How Should We Validate University Admissions Tests?

Stephen G. Sireci
University of Massachusetts Amherst

Presentation a la Seminario DEMRE
Transparencia y Validez en la Selección en Educación Superior
6 de diciembre de 2017
Universidad de Chile
Purposes of presentation

1. Define validity
   - For testing in general
2. Discuss validation
   - For college admissions tests
3. Provide a framework for validating
   - Educational tests in general
   - College admissions tests
Defining validity

- What is the common interpretation of this term?
- Let’s look at the definition in the dictionary
What is validity?

According to Webster's Dictionary:

Validity:

1. the state or quality of being valid; specifically, (a) strength or force from being supported by fact; justness; soundness; (b) legal strength or force.

2. strength or power in general

3. value (rare)

http://www.merriam-webster.com/dictionary/validity
How do psychometricians describe validity?

- OLDER notions:
  
  “a test is valid for anything with which it correlates”  --Guilford (1946)

  “the validity of a test is the correlation of the test with some criterion”  --Gulliksen (1950)
How do psychometricians describe validity?

- OLDER notions:
  
  “a test is valid for anything with which it correlates” — Guilford (1946)
  
  “validity of a test is the correlation of the test with some criterion” — Gulliksen (1950)
APA, AERA, & NCME

“Standards”

- 1954
- 1966
- 1974
- 1985
- 1999
- 2014
(Current) Psychometric definition of validity

“Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests”

--AERA, APA, NCME Standards (2014, p. 11)
Current definition of Validity

“Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests”

--AERA/APA/NCME Standards (2014, p. 11)
Understanding validity

- Validity is **NOT** a property of a test.
- Validity refers to inferences derived from test **scores**.
  - What we seek to validate are the uses (decisions) of test scores.
- Validity must be evaluated **with respect to a specific testing purpose**. Thus, a test may be appropriate for one purpose, but not for another.
So, what does this 21st century definition mean for the validation of college admissions tests?

- In the USA?
- In Mexico?
- In Chile?
- In Sweden?

**ANSWER:** The same thing it means for ANY test.

We must begin by specifying the intended PURPOSE and USE of the test scores.
Therefore,

Before we talk about:
- Methods
- Statistics
- Validity theory
- Validity terminology

We must identify the purpose of the *Prueba de Selección Universitaria* (PSUs)
and,

Before we talk about
- Methods
- Statistics
- Validity theory
- Validity terminology

We must ask “What are the uses of PSU scores?”
PSU purposes

- Provide information for universities to use in selecting students
- Assess (measure)
  - “Secondary school curriculum”
  - Language skill
  - Math reasoning
  - Geography, economics…
  - Physics, biology, chemistry
PSU uses

- University admissions
- Student scholarships
- University funding
  - aporte fiscal indirecto (AFI)
  - ya no!

- Accountability?
  - Colegios?
  - Escuelas?
  - Universidades?
The AERA et al. (2014) *Standards* define validity as,

“Validity refers to the degree to which evidence and theory support the interpretations of test scores for proposed uses of tests” (p. 11).
What guidance does the *Standards* give us for validation research?

Five “sources of evidence that might be used in evaluating the validity of a proposed interpretation of test scores for a particular use” (pp. 13).
Standards’ Validation Framework

5 Sources of Validity evidence:

1. Test content
2. Response processes
3. Internal structure
4. Relations to other variables
5. Testing consequences
Standards’ 5 Sources of Validity evidence:

1. Validity evidence based on test content
   a) Domain definition
   b) Domain relevance
   c) Domain representation
   d) Appropriate test construction procedures

Sireci (1998), Sireci & Faulkner-Bond (2014)
Validity evidence based on test content

a) Domain definition
   - How is the domain of content being measured defined?
   - Would most experts and stakeholders agree with this definition?
   - Do we have consensus that the knowledge and skill domain measured is consistent with the test purpose?
   - PSU: Content defined by high school curriculum? Content defined by skills needed for success in college?
Validity evidence based on test content

b) Domain relevance

- Are all items on the test relevant to the content domain?
- Is test content relevant to success in college?
Validity evidence based on test content

c) Domain representation

- Does the test fully represent (measure all aspects of) the intended content domain?
- Does test content represent skills needed for success in college?
- Does test content represent escuela curriculum?
Three Components in the Educational Process:

- Curriculum
- Instruction
- Assessment
Aligning Curriculum, Instruction, & Assessment

- Curriculum
- Test Specifications
- Instruction
- Assessment
Messick (1989)

“Tests are imperfect measures of constructs because they either leave out something that should be included...or else include something that should be left out, or both” (p. 34)

Validity evidence based on test content is needed to ensure construct representation, and the absence of irrelevant material
Validity evidence based on test content

d) Appropriate test construction procedures
   - Expert consensus define domain
   - Content reviews of items
   - Sensitivity reviews of items
   - Statistical reviews of items
   - Other quality control procedures
Standards’ 5 Sources of Validity evidence:

2. Validity evidence based response processes
   The degree to which test items measure the intended cognitive processes. Are the cognitive skills intended to be measured actually being measured?
   – Are we only measuring “test wiseness”
   – How do students solve items?
   – Does the test really measure “higher-order” thinking?
Validity evidence based on response processes

Examples:
- Cognitive interviews (Padilla, Benitez)
- Think-aloud protocols
- Computer-based testing: analysis of item response time, “Semi-amorphous data”
- Analysis of eye-movements
Standards’ 5 Sources of Validity evidence:

3. Validity evidence based on “internal structure”
   - Dimensionality analyses
   - DIF (item bias) analyses
   - Equating invariance
   - Analysis of measurement precision
     - Internal consistency, other rel. analyses
     - Test information, SEM/CSEM
     - Decision consistency/accuracy
     - G-studies, D-studies
Validity Evidence Based on Internal Structure

- **Dimensionality analyses**
  - How many dimensions are being measured? Are these the hypothesized/intended dimensions?
  - Statistical procedures:
    - IRT model-data (residual) analyses
    - Factor analyses: exploratory, confirmatory, non-linear
    - Multidimensional scaling, etc.

- Can focus on entire population, or subpopulations
Standards’ 5 Sources of Validity evidence:

4. Validity evidence based on relations to other variables
   - How well do test scores...
     - Predict?
     - Relate?
     - Distinguish?
     - Confirm?
Validity evidence based on relations to other variables

- Many types
  - Concurrent Validity
  - Predictive Validity (formerly criterion-related validity)
  - “Differential” predictive validity
  - Multitrait-Multimethod matrix studies
  - Experimental studies, comparisons of groups
<table>
<thead>
<tr>
<th>Test</th>
<th>Benchmark</th>
<th>Criterion</th>
<th>Comments/Citations</th>
</tr>
</thead>
<tbody>
<tr>
<td>ACT English</td>
<td>18</td>
<td>.75 probability of C and .50 probability of B</td>
<td>Allen &amp; Sconing (2005)</td>
</tr>
<tr>
<td>ACT Reading</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>ACT Math</td>
<td>22</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT Composite</td>
<td>1550</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT-Math</td>
<td>610-630</td>
<td>.65 probability of B- in first-year GPA</td>
<td>Wyatt et al. (2013)</td>
</tr>
<tr>
<td>SAT-Reading</td>
<td>500</td>
<td></td>
<td></td>
</tr>
<tr>
<td>SAT-Writing</td>
<td>470</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Advanced Placement (AP)</td>
<td>3</td>
<td>Standard setting and benchmarks (.65 p. of B- in class)</td>
<td>Calculus AB, BC, English Language &amp; Comp, English Lit &amp; Comp, Statistics</td>
</tr>
<tr>
<td>COMPASS</td>
<td>77 English 52 Math</td>
<td></td>
<td>ACT (2010)</td>
</tr>
<tr>
<td>PLAN</td>
<td>15 English 19 Math</td>
<td></td>
<td>ACT (2010)</td>
</tr>
<tr>
<td>PSAT Total</td>
<td>145</td>
<td></td>
<td>Proctor, Wyatt, &amp; Wiley (2010). These benchmarks are for grade 10 students; grade 11 are higher.</td>
</tr>
<tr>
<td>PSAT Reading</td>
<td>49</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSAT Math</td>
<td>47</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PSAT Writing</td>
<td>48</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Validity evidence based on relations to other variables

- **Concurrent:** Students take PSU and external assessments (or courses) around same time

- **Predictive:** Students’ university GPA or other criteria gathered later (retrospective analysis)

- **Linking studies:** PSU items embedded in external assessments and/or vice-versa

- **Projection:** Map cut-score from external assessment onto PSU test scale using population and sampling assumptions
Issues in Validating Admissions Tests Using External Criteria

- Defining “Success” in University
- Finding relevant external criteria
- Validating external criteria
- Deciding on research design(s)
- Defining probability of success criterion
Defining Success in University

- First-year grades (GPA)?
  - University of Chile GPA = Catholic University of Chile GPA?
  - Pre-med GPA = Psychology GPA?
How Should we Define “Success” in College

- First-year GPA?
- GPA in specific courses?
- Course completion?
- Number of credits?
- Graduation?
- Persistence?
Standards’ 5 Sources of Validity evidence:

5. Validity Evidence based on testing consequences
   - The AERA et al. *Standards* stress the importance of evaluating consequences, but do not do a good job of defining this source of evidence.
   - However, evaluating consequences is the most important aspect of test evaluation because testing has consequences.
Testing has consequences

- Intended consequences
  - Purpose of test
  - Intended positive consequences

- Unintended consequences
  - Negative
  - Positive
Evaluating consequences of admissions tests

Positive

Do admissions tests promote access to university?

Do admissions tests promote success in university?

Do admissions tests improve instruction so students are better prepared for university?
Evaluating consequences of admissions tests

- Negative

Do admissions tests *prevent* students from reaching their potential?

Do admissions tests *discourage* students from applying to university?

Do admissions tests *dilute* secondary school instruction?
So, how *should* we validate university admissions tests?

- Use the AERA et al. (2014) *Standards* as a validation framework.
The *Standards* as a validation framework:

- Provide a system for categorizing validity evidence so a coherent argument can be developed.
- Provide a way of standardizing the reporting of validity evidence.
- Focus on both test construction and test score validation.
- Emphasize the importance of evaluating consequences.
Sireci (2012, 2013)

- Validation can be viewed as a 5-step process.
Validation Steps (1)

1. Identify testing purposes
   - Should not be hard to do—they are explicitly stated in technical manuals and official documents/web sites of testing agencies!

2. Identify potential test misuse

3. Prioritize validity questions based on explicit purposes and potential misuse
Validation Steps (2)

4. Determine sources of evidence needed to answer each question
5. Cross validity questions with sources of evidence
The following slide shows how we applied this framework:

- **To the “Massachusetts Adult Proficiency Tests”**
  - Reading and Math tests for adult education students in Massachusetts
  - Designed to measure students' mastery of curriculum frameworks
  - And to measure students' educational gains for Federal and State accountability
<table>
<thead>
<tr>
<th>Validity Question</th>
<th>Content</th>
<th>Structure</th>
<th>Ext. Variables</th>
<th>Processes</th>
<th>Testing</th>
<th>Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does the MAPT measure the correct skills?</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are the tests congruent with the curriculum frameworks?</td>
<td></td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are the scores accurate?</td>
<td></td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Do they adequately measure progress?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do they meet Federal requirements?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Are they useful for program evaluation?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Inappropriate diagnostic use?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Inappropriate placement?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
</tr>
<tr>
<td>Positive effect on instruction?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
</tr>
</tbody>
</table>

Test Purpose

Check marks indicate where evidence is needed.
<table>
<thead>
<tr>
<th>Validity Question</th>
<th>Content</th>
<th>Structure</th>
<th>Variables</th>
<th>Processes</th>
<th>Testing</th>
<th>Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Does the MAPT measure the correct skills?</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are the tests congruent with the curriculum frameworks?</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are the scores accurate?</td>
<td></td>
<td>√</td>
<td>√</td>
<td>√</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do they adequately measure progress?</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Do they meet Federal requirements?</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Are they useful for program evaluation?</td>
<td>√</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inappropriate diagnostic use?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Inappropriate placement?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Positive effect on instruction?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>√</td>
<td></td>
</tr>
<tr>
<td>Purpose/Validity Question</td>
<td>Source of Validity Evidence</td>
<td>Content</td>
<td>Internal Structure</td>
<td>Relations w/ Ext. Variables</td>
<td>Response Processes</td>
<td>Testing Consequences</td>
</tr>
<tr>
<td>----------------------------------------------</td>
<td>-----------------------------</td>
<td>---------</td>
<td>-------------------</td>
<td>-----------------------------</td>
<td>--------------------</td>
<td>----------------------</td>
</tr>
<tr>
<td>Measure correct skills?</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Congruent w/ frameworks?</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Accurate?</td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Measure progress?</td>
<td></td>
<td>✓</td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Meet Federal requirements?</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Useful for program Evaluation?</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inappropriate diagnostic use?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Inappropriate placement?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Effect on instruction?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
By reviewing validity evidence for the MAPT, we can see

- No validity evidence based on testing consequences
- No validity evidence to evaluate potential negative effects or positive effect on instruction
- Are we proud of this?
- No, but we know what our next steps are.
- Point is not to evaluate MAPT, but to demonstrate validation approach
Can we apply this approach to university admissions tests like the PSU?

Por supuesto!
<table>
<thead>
<tr>
<th>Purpose/Validity Question</th>
<th>Content</th>
<th>Internal Structure</th>
<th>Relations w/ Ext. Variables</th>
<th>Response Processes</th>
<th>Testing Consequences</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure escuela curriculum?</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Measure university skills?</td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
<td></td>
</tr>
<tr>
<td>Make admissions decisions?</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Provide scholarships?</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Improve instruction?</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Dilute instruction?</td>
<td></td>
<td>✓</td>
<td></td>
<td></td>
<td>✓</td>
</tr>
<tr>
<td>Promote “dropout?”</td>
<td></td>
<td></td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
<tr>
<td>Increase inequities?</td>
<td></td>
<td>✓</td>
<td></td>
<td>✓</td>
<td>✓</td>
</tr>
</tbody>
</table>
Discussion

- We know there are limitations to any university admissions tests
- However, we also know there are some fundamental requirements that should be in place for admissions tests to be defensible.
Discussion (2)

- The short story is we need a predominance of evidence to support the use of a test for each specific purpose
  - Intended purposes should be clearly articulated
  - Theory underlying test development, and empirical evidence, should support test use.
21\textsuperscript{st} Century Validation

1. Focuses on test USE.
2. Requires evidence test measures what it “purports” to measure
3. Requires evidence of test “utility.”
4. Requires evidence test is doing more good than harm.
3 Minimum Requirements for Valid Admissions Testing Programs

1. Validity evidence based on test content:
   - content of assessments should reflect academic aspects of university success

2. Validity evidence based on relations to other variables
   - Students’ test scores should be positively related to other measures of academic achievement
Requirements for Valid Admissions Programs (cont.)

3. Validity evidence based on testing consequences
   - Evidence that the use of admissions test scores are having intended effects
   - And are not presenting a barrier to students who may otherwise be successful in university
Conclusions

- By using the validation frameworks provided by the AERA et al. Standards, we can gather, analyze, and report the evidence we need to defend the validity of PSU and other admissions tests (if warranted!).

- By developing a research agenda around interpreting and reporting admissions test scores, we can avoid negative consequences.
Thanks to Directora Varas y DEMRE for the invitation!

And to you for your attention.

Questions or Comments
Sireci@acad.umass.edu