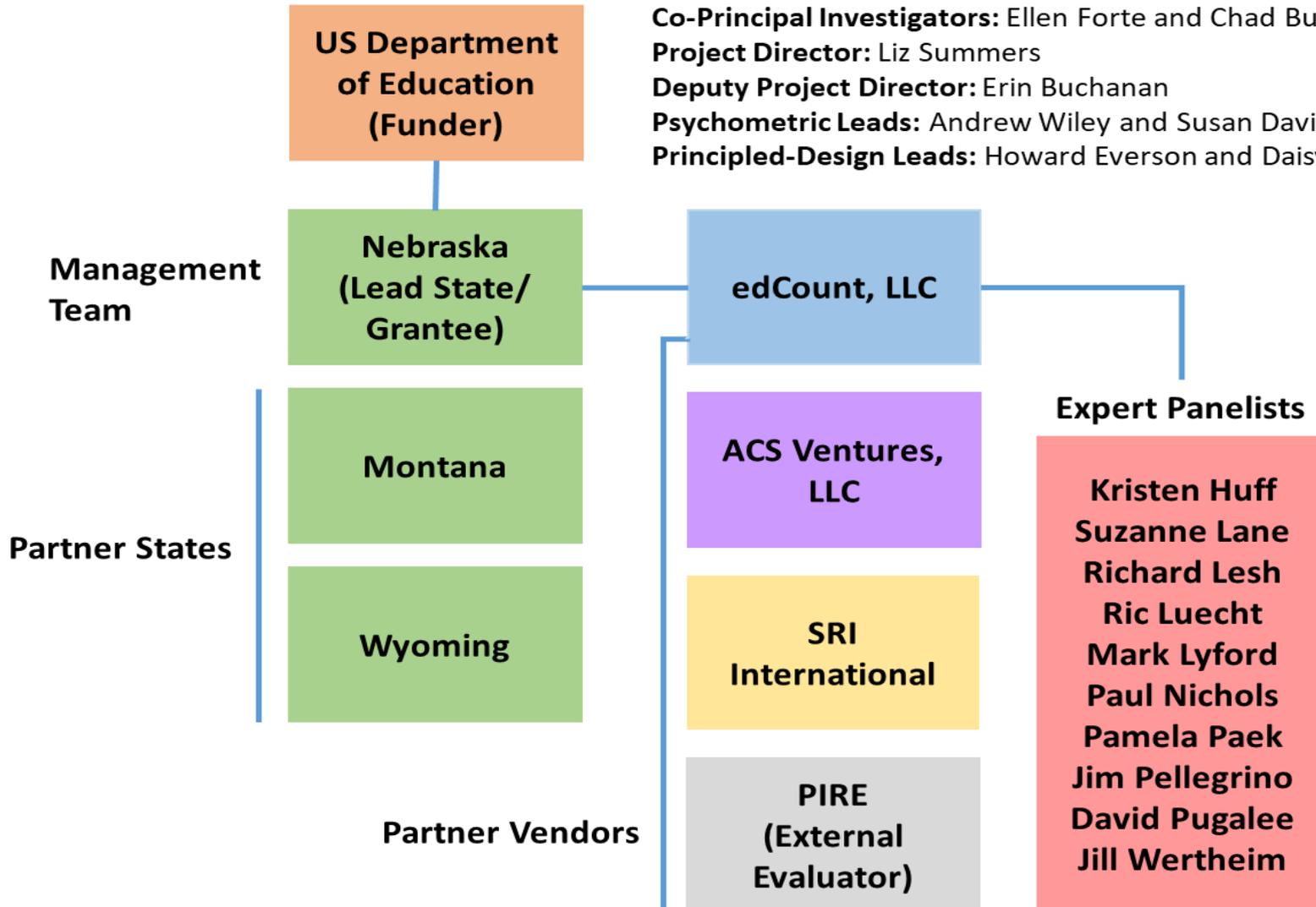


- **Strengthening Claims-based Interpretations and Uses of Local and Large-scale Science Assessment Scores**
- One of two projects funded by the US Department of Education’s Enhanced Assessment Instruments Grant Program (EAG), announced in December 2016
- Four-year timeline (April 2017 – December 2020)
- Collaborative partnership including three states, four organizations, and 10 expert panel members
- Nebraska is the grantee and lead state; Montana and Wyoming are partner states

SCILLSS Partner States, Organizations, and Staff



Co-Principal Investigators: Ellen Forte and Chad Buckendahl
Project Director: Liz Summers
Deputy Project Director: Erin Buchanan
Psychometric Leads: Andrew Wiley and Susan Davis-Becker
Principled-Design Leads: Howard Everson and Daisy Rutstein

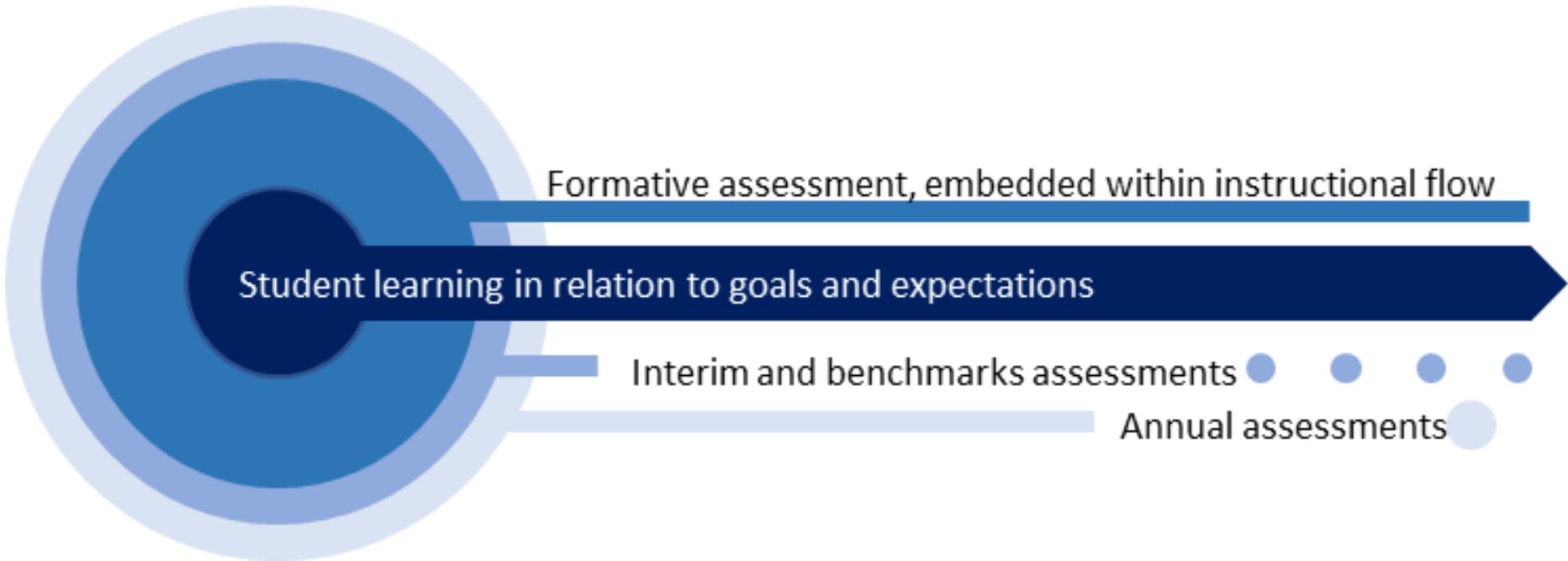


SCILLSS Project Goals



- Create a science assessment design model that **establishes alignment with three-dimensional science standards by eliciting common construct definitions** that drive curriculum, instruction, and assessment
- Strengthen a shared knowledge base among instruction and assessment stakeholders for using **principled-design approaches** to create and evaluate science assessments that **generate meaningful and useful scores**
- Establish a means for state and local educators to **connect statewide assessment results with local assessments and instruction** in a coherent, standards-based system

Assessment of Student Learning



SCILLSS Resources and Student Learning



SCILLSS Goal 1, Coherence: Establish a means for states to strengthen the meaning of statewide science assessment results and to connect those results with local science curriculum, instruction, and assessment

SCILLSS Goal 2, Support Implementation of Principled-Design: Strengthen the knowledge base and experience among stakeholders in using principled-design approaches to create and evaluate quality science assessments that generate meaningful and useful scores

Student learning in relation to goals and expectations

Interim and other classroom assessments ● ● ● ●

Annual assessments ●

SCILLSS resources

A Principled-Design Approach to Creating PLDs and Building Score Scales

Purpose: To explain how and why to develop PLDs and score scales using a principled-design approach

Audience: State and local educators; vendors

Format: White paper

Guide to Developing Three-Dimensional Science Tasks for Large-Scale Assessments

Purpose: To guide implementation of principled-approaches for developing three-dimensional tasks aligned to NGSS-like standards for large-scale science assessments

Audience: State administrators; vendors

Format: Guidebook; templates; tasks; exemplars

Guide to Developing Three-Dimensional Science Tasks for Classroom Assessments

Purpose: To guide implementation of principled-approaches for developing three-dimensional tasks aligned to NGSS-like standards for use within classrooms

Audience: Local educators and administrators

Format: Guidebook; templates; tasks; exemplars

Professional Learning Sessions on Using a Principled Approach to Designing Classroom Assessment Tasks

Purpose: To support local educators in applying principled-design in the development of classroom assessment tasks that link to curriculum and instruction

Audience: Local educators and administrators

Format: Workbook; templates; PPT slides; guiding questions

Assessment Fundamentals

Self-Evaluation Protocols

Purpose: To support educators in evaluating the quality of the assessments in their assessment systems

Audience: State and local educators; vendors

Format: Protocol

Assessment Literacy Workbook

Purpose: To strengthen educators' understanding of and ability to make good decisions about assessments

Audience: State and local educators; vendors

Format: Digital workbook

Theory of Action Principles

Assessment systems are developed such that they can inform improvements to curriculum and instruction

Assessments are equitable, accessible, and culturally relevant for widest range of students

Educators use student performance data appropriately to monitor progress toward CCR and to inform teaching

State assessments connect coherently to local C-I-A in a way that provides comprehensive coverage of the standards

Stakeholders collaborate to effectively coordinate alignment of C-I-A systems



Assessment Literacy and Validity Evidence

Purpose and Use of Assessment Scores

The Life Cycle of a Test

Four Fundamental Validity 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.

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 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?
 - What do students already know about what I'm about to teach?
 - How well are students understanding this lesson so far?
- Why do you want to know this and what will you do with this information?
 - 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

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.

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

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

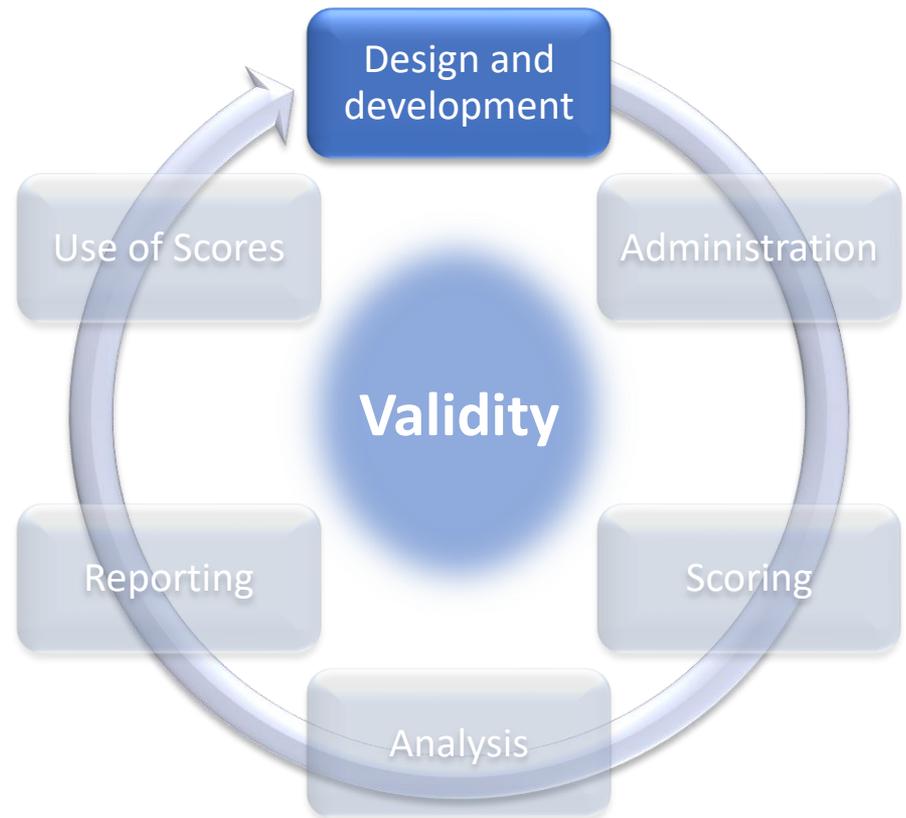
The Life Cycle of a Test



The Life Cycle of a Test

Key questions for Design and Development

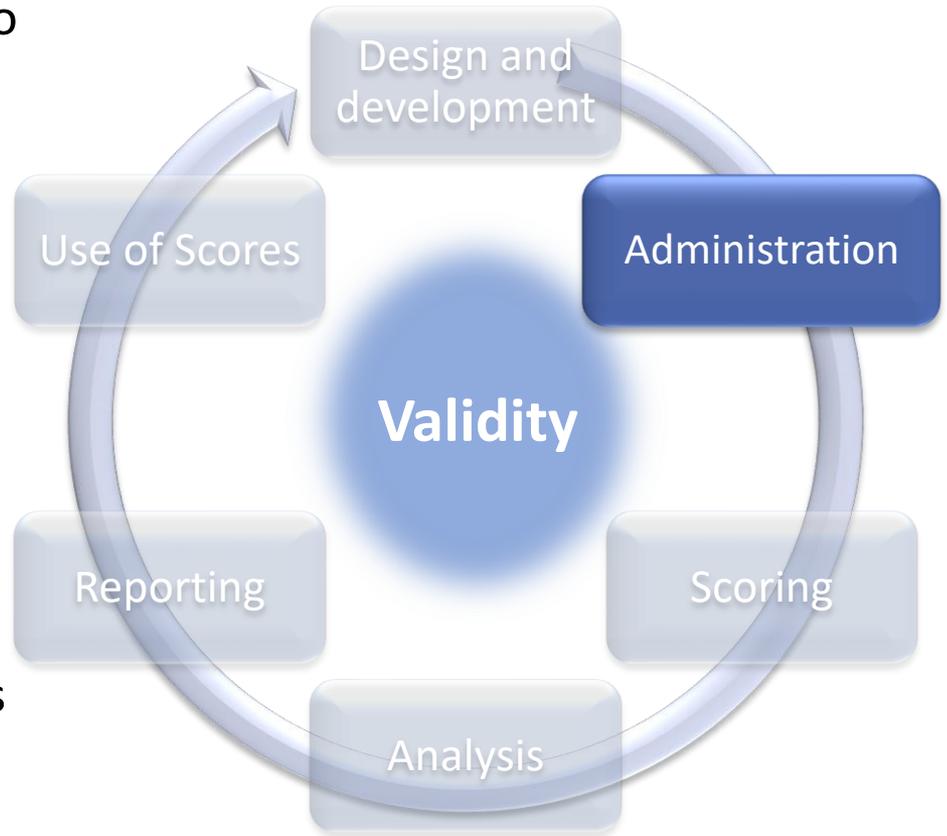
- How will the scores be used?
- What is the test intended to measure? These are the measurement targets.
- How can these targets best be measured?
- What questions or tasks will work best and how should they be combined into a test?
- How will students interact with the questions and record their responses?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

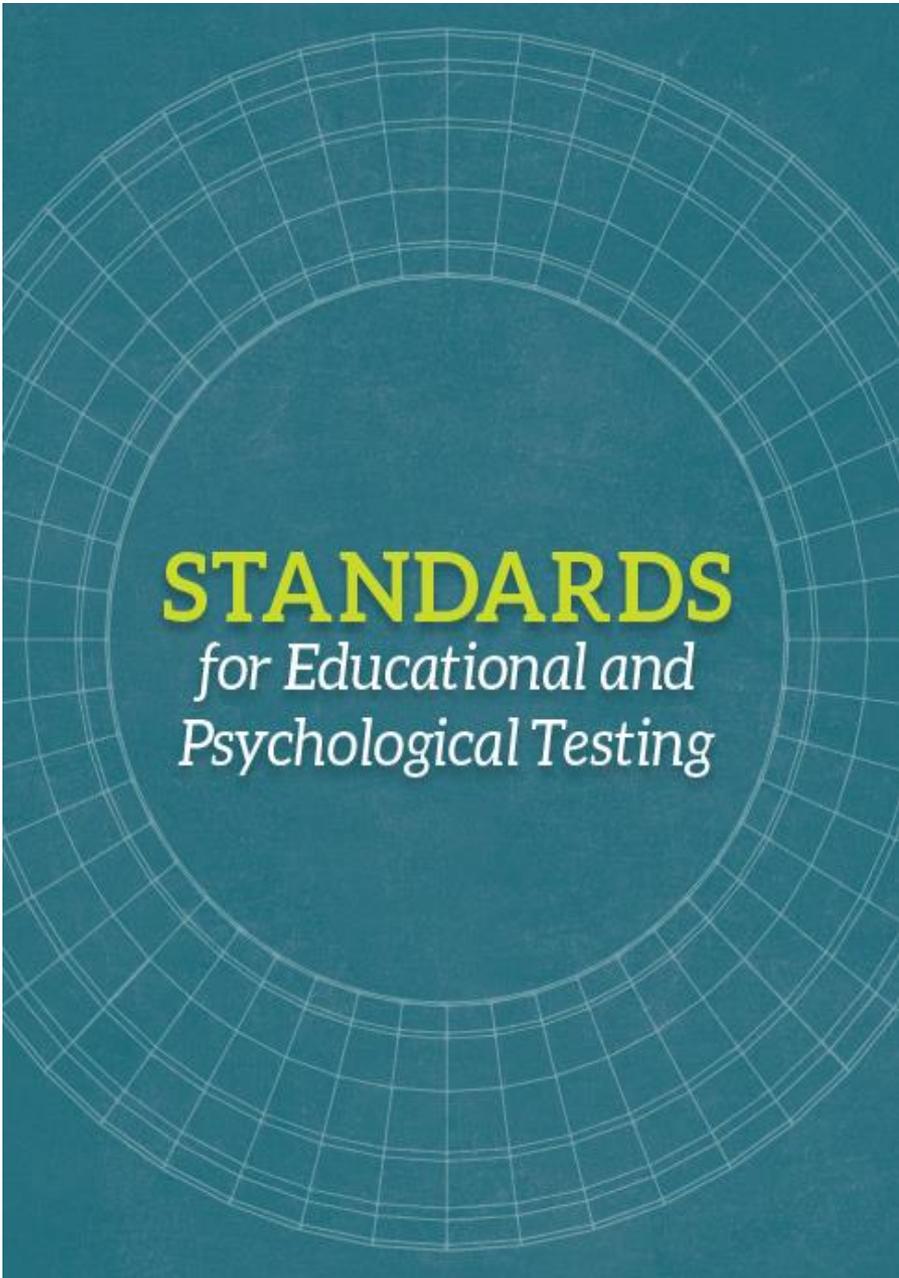


Validity in Assessments

- *The degree to which evidence and theory supports the interpretations of test scores for the proposed uses of the tests.*
- 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.

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.

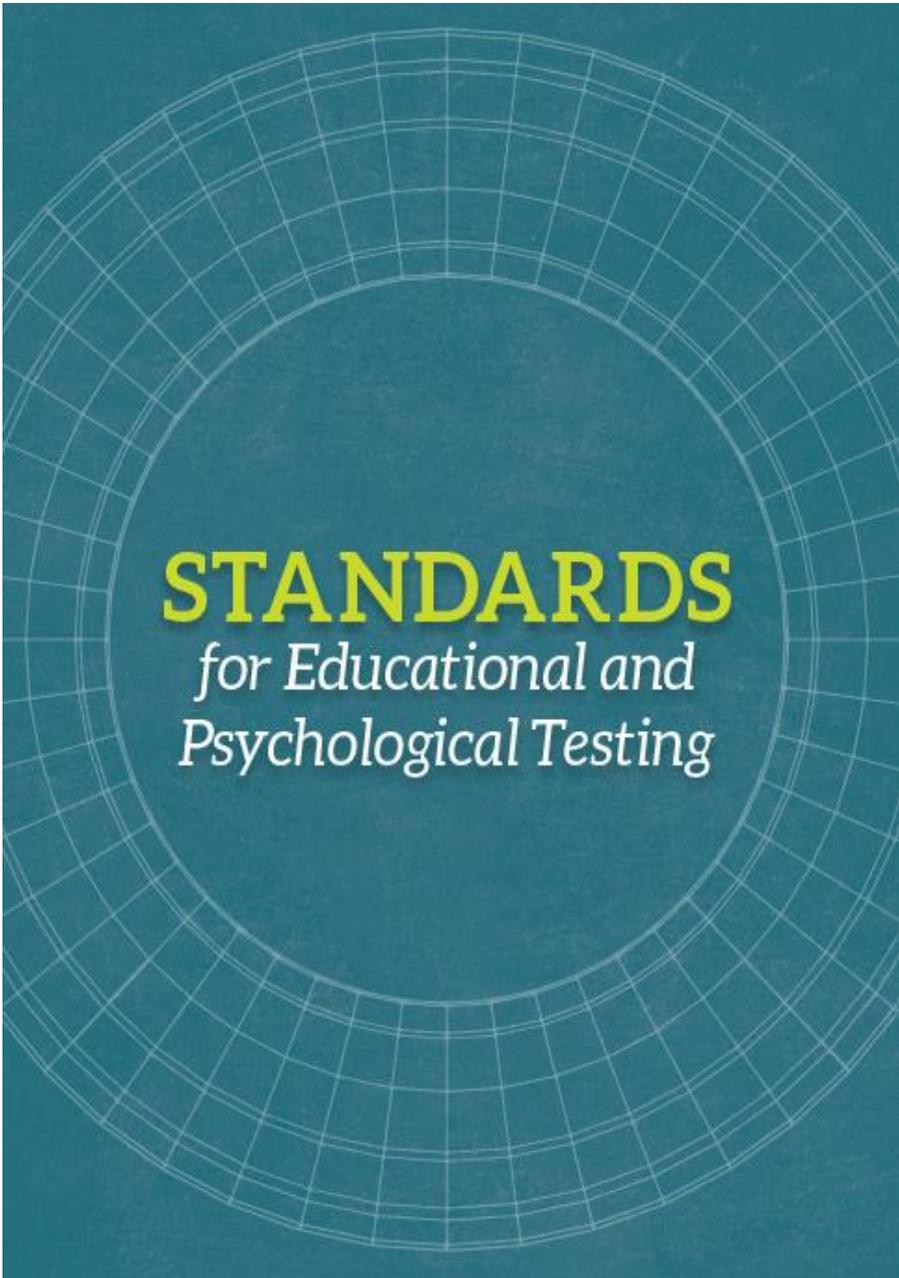
(AERA, APA, & NCME, 2014, p. 23)



STANDARDS
*for Educational and
Psychological Testing*

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)



STANDARDS
*for Educational and
Psychological Testing*

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?

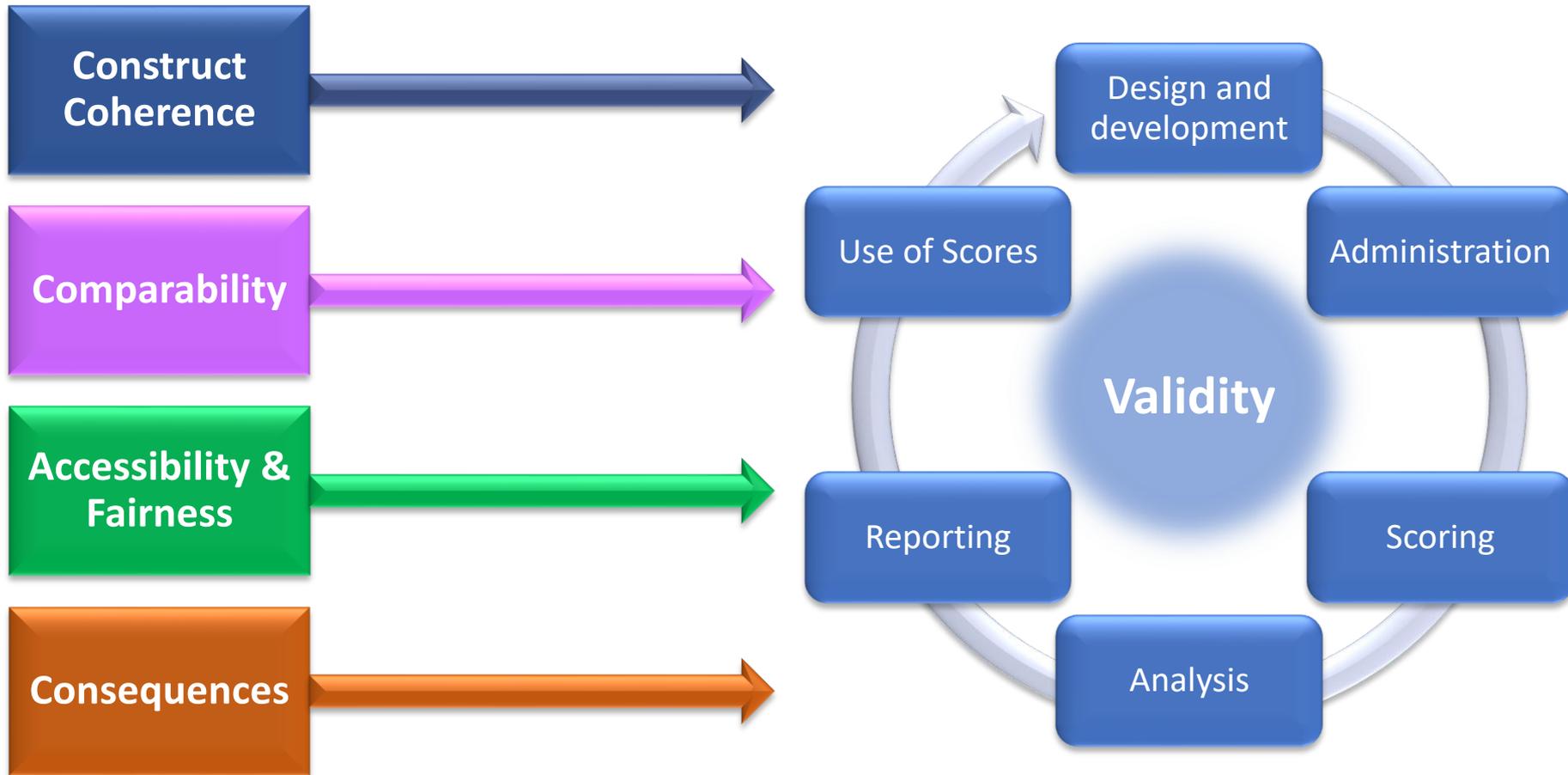
Accessibility & Fairness

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?

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)?

Construct:

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

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



Overview of the Local and State Self-Evaluation Protocols

Purpose, Audience, and Structure



Self-evaluation Protocols Purposes

The local and state self-evaluation protocols are designed to provide frameworks for how local schools or districts can consider how to best implement their local assessment program, and how states can evaluate their options for their statewide assessment program.

Local or District	State
<ul style="list-style-type: none">• Designed to focus on assessments used within the classroom• Typically used for lower-stakes decisions<ul style="list-style-type: none">– Helps guide curriculum– Provides information to teachers on the status of students' progression through key topics or curriculum items	<ul style="list-style-type: none">• Focused on higher-stakes assessments• Typically used within accountability models

Self-evaluation Protocols Audience



The self-evaluation protocols are intended for a specific audience:

- State and district administrators who may be—
 - Instructional leaders
 - Content specialists
 - Assessment specialists
 - Decision-makers regarding state or local assessments
- These educational leaders will strengthen their assessment literacy by building their knowledge base, understanding the nuances of validity and reliability, and applying their knowledge in the evaluation of their own systems.

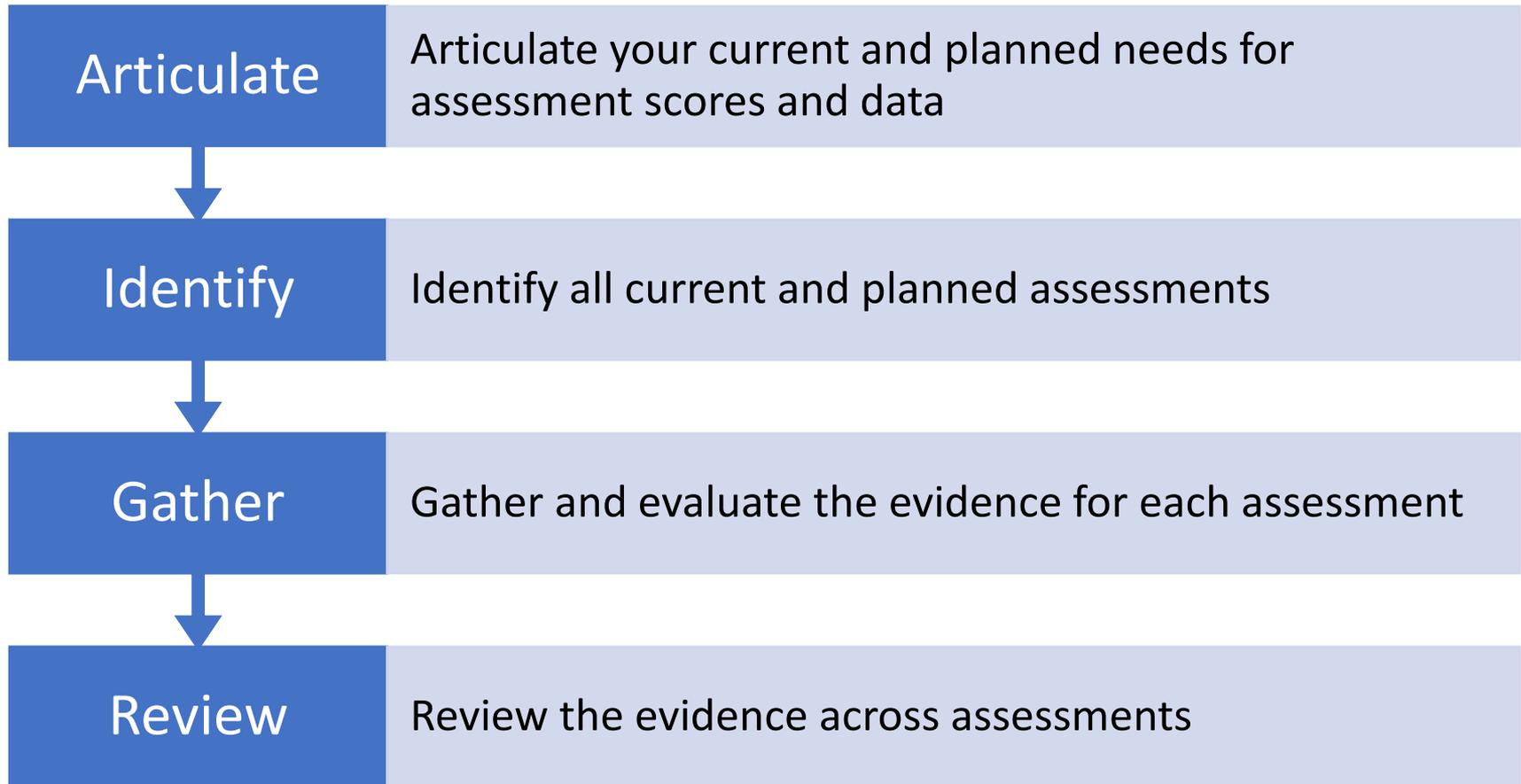
Self-evaluation Protocols

Goals and Objectives

- Identify intended use(s) of test scores
- Foster an internal dialogue to clarify the goals and intended uses of an assessment program
- Evaluate whether given test interpretations are appropriate
- Identify gaps in assessment programs
- Identify overlap across multiple assessment programs

Self-evaluation Protocol Structure

4 Steps



Step 1: Articulate Current Needs

- What is the purpose of the given assessment and how will it be used?
- What are the stakes associated with your assessment?
 - Low stakes – Instructional guidance
 - High stakes – Evaluate programs or services
- What questions are you trying to answer about student achievement?
- What information is provided from your assessment and how does it address your questions about student achievement?
 - Is there additional information that the assessment could provide to help address these questions?

Step 2: Identify Current Assessments

- Gather all current and planned assessments
- Assessments can be organized multiple ways
 - All assessments within a grade level
 - All assessments within a given content area
- Identify areas where overlap in information has occurred
- Identify areas where a gap appears to exist



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
<i>Evaluate science curricula</i>	<i>ISTEP+: science in grades 4, 6, and 10</i>
<i>Monitor learning and guide instruction in math</i>	<i>MAP Growth in grades K-10</i>
<i>Monitor reading development</i>	<i>Edmentum Reading Eggs</i>

Step 3: Gather and Evaluate Validity Evidence



FOUR FUNDAMENTAL VALIDITY QUESTIONS

1. To what extent does the assessment as designed capture the knowledge and skills defined in the target domain?
2. To what extent does the assessment as implemented yield scores that are comparable across students, sites, time, forms?
3. To what extent are students able to demonstrate what they know and can do in relation to the target knowledge and skills on the test in a manner that can be recognized and accurately scored?
4. To what extent does the test yield information that can be and is used appropriately within a system to achieve specific goals?

Construct Coherence

Comparability

Accessibility and Fairness

Consequences



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 <u>(7-10)</u>	Strong (11-14)
Construct Coherence: _____		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Comparability & Reliability: _____		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Fairness & Accessibility: _____		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
Consequences & Use: _____		<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>

**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 <input checked="" type="checkbox"/>	To evaluate learning for calculating grades <input type="checkbox"/>	To evaluate teachers <input type="checkbox"/>
To evaluate instruction <input type="checkbox"/>	To determine eligibility for program entry or exit <input type="checkbox"/>	To evaluate schools or districts <input type="checkbox"/>
To evaluate curriculum <input type="checkbox"/>	To diagnose learning difficulties <input type="checkbox"/>	To evaluate programs or services <input type="checkbox"/>
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?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> 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?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
3. How are items reviewed and evaluated during the development process to ensure they appropriately address the intended measurement target(s) and <u>not</u> 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?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> 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?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> 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)?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
6. What independent evidence supports the alignment of the assessment items and forms to the measurement target(s)?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> 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?			<input type="checkbox"/> Adequate <input type="checkbox"/> Incomplete <input type="checkbox"/> Lacking
Number of Adequate ratings: ____ X 2 =			
Number of Incomplete ratings: ____ X 1 =			
Number of Lacking ratings: ____ X 0 =			
Construct Coherence Total =			

Step 4: Review and Evaluate Current Assessment Program



- Review each assessment's purpose and use
- Identify areas with adequate evidence for test scores
- Identify areas where the degree of data available is either incomplete or lacking



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	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>
	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>	<input type="checkbox"/>



Overview of the Digital Workbook on Educational Assessment Design and Evaluation

Purpose, Audience, and Structure

Digital Workbook Purpose

The digital workbook includes five assessment literacy chapters designed to:

- Inform state and local educators and other stakeholders on the purposes of assessments
- Ensure a common understanding of the purposes and uses of assessment scores, and how those purposes and uses guide decisions about test design and evaluation
- Complement the needs assessment by providing background information and resources for educators to grow their knowledge about foundational assessment topics
- Address construct coherence, comparability, accessibility and fairness, and consequences

Digital Workbook Audience

The digital workbook is intended for a specific audience:

- State and district administrators who may be—
 - Instructional leaders
 - Content specialists
 - Assessment specialists
 - Decision-makers regarding state or local assessments
 - Responsible for implementing State and Local Self-Evaluation Protocols
- These educational leaders will strengthen their assessment literacy by building their knowledge base, understanding the nuances of validity and reliability, and applying their knowledge in the evaluation of their own systems.

Digital Workbook Series Overview

Chapter 1	Validity, validity evidence, and the assessment life cycle (design and development, administration, scoring, analysis, reporting, score use)
Chapter 2	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?
Chapter 3	Comparability: To what extent are the test scores reliable and consistent in meaning across all students, classes, and schools?
Chapter 4	Accessibility and Fairness: To what extent does the test allow all students to demonstrate what they know and can do?
Chapter 5	Consequences: To what extent are the test scores used appropriately to achieve specific goals?

Accessing the Digital Workbook

- The local and state needs assessments and first two chapters in the digital workbook can be found at this link: <http://www.scillspartners.org/scillss-resources/>
- The remaining three chapters of the digital workbook are currently being developed and will be made available by March 2019.

Key Take-aways: Purpose

- Every test must have a purpose.
- Validity relates to the interpretation and use of test scores for a specific purpose and not to the test itself.
- Test scores may be used for more than one purpose, but validity evidence is necessary for each purpose.

Key Take-aways: Validity Evidence

- Validity evidence should be gathered in relation to key questions related to:
 - Construct coherence
 - Comparability
 - Accessibility and fairness
 - Consequences
- Validity evidence should gathered throughout the life cycle of a test.

“If the purpose for learning is to score well on a test, we've lost sight of the real reason for learning.”

Jeannie Fulbright

#WhatSchoolCouldBe

Comments and Questions



Liz Summers

202-297-8629

LSUMMERS@EDCOUNT.COM