Validity and Educational Testing: Purposes and Uses of Educational Tests

Jennifer Lewis & Stephen G. Sireci

University of Massachusetts Amherst

Learning Objectives

1. Describe the three steps of test development that maximize the benefits of tests.
2. Define validity
3. Define and illustrate the five sources of validity evidence.
4. Provide examples of documenting the five sources of validity evidence.

Other Useful Resources

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education (2014). *Standards for educational and psychological testing,* American Educational Research Association.

Fremer, J. & Wall, J. (2004). Why use tests and assessments? In J. Wall and G. Walz (Eds.), *Measuring up: Assessment issues for teachers, counselors, and administrators* (pp. 3-19), CAPS Press.

Furr, R.M. (2011). Evaluating Psychometric Properties: Dimensionality & Reliability. In *Scale Construction and Psychometrics for Social & Personality Psychology*. (pp. 25-51), SAGE Publications, Ltd.

Miller, M. D., Linn, R. L., & Gronlund, N. E. (2009). *Measurement and assessment in teaching* (10th edition). Merrill.

Sireci, S. G. (2007). On test validity theory and test validation. *Educational Researcher*, 36(8), 477-481.

Sireci, S. G. (2020) "De-“Constructing test validation. *Chinese/English Journal of Educational Measurement and Evaluation*, 1. 教育测量与评估双语季刊: Available at: https://www.ce-jeme.org/journal/vol1/iss1/3

Sireci, S. G., & Faulkner-Bond (2014). Validity evidence based on test content. *Psychometrika*, 26, 100-107. doi: https://10.7334/psicothema2013.256.

Rios, J., & Wells, C. (2014). Validity evidence based on internal structure. *Psicothemia*, 26(1), 108–116. doi: 10.7334/psicothema2013.260

Zenisky, A. L., Sireci, S. G., Lewis, J., Lim, H., O’Donnell, F., Wells, C. S., Padellaro, F., Jung, H. J., Banda, E., Pham, D., Hong, S. E., Park, Y., Botha, S., Lee, M.,& Garcia, A. (2018). Massachusetts Adult Proficiency Tests - College and Career Readiness (MAPT-CCR) Technical Manual. *Center for Educational Assessment Research Report No. 974.* Amherst, MA: University of Massachusetts, Center for Educational Assessment.

Referenced ITEMS Modules

Module Citation

Lewis, J. & Sireci, S. G.(2022). Validity and educational testing: Purposes and uses of educational tests [Digital ITEMS Module 30]. *Educational Measurement: Issues and Practice, 41*(4), 81-82. https://doi.org/10.1111/emip.12533

Purposes and Uses of Educational Tests

Section Learning Objectives

1 Purposes and Uses of Educational Tests

Gain a deeper understanding of the purposes of tests

Examine the benefits and criticisms of tests

Understand the steps of developing educational tests of high quality

Test Purpose

What questions can be answered with assessment results?

- Context matters
  - Educational assessments
    - Classroom assessments
    - Interim/benchmark assessments
    - Summative assessments
  - Other assessments
    - Credentialing and Licensure exams
    - Entrance exams
- This list is NOT exhaustive

Test Purpose

What questions can be answered with **classroom assessment** results?

- How effective is the instruction?
- Are students meeting expectations/learning goals?
  - If not, then what review is needed to move students closer to their learning goals?
- What do students already know about this topic?

Test Purpose

What questions can be answered with **other assessment** results?

- Does a candidate have the skills and knowledge required for licensure?
- Does a student have the skills and knowledge to succeed in college?
- Are 5th grade students across the United States meeting expectations?
- What trends in educational achievement do we see over time?

Benefits of Tests

- Tests are a tool to gather information
  - This information can be used to:
    - make decisions
    - demonstrate competence, mastery, or excellence
    - identify changes over time
    - evaluate the effectiveness of instruction or intervention
    - encourage and motivate

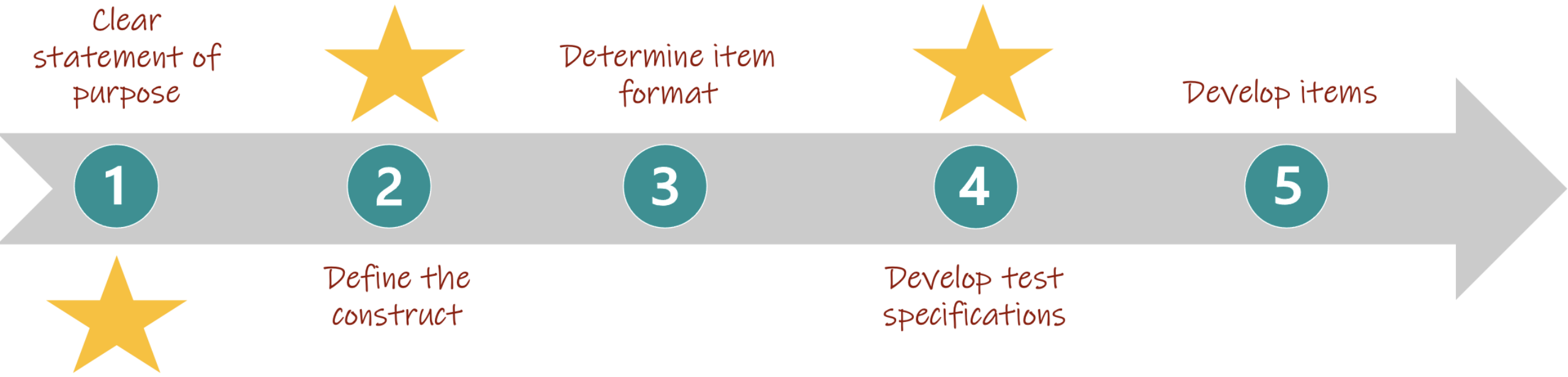
Criticisms of Tests

- Test results lead to bad decisions
- Test results discourage students and/or teachers
- Testing takes away from instruction time
- Tests provide misleading information
- Tests are unfair and biased

Maximizing the Benefits of Tests

Quality test development can maximize the benefits and minimize the weaknesses of educational tests

Steps for developing educational tests of high quality



There are more steps in the test development process, but these five are important for our goals today, and we will focus on steps 1, 2, and 4.

Step 1: Define a clear test purpose

A statement of purpose should include answers to the following question:

- Why do we need the test?
- What do the test developers say about the test?
- What information can we get from the test scores?
- How are the test scores intended to be used (i.e., make decisions)?
- How are the test scores NOT supposed to be used?

Example 1

Massachusetts Adult Proficiency Test

Purpose of the MAPT

"The purposes of the MAPT are to measure adult education learners' knowledge and skills in mathematics and reading so that their progress in meeting educational goals can be evaluated. "

Primary Purpose

To measure learners' educational gains for the purposes of state monitoring and accountability.

Secondary Purposes

Learners' MAPT scores and score gains can be aggregated to provide meaningful summative measures of program effectiveness.

Although the MAPT is not designed for diagnostic purposes, learners' performance on the MAPT can be used, in conjunction with other measures of their knowledge and skills, to better understand their strengths and weaknesses.

(Zenisky et al., 2018, p. 10)

Questions the Intended Purpose Should Answer

Q: Why do we need the test?

A: measure adult learners' knowledge and skills in mathematics and reading so their progress in meeting educational goals can be evaluated, and to fulfill accountability requirements

Q: What do the test developers say about the test?

A: measure learners' educational gains for the purposes of state monitoring and accountability

Q: What information can we get from the test scores?

A: scores and score gains can be aggregated to provide meaningful summative measures of program effectiveness

Q: How are the test scores intended to be used (i.e., make decisions)?

A: learners' performance on the MAPT can be used, in conjunction with other measures of their knowledge and skills, to better understand their strengths and weaknesses

(Zenisky et al., 2018, p. 10)

Example 2

Adding Fractions Test

Purpose of the Adding Fractions Test

Primary Purpose

The purpose of the test is to evaluate a student's numeracy skills as they are related to adding fractions.

- The test is designed to measure a student's ability to appropriately identify fractions, accurately find the sum of two fractions, and write the steps that they followed to find the sum.
- This assessment will draw upon all numeracy skills the student has had the opportunity to learn.
- The results of the test can be used to group students for instruction and identify student instructional needs.

Questions the Intended Purpose Should Answer

Q: Why do we need the test?

A: To assess a student's numeracy skills as they are applied to adding fractions.

Q: What do the test developers say about the test?

A: The test measures students' ability to appropriately identify fractions, accurately find the sum of two fractions, and write the steps they followed to find the sum. The test is designed to draw upon all numeracy skills students have had the opportunity to learn.

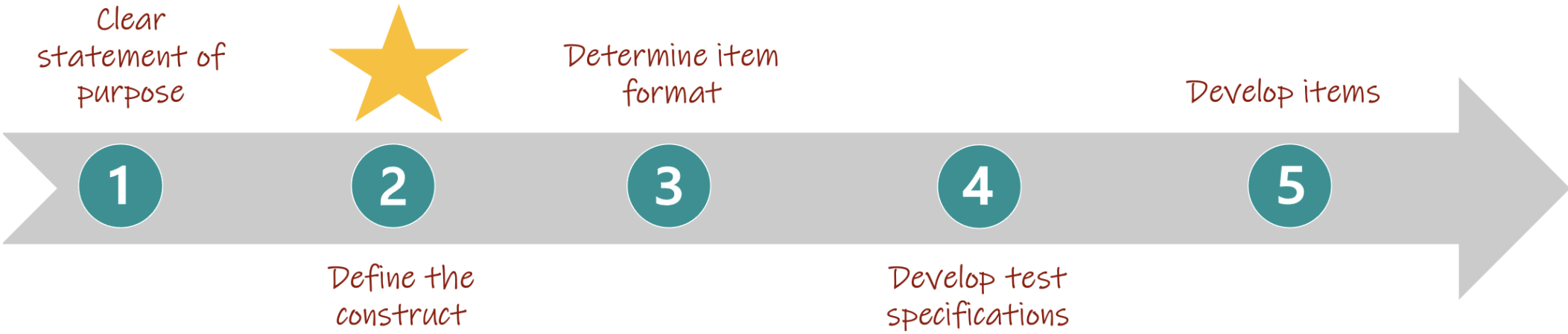
Q: What information can we get from the test scores?

A: How well students can appropriately identify fractions, accurately find the sum of two fractions, and write the steps that they followed to find the sum.

Q: How are the test scores intended to be used?

A: The results of the test can be used to group students for instruction, as well as to identify student instructional needs.

Step 2: Define the Construct



A construct:

- Is the concept or characteristic that a test is designed to measure.
- In educational testing, is often thought of as a domain of knowledge and skill (e.g., mathematics proficiency)
- can be made up of more than one attribute or **dimension**.
  - Dimensions are naturally occurring (e.g., curriculum frameworks, major content areas).
  - Common dimensions are:
    - Content Areas being measured
    - Cognitive Levels (process areas) being measured

Construct Definition: MAPT for Reading

Group	Topics	Text Type	
		Literary	Informational
Key Ideas and Details (KID)	Identifying words		
	Using general academic vocabulary		
	Locating explicit information in text		
	Determining central ideas/themes		
	Summarizing key supporting details and ideas		
	Identifying and analyzing connections in text		
	Key Ideas and Details Total:		
Craft and Structure (C&S)	Understanding figurative language, word relationships, and nuances in word meanings		
	Understanding author's purpose and organization		
	Identifying and analyzing literary structures, techniques, and styles		
	Craft and Structure Total:		
Integration of Knowledge and Ideas (IKI)	Using information & ideas from diverse media and formats		
	Evaluating content and claims		
	Combining and comparing/contrasting themes, ideas, points of view, claims		
	Integration of Knowledge and Ideas Total:		
TOTAL			

(Zenisky et al., 2018, p. 38)

Communicating the Construct Definition

Group	Topics	Text Type	
		Literary	Informational
Key Ideas and Details (KID)	Identifying words		
	Using general academic vocabulary		
	Locating explicit information in text		
	Determining central ideas/themes		
	Summarizing key supporting details and ideas		
	Identifying and analyzing connections in text		
	Key Ideas and Details Total:		
Craft and Structure (C&S)	Understanding figurative language, word relationships, and nuances in word meanings		
	Understanding author's purpose and organization		
	Identifying and analyzing literary structures, techniques, and styles		
	Craft and Structure Total:		
Integration of Knowledge and Ideas (IKI)	Using information & ideas from diverse media and formats		
	Evaluating content and claims		
	Combining and comparing/contrasting themes, ideas, points of view, claims		
	Integration of Knowledge and Ideas Total:		
TOTAL			

(Zenisky et al., 2018, p. 38)

Communicating the Construct Definition

Topic Subset	Text Type	
	Literary Text	Informational Text
Identifying words	Identify words from literary text	Identify words from informational text
Using general academic vocabulary	Use general academic vocabulary based on literary text	Use general academic vocabulary based on informational text
Locating explicit information in text	Locate explicit information in literary text	Locate explicit information in informational text

Construct Definition: Adding Fractions Test

CONTENT AREAS

- parts of a fraction
- equivalent fractions
- reducing fractions
- adding fractions with like denominators
- adding fractions with unlike denominators

LEVELS OF COMPREHENSION

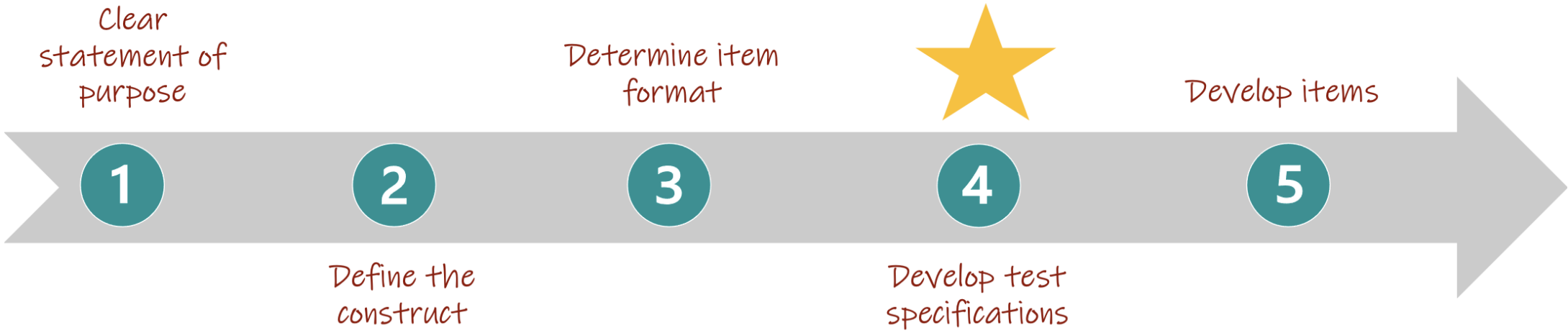
- recall of math facts
- application of arithmetic
- explanatory skill

Communicating the Construct Definition

The “dimensions” of a construct are often represented by cross-tabulating them.
This cross-tabulation illustrates the specific skills at the intersection of the dimensions.

Content Area	Levels of Comprehension		
	Recall of Math Facts	Application of Arithmetic	Explanation of Steps
Parts of a Fraction	Identify the numerator and denominator of a fraction.	n/a	Explain that the numerator is the number of parts that you have and the denominator is the number of parts in the whole.
Equivalent Fractions	Identify equivalent fractions.	Calculate equivalent fractions.	Explain how someone would calculate four equivalent fractions of ½.
Reducing Fractions	Identify fractions that need to be reduced.	Reduce fractions (when necessary) to their simplest form.	Explain the steps to reduce a fraction.
Adding Fractions with Like Denominators	Identify fractions with like denominators.	Add two fractions that have the same denominators (i.e., 1/3 + 1/3 = 2/3).	Explain the steps to add two fractions that have the same denominator.
Adding Fractions with Unlike Denominators	Identify fractions with unlike denominators.	Add two fractions that have different denominators (i.e., 1/6 + 1/3 = 1/6 + 2/6 = 3/6 = 1/2	Explain the steps to add two fractions that have unlike denominators.

Step 4: Develop Test Specifications



- Include each of the specific dimensions and the levels of each dimension
- Often displayed as a table to identify the proportion of the test that is made up by each dimension
- ”Evidence-centered” or “principled assessment” designs are also used.

Adding Fractions Test: Test Specifications

Content Area	Levels of Comprehension			
	Recall of Math Facts	Application of Arithmetic	Explanation of Steps	**Total for Content Dimensions**
Parts of a Fraction	10%	n/a	10%	20%
Equivalent Fractions	5%	5%	10%	20%
Reducing Fractions	5%	5%	10%	20%
Adding Fractions with Like Denominators	5%	5%	10%	20%
Adding Fractions with Unlike Denominators	5%	5%	10%	20%
Total for Levels of Comprehension	30%	20%	50%	100%

Example: MAPT for Reading

Test Specifications for MAPT for Reading Level 2

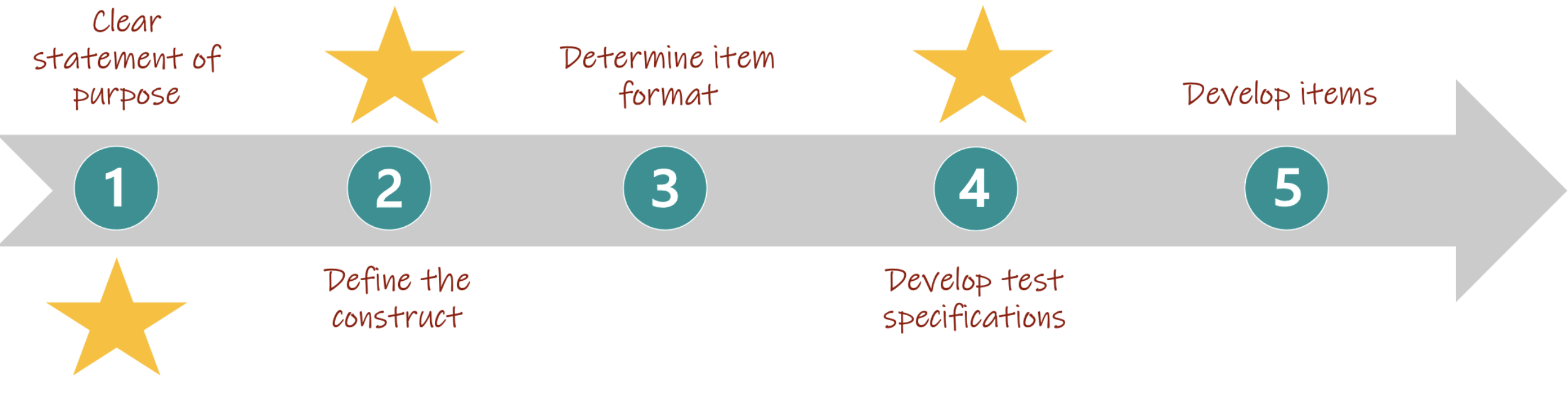
(Educational Functioning Level Beginning Basic (GLE 2-3))

Group	Topics	Text Type	
		Literary	Informational
Key Ideas and Details (KID)	Locating explicit information in text	27.5%	17.5%
	Determining central ideas/themes		
	Summarizing key supporting details and ideas		
	Identifying and analyzing connections in text		
	Key Ideas and Details Total: 45%		
Craft and Structure (C&S)	Identifying words and using general academic vocabulary	22.5%	12.5%
	Understanding figurative language, word relationships, and nuances in word meanings		
	Understanding author's purpose and organization		
	Identifying and analyzing literary structures, techniques, and styles		
	Craft and Structure Total: 35%		
Integration of Knowledge and Ideas (IKI)	Using information & ideas from diverse media and formats	0%	20%
	Evaluating content and claims		
	Combining and comparing/contrasting themes, ideas, points of view, claims		
	Integration of Knowledge and Ideas Total: 20%		
TOTAL		**50%**	**50%**

(Zenisky et al., 2018, p. 41)

Section 1 Summary

- Test Purpose
- Benefits and Criticisms of Tests
- Steps in Test Development

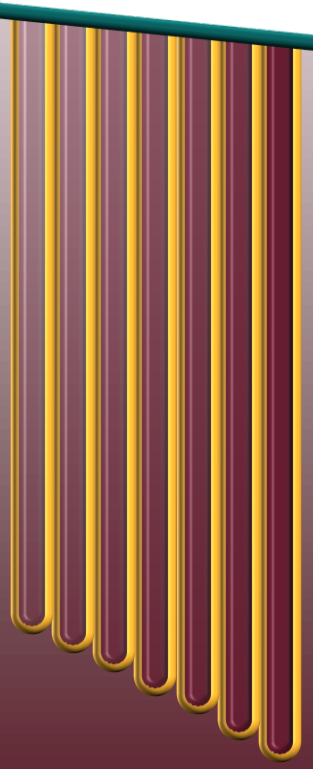


Thank you

You have reached the end of the first section of the **Validity and Educational Testing ITEMS Module, Purposes and Uses of Educational Tests**.

Please join me for the remaining sections of this module!

Understanding Validity



2

Section Learning Objectives

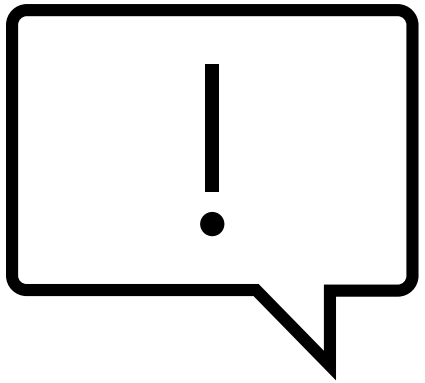
2 Understanding Validity

Define validity

Identify appropriate and inappropriate uses of tests

List the sources of validity evidence

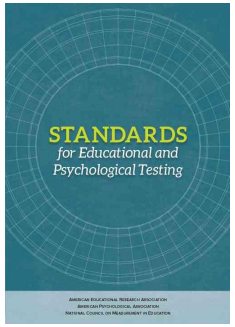
Things to remember about test development



1. The importance of the statement of purpose

2. The explicit definition of a construct

What is validity?



The Standards for Educational & Psychological Testing (2014) define validity as:

"the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."

(AERA, NCME, & APA, 2014, p. 11)

Unpacking the definition of validity

"the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."

- "…proposed uses of tests."
- Remember, why we use tests?
  - Stating the purpose of the test is the first step in test development
  - Purpose statement answers five questions

Questions to Consider Given a Test Purpose

"the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."

1. Why we need it?
2. What test developers say about it?
3. What information do we get from scores?
4. How to use the scores?
5. How NOT to use the scores?

Unpacking the definition of validity

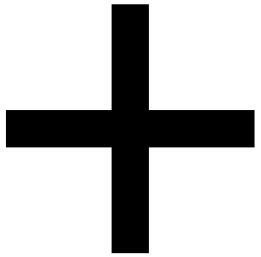
"the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."

- These interpretations should come directly from the purpose of the test.
- The purpose should state how test scores are *intended* to be used.
- The purpose statement should also guard against any potential *misuses* (i.e., how not to use the test scores).

Appropriate and Inappropriate Test Use

Adding Fractions Test

- The intended use is described as “the results of the test can be used to group students for instruction and identify student instructional needs.”
  - Uses that align with this description would be appropriate for this test.
  - This test should not be used to place students in a math intervention, to evaluate overall math ability, or determine graduation/classification of any kind.



Appropriate and Inappropriate Test Use

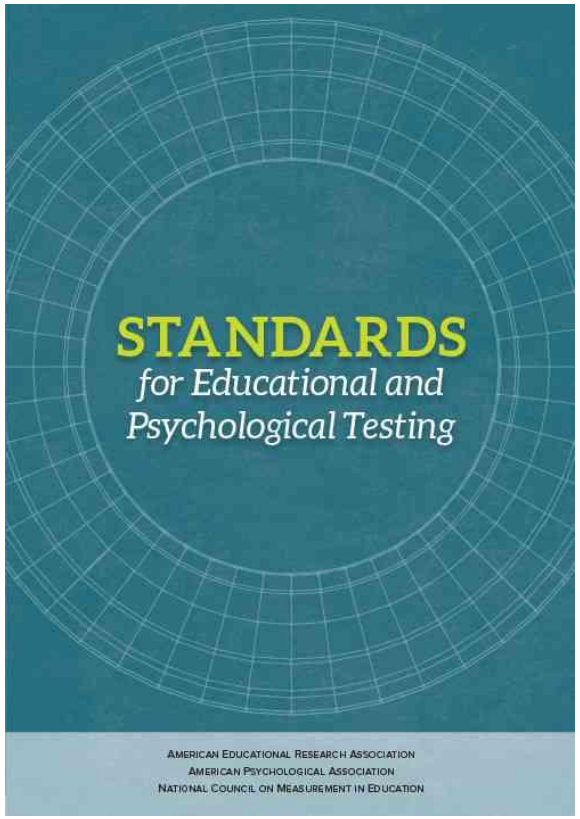
MAPT for Reading

- The intended use is described as "scores and score gains can be aggregated to provide meaningful summative measures of program effectiveness."
  - Uses that align with this description would be appropriate for this test.
  - This test should **not** be used as a measure of teacher effectiveness.
- In addition, "learners' performance on the MAPT can be used, in conjunction with other measures of their knowledge and skills, to better understand their strengths and weaknesses."
  - Teachers can use MAPT scores along with other indicators of knowledge and skills to determine learner strengths and weaknesses.
  - Teachers should not use the MAPT scores as stand-alone diagnostics of student strengths and weaknesses.

Keep unpacking…

"the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."

Five Sources of Validity Evidence



1. Test content
2. Response processes
3. Internal structure
4. Relations to other variables
5. Consequences of testing

Thank you

You have reached the end of the second section of the **Validity and Educational Testing ITEMS Module, Understanding Validity**.

Please join me for the next section of this module!

The Five Sources of Validity Evidence

3

3 The Five Sources of Validity Evidence

Section Learning Objectives

Define the five sources of validity evidence as described in the Standards

Use examples to illustrate the five sources of validity evidence.

Understand how to create a validity argument

Review

Validity is defined in the Standards (2014) as **"the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."**

Five Sources of Validity Evidence

1. Test content
2. Response processes
3. Internal structure
4. Relations to other variables
5. Consequences of testing

Validity Evidence based on Test Content

1. Test content
2. Response processes
3. Internal structure
4. Relations to other variables
5. Consequences of testing

"the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."

Five Sources of Validity Evidence

1 Test content

- The content of a test reflects the construct, or the characteristic, the test is designed to measure
- All dimensions of the construct (behaviors, knowledge, skills, abilities, etc.) described in the construct definition should be included in the test.
  - That is, the test needs to *represent* the construct
- For educational tests, the test questions should ALIGN with the curriculum or instructional framework established by the teacher or state.

(Sireci & Faulkner-Bond, 2014)

Five Sources of Validity Evidence

1 Test content

- Test developers gather evidence based on test content by conducting content validity, or "alignment" studies.
  - Digital ITEMS module #26: Content Alignment in Standards-based Educational Assessment
- **Subject matter experts (SMEs)**, such as teachers or other practitioners, are recruited and trained to review and rate test items with respect to their relevance and representativeness of the intended construct and test specifications.
  - SMEs evaluate whether test questions measure what they are supposed to measure at the appropriate level.

Five Sources of Validity Evidence

1 Test content

Item Number	Standard	How well does the item measure its standard? (1=Not at all; 6 = Verry well)						Comments
Item_1	Determine the meaning of words and phrases in a text as it relates to a topic or subject area.	1	2	3	4	5	6	
Item_2	Use text features and search tools to locate information as it relates to a topic.	1	2	3	4	5	6	

(adapted from the content validity item rating sheet in Zenisky et al., 2018, p. 148)

Five Sources of Validity Evidence

1 Test content

- Key questions to ask when gathering validity evidence based on test content:
  - Does the test fully represent all construct components?
  - Are there any components present that should ***not*** be on the test?
  - Do the proportions of items meet the proportions outlined in the test specifications?
  - How well do the test items measure the intended content standards (or knowledge and skill domain in credentialing exams)?
  - Do the test specifications align with the curriculum?

Are all construct components and levels of comprehension represented in the test?

Content Area	Levels of Comprehension		
	Recall of Math Facts	Application of Arithmetic	Explanation of Steps
Parts of a Fraction	☑	NA	?
Equivalent Fractions	☑	☑	☑
Reducing Fractions	☑	☑	☑
Adding Fractions with Like Denominators	☑	☑	☑
Adding Fractions with Unlike Denominators	☑	☑	☑

Are there any components that should not be on the test?

The following item does not align with any of the content components identified in the test specifications.

$$\frac{8}{7} + \frac{1}{7} =$$

Do the proportions of items meet those outlined in the test specifications?

Content Area	Levels of Comprehension			
	Recall of Math Facts	Application of Arithmetic	Explanation of Steps	**Total for Content Dimensions**
Parts of a Fraction	10%	n/a	10%	20%
Equivalent Fractions	5%	5%	10%	20%
Reducing Fractions	5%	5%	10%	20%
Adding Fractions with Like Denominators	5%	5%	10%	20%
Adding Fractions with Unlike Denominators	5%	5%	10%	20%
Total for Levels of Comprehension	30%	20%	50%	100%

Do the test specifications align with the curriculum?

Content Area	Levels of Comprehension			
	Recall of Math Facts	Application of Arithmetic	Explanation of Steps	**Total for Content Dimensions**
Parts of a Fraction	10%	n/a	10%	20%
Equivalent Fractions	5%	5%	10%	20%
Reducing Fractions	5%	5%	10%	20%
Adding Fractions with Like Denominators	5%	5%	10%	20%
Adding Fractions with Unlike Denominators	5%	5%	10%	20%
Total for Levels of Comprehension	30%	20%	50%	100%

Validity Evidence based on Response Processes

1. Test content
2. Response processes 
3. Internal structure
4. Relations to other variables
5. Consequences of testing

"the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."

Five Sources of Validity Evidence

2 Response processes

The Standards for Educational and Psychological Testing (2014) describe validity evidence based on response processes as

the alignment between the construct being measured and the actual response of the test taker.

The expectation is that test takers are using problem solving, critical thinking, and reasoning skills to solve each problem.

Five Sources of Validity Evidence

2 Response processes

1. What processes are students using to answer the questions?
   - *Are these process the ones intended to be measured?*
2. Do differences in how a student answers a question help us understand the scores?
3. Are there differences in how students answer questions based on an unrelated trait (i.e., test taking strategies)?

Five Sources of Validity Evidence

2 Response processes

- Evidence based on response processes is gathered by
  - evaluating individual item responses
  - conducting interviews with students about how they answered questions after taking a test or while they are answering questions (think-aloud protocol)
    - For more information regarding think-aloud interviews and cognitive labs, check out Digital Module 12: Think-aloud Interviews and Cognitive Labs
  - analyzing additional data extracted from the testing occasion (i.e., item response times, use of calculator or other digital tools, etc.)

Five Sources of Validity Evidence

2 Response processes

- One main use for information about response processes is to identify any differences in response processes for specific groups of students
    - Is there a feature of the item that is creating this difference?
        - If yes, then revising the item or test to promote similarity across groups.
    - Did one group have more opportunities to learn? Different instruction received?

Example

Sample Responses to Item A and Item B

Item A:
What is the product of 12 and 10?

a) 12 b) 22 c) 120 d) 1200

Response A: c

Item B:
What is the product of 5 and 12?

a) 17 b) 55 c) 60 d) 70

Response B: d

Evaluating individual item responses

Interpretation
One of the following could be happening for this student: 1. They memorized their multiplication tables and are using pure recall and recalled the product of 5 and 12 incorrectly. 2. They added a zero to 12 because when you multiply any number times 10 you can just add a zero and they don't know the product of 5 and 12. 3. They found the sum of (2 x 10) and (10 x 10) and used the same process on the Item B and made an error when adding the two products.

Example

Sample Responses to Item A and Item B

Item A:
What is the product of 12 and 10?

a) 12 b) 22 c) 120 d) 1200

Response A: c

Item B:
What is the product of 5 and 12?

a) 17 b) 55 c) 60 d) 70

Response B: d

Conducting interviews with students about how they answered questions after taking a test

Interpretation
Asking this test taker questions about their responses to Item A and B could help determine which of the previous interpretations are more representative of this test taker.

Example

Sample Responses to Item A and Item B

Item A:
What is the product of 12 and 10?

a) 12 b) 22 c) 120 d) 1200

Response A: c

Item B:
What is the product of 5 and 12?

a) 17 b) 55 c) 60 d) 70

Response B: d

Conducting interviews with students while they are answering questions using a think-aloud protocol

Interpretation
During an interview, the test taker would have the chance to talk through each of the above responses. This information would provide insight in real time as the student is working through each item. In addition, the data gathered using a think-aloud protocol could highlight areas where the item could be revised to elicit the intended response process or removed.

Example

Sample Responses to Item A and Item B

Item A:
What is the product of 12 and 10?

a) 12 b) 22 c) 120 d) 1200

Response A: c

Item B:
What is the product of 5 and 12?

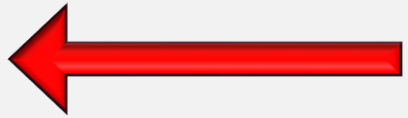
a) 17 b) 55 c) 60 d) 70

Response B: d

Analyzing additional data extracted from the testing occasion (i.e., item response times, use of calculator or other digital tools, etc.)

Interpretation
The comparison across response times for correct and incorrect responses could help identify which of these responses is more indicative of what this test taker really knows and can do. If there was a calculator option on these two items, then maybe the incorrect response was due to a mistake when using the calculator.

Validity Evidence based on Internal Structure

1. Test content
2. Response processes
3. Internal structure 
4. Relations to other variables
5. Consequences of testing

"the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."

Five Sources of Validity Evidence

3 Internal structure

Validity evidence based on internal structure should:

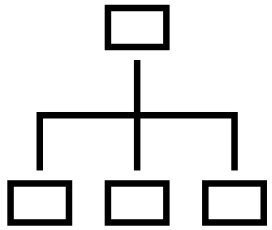
- Provide support for the alignment between the dimension(s) being measured on the test and the intended dimension(s) of the test and the interpretation of test scores.
  - This is known as evaluating the **dimensionality**

- Provide support for similar functioning of all items on the test for all identifiable subgroups of test takers.
  - The degree to which items are not functioning the same for all identifiable subgroups of test takers is known as **differential item functioning (DIF)** and should be evaluated.

- Provide support for the consistency of scores when there are repeated or multiple testing occasions.
  - This is known as evaluating the reliability. You can learn more about reliability from the ITEMS Digital Module 01: Reliability in Classical Test Theory

Five Sources of Validity Evidence

3 Internal structure



Does the dimensionality of the **test** data match the intended dimensionality of the construct?

Do all test **items** function the same for all students, regardless of gender, race, age, culture, socioeconomic status, etc.,?

Five Sources of Validity Evidence

3 Internal structure → Dimensionality

- **Dimensionality** refers to the number of constructs/skills/traits a test is designed to measure.

- A test for dimensionality investigates the intended relationships between item responses and the construct(s) that the items were designed to measure.
  - There are many approaches and techniques for evaluating dimensionality and selection of the approach should be driven by the characteristics of the test data.

- **Ideally, the results of the dimensionality analysis provides evidence to support the intended dimensionality of the test.**

Five Sources of Validity Evidence

3 Internal structure → Differential Item Functioning

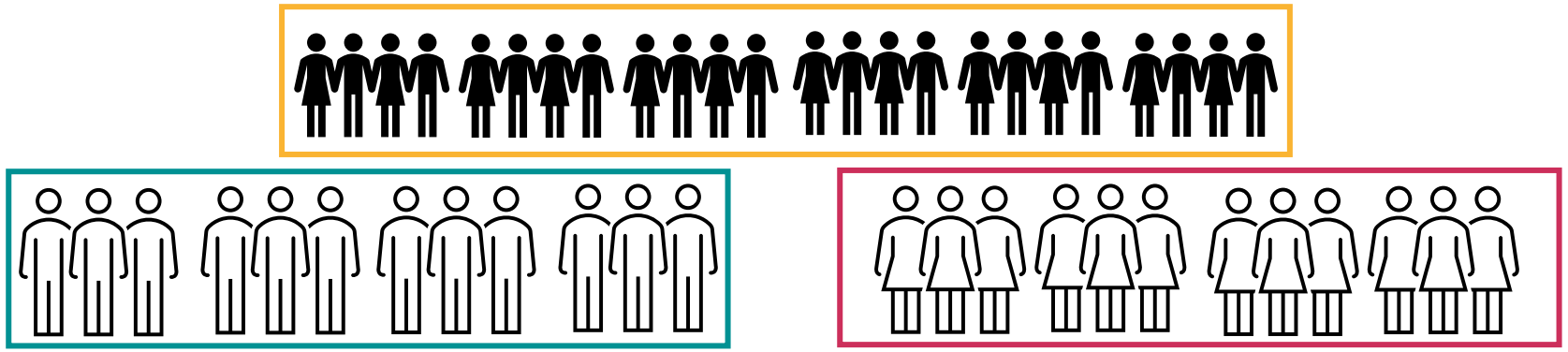
- Differential Item Functioning (DIF) is detected when an identified group is performing better than another group.
- The presence of DIF would indicate test takers of the **same ability** that belong to different groups have different expectations or probabilities of answering an item correctly.
- When DIF is present, the item is flagged for review. An item exhibiting DIF is not inherently a biased item.

Five Sources of Validity Evidence

3 Internal structure → Differential Item Functioning

Are the items functioning the same for males and females?

The adding fractions test was administered to the entire fifth grade cohort.



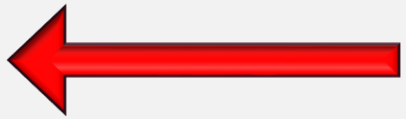
The probability of a correct response for each item is compared across students with similar ability.

The probability of answering item 5 correctly is different for students with similar ability depending on gender.

Female students have a significantly higher chance of getting this item correct.

This item should be flagged and reviewed; however, the identification of DIF does not automatically mean the item is biased.

Validity Evidence based on Relations to Other Variables

1. Test content
2. Response processes
3. Internal structure
4. Relations to other variables 
5. Consequences of testing

"the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."

Five Sources of Validity Evidence

4 Relations to other variables

- How well do test scores relate to scores on another test that is measuring a *similar* construct?

- How well do test scores relate to scores on another test that is measuring a *different* construct?

Five Sources of Validity Evidence

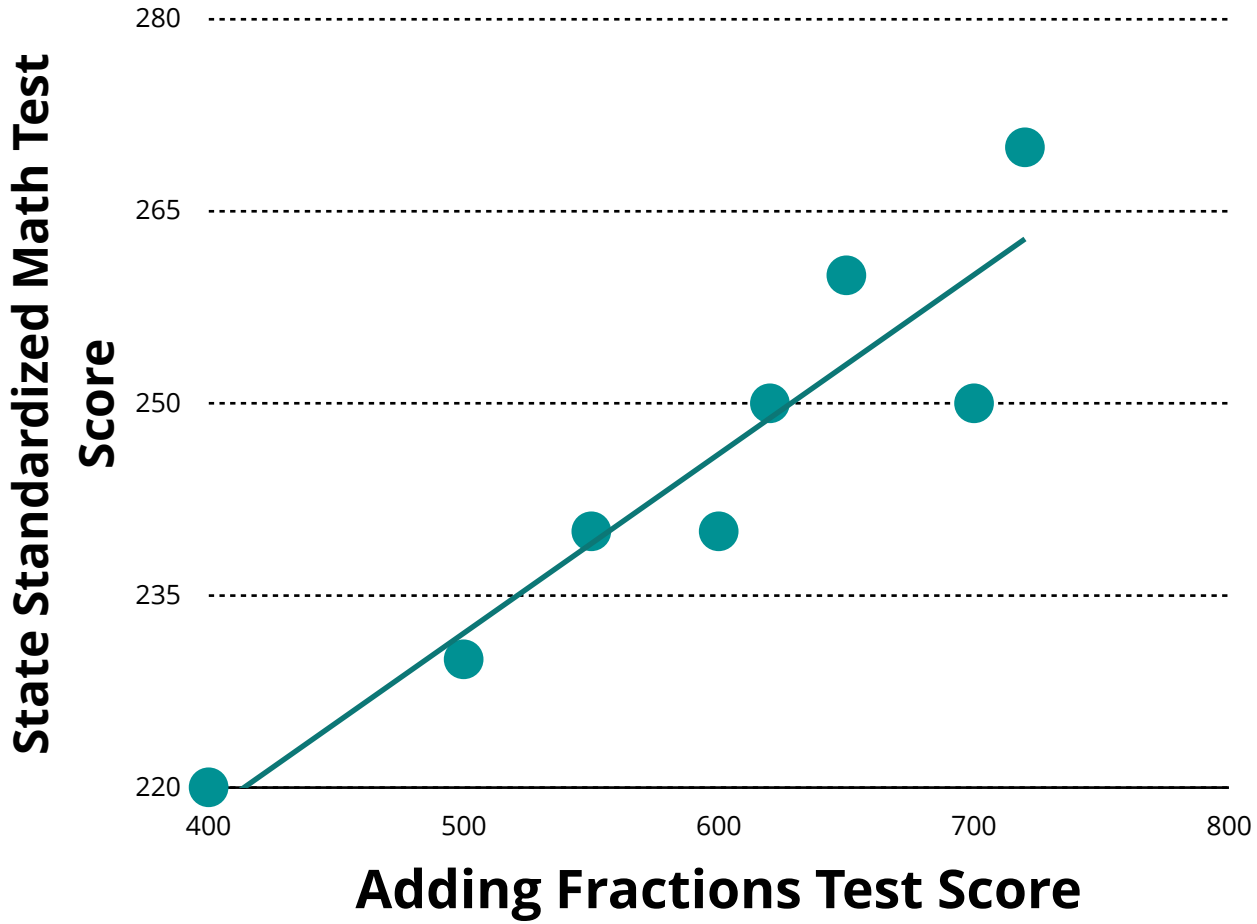
4 Relations to other variables

- Evidence based on relations to other variables is gathered **after** the test is administered and scores are generated.

- Test scores are compared to scores on other measures to see if the scores are **similar to** assessments measuring similar constructs and **dissimilar to** assessments measuring different constructs.

Five Sources of Validity Evidence

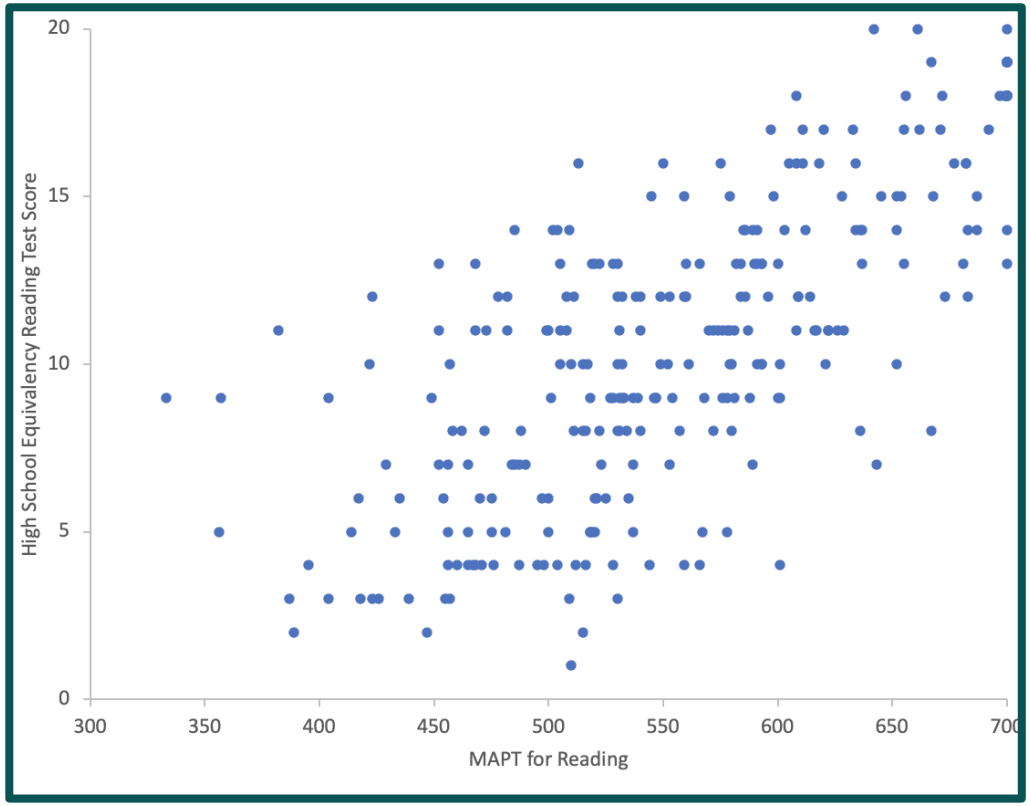
4 Relations to other variables



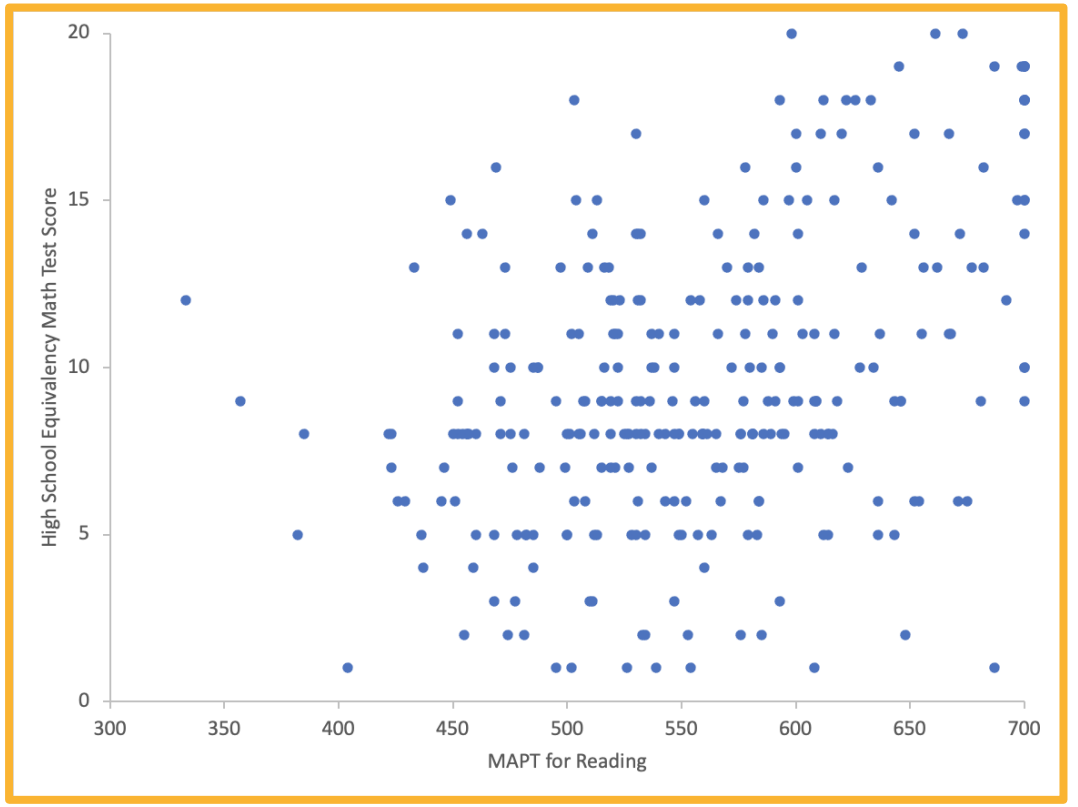
Five Sources of Validity Evidence

4 Relations to other variables

Within-subject correlation (r = .73)



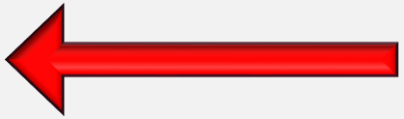
Across subject correlation (r = .47)



This pattern of correlations supports the **convergent** (relatively higher correlations across measures of similar constructs) and **discriminant** (relatively lower correlations among measures of dissimilar constructs) validity.

(Zenisky et al., 2018)

Validity Evidence based on Consequences of Testing

1. Test content
2. Response processes
3. Internal structure
4. Relations to other variables
5. Consequences of testing 

"the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests."

Five Sources of Validity Evidence

5 Consequences of testing

- **Intended consequences** are directly related to the interpretation of the test scores for specific uses identified by the test developer.

- **Unintended consequences** are those that are not expected, **good or bad**, but that occur as a result of the test.

Five Sources of Validity Evidence

5 Consequences of testing

Low Stakes Example: The Adding Fractions Test

- Intended consequences of the test are identifying student knowledge and understanding, grouping students for instruction, and identifying student instructional needs.
- Unintended consequences could be decreasing student motivation.

High Stakes Example: A High School Graduation Exam

- Intended consequences of the test are to illustrate mastery of the academic skills necessary to graduate high school.
- Unintended consequences of a high school graduation exam include the possibility of increased drop out rates for individuals with undiagnosed learning disabilities.

Five Sources of Validity Evidence

5 Consequences of testing

Analysis of Adverse Impact

- Are the passing rates similar for students with different cultural backgrounds?

Teacher Perspectives (surveys/interviews)

- Do teachers find the test results useful for instruction?
- What is the impact of the test on instruction/curriculum?

Student Perspectives (surveys/interviews)

- What is the impact of test results on student motivation/learning?
- Do students find the results helpful?

Thank you

You have reached the end of the third section of the **Validity and Educational Testing ITEMS Module, Five Sources of Validity Evidence**.

Please join me for the final section of this module!

Summarizing and Documenting the Validity Argument

4

4 Summarizing and Documenting the Validity Argument

Section Learning Objectives

- Define the “Validity Argument”
- Exemplify documenting validity evidence

Summarizing the "Validity Argument"

- As defined earlier, validation is the degree to which evidence and theory support the use of a test for a particular purpose.

- A "validity argument" is a summary of the gathered evidence and theory that supports the use of a test for the intended purpose.

Documenting Validity Evidence

1 Test content

- Does the test fully represent all construct components?
- Are there any components present that should ***not*** be on the test?
- Do the proportions of items meet the proportions outlined in the test specifications?
- How well do the test items measure the intended content standards (or knowledge and skill domain in credentialing exams)?
- Do the test specifications align with the curriculum?

Five Sources of Validity Evidence

1 Test content

Key Question to Answer	Yes/No	Description of the Evidence
Does the test fully represent all construct components?	Yes	On average the SMEs rated the alignment of the items to each of the construct components a five out of six representing "appropriate representation".
Are there any components present that should ***not*** be on the test?	Yes	One item measured a skill that was not included in the test specifications and was removed from the test.
Do the proportions of items meet the proportions outlined in the test specifications?	Yes	The evidence would include the operational proportional distribution of items across the test specifications.
Do the test items measure the intended content standards?	Yes	On average the SMEs rated the alignment of the items with the intended content standards a five out of six representing "appropriately aligned".
Do the test specifications align with the curriculum?	Yes	A survey of the curriculum was completed and compared to the test specifications.

The evidence provided in this table is fictional and was created as an example of what could be included as validity evidence based on test content.

2 Response processes

A think-aloud protocol was followed with a subset of items from the Adding Fractions Test to evaluate the extent to which the intended response processes were being applied by test takers.

- The *results* indicated the test takers did not consistently understand the word identify in items such as "Identify the numerator in the following fraction."
- *Resolution:* The items requesting test takers to *identify* were changed to "Which of the…?"

Response times were analyzed along with other test taker characteristics, specifically whether the test taker was participating in an after-school reading comprehension intervention.

- The *results* revealed that the test takers in the after-school intervention were responding to items more quickly and accurately than those not in the after-school intervention.
- *Resolution*: The items were reviewed and revised to ensure the reading level was appropriate for all test takers.

Five Sources of Validity Evidence

3 Internal structure

Differential Item Functioning

The documentation for differential item functioning should include the items flagged, and the actions taken to resolve or an explanation for any flags that weren't resolved.

Dimensionality

The results of the evaluation of dimensionality should include an appropriate level of detail to support the reporting of scores on a unidimensional scale or subscores based on a multidimensional construct.

A complete explanation of the hypothesized dimensionality, an explanation of the model selected to determine the adequacy of the expected dimensionality, and the results of the fit analysis (i.e., parameter estimates and model fit indices).

Five Sources of Validity Evidence

4 Relations to other variables

Appropriate documentation of relations to other variables could include:

- Reporting the results of the relationship between the test score of interest and a test measuring a similar construct.
  - You expect *higher* correlations across measures of similar constructs as evidence of **convergent validity**

- Reporting the results of the relationship between the test score of interest and a test measuring a different construct.
  - You expect *lower* correlations among measures of dissimilar constructs as evidence of **discriminant validity**

For example, the relationship between the scores from two math tests is stronger (r = .73) when compared to the relationship between the scores from a math test and reading test (r = .47).

Five Sources of Validity Evidence

5 Consequences of testing

Consequences of testing aren't always easy to evaluate. Here are some examples of approaches to gathering data with questions that can be answered based on those data.

Analysis of Adverse Impact

Teacher Perspectives (surveys/interviews)

Student Perspectives (surveys/interviews)

Conclusion

Learning objectives:

- Gain a deeper understanding of the purposes of tests
- Examine the benefits and criticisms of tests
- Understand the steps of developing educational tests of high quality
- Define validity
- List the sources of validity evidence
- Identify appropriate and inappropriate test use
- Define and illustrate the five sources of validity evidence
- Understand how to create a validity argument
- Define “validity argument”
- Provide examples of documenting validity evidence