Validity

Definitions

- Validity is the extent to which the inferences made from test scores are accurate
  - Variation in the underlying construct causes variation in the measurement process
  - Establishing causality in measurement is no different than establishing causality for any research question
  - Must show evidence to support the inference
  - Must rule out alternative explanations

Examples

- People with higher levels of depression will score higher on the Beck depression inventory
- A person with this score on the LSAT will do well in law school.
- A person with these scores on the SVIB will be happy as an engineer.
- A person with these scores on the REID report will likely steal from an employer.

General Issues

- Validity statements should address particular interpretations or types of decisions
  - Ex: a test for general intelligence (Wonderlic) may discriminate well for the general population but not very well for college grads
- Validation is a process of hypothesis testing:
  - Someone who scores high on this measure will also do well in situation A
- In validity assessment the aim is inferential
  - Ex: a person who does well on rheumatology exam can be expected to know more about rheumatic disease, or manage patient with rheumatic disease appropriately.

Sources of Validity Evidence

- In the beginning, there were types of validity (e.g., content, criterion-related, construct)
- Now, All validity evidence is construct validity evidence
- Sources of Validity Evidence
  - Content Evidence
  - Criterion-Related Evidence
  - Discriminant Groups Evidence
  - Multi- Trait Multi- Method Correlations
  - Face Validity
  - Consequential Evidence?
  - Internal Structure Evidence

Definitions

- Validation: the process of gathering and evaluating information to support the desired inference
- You don't validate a test!
- Instead, you validate inferences or decisions based on test scores
- Validation determines the degree of confidence that decision makers can place in inferences we make about people based on their test scores
Content Validity

- Degree to which test taps into domain or “content” of what it is supposed to measure

Content Validity

- Content validity is judgment concerning how adequately a test samples behavior representative of the universe of behavior the test was designed to sample

- Content validity draw an inference from test scores to a large domain of items similar to those on the test

Content validity as representativeness

- Content validity is concerned with sample-population measurement representation
  
  - Ex: The knowledge, skills, abilities, and personal characteristics (KSPAs) covered by the test items should be representative to the entire domain of KSPAs
  
- A test that includes a more representative sample of the target behavior lends itself to more accurate inferences
  
  - that is, inferences which hold true under a wider range of circumstances
  
  - If important aspects of the outcome are missed by the scales, then some inferences which will prove to be wrong, then inferences that the test are invalid

Content experts

- Content validity is usually established by content or subject matter experts (SMEs).

- In content validity evidence is obtained by looking for agreement in judgments by experts panel

Quantification of Content validity? (Lawshe)

- Each member of panel of experts responds to the question “is the skill or knowledge measured by this item…
  
  - Essential versus
  
  - Useful but not essential versus
  
  - Not necessary
  
  - …to the behavioral domain?”

Quantification of Content validity? (Lawshe)

- For each item, the number of panelists stating that item is essential is noted. If more than half the panelists indicate that an item is essential, that item has at least some content validity

- Greater levels of content validity exist as larger numbers of panelists agree that a particular item is essential

- Drawbacks:
  
  - Experts tend to take their knowledge for granted and forget how little other people know. Some tests written by content experts are extremely difficult.
  
  - Content experts often fail to identify the learning objectives of a subject.
**Content Relevance & Coverage**

- Messick (1980)
- **Content relevance**: each item on the test should relate to one dimension of the domain.
- **Content coverage**: each domain dimension should be represented by one or more item

---

**Specification Chart**

<table>
<thead>
<tr>
<th>Question</th>
<th>Content Area</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Physiology</td>
</tr>
<tr>
<td>1</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td></td>
</tr>
<tr>
<td>4</td>
<td></td>
</tr>
<tr>
<td>5</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td></td>
</tr>
<tr>
<td>7</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td></td>
</tr>
<tr>
<td>10</td>
<td></td>
</tr>
</tbody>
</table>

---

**Criterion-Related Validity**

- A judgment regarding how adequately a test score can be used to infer an individual's most probable standing on a criterion (e.g., performance) of interest.
- Indexed by the correlation between scores on a measure of the construct and a measure of the criterion of interest.
  - Validity Coefficient
- Correlation is estimated in one of two ways:
  - Concurrent validity estimate
  - Predictive validity estimate

---

**Concurrent Validity**

- Concurrent validity refers to the form of criterion-related validity that is an index of the degree to which a test score is related to some criterion measure obtained at the same time.
- Statements of concurrent validity indicate the extent to which test scores may be used to estimate an individual’s present standing on a criterion.
  - Must involve current employees, which results in range restriction & non-representative sample.
  - Current employees will not be as motivated to do well on the test as job seekers.

---

**Predictive Validity**

- Predictive validity refers to the form of criterion-related validity that is an index of the degree to which a test score predicts some criterion measure obtained at a future time.
  - Example: clerkship scores of a medical student as predictor of physician performance after graduation as criterion.
- Drawbacks:
  - Will have range restriction unless all applicants are hired.
  - Must wait several months for job performance (criterion) data.

---

**Criterion Related Validity**

- Must have a criterion measure to use this form of validity.
- If a good criterion measure already exists, why use another test?
- Because in many situations the criterion measurement
  - Is impractical
  - Is expensive
  - Is time consuming
  - Associated with delayed outcome.
**Expectancy Table**

- If you have a validity coefficient, you can form a chart to communicate the expected performance gains associated with basing decisions on the predictor.
- Expectancy tables illustrate the likelihood that the test taker will score within some interval of scores on a criterion measure:
  - Show the % of people within specified test-score intervals who were placed in various categories of the criterion.

**Taylor-Russell Table**

**Cronbach & Gleser Decision Theory**

- A classification of decision problems, various selection strategies ranging from single stage processes to sequential analysis;
- A quantitative analysis of the relationship between test utility, the selection ratio, cost of the testing program, and expected value of the outcome;
- A recommendation that in some instances job requirements be tailored to the applicant’s ability instead of the other way around.

**Decision Theory Terminology**

- **Base rate**: the extent to which a particular characteristic or attribute exist in the population.
- **Hit rate**: the proportion of people a test accurately identifies as possessing a particular characteristic or attribute.
- **Miss rate**: the proportion of people the test fails to identify as having-or not having-a particular characteristic or attribute.
- **False positive**: a miss wherein the test predicted that the test taker did possess the particular characteristic or attribute being measured.
- **False negative**: a miss wherein the test predicted that the test taker did not possess the particular characteristic or attribute being measured.

**Decision Theory Graph**

Patients with disease are shown in the upper distribution. Patients without disease are shown in the lower distribution.
Construct Validity Evidence

- Concerned with the theoretical relationships among constructs
- And
- The corresponding observed relationships among measures

Decision Theory

Constructs are interrelated

- concepts are not mutually exclusive
- they exist in a web of overlapping meaning
- to enhance construct validity, we must show where the construct is in its broader network of meaning
Could show that...
the construct is slightly related to the other four...

other construct: A
other construct: B
other construct: C
other construct: D

Could show that...
...and, constructs A & C and constructs B & D are related to each other...

other construct: A
other construct: B
other construct: C
other construct: D

Could show that...
...and, constructs A & C are not related to constructs B & D

other construct: A
other construct: B
other construct: C
other construct: D

Example: Want to Measure...

self esteem

Example: Distinguish From...

self worth
self disclosure
self esteem
confidence
openness

To Establish Construct Validity

• Must set the construct within a semantic (meaning) net
• Evidence that you control the operationalization of the construct (that your theory has some correspondence with reality)
• Must provide evidence that your data support the theoretical structure
Convergent and Discriminant Validity

**The Convergent Principle**

Measures of constructs that are related to each other should be strongly correlated.

**How it works**

**Theory**

1. You theorize that the items all reflect self-esteem.
2. The items are self-esteem construct.
3. The correlations provide evidence that the items all converge on the same construct.

**Observation**

<table>
<thead>
<tr>
<th>Item 1</th>
<th>Item 2</th>
<th>Item 3</th>
<th>Item 4</th>
</tr>
</thead>
<tbody>
<tr>
<td>.83</td>
<td>.89</td>
<td>.81</td>
<td>.89</td>
</tr>
<tr>
<td>.81</td>
<td>1.00</td>
<td>.85</td>
<td>.84</td>
</tr>
<tr>
<td>.89</td>
<td>.85</td>
<td>1.00</td>
<td>.86</td>
</tr>
<tr>
<td>.81</td>
<td>.84</td>
<td>.86</td>
<td>1.00</td>
</tr>
</tbody>
</table>

**The Discriminant Principle**

Measures of different constructs should *not* correlate highly with each other.

**How it works**

**Theory**

1. You theorize that the items pertain to different constructs.
2. The items are different constructs.
3. The correlations provide evidence that the items do not discriminate.

**Observation**

<table>
<thead>
<tr>
<th>PPSS1</th>
<th>FK1</th>
<th>PPSS2</th>
<th>FK2</th>
</tr>
</thead>
<tbody>
<tr>
<td>.12</td>
<td>.09</td>
<td>.04</td>
<td>.11</td>
</tr>
</tbody>
</table>

**Putting It All Together**

Convergent and Discriminant Validity
We have two constructs we want to measure, problem solving and factual knowledge. For each construct we develop three scale items; our theory is that items within construct will converge, across constructs will discriminate.

The correlations support both convergence and discrimination and therefore, construct validity.

The Nomological Network

- What is it?
- nomological is derived from Greek and means "lawful"
- links interrelated theoretical ideas with empirical evidence
What is the Nomological Net?

**Theoretical Level:** Concepts, Ideas

**Observed Level:** Measures, Programs

Principles

- Scientifically, to make clear what something is means to set forth the laws in which it occurs.
- This interlocking system of laws is the Nomological Network.

The laws in a nomological network may relate...

- Observable properties or quantities to each other
- Different theoretical constructs to each other
- Theoretical constructs to observables

"Learning more about" a theoretical construct is a matter of elaborating the nomological network in which it occurs...

...or of increasing the definiteness of its components.
The Main Problem with the Nomological Net

...it doesn't tell us how we can assess the construct validity in a study.

The Multitrait-Multimethod Matrix

What is the MTMM Matrix?

- A matrix (table) of correlations arranged to facilitate the assessment of construct validity
- Integrates both convergent and discriminant validity

What is the MTMM Matrix?

- Assumes that you measure each of several concepts (trait) by more than one method
- Very restrictive - ideally you should measure each concept by each method
- Arrange the correlation matrix by concepts within methods

Principles

- Convergence: Things which should be related are
- Divergence/Discrimination: Things which shouldn't be related aren't

A Hypothetical MTMM Matrix

<table>
<thead>
<tr>
<th>Traits</th>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A</td>
<td>A1</td>
<td>A2</td>
<td>A3</td>
</tr>
<tr>
<td>B</td>
<td>B1</td>
<td>B2</td>
<td>B3</td>
</tr>
<tr>
<td>C</td>
<td>C1</td>
<td>C2</td>
<td>C3</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method 1</th>
<th>B1</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>.67</td>
<td>.58</td>
<td>.11</td>
<td>.67</td>
</tr>
<tr>
<td>A2</td>
<td>.57</td>
<td>.77</td>
<td>.13</td>
<td>.51</td>
</tr>
<tr>
<td>A3</td>
<td>.72</td>
<td>.75</td>
<td>.15</td>
<td>.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method 2</th>
<th>B1</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>.67</td>
<td>.58</td>
<td>.11</td>
<td>.67</td>
</tr>
<tr>
<td>A2</td>
<td>.57</td>
<td>.77</td>
<td>.13</td>
<td>.51</td>
</tr>
<tr>
<td>A3</td>
<td>.72</td>
<td>.75</td>
<td>.15</td>
<td>.51</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Method 3</th>
<th>B1</th>
<th>C1</th>
<th>C2</th>
<th>C3</th>
</tr>
</thead>
<tbody>
<tr>
<td>A1</td>
<td>.67</td>
<td>.58</td>
<td>.11</td>
<td>.67</td>
</tr>
<tr>
<td>A2</td>
<td>.57</td>
<td>.77</td>
<td>.13</td>
<td>.51</td>
</tr>
<tr>
<td>A3</td>
<td>.72</td>
<td>.75</td>
<td>.15</td>
<td>.51</td>
</tr>
</tbody>
</table>
### Interpreting the MTMM Matrix

<table>
<thead>
<tr>
<th>Method 1</th>
<th>Method 2</th>
<th>Method 3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Traits</td>
<td>Traits</td>
<td>Traits</td>
</tr>
<tr>
<td>A1</td>
<td>B1</td>
<td>C1</td>
</tr>
<tr>
<td>A2</td>
<td>B2</td>
<td>C2</td>
</tr>
<tr>
<td>A3</td>
<td>B3</td>
<td>C3</td>
</tr>
</tbody>
</table>

#### Reliability
- Should be highest coefficients

#### Convergent
- Validity diagonals should have strong r's

#### Discriminant
- A variable should have higher r with another measure of the same trait than with different traits measured by the same method

### Advantages
- Addresses convergent and discriminant validity simultaneously
- Addresses the importance of method of measurement
- Provides a rigorous standard for construct validity
Disadvantages

- hard to implement
- no known overall statistical test for validity
- requires judgment call on interpretation

Additional Representations of Validity

- Face Validity – degree to which a test appears to measure what it purports to measure; i.e., do the test items appear to represent the domain being evaluated?
  - important because lack of it could contribute to a lack of confidence with respect to perceived effectiveness of the test
- Physical Fidelity – do physical characteristics of test represent reality?
- Psychological Fidelity – do psychological demands of test reflect real-life situation?

Threats to Construct Validity

Inadequate Preoperational Explication of Constructs

- preoperational = before translating constructs into measures or treatments
- in other words, you didn't do a good enough job of defining (operationally) what you mean by the construct

Mono-operation Bias

- pertains to the treatment or program
- used only one version of the treatment or program

Mono-method Bias

- pertains especially to the measures or outcomes
- only operationalized measures in one way
- for instance, only used paper-and-pencil tests
**Restricted Generalizability Across Constructs**

- you didn't measure your outcomes completely
- or, you didn't measure some key affected constructs at all (i.e., unintended effects)

**Consequential Validity**

- Messick (1995)
- the aspect of construct validity that appraises the value implications of score interpretation as a basis for action as well as the actual and potential consequences of test use, especially in regard to sources of invalidity related to issues of bias, fairness, and distributive justice

**Discriminant Groups Evidence**

- A measure of a construct should be sensitive to differences between groups known to differ on the construct
  - Alcoholism
  - Gender attitudes
  - MMPI

**Internal Structure Evidence**

- Factor Analysis! :)