Aspects of Validity

An Argument-Based, Systematic Framework to Study Validity and Reliability of Unit and Program Assessments

Nancy Wellenzohn, EdD
Associate Dean & Director of Accreditation
CAEP Coordinator
WHY ANALYZE VALIDITY?

• Assessments are instruments that demonstrate that goals and objectives are being met.
• Goals and objectives are established using standards, current research in best practice, conversations with field partners, and other relevant sources.
• A validity study provides legitimacy to the program assessments.
WHY ANALYZE VALIDITY?

• Connecting assessments and curriculum to standards, best practices, and needs of the field is something that has always been done.
• The new CAEP processes now want us to prove it.
• Validity is more than just statistical validity.
CAEP REQUIREMENTS

- CAEP Evidence Guide provides a broad discussion of what makes an assessment valid and reliable.
- CAEP White Paper “Principles for measures used in CAEP Accreditation Process” (Ewell, 2013) provides relevant insights.
- Supporting literature provides additional guidance.
- Informal perceptions of validity are no longer enough.
CAEP REQUIREMENTS

CAEP 5.2 says “provider’s quality assurance system relies on relevant, verifiable, representative, cumulative, and actionable measures, and produces empirical evidence that interpretations of data are valid and consistent.”
 VALIDITY LITERATURE

• Messick (1995) defined validity as “nothing less than an evaluative summary of both the evidence for and the actual as well as potential consequences of score interpretation and use.”

• Need to look at the validity of the instrument and the validity of the data.
ASPECTS OF VALIDITY

• Need a clear and practical way to systematically study validity.
• Messick separated the concept of validity into six separate aspects.
• These aspects provide a good place to start.
ASPECTS OF VALIDITY

Instrument Aspects

• Content
• Structural
• Consequential
ASPECTS OF VALIDITY

Results Aspects

- Generalizability
- External
- Substantive
CONTENT ASPECT OF VALIDITY

- Evidence of “content relevance, representativeness, and technical quality” (Messick 1995, p.6)
- Can be supported by content and performance standards
- Topic of assessment can be found in professional domain
STRUCTURAL ASPECT OF VALIDITY

• Instrument “appraised the extent to which the internal structure of the assessment is consistent with the construct domain.” (Messick, 1995, p. 6)

• Are we asking the right question?
CONSEQUENTIAL ASPECT OF VALIDITY

• “Appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use…” (Messick, 1995, p.6)

• Does the instrument lead to results, positive or negative, that are meaningful?
GENERALIZABILITY ASPECT OF VALIDITY

• “Extent to which score properties and interpretations generalize to and across population groups, setting, and tasks” (Messick, 1995, p.6)
• Are the data consistent between groups, over time, and consistent with best practice in the field?
• Are the data predictive?
EXTERNAL ASPECT OF VALIDITY

• Includes “convergent and discriminant evidence from multi-trait and multi-method comparisons as well as evidence of criterion relevance and applied utility” (Messick 1995, p. 6)

• Does the data correlate with other variables? Are the results consistent with other assessments? Are conclusions made considering results of multiple assessments?
SUBSTANTIVE ASPECT OF VALIDITY

• “Theoretical rationales for the observed consistencies in test responses” (Messick, 1995, p.6)

• Are the candidates taking the right actions, meaning similar to those in the field?
CAEP WHITE PAPER
PETER EWELL
“PRINCIPLES FOR MEASURES USED IN THE CAEP ACCREDITATION PROCESS”

- Validity and Reliability
- Relevance
- Verifiability
- Representativeness
- Cumulativeness
- Fairness
- Stakeholder Interest
- Benchmarks
- Vulnerability to Manipulation
- Actionability

OVERLAP BETWEEN THE EWELL AND MESSICK CONCEPTS IS APPARENT
CONCEPT OF UNITARY VALIDITY
(MESSICK 1989)

• The standard for studying validity for years had been to consider content, construct, and criterion validity.
• Messick said “an ideal validation includes several different types of evidence that spans all three of the categories.” (Messick 1989)
• This allows for the consideration of different types of evidence rather than separately studying different types of validity.
AN ARGUMENT-BASED APPROACH TO VALIDATION (KANE, 2013)

• “Under the argument-based approach to validity, test-score interpretations and uses that are clearly stated and are supported by appropriate evidence are considered to be valid.” (Kane, 2013)

• This means that programs can validate their assessments by providing multiple pieces of evidence that lend support to the notion that an assessment is valid
TAKEAWAYS FROM THE LITERATURE

• Messick’s aspects of validity provide a useful framework for analysis
• Ewell’s principles for measures used in CAEP accreditation are supportive of and related to the aspects of validity
• Kane suggests that programs can make arguments that assessments are valid
• Messick’s unitary theory suggests that multiple factors can be considered
Aspects of Validity Plot Graph
Composite Scores for Instrument and Result Validity

Quadrants:
I - High Instrument and High Result Validity
II - High Instrument and Low Result Validity
III - Low Instrument and Low Result Validity
IV - Low Instrument and High Result Validity

Key:
- Results are Unacceptable
- Acceptable but Improvement
- Results are Acceptable or Target
- No Fill
# Aspects of Validity Review - Instrument

<table>
<thead>
<tr>
<th>Aspect of Construct Validity</th>
<th>Unacceptable 1</th>
<th>Acceptable 2</th>
<th>Target 3</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Content Aspect of Construct Validity: Evidence of content relevance</strong></td>
<td>Assessment Content does not meet at least two of the following:</td>
<td>Assessment Content meets at least two of the following:</td>
<td>Assessment Content meets all of the following:</td>
</tr>
<tr>
<td></td>
<td>Aligned with national or state standards</td>
<td>Aligned with national or state standards</td>
<td>Aligned with national or state standards</td>
</tr>
<tr>
<td></td>
<td>Developed with input from external partners.</td>
<td>Developed with input from external partners.</td>
<td>Developed with input from external partners.</td>
</tr>
<tr>
<td></td>
<td>Measure is relevant and demonstrably related to an issue of importance</td>
<td>Measure is relevant and demonstrably related to an issue of importance</td>
<td>Measure is relevant and demonstrably related to an issue of importance</td>
</tr>
<tr>
<td><strong>Structural Aspect of Construct Validity: Observed consistency in responses</strong></td>
<td>Assessment Structure does not meet at least 2 of the following:</td>
<td>Assessment Structure meets at least 2 of the following:</td>
<td>Assessment Structure meets at least 3 of the following:</td>
</tr>
<tr>
<td></td>
<td>Data are verifiable; can be replicated by third parties</td>
<td>Data are verifiable; can be replicated by third parties</td>
<td>Data are verifiable; can be replicated by third parties</td>
</tr>
<tr>
<td></td>
<td>Measure is typical of underlying situation, not an isolated case</td>
<td>Measure is typical of underlying situation, not an isolated case</td>
<td>Measure is typical of underlying situation, not an isolated case</td>
</tr>
<tr>
<td></td>
<td>Data are compared to benchmarks such as peers or best practices.</td>
<td>Data are compared to benchmarks such as peers or best practices.</td>
<td>Data are compared to benchmarks such as peers or best practices.</td>
</tr>
<tr>
<td></td>
<td>Program considers sources of potential bias</td>
<td>Program considers sources of potential bias</td>
<td>Program considers sources of potential bias</td>
</tr>
<tr>
<td><strong>Consequential Aspect of Construct Validity: Positive and negative consequences, either intended or unintended, are observed and discussed</strong></td>
<td>Assessment Consequences are reviewed but do not ensure at least 2 of the following:</td>
<td>Assessment Consequences are reviewed to ensure at least 2 of the following:</td>
<td>Assessment Consequences are reviewed to ensure all of the following:</td>
</tr>
<tr>
<td></td>
<td>Measure is free of bias</td>
<td>Measure is free of bias</td>
<td>Measure is free of bias</td>
</tr>
<tr>
<td></td>
<td>Measure is justly applied</td>
<td>Measure is justly applied</td>
<td>Measure is justly applied</td>
</tr>
<tr>
<td></td>
<td>Data are reinforced by reviewing related measures to decrease vulnerability to manipulation</td>
<td>Data are reinforced by reviewing related measures to decrease vulnerability to manipulation</td>
<td>Data are reinforced by reviewing related measures to decrease vulnerability to manipulation</td>
</tr>
<tr>
<td>ASPECTS OF VALIDITY REVIEW - RESULTS</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>--------------------------------------</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>Unacceptable 1</strong></td>
<td><strong>Acceptable 2</strong></td>
<td><strong>Target 3</strong></td>
<td></td>
</tr>
<tr>
<td><strong>Substantive Aspect of Construct Validity: Observed consistency in the test responses/scores</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment Substance does not meet at least 2 of the following:</td>
<td>Assessment Substance meets at least 2 of the following:</td>
<td>Assessment Substance meets at least 3 of the following:</td>
<td></td>
</tr>
<tr>
<td>Measure has been subject to independent verification</td>
<td>Measure has been subject to independent verification</td>
<td>Measure has been subject to independent verification</td>
<td></td>
</tr>
<tr>
<td>Measure is typical of situation, not an isolated case, and representative of entire population</td>
<td>Measure is typical of situation, not an isolated case, and representative of entire population</td>
<td>Measure is typical of situation, not an isolated case, and representative of entire population</td>
<td></td>
</tr>
<tr>
<td>Data are compared to benchmarks such as peers or best practices</td>
<td>Data are compared to benchmarks such as peers or best practices</td>
<td>Data are compared to benchmarks such as peers or best practices</td>
<td></td>
</tr>
<tr>
<td>The program considers whether data are vulnerable to manipulation</td>
<td>The program considers whether data are vulnerable to manipulation</td>
<td>The program considers whether data are vulnerable to manipulation</td>
<td></td>
</tr>
<tr>
<td><strong>Generalizability Aspect of Construct Validity: Results generalize to and across population groups</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Assessment Generalizability does not include at least 2 of the following:</td>
<td>Assessment Generalizability includes at least 2 of the following:</td>
<td>Assessment Generalizability includes all of the following:</td>
<td></td>
</tr>
<tr>
<td>The results can be subject to independent verification and if repeated by observers it would yield similar results</td>
<td>The results can be subject to independent verification and if repeated by observers it would yield similar results</td>
<td>The results can be subject to independent verification and if repeated by observers it would yield similar results</td>
<td></td>
</tr>
<tr>
<td>Measure is free of bias able to be justly applied by any potential user or observer</td>
<td>Measure is free of bias able to be justly applied by any potential user or observer</td>
<td>Measure is free of bias able to be justly applied by any potential user or observer</td>
<td></td>
</tr>
<tr>
<td>Measure provides specific guidance for action and improvement</td>
<td>Measure provides specific guidance for action and improvement</td>
<td>Measure provides specific guidance for action and improvement</td>
<td></td>
</tr>
<tr>
<td><strong>External Aspect of Construct Validity: Correlations with external variables exist</strong></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>External Aspect of the assessment does not include at least 2 of the following:</td>
<td>External Aspect of the assessment includes at least 2 of the following:</td>
<td>External Aspect of the assessment includes 3 of the following:</td>
<td></td>
</tr>
<tr>
<td>Data are correlated with external data.</td>
<td>Data are correlated with external data.</td>
<td>Data are correlated with external data.</td>
<td></td>
</tr>
<tr>
<td>If repeated by third parties, results would replicate.</td>
<td>If repeated by third parties, results would replicate.</td>
<td>If repeated by third parties, results would replicate.</td>
<td></td>
</tr>
<tr>
<td>Measure is combined with other measures to increase cumulative weight of results</td>
<td>Measure is combined with other measures to increase cumulative weight of results</td>
<td>Measure is combined with other measures to increase cumulative weight of results</td>
<td></td>
</tr>
<tr>
<td>Student work is created with some instructor support</td>
<td>Student work is created with some instructor support</td>
<td>Student work is created without instructor support</td>
<td></td>
</tr>
</tbody>
</table>
RESULTS CAN BE GRAPHAED

• The instrument aspect scores (content, structural, and consequential aspects) can be averaged.
• The results aspects scores (substantive, generalizability, and external aspects) can be averaged.
• Resulting scores can be plotted.
• This can be done for a collection of assessments in a program to provide a visual representation of how the program is doing overall.
Aspects of Validity Plot Graph
Composite Scores for Instrument and Result Validity

Quadrants:
I - High Instrument and High Result Validity
II - High Instrument and Low Result Validity
III - Low Instrument and Low Result Validity
IV - Low Instrument and High Result Validity

Key:
- Results are Unacceptable
- Acceptable but Improvement
- Results are Acceptable or Target
- No Fill

Results Aspects
IMPLEMENTATION

• This is a peer-review process.
• Directors/Chairs make arguments that their assessments are valid.
• They submit evidence for each argument. The Director/Chair self-scores the evidence in a software system.
• A 2-3 person panel reviews the arguments and applies the rubric. They meet to form a consensus for scoring.
## Content Aspect of Construct Validity

<table>
<thead>
<tr>
<th>Director:</th>
<th>Target</th>
<th>Reviewer:</th>
<th>Acceptable</th>
</tr>
</thead>
</table>

**Director's Explanation:**
This assessment is aligned with ELCC Standards 1.1 to 1.4. The ELCC is the recognition body approved by NCATE/CAEP. (See ELCC Standard 1) (EN 01)

The standards were developed in consultation with external partners, namely members of the Interstate School Leaders Licensure Consortium which is made up of both school and district leaders as well as faculty from EPPs. State officials, such as Dr. King from New York State Education Department are involved in the review process, representing the students, schools, and districts of New York State. (EN 02)

Effective leadership has been tied to student performance in resources provided by NYSED at https://www.engageny.org/about. (EN 03)

[Need rationale for how assignment was developed with input from partners]

---

**Table:**

<table>
<thead>
<tr>
<th>Content Aspect of Construct Validity: Evidence of content relevance</th>
<th>Assessment Content does not meet at least two of the following:</th>
<th>Assessment Content meets at least two of the following:</th>
<th>Assessment Content meets all of the following:</th>
</tr>
</thead>
<tbody>
<tr>
<td>Content Aspect of Construct Validity: Evidence of content relevance</td>
<td>Aligned with national or state standards</td>
<td>Aligned with national or state standards</td>
<td>Aligned with national or state standards</td>
</tr>
<tr>
<td>Assessment Content does not meet at least two of the following:</td>
<td>Developed with input from external partners</td>
<td>Developed with input from external partners</td>
<td>Developed with input from external partners</td>
</tr>
<tr>
<td>Aligned with national or state standards</td>
<td>Measure is relevant and demonstrably related to an issue of importance</td>
<td>Measure is relevant and demonstrably related to an issue of importance</td>
<td>Measure is relevant and demonstrably related to an issue of importance</td>
</tr>
</tbody>
</table>
EDUCATIONAL LEADERSHIP EXAMPLE

Substantive Aspect of Construct Validity

**Director:** Target  | **Reviewer:** Target

**Director's Explanation:**
Data are collected in a system called Taskstream. Students submit their work to a dropbox in the system and the instructors evaluate work that is submitted. The reports are generated by a disinterested third party, the SEHS program coordinator. Multiple people have held this role and results generated have been consistent. (EN 4)

The process of instructional improvement planning is typical of most schools and districts. Schools in New York are required to engage in data-driven improvement planning. This is highlighted at NYSED's website: http://www.p12.nysed.gov/accountability/School_Improvement/home.html. (EN 10)

Students are encouraged to rely on publically generated data that are benchmarked across schools and districts. This includes school report cards that can be found here: https://reportcards.nysed.gov/. (EN 6)

Improvement planning is a typical activity. Students are able to view many different types of improvement plans at the NYSED website: http://www.p12.nysed.gov/accountability/T1/titlea/sig1003g/home.html. (EN 11)

The data from this assessment and the NYS SDL examinations demonstrate proficiency of ELCC SDL 1.1 to 1.4. The results are internally consistent with each other. The SDL examination allows for comparison to peers. The program exceeds the required 80% pass rate to be in good standing in New York. The two data sources were described favorably in the Comments from SPA Reviewer on Assessments Meeting SDL 1.1 to 1.4. (EN 12)

<table>
<thead>
<tr>
<th>Substantive Aspect of Construct Validity: Observed consistency in the test responses/scores</th>
<th>Assessment Substance does not meet at least 2 of the following:</th>
<th>Assessment Substance meets at least 2 of the following:</th>
<th>Assessment Substance meets at least 3 of the following:</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Measure has been subject to independent verification</td>
<td>Measure has been subject to independent verification</td>
<td>Measure has been subject to independent verification</td>
</tr>
<tr>
<td></td>
<td>Measure is typical of situation, not an isolated case, and representative of entire population</td>
<td>Measure is typical of situation, not an isolated case, and representative of entire population</td>
<td>Measure is typical of situation, not an isolated case, and representative of entire population</td>
</tr>
<tr>
<td></td>
<td>Data are compared to benchmarks such as peers or best practices</td>
<td>Data are compared to benchmarks such as peers or best practices</td>
<td>Data are compared to benchmarks such as peers or best practices</td>
</tr>
<tr>
<td></td>
<td>The program considers whether data are vulnerable to manipulation</td>
<td>The program considers whether data are vulnerable to manipulation</td>
<td>The program considers whether data are vulnerable to manipulation</td>
</tr>
</tbody>
</table>
RESULT OF VALIDITY REVIEW

COMPOSITE SCORE – ASPECTS OF VALIDITY

PROGRAM: EDUCATIONAL LEADERSHIP

Aspects of Validity Plot Graph
Composite Scores for Instrument and Result Validity

Quadrants:
I - High Instrument and High Result Validity
II - High Instrument and Low Result Validity
III - Low Instrument and Low Result Validity
IV - Low Instrument and High Result Validity
REFERENCES

Council for the Accreditation of Educator Preparation, CAEP Evidence Guide, February 2014


QUESTIONS?
wellenzn@Canisius.edu