

**Assessment Power
Session:**

**Unveiling the
Mysteries of
Validity and
Reliability**

**SACSCOC Annual
Conference**

December 8, 2019

Houston, Texas

Timothy S. Brophy

**Professor and Director,
Institutional Assessment**

Office of the Provost

University of Florida

**Gainesville, Florida,
USA**

Today's Goals

To introduce and reinforce the concept of validity as it applies to assessment in higher education

To introduce and reinforce the concept of reliability as it applies to assessment in higher education

Validity

What it is, and how we examine it

Validity defined

- Validity refers to the *degree to which evidence and theory support the interpretations of the test scores for proposed uses of tests.* (p. 11)

Source:

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

The Importance of Validity

Validity is, therefore, *the most fundamental consideration in developing tests and assessments.*

The process of validation involves accumulating relevant evidence to provide a sound scientific basis for the proposed score and assessment results interpretations. (p. 11)

Source:

AERA, APA, & NCME. (2014). Standards for educational and psychological testing. Washington, DC: AERA

How Do We Examine Validity in Higher Education?

- Most often this is qualitative; colleagues are a good resource
- Review the evidence
- Common sources of validity evidence
 - Test Content
 - Construct (the idea or theory that supports the assessment)
 - The validity coefficient

Important distinction

It is not the test or assessment itself that is validated, but the *inferences one makes from the measure based on the context of its use.*



Therefore it is not appropriate to refer to the ‘validity of the test’; instead, we refer to the validity of the interpretation of the results for the test’s intended purpose.

Validation Process

How do we establish that the interpretation and use of results of our assessments are valid for their intended purpose?

Validity
Rationale:
Building an
*Interpretation
and Use
Argument*

When developing a validity rationale, it's important to state that the assessment covers relevant knowledge and skills (content validity).

However, validity rationales should also include support for the degree to which *inferences* you make from your results *match your assessment's intended purpose*

Building an Interpretation and Use Argument: **Purpose**

Describe the purpose of your assessment



What is its relationship to your curriculum?



Why this assessment is important to your program?

Building an Interpretation and Use Argument: **Content**


Does the measure adequately cover the content and skills students are expected to know and demonstrate?



Does the assessment measure what you've covered?

Building an Interpretation and Use Argument: **Context**


Is the assessment implementation strategy consistent with the purpose of the assessment?



Is the way the assessment is implemented going to allow students to yield their best performance?

Building an Interpretation and Use Argument:
Rubric Achievement Levels
(when appropriate)

How do you know that the levels of achievement you developed for the rubric are appropriate for this assessment?




Do they adequately cover the potential responses you might hear?




Could any performance fall outside or in between your rubric levels?

Building an Interpretation and Use Argument: Inferences

How do you know that a rubric level matches your expectations for that level of performance?



How do you know that your use of test scores appropriately reflects the test's intended purpose?



Does your assignment of an achievement level or score *infer* that student has met the criterion at that level?

Review of the
Validation
Process – *The
Interpretation
and Use
Argument*

- Establish the purpose of your test/assessment
- Develop content evidence
- Align the test/assessment context with its purpose
- Align the achievement levels and/or scores with their intended uses
- Establish the degree to which your assessment results infer the student's attainment of your established levels of achievement

Small Group Discussions

How would you establish validity in this situation?

Validity – Interpretation and Use Argument 1

- Your Biology faculty have established student learning outcomes for their program. Their content outcome is that students will:

Identify, describe and explain the basic terminology, concepts, methodologies and theories used within the biological sciences.

- They have selected the *Educational Testing Service (ETS) Biology Major Field Test* to measure this content outcome.
- This is a 3rd party exam used as a measure of a programmatic student learning outcome.
- What would be your interpretation and use argument for using this Field Test instead of a faculty-developed exam?

Validity – Interpretation and Use Argument 2

- Your Animal Sciences faculty have established student learning outcomes for their program. Their communication outcome is that students will:

Effectively communicate in written form in a manner appropriate for the animal sciences.

- They have decided to use the course grade in AEC3033, Research and Business Writing in Agricultural and Life Sciences
- This is a course grade, not an assignment grade.
- What would be your interpretation and use argument for using this course grade as a measure of this SLO?

Reliability

What it is, and how we examine it

Reliability defined

- The general notion of reliability/precision is defined in terms of *consistency over replications* of the testing procedure. Reliability/precision is high if the scores/results for each person are consistent over replications of the testing procedure and is low if the scores are not consistent over replications. (p. 35, emphasis added)

Source:

- *AERA, APA, & NCME. (2014). Standards for educational and psychological testing. Washington, DC: AERA*

Important: The Relationship of Validity and Reliability

Reliability: *Replications* are defined to reflect a particular interpretation and use of test scores/assessment results

Validity: A test or assessment with high reliability for a particular interpretation and use of its results may not be interpreted similarly for a different purpose:

Validity must be established for each intended interpretation and use of the test or assessment results.

Basic Concepts
Central to
Establishing
Reliability:
Correlation

- **Correlation coefficient** – a statistical estimate of the strength and direction of the relationship between two continuous variables
- For every observed change in one variable, there is a related observed change in the other – they vary together
- These can be positive or negative, and the values range from -1 to +1
- We consider a test to be reliable if the coefficient is .70 or higher

Basic Concepts Central to Establishing Reliability: True Score Theory

- **True Score theory** – every person's **observed score** is a combination of the person's **true score** and some **measurement error**

- **True Score** formula:

Observed Score (X) = True Score (T) + Measurement Error (E)

- **True scores** are conceptualized as those obtained after *infinite* replications of the test/assessment
- We define the reliability coefficient as the **correlation** between **true scores** and **observed scores**
- Since we can never know a person's **true score**, we *approximate reliability* with a reliability coefficient – the **correlation between observed scores across repetitions**
- **Internal consistency** is typically a measure based on the *correlations* between different items on the same test

Basic Concepts
Central to
Establishing
Reliability:
Types of
Reliability
Estimates

- This number is estimated in several ways
- Parallel forms – the Pearson *product-moment correlation coefficient*
- Test-retest reliability – correlate scores on two administrations of the same test
- Split-halves reliability- Flanagan's Formula
- Kuder-Richardson Formulae 20 and 21 (KR_{20} and KR_{21}) – for tests with dichotomous items

A Common
Approach to
Estimating
Reliability in
Instructor-Made
Tests:
**Split-Half
Reliability**

- First, split the test questions evenly into two parts – part a and part b , and calculate the subscores for each student on each part. Then, calculate the variances of subscores on the two parts and then the variance of the total scores. Enter these numbers into this formula and calculate:

$$r = 2 \left(1 - \frac{S_a^2 + S_b^2}{S^2} \right)$$

- Key:
 - S_a^2 is the variance of part a of the test
 - S_b^2 is the variance of part b of the test
 - S^2 is the variance of scores of the entire test

Our example – History quiz results (max score = 10)

<i>Name</i>	<i>Quiz part a</i>	<i>Quiz part b</i>	<i>Total Score</i>
Joe	5	5	10
Mary	3	2	5
Cheryl	5	4	9
Charlene	4	4	8
Chris	2	3	5
Brian	3	4	7
LaTerrance	5	4	9
Kim	4	3	7
Robbie	3	5	8
Anwar	5	3	8

Calculating the Variances

To try this, use the [standard deviation online calculator](#), and click “sample.” Enter the scores for part *a*, part *b*, and the *total*, one at a time. You will get three variances.

Here are the results:

Part *a* = 1.21

Part *b* = .9

Total = 2.71

Next, we place these figures into the formula.

Calculating the Split-Half reliability

- Flanagan's Formula:

$$r = 2 \left(1 - \frac{S_a^2 + S_b^2}{S^2} \right)$$

- Our data:

$$r = 2 \left(1 - \frac{1.21 + .9}{2.71} \right) = 2 \left(1 - \frac{2.11}{2.71} \right)$$

$$= 2(1 - .78) = 2(.22) = \underline{\underline{.44}}$$

Interpreting the coefficient


The Split-Half reliability is .44

Reliability coefficients should be at .70 or higher for a test to be considered reliable

This quiz is not reliable. What should the instructor do?

Another Approach to Reliability: Inter-rater Agreement

For open-ended decisions (e.g., rubrics, qualifying exams), have two readers score the results, and examine the percentage of agreement



If agreement is low ($< 70\%$, or $.70$), discuss where the disagreements occur and adjust scoring



Agreement cannot be at the cost of validity or interpretability

Student	Rater 1	Rater 2	Agreement
1	6	6	1
2	7	6	0
3	4	4	1
4	3	2	0
5	4	4	1
			$3/5 =$ 60% agreement

An Example – Two raters

- The basic measure for inter-rater reliability is a **percent agreement between raters.**

Small Group Discussions

How would you interpret reliability coefficients in these situations?

Reliability –
How would you interpret these coefficients?

An instructor has given a test in a Calculus 1 class. The instructor used the split-half reliability formula to calculate a reliability coefficient of .85. What does this tell you? What would you advise this instructor?

Two raters review a set of student research papers. The percentage of agreement is 55%, or .55. What would you advise the instructor?

Summary

- Validity refers to the *degree to which evidence and theory support the interpretations of the test scores for proposed uses of tests.*
- Primary areas to consider when building an interpretation and use argument:
 - Purpose
 - Content
 - Context
 - When appropriate, rubric achievement levels
- The general notion of **reliability**/precision is defined in terms of *consistency over replications* of the testing procedure. Reliability/precision is high if the scores/results for each person are consistent over replications of the testing procedure and is low if the scores are not consistent over replications.
 - *Replications* are defined to reflect a particular interpretation and use of test scores/assessment results

THANK YOU!

Timothy S. Brophy, Ph.D.

Professor and Director, Institutional Assessment

tbrophy@aa.ufl.edu

+1-352-273-4476

Resources:

UF Institutional Assessment website – assessment.aa.ufl.edu (Faculty Resources)

Brophy, T. S. (2017). Case study: The University of Florida assessment system. In T. Cumming and M. D. Miller (Eds.), *Enhancing assessment in higher education: Putting psychometrics to work* (pp.184-202). Sterling, VA: Stylus.

