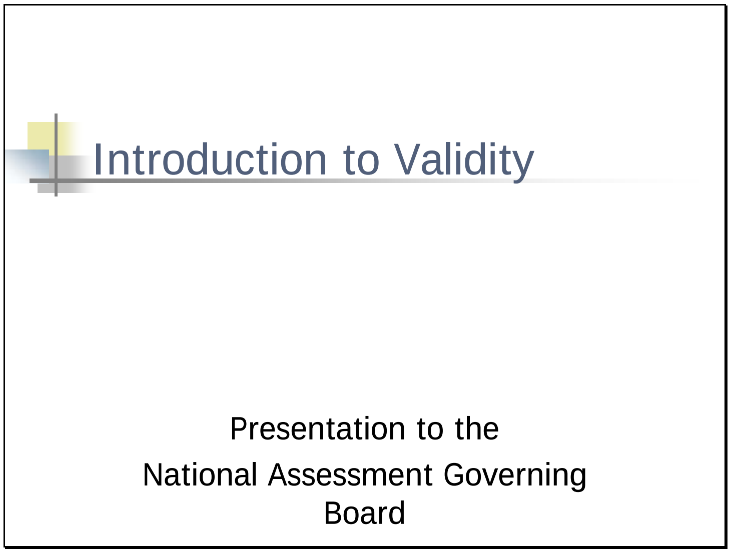
Slide 1

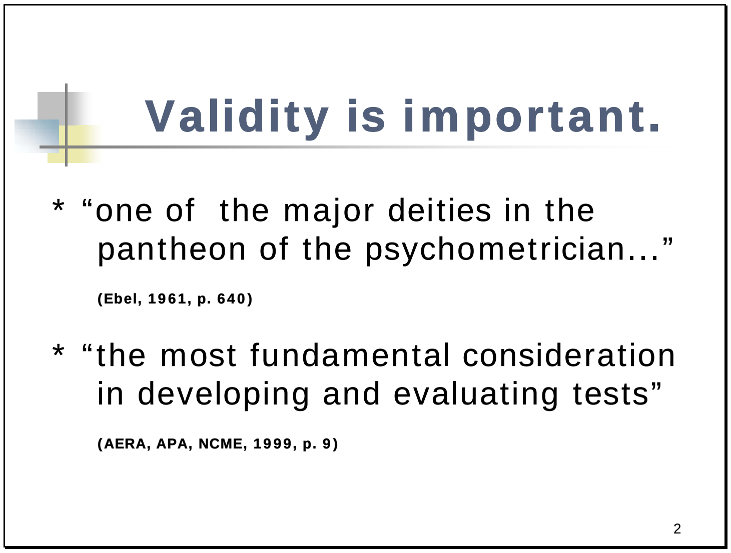


Introduction to Validity

Presentation to the
National Assessment Governing Board

This presentation addresses the topic of validity. It begins with some first principles of validity.

Slide 2



Validity is important.

* "one of the major deities in the pantheon of the psychometrician..."

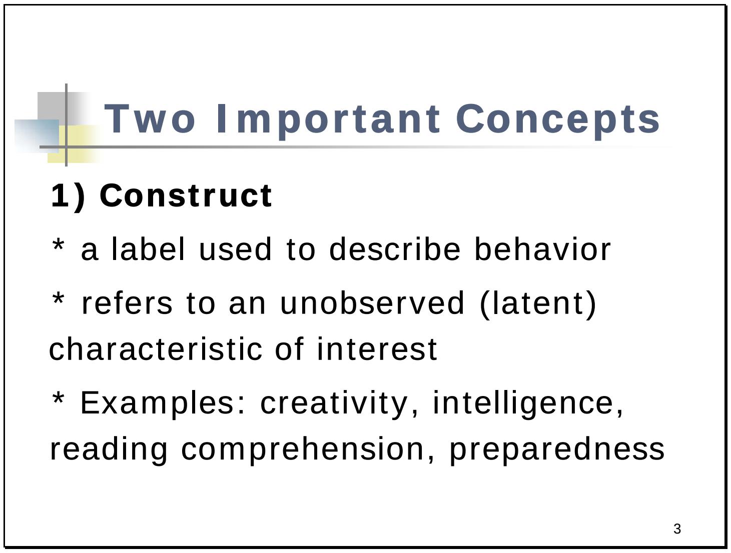
(Ebel, 1961, p. 640)

* "the most fundamental consideration in developing and evaluating tests"

(AERA, APA, NCME, 1999, p. 9)

These two quotes highlight the importance of validity in measurement and assessment. The first of these quotes is by a renowned psychometrician, Robert Ebel. The second quote is from the Standards for Educational and Psychological Testing.

Slide 3



Two Important Concepts

1) Construct

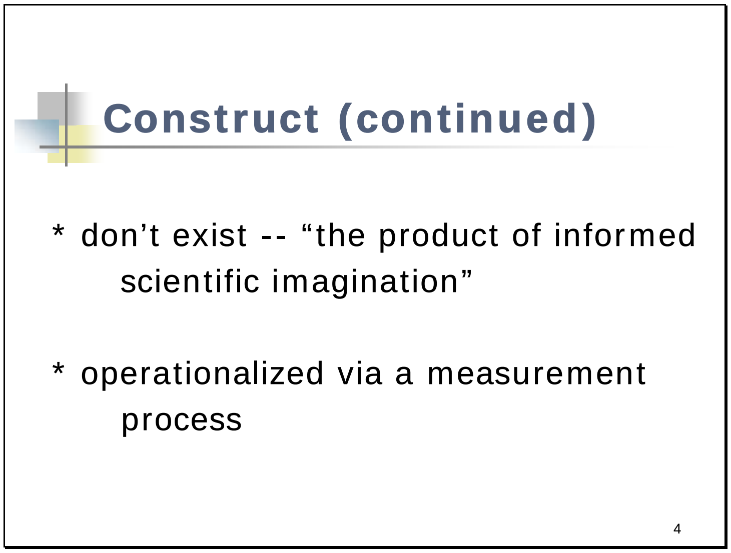
* a label used to describe behavior

* refers to an unobserved (latent) characteristic of interest

* Examples: creativity, intelligence, reading comprehension, preparedness

There are two important concepts to validity. The first one is the notion of a construct. Constructs can be defined as labels used to describe an unobserved behavior, such as honesty. Such behaviors vary across individuals; although they may leave an impression, they cannot be directly measured. In the social sciences virtually everything is a construct.

Slide 4



Construct (continued)

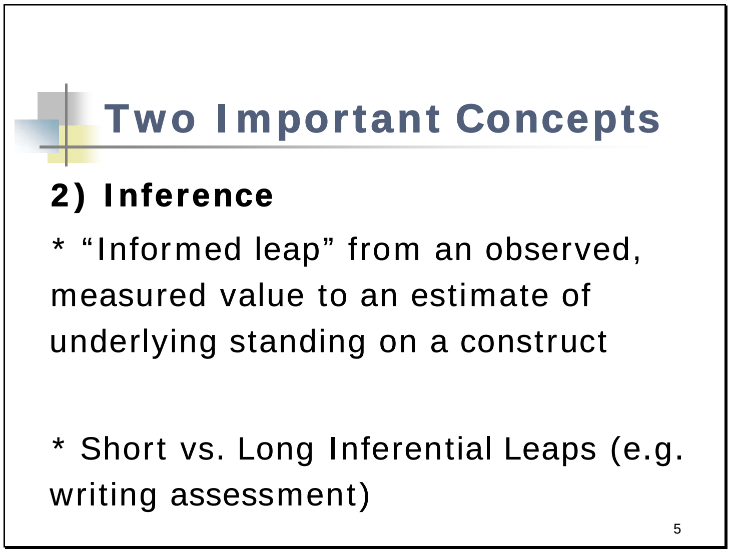
* don't exist -- "the product of informed scientific imagination"

* operationalized via a measurement process

4

Linda Crocker and James Algina described constructs as "the product of informed scientific imagination." Constructs represent something that is abstract, but can be operationalized through the measurement process. Measurement can involve a paper and pencil test, a performance, or some other activity through observation which would then produce variables that can be quantified, manipulated, and analyzed.

Slide 5



Two Important Concepts

2) Inference

* “Informed leap” from an observed, measured value to an estimate of underlying standing on a construct

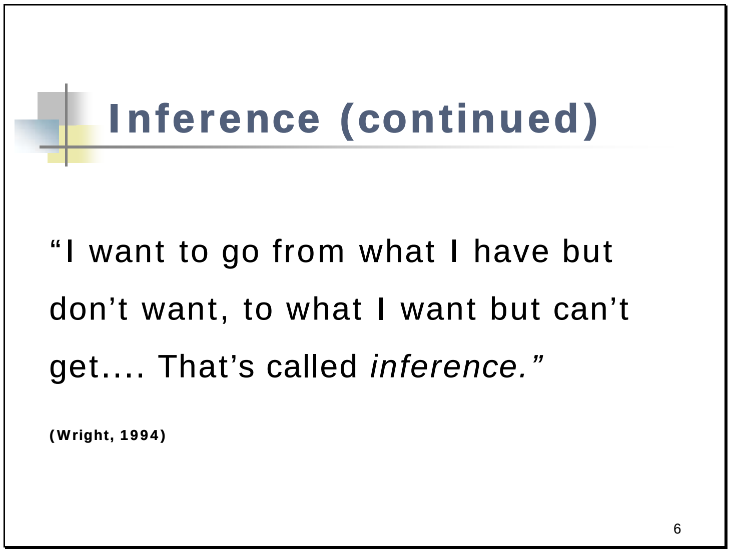
* Short vs. Long Inferential Leaps (e.g. writing assessment)

5

The second important concept is inference. An inference is an “informed leap” from the observed or measured value to an estimate of some underlying variable. This is called an inferential leap because we cannot directly observe what is being measured, but we can observe its manifestations. Through observation of the results of the measurement process, an inference can be made about the underlying characteristics and its nature or status.

Short and long inferential leaps are possible. For a long time, writing tests were multiple choice tests of grammar, usage, vocabulary, and mechanics, (long inferential leap). Writing assessments have changed to be a shorter inferential leap of student writing ability by actually asking a student to produce writing samples. Long or short, the leap still exists. The scenario of producing a writing sample is contrived–a more realistic measure of a student’s writing ability is when students write for their own purposes. We probably feel more comfortable with shorter inferential leaps. However, longer inferential leaps may be preferred by policymakers for cost reasons; it is cheaper to do the proxy measure, the more distal inference, than to do the shorter inferential leap in many cases.

Slide 6



Inference (continued)

"I want to go from what I have but don't want, to what I want but can't get.... That's called *inference."*

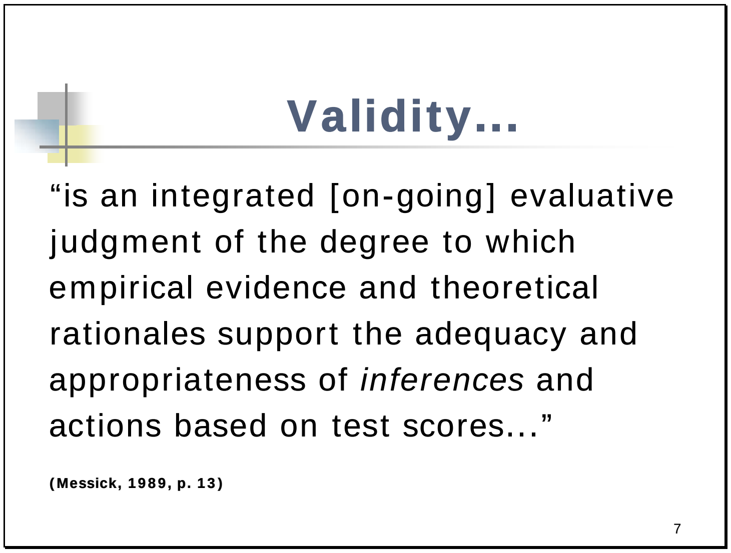
(Wright, 1994)

6

In this quote, Benjamin Wright, an item response theory psychometrician, is saying that a typical standards–referenced test, such as any of the NAEP subject tests, provides a sample of a student's behavior. The test reveals which questions the student answered right or wrong. Correct or incorrect responses to specific test items, according to Wright, are not what we really care about. What we want to do is to generalize, or make inferences, about the student's broader universe of skills or knowledge represented by these responses.

So, the problem is exactly what Ben Wright described, going from what we have but do not really want, to what we want but really cannot get easily. And that's the challenge of inference. Validation is the process that helps us make that inference stronger, to be more confident in saying, "Here is what I have, and here is what it means." That is the process of validation; validating the inferences.

Slide 7



Validity…

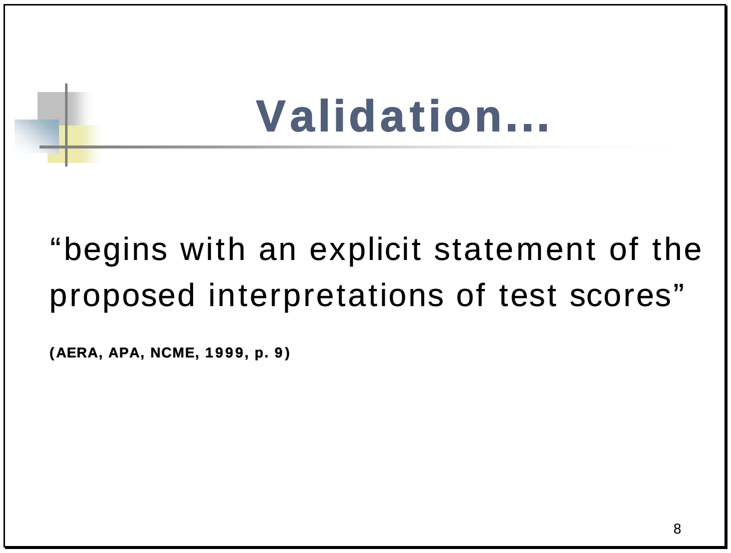
“is an integrated [on-going] evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of *inferences* and actions based on test scores…”

(Messick, 1989, p. 13)

7

In 1989, Samuel Messick wrote a chapter in a standard reference text called Educational Measurement. In that chapter, he defined validity as an “integrated, evaluative judgment of the degree to which empirical evidence and theoretical rationale support the adequacy and appropriateness of inferences and actions based on test scores.” I added the term, “on-going,” to his definition because it is important to emphasize that validity is not a one–time judgment. Instead, it is a matter of continual collection of evidence to support the intended inference.

Slide 8



Validation…

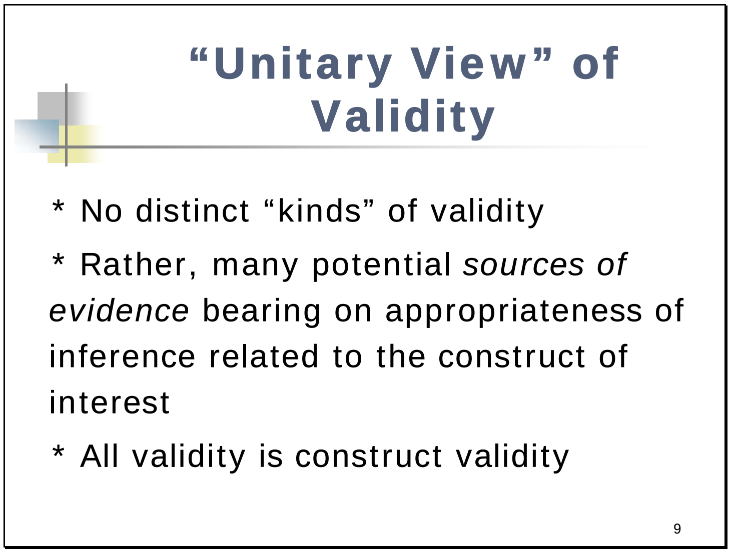
“begins with an explicit statement of the proposed interpretations of test scores”

(AERA, APA, NCME, 1999, p. 9)

Another commonly accepted principle is that validity does not apply to tests at all. Instead, it applies only to the inferences we want to make from the test results. So, one should never say “this test is valid.” Rather, one should say “these inferences are valid.” For example, if a student scores high on a reading test, it does not validate the reading test. However, it may be valid to infer that a student who has a high score on the reading test has a high degree of reading ability.

The most important thing about validation is the purpose or the intended interpretation. Validity begins with an explicit, clear statement about the intended interpretation or inferences to be made.

Slide 9



“Unitary View” of Validity

* No distinct “kinds” of validity

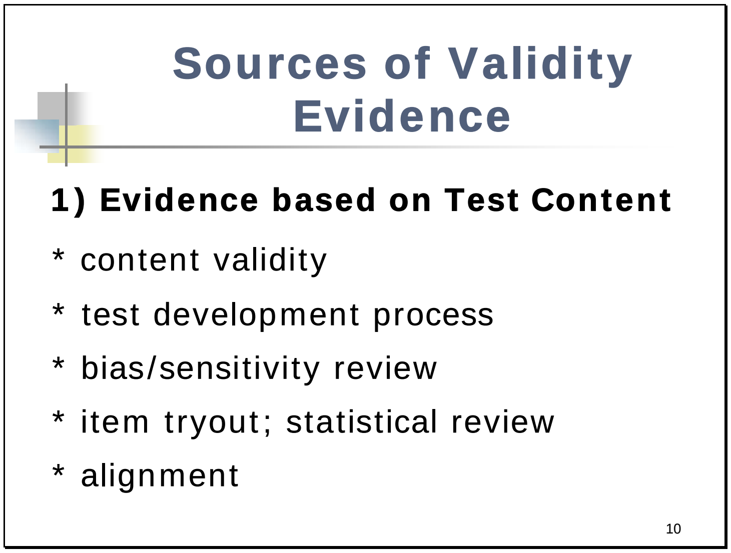
* Rather, many potential *sources of evidence* bearing on appropriateness of inference related to the construct of interest

* All validity is construct validity

Modern validity theory is considered unitary and can be traced back to Lee Cronbach. In contrast to modern validity theory, older validity theory described different kinds of validity: content validity, construct validity, and criterion validity. Modern validity theory posits that all validation is singly focused on providing evidence to support the interpretation or the inference. All validation bears on validation of the claims or the inferences we want to make with respect to the construct; all validity is essentially construct validity.

Instead of discussing different kinds of validity, we now focus on potential sources of evidence to support the inference. The most relevant source of evidence is that which is directly tied to the purpose of the inference. For example, in the case of a typical standards–referenced test, the objective is to make a claim about some set of content standards, or some set of knowledge or skills.

Slide 10



Sources of Validity Evidence

1) Evidence based on Test Content

* content validity

* test development process

* bias/sensitivity review

* item tryout; statistical review

* alignment

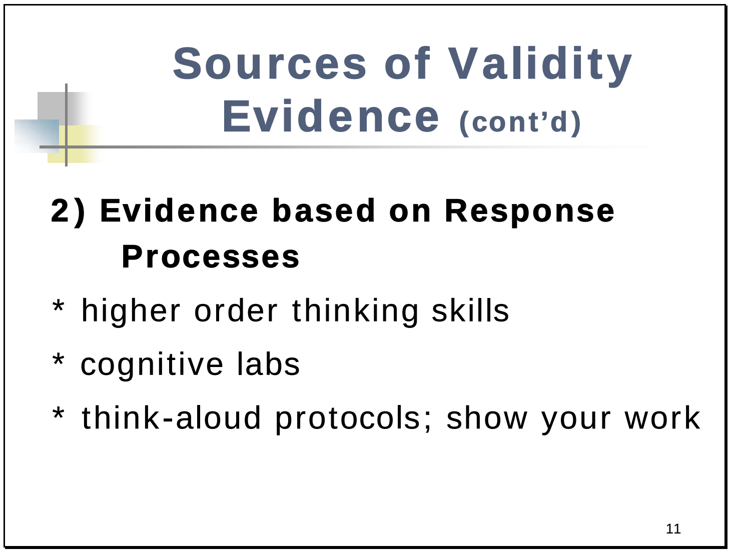
10

Professional organizations, such as the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) provide standards for sources of validity evidence in educational testing.

The first source of validity evidence is based on test content, which was previously called content validity. Some of the things that makes the NAEP assessments so solid are the test development process; how the frameworks are developed and the qualifications of the individuals involved; how items are developed and the methods through which the assessments are vetted; and how the assessments are shared with various audiences for review and comment. This is all part of NAEP’s content validity evidence.

Content validity is also derived from the item writing procedures, the item tryouts, the bias and sensitivity review procedures, the statistical reviews, the alignment studies, and other related processes. These all constitute validity evidence based on test content.

Slide 11



Sources of Validity Evidence (cont'd)

2) Evidence based on Response Processes

* higher order thinking skills

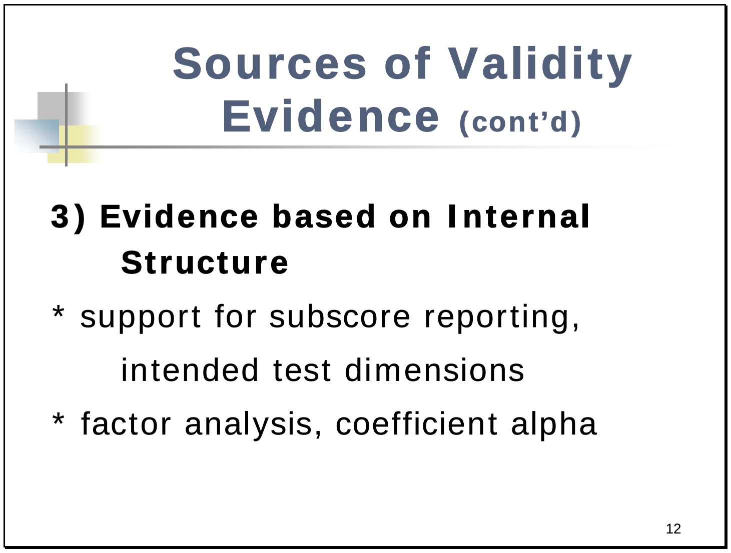
* cognitive labs

* think-aloud protocols; show your work

The second source of validity evidence is based on response processes. Suppose a test is developed that claims to measure an individual's level of cognition (e.g., higher–level thinking skills). To deliver solid inferences based on that claim, it would need to be demonstrated that when individuals are responding to a task or to an item they are not just recalling something from memory, but actually engaging in some higher–order cognitive process (e.g., synthesizing various sources of information and coming to a unique conclusion).

One method to identify response processes is to actually sit down with test takers and ask them questions such as: What are you doing? Why are you doing this? Similarly, math teachers attempt to understand their students' thought processes by asking them to solve math problems and show their work.

Slide 12



Sources of Validity Evidence (cont'd)

3) Evidence based on Internal Structure

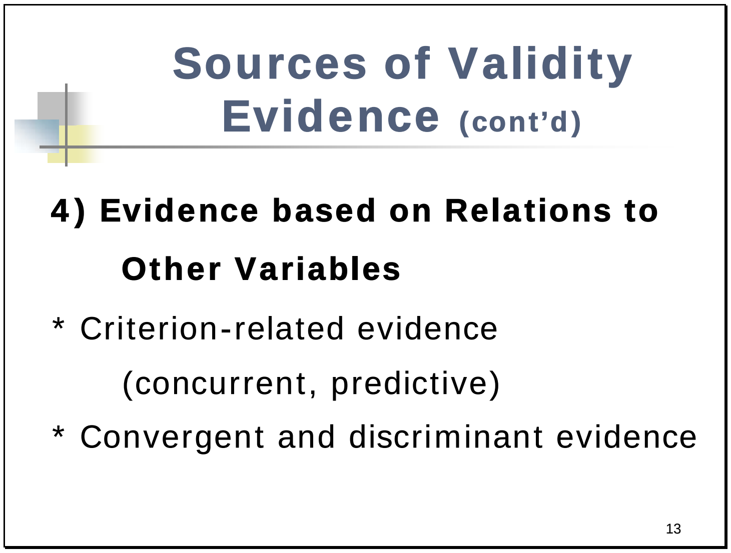
* support for subscore reporting, intended test dimensions

* factor analysis, coefficient alpha

A third source of validity evidence is evidence based on internal structure. There are ways of looking at tests and data to determine what is going on underneath the surface of observed responses. One of the common applications of evidence based on test structure is to support sub-score reporting. This is very typical in mathematics.

The National Council of Teachers of Mathematics suggests that math tests report sub-scores in strands, (e.g., numbers and operations, algebra, geometry, measurement) to determine how a student is performing in a particular skill area. This requires that items in the numbers and operations strand measure something distinct from the items in the algebra strand. An analysis can determine whether a collection of items supports inferences about a student's distinct knowledge or skill. This analysis is called an assessment of dimensionality, as it attempts to discern how many constructs are being measured. Typical statistical methods used to evaluate the dimensionality and homogeneity of a set of items are factor analysis and coefficient alpha. If the items seem to form a homogeneous subset, then that would support reporting scores based on distinct strands.

Slide 13



Sources of Validity Evidence (cont'd)

4) Evidence based on Relations to Other Variables

* Criterion-related evidence (concurrent, predictive)

* Convergent and discriminant evidence

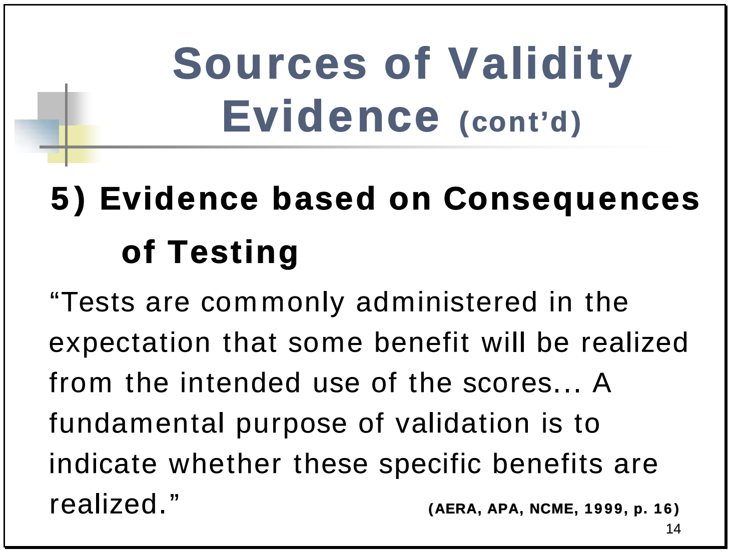
13

The fourth source of validity evidence is based on relations to other variables. Evidence of relationships to other variables is commonly found by using correlations. If observed relationships match predicted relationships, then the evidence supports the validity of the interpretation.

A criterion is a dependent variable about which we want to make a statement. Criterion–related evidence takes two forms: concurrent or predictive, based on how far into the future the criterion variable is measured. Criterion–related concurrent evidence requires that both variables are captured at one point in time. For example, two measures of fourth grade reading are collected at the same time, and then a correlation is observed. On the other hand, predictive criterion–related evidence is based on two sources of data collected at different points in time. For example, in the case of fourth and eighth grade reading performance, if one is good in reading at the fourth grade, to some extent, one's performance at eighth grade should also be good. A positive relationship would show evidence of predictive validity of fourth grade reading scores.

The second grouping is convergent and discriminant evidence. Convergent is when two measures are converging on the same construct and thus, should be strongly related. A fourth grade reading test, for example, should correlate very strongly with another fourth grade reading test. Discriminant is when two measures claim to be measuring different things and thus, should not be highly related. For example, the fourth grade reading test should correlate moderately with a fourth grade mathematics test and very poorly with a personality test measuring levels of introversion and extroversion.

Slide 14



Sources of Validity Evidence (cont'd)

5) Evidence based on Consequences of Testing

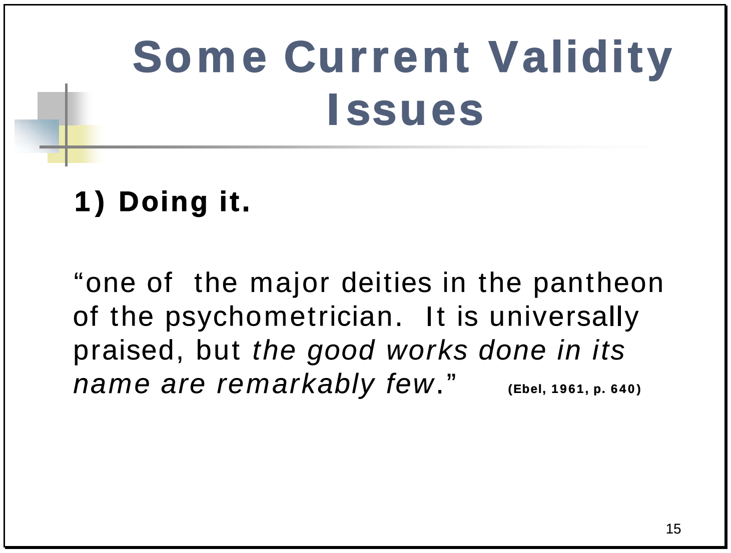
"Tests are commonly administered in the expectation that some benefit will be realized from the intended use of the scores... A fundamental purpose of validation is to indicate whether these specific benefits are realized." **(AERA, APA, NCME, 1999, p. 16)**

14

The fifth and final source of validity evidence is evidence based on consequences of testing. This source of validity evidence is the most controversial source mentioned in the AERA, APA, and NCME standards.

Tests are commonly administered in the expectation that some benefit will be realized from the intended use of the scores. This is referred to as consequential validity. The consequences of tests and test scores are clearly important and can be both positive and negative. To give you an example, suppose that scientists developed a new test for detecting a type of cancer. And this test was very accurate. Suppose that the test began to be used widely, and it was noticed that many people who had a positive test result were committing suicide. It is obvious that there is an unintended negative consequence. But the consequence has absolutely no bearing on the accuracy of the test. The test is still accurate in detecting the cancer. It is incumbent on the test developer to check for consequences. But it should be clear that consequences are not a part of the inference at all, and therefore, consequences have no part in validity.

Slide 15



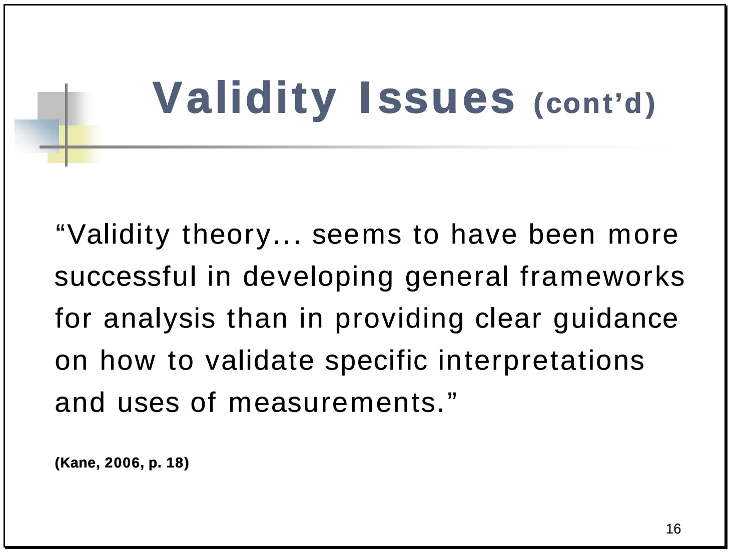
Some Current Validity Issues

1) Doing it.

“one of the major deities in the pantheon of the psychometrician. It is universally praised, but *the good works done in its name are remarkably few*.”   **(Ebel, 1961, p. 640)**

Validity is the most fundamental concern in testing. But, as related by Ebel’s quote, it is difficult to gather data on multiple variables, to design the kind of studies that would provide convincing evidence, to refute competing claims or competing interpretations, and to search out both positive and negative evidence to reach a conclusion. However, NAEP is certainly a leader in its efforts to validate inferences arising from its assessments.

Slide 16



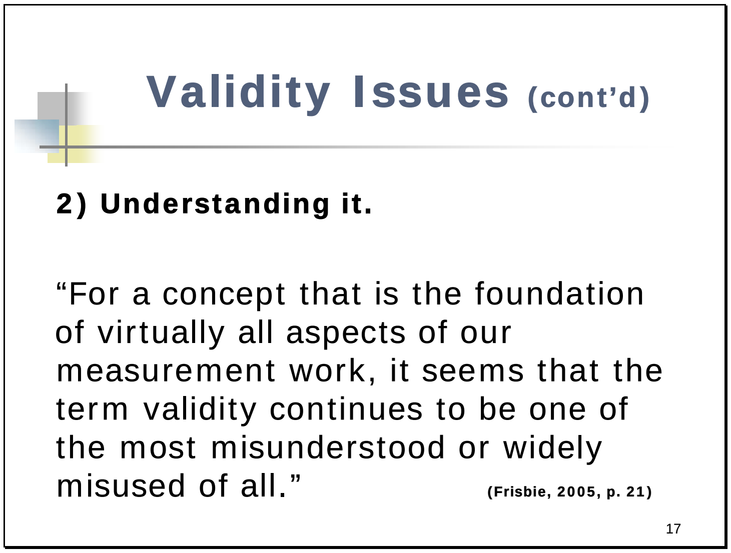
Validity Issues (cont'd)

"Validity theory… seems to have been more successful in developing general frameworks for analysis than in providing clear guidance on how to validate specific interpretations and uses of measurements."

(Kane, 2006, p. 18)

Michael Kane's quote above is supported by noting that most measurement students can repeat the five sources of validity evidence related in this presentation, but we all agree that the application of these principles is far more difficult.

Slide 17



Validity Issues (cont'd)

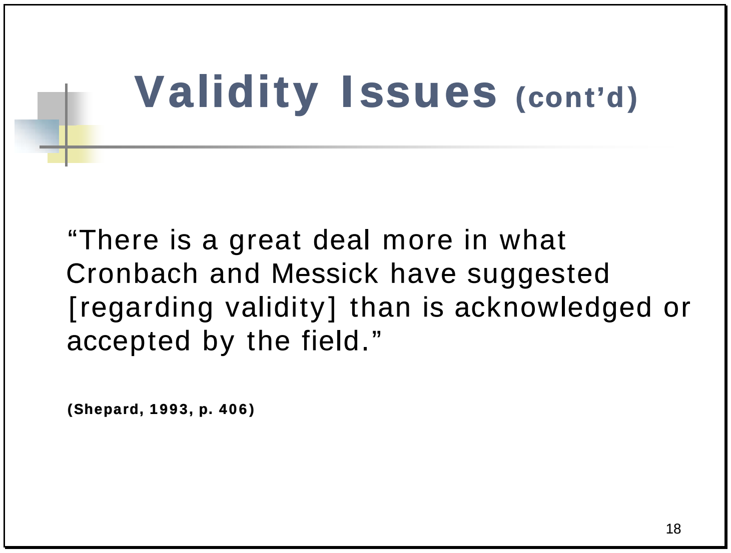
2) Understanding it.

"For a concept that is the foundation of virtually all aspects of our measurement work, it seems that the term validity continues to be one of the most misunderstood or widely misused of all." **(Frisbie, 2005, p. 21)**

17

Along with the difficulty in applying the principles of validity—"Doing It"—the second issue is "Understanding It." It is remarkable how many misunderstandings there are about validity. According to Dave Frisbie's presidential address to the National Council on Measurement in Education, …. validity continues to be the most widely misunderstood or misused term of all.

Slide 18



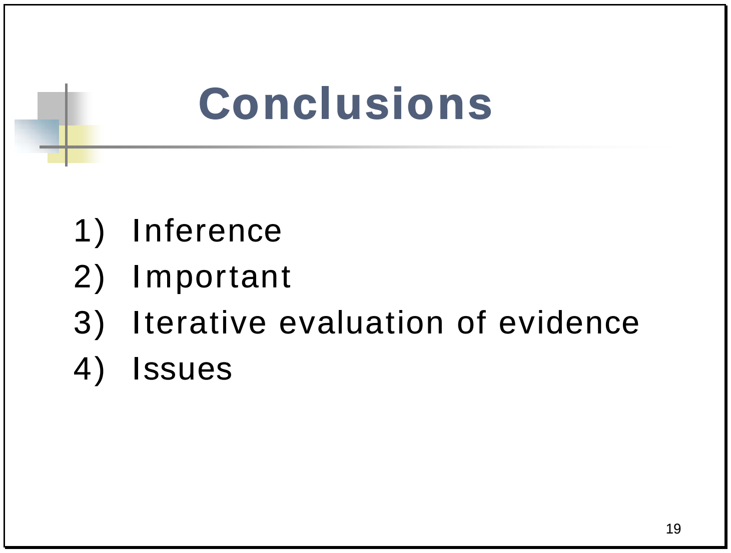
Validity Issues (cont'd)

"There is a great deal more in what Cronbach and Messick have suggested [regarding validity] than is acknowledged or accepted by the field."

(Shepard, 1993, p. 406)

The exact interpretation for this quote is not certain. Lorrie Shepard might have been saying Lee Cronbach and Samuel Messick were obtuse. But the most optimistic reading is that the field generally has not caught up with some of the very basics of modern validity theory.

Slide 19



Conclusions

1) Inference
2) Important
3) Iterative evaluation of evidence
4) Issues

In conclusion, there are four statements I would like to make about validity.

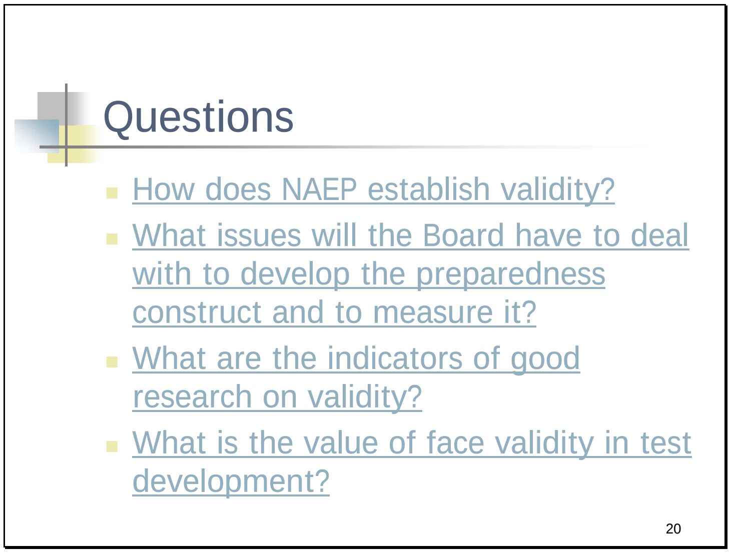
The first one is that validity is all about inferences. All validation is singly focused on providing evidence to support the interpretation or the inference. It is the purpose or intended inference that grounds our work. Whenever there is an issue with regard to validation of any NAEP assessment, the first questions should be: What is the purpose of this test? What is the intended inference one wants to make from scores on this test?

Second, validity is probably the most important thing we do in our field. Anybody can develop and distribute a test, but whether the scores on that test are meaningful and useful is the question to answer.

Third, validity is an ongoing process that requires gathering and synthesizing evidence. Evidence should continually be gathered to support or refute what is being claimed about the meaning of a test score.

Finally, like every field, validity has issues, and we have touched on several of these in this presentation.

Slide 20



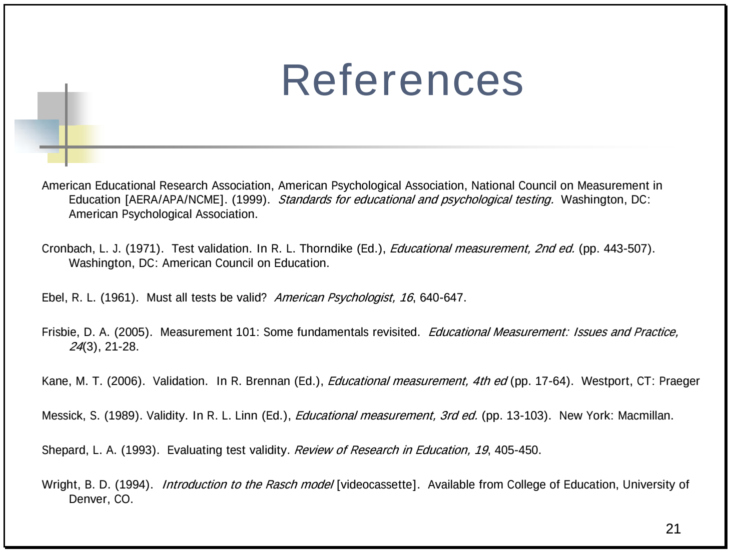
Questions

- How does NAEP establish validity?
- What issues will the Board have to deal with to develop the preparedness construct and to measure it?
- What are the indicators of good research on validity?
- What is the value of face validity in test development?

20

Questions asked following the presentation are listed here. Click on the question to receive the response.

Slide 21



References

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing.* Washington, DC: American Psychological Association.

Cronbach, L. J. (1971). Test validation. In R. L. Thorndike (Ed.), *Educational measurement, 2nd ed.* (pp. 443-507). Washington, DC: American Council on Education.

Ebel, R. L. (1961). Must all tests be valid? *American Psychologist, 16*, 640-647.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice, 24*(3), 21-28.

Kane, M. T. (2006). Validation. In R. Brennan (Ed.), *Educational measurement, 4th ed* (pp. 17-64). Westport, CT: Praeger

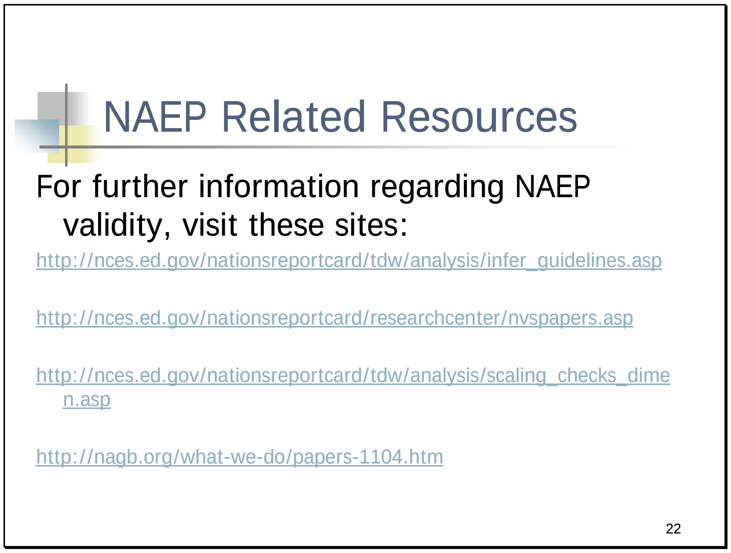
Messick, S. (1989). Validity. In R. L. Linn (Ed.), *Educational measurement, 3rd ed.* (pp. 13-103). New York: Macmillan.

Shepard, L. A. (1993). Evaluating test validity. *Review of Research in Education, 19*, 405-450.

Wright, B. D. (1994). *Introduction to the Rasch model* [videocassette]. Available from College of Education, University of Denver, CO.

These are the resources used to prepare the slide presentation.

Slide 22



NAEP Related Resources

For further information regarding NAEP validity, visit these sites:

http://nces.ed.gov/nationsreportcard/tdw/analysis/infer_guidelines.asp

http://nces.ed.gov/nationsreportcard/researchcenter/nvspapers.asp

http://nces.ed.gov/nationsreportcard/tdw/analysis/scaling_checks_dimen.asp

http://nagb.org/what-we-do/papers-1104.htm

These websites offer further information on validity as it relates to NAEP. The last resource deals specifically with validity issues for grade 12 NAEP preparedness.