AERA Mini-course: Using the *AERA/APA/NCME Standards for Educational and Psychological Testing* to Improve the Quality of Educational Research

Chapter 1, Validity

AERA 2017 Annual Meeting

Wayne J. Camara

ACT®

Organization of Validity Presentation

- Background and Context
- Validity Chapter – Standards and Comments
- Review of Scenarios

- 25 Standards – not all may be applicable to every scenario, research study, etc.

Three Clusters for Validity Standards

1. Establishing Intended Uses and Interpretations
2. Issues Regarding Samples and Settings Used in Validation
3. Specific Forms of Validity Evidence

Overarching Validity Standard

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

Cluster 1 – Establishing Intended Uses and Interpretations

Seven standards

Standard 1.1

The test developer should set forth clearly how test scores are intended to be interpreted and consequently used. The population(s) for which a test is intended should be delimited clearly, and the construct or constructs that the test is intended to assess should be described clearly.

Comments:

- It is incorrect to refer to the validity of a test or a valid/invalid test.
- Each interpretation for a given use requires a separate line of evidence.
- Developer should identify intended interpretations, population, limitations, and construct.
- Attend to construct relevant and irrelevant evidence (reading in math)

Standard 1.2

A rationale should be presented for each intended interpretation of test scores for a given use, together with a summary of the evidence and theory bearing on the intended interpretation.

Comments:

- Rationale for each intended score use needs to be given – it can include logical analysis and empirical evidence.
- Various types of evidence can support a validity argument – but no specific type is more rigorous and best.
- Test user bears ultimate responsibility for evaluating the quality of evidence and appropriateness for their local use(s).

Standard 1.3

If validity for some common or likely interpretation for a given use has not been evaluated, or if such an interpretation is inconsistent with available evidence, that fact should be made clear and potential users should be strongly cautioned about making unsupported interpretations.

Standard 1.4

If a test score is interpreted for a given use in a way that has not been validated, it is incumbent on the user to justify the new interpretation for that use, providing a rationale and collecting new evidence, if necessary.

Standard 1.5

When it is clearly stated or implied that a recommended test score interpretation for a given use will result in a specific outcome, the basis for expecting that outcome should be presented, together with relevant evidence.

Standard 1.6

When a test use is recommended on the grounds that testing or the testing program itself will result in some indirect benefit, in addition to the utility of information from interpretation of the test scores themselves, the recommender should make explicit the rationale for anticipating the indirect benefit. Logical or theoretical arguments and empirical evidence for the indirect benefit should be provided. Appropriate weight should be given to any contradictory findings in the scientific literature, including findings suggesting important indirect outcomes other than those predicted.

Standard 1.7

If test performance, or a decision made therefrom, is claimed to be essentially unaffected by practice and coaching, then the propensity for test performance to change with these forms of instruction should be documented.

Comments:

- Often claims may be made that retesting increases scores, or has no impact on scores, but that evidence may be based on the entire distribution of scores and not generalize to scores at the end points (high/low).
- Other claims may be made that test prep or coaching improves scores – evidence required to support claim (gain over no preparation needed rather than simple measure of change).
- Reports should advise on appropriate level of preparation, practice and familiarization that is beneficial. Familiarity with device, tools, item types/formats…)

Cluster 2 – Issues Regarding Samples and Settings Used in Validation

Three standards

Standard 1.8

The composition of any sample of test takers from which validity evidence is obtained should be described in as much detail as is practical and permissible, including major relevant sociodemographic and developmental characteristics.

Comments:

- When a sample is intended to represent a population, both the sample and population should be described in sufficient detail to allow comparisons based on factors that might reasonably impact representativeness.
- When a sample does not represent a population in one or more aspects (age, ethnicity, SES, geographic area, linguistic ability, motivation, experience, self-selection, institutional type) that should be documented in technical materials.
- Missing data should be noted, as well as any methods for handing such data.

Standard 1.9

When a validation rests in part on the opinions or decisions of expert judges, observers, or raters, procedures for selecting such experts and for eliciting judgments or ratings should be fully described. The qualifications and experience of the judges should be presented. The description of procedures should include any training and instructions provided, should indicate whether participants reached their decisions independently, and should report the level of agreement reached. If participants interacted with one another or exchanged information, the procedures through which they may have influenced one another should be set forth.

Comments:

- Content validation approaches often rely on SMEs to determine appropriate representation of content, establish cut scores).
- Report levels of agreement, types of backgrounds and expertise represented.

Standard 1.10

When validity evidence includes statistical analyses of test results, either alone or together with data on other variables, the conditions under which the data were collected should be described in enough detail that users can judge the relevance of the statistical findings to local conditions. Attention should be drawn to any features of a validation data collection that are likely to differ from typical operational testing conditions and that could plausibly influence test performance.

Comments:

- If data collection in such analyses differ from operational conditions – those differences need to be called out. Interpretations need to consider the influence of such conditions.
- Conditions include – motivation, device, timing, administrative conditions, scoring, test specifications.

Cluster 3 – Specific Forms of Validity Evidence 15 Standards

a) Content-Oriented - 1

b) Evidence regarding cognitive processes -1

c) Evidence regarding internal structure - 3

d) Evidence regarding relationships with conceptually related constructs - 1

e) Evidence regarding relationships with criteria - 8

f) Evidence based on consequences of tests - 1

Standard 1.11 - CONTENT

When the rationale for test score interpretation for a given use rests in part on the appropriateness of test content, the procedures followed in specifying and generating test content should be described and justified with reference to the intended population to be tested and the construct the test is intended to measure or the domain it is intended to represent. If the definition of the content sampled incorporates criteria such as importance, frequency, or criticality, these criteria should also be clearly explained and justified.

Comments:

- Alignment and item mapping processes should be documented.
- Areas of the content domain not included in the test blueprint should be identified.
- Document the basis for processes, logical structure, methods.

Standard 1.12 – COGNITIVE PROCESSES

If the rationale for score interpretation for a given use depends on premises about the psychological processes or cognitive operations of test takers, then theoretical or empirical evidence in support of those premises should be provided. When statements about the processes employed by observers or scorers are part of the argument for validity, similar information should be provided.

Comments:

- If specific cognitive processes are specified then evidence is needed to verify items do in fact tap those processes and not other facets (reasoning vs computation).

Standard 1.13- 1.15 – INTERNAL STRUCTURE

1.13 Evidence of internal structure of the test

- Comments: Unidimensionality, multivariate analysis, factor analysis, interrelationships of factors and scores.

1.14 Subscores, profiles – require evidence. Explain the basis and rationale for composite scores.

- Comments: Distinctiveness and reliability of separately reported scores is required. Evidence to support interpretations of separate scores is required. Explanation needed for rationale and computation of composite scores

1.15 Interpretations of individual items (Exemplar items)

- Comments: Documentation required to discourage over interpretations of isolated items.

Standard 1.16 - CONSTRUCTS

When validity evidence includes empirical analyses of responses to test items together with data on other variables, the rationale for selecting the additional variables should be provided. Where appropriate and feasible, evidence concerning the constructs represented by other variables, as well as their technical properties, should be presented or cited. Attention should be drawn to any likely sources of dependence (or lack of independence) among variables other than dependencies among the construct(s) they represent.

Comments:

- Statistical relationships among scores and other variables should be consistent with external research and theory (parental education, rigor of coursework, persistence, achievement).
- Examine relationship with other measures of similar construct.
- Guard against spurious relationships and dependencies.

Standard 1.17- 1.24 – CRITERIA

1.17 Technical information on criterion variable – when evidence relies primarily on relationship to criteria.

1.18 Specify level of criterion performance when the claim is a test adequately predicts performance (College and Career Readiness)

1.19 – When test scores used with other variables to predict outcomes – statistical models should include relevant variables.

1.20 – When relationships are reported between test scores and outcomes indices of degree of uncertainty should be reported (confidence intervals, standard errors).

Standard 1.17- 1.24 – CRITERIA

1.21 – When statistical adjustments (restriction of range or attenuation) are made, both the adjusted and unadjusted coefficients and procedures used should be reported.

1.22 – If meta analysis used as evidence of test-criterion relationship, the test and criterion variables in local situation should be comparable to those in the studies.

1.23 – Any meta-analytic evidence used to support a score interpretation should describe each study, permitting independent evaluation.

1.24 – Tests used for assignment to different programs should report evidence of different outcomes resulting from programs (treatments) when feasible.

Standard 1.25 – CONSEQUENCES

1.25 When unintended consequences result from test use, an attempt should be made to investigate whether such consequences arise from the test's sensitivity to characteristics other than those it is intended to assess or from the test's failure to fully represent the intended construct.

Comments:

- Is the score limited by construct-irrelevant components or construct-underrepresentation?

SCENARIO 1

College Z requires entering students to either have a grade of B in their high school Algebra II course or a score in the top 60% on their local placement test to be placed in a college credit math course. Students who do not meet this requirement must complete a non-credit developmental math course. Overall, ¼ of all freshmen do not meet either requirement and must take a developmental course. Over half of all students placed into developmental math either do not complete the course or drop out during their freshmen year. In comparison, less than 20% of students who barely meet this requirement will drop out during their freshmen year. First-generation students and underrepresented minorities are twice as likely to be placed in the developmental courses the drop-out rates for these groups are three times as high. Math faculty only imposed this requirement recently because they believed standards were too low, however, no research was conducted on the placement policy, cut scores, or effectiveness of developmental courses.

Which standards are most relevant? What type of validity evidence is required to support this use? Is the validity evidence required at that specific university or could other research support such use?

SCENARIO 2

A 30-question middle school math test provides 6 subscores on specific math competencies based on percentiles from a large sample of students attending parochial schools in NYC. The test is being used in Tulsa, OK to determine if students are prepared to advance to the next level in math. An educational vendor will provide mandatory on-line instruction during the summer for any student scoring in the bottom 20% on any subtest. A research study has been cited that shows students at the bottom 20% of the distribution on 3 subtests have a 50% probability of failing math during middle school, but requests for details on the study have not been provided to teachers and parents.

Which standards are most relevant? What type of validity evidence is required to support this use? What is the responsibility of the district and vendor related to the required intervention and details of the research study? What other concerns do you have?

Wayne.camara@act.org