Is the test valid?"

Jum Nunnally (one of the founders of modern psychometrics) claimed this was "silly question"! The point wasn’t that tests shouldn’t be “valid” but that a test’s validity must be assessed relative to...

- the specific construct(s) it is intended to measure
- the population for which it is intended (e.g., age, level)
- the application for which it is intended (e.g., for classifying folks into categories vs. assigning them quantitative values)

So, the real question is, “Is this test a valid measure of this construct for this population for this application?” That question can be answered!

Criterion-related Validity – 3 classic types

- does test correlate with “criterion”? -- has three major types
- predictive -- test taken now predicts criterion assessed later
  - most common type of criterion-related validity
  - e.g., your GRE score (taken now) predicts how well you will do in grad school (criterion -- can’t be assessed until later)
- concurrent -- test “replaces” another assessment (now)
  - often the goal is to substitute a “shorter” or “cheaper” test
  - e.g., the written drivers test is a replacement for driving around with an observer until you show you know the rules
- postdictive -- least common type of criterion-related validity
  - can I test you now and get a valid score for something that happened “earlier” -- e.g., adult memories of childhood feelings

The advantage of criterion-related validity is that it is a relatively simple statistically based type of validity!

- If the test has the desired correlation with the criterion, then you have sufficient evidence for criterion-related validity.

There are, however, some limitations to criterion-related validity...

- It is dependent upon your having a criterion
  - Sometimes you don’t have a criterion variable to use -- e.g., first test of construct that is developed
- It is dependent upon the quality of the criterion variable
  - Sometimes there are “limited” or “competing” criteria
- “Correlation is not equivalence”
  - your test that is correlated with the criterion might also be correlated with several other variables -- what does it “measure”?

<table>
<thead>
<tr>
<th>Predictive Validity</th>
<th>Concurrent Validity</th>
<th>Postdictive Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>Measure the Criterion</td>
<td>Administer test</td>
<td>Measure the Criterion</td>
</tr>
<tr>
<td>Administer test</td>
<td>Measure the Criterion</td>
<td>Administer test</td>
</tr>
</tbody>
</table>

Past Present Future
Conducting a Predictive Validity Study

example -- test designed to identify qualified “front desk personnel” for a major hotel chain
-- 200 applicants - and 20 position openings

Conducting the “proper study”
- give each applicant the test (and “seal” the results)
- give each applicants a job working at a front desk
- assess work performance after 6 months (the criterion)
- correlate the test (predictor) and work performance (criterion)

Anybody see why the chain might not be willing to apply this design?
Here are two designs often substituted for this proper design.

Substituting concurrent validity for predictive validity

- assess work performance of all folks currently doing the job
- give them each the test
- correlate the test (predictor) and work performance (criterion)

Problems?
- Not working with the population of interest (applicants)
- Range restriction -- work performance and test score variability are “restricted” by this approach
  - current hiring practice probably not “random”
  - good workers “move up” -- poor ones “move out”
- Range restriction will artificially lower the validity coefficient (r)

What happens to the sample ...

Applicant pool -- target population
Selected (hired) folks
  • assuming selection basis is somewhat reasonable/functional
Sample used in concurrent validity study
  • worst of those hired have been “released”
  • best of those hired have “changed jobs”

What happens to the validity coefficient -- r

Applicant pool
\[ r = .75 \]
Hired Folks
Sample used in validity study
\[ r = .20 \]
Using and testing predictive validity simultaneously

• give each applicant the test
• give those applicants who “score well” a front desk job
• assess work performance after 6 months (the criterion)
• correlate the test (predictor) and work performance (criterion)

Problems?
• Not working with the population of interest (all applicants)
• Range restriction -- work performance and test score variability are “restricted” by this approach
• only hired good those with “better” scores on the test
• (probably) hired those with “better” work performance
• Range restriction will artificially lower the validity coefficient (r)
• Using a test before its validated can have “legal ramifications”

Thinking about the procedures used to assess criterion related validity …

All the types of criterion related validity involved correlating the new measure/instrument with some selected criterion
• large correlations indicate criterion related validity (.5-.7)
• smaller correlations are interpreted to indicate the limited validity of the instrument

(As mentioned before) This approach assumes you have a criterion that really is a “gold standard” of what you want to measure.
• Even when such a measure exists it will itself probably have limited validity and reliability
• We will consider each of these and how they limit the conclusions we can draw about the criterion related validity of our instrument from correlational analyses

Let’s consider the impact of limited validity of the criterion upon the assessment of the criterion related validity of the new instrument/measure

• let’s assume we have a “perfect” measure of the construct
• if the criterion we plan to use to validate our new measure is “really good” it might itself have a validity as high as, say .8 -- shares 64% of its variability with perfect measure
• here are two hypothetical “new measures” - which is more valid?
  • Measure 1 -- r with criterion = .70 (49% overlap)
  • Measure 2 -- r with criterion = .50 (25% overlap)

Measure 1 has the higher validity coefficient, but the weaker relationship with the perfect measure
Measure 2 has the stronger relationship with the perfect measure, but looks bad because of the choice of criterion

So, the meaningfulness of a validity coefficient is dependent upon the “quality” of the criterion used for assessment

Best case scenario …
• criterion is objective measure of the specific behavior of interest
• when the measure IS the behavior we are interested in, not some “representation”
  • e.g., graduate school GPA, hourly sales, # publications

Tougher situation …
• objective measure of behavior “represents” construct of interest, but isn’t the specific behavior of interest
  • e.g., preparation for the professorate, sales skill, contribution to the department
  • notice each of the measures above is an incomplete representation of the construct listed here

Horror show …
• subjective (potentially biased) rating of behavior or performance
  • advisor’s eval, floor manager’s eval, Chair’s evaluations
Local Validity
Explicit check on validity of the test for your population and application.

Sounds good, but likely to have the following problems
• Sample size will be small (limited to your “subject pool”)
• Study will likely be run by “semi-pros”
• Optimal designs probably won’t be used (e.g., predictive validity)
• Often (not always) this is an attempt to “bend” the use of an established test to a population/application for which it was not designed nor previously validated

Other kinds of criterion-related validity...

Incremental Validity
Asks if the test “improves on” the criterion-related validity of whatever tests are currently being used.

Example. I claim that scores from my new structured interview will lead to more accurate selection of graduate students. I’m not suggesting you stop using what you are using, but rather that you ADD my interview.

Demonstrating Incremental Validity requires we show that the “new test + old tests” do better than “old tests alone”. R² Δ test...

R² grad. grea, grev, greq = .45
R² grad. Grea, grev, greq, interview = .62
Incremental validity is .17 (or 38% increase)

Experimental Validity
A study designed to show that the test “reacts” as it should to a specific treatment.

In the “usual experiment,” we have confidence that the DV measures the construct in which we are interested, and we are testing if the IV is related to that DV (that we trust).

In Experimental Validity, we have confidence in the IV (treatment) and want to know if the DV (the test being validated) will respond “as it should” to this treatment.

Example: I have this new index of social anxiety. I know that a particular cognitive-behavioral treatment has a long, successful history of treating social anxiety. My experimental validity study involves pre- and post-testing 50 participants who receive this treatment -- experimental criterion-related validity would be demonstrated by a pre-post score difference (in the right direction)

Now let’s consider the relationship between reliability & validity...

“reliability is a precursor for validity”
• conceptually -- how can a measure be consistently accurate (valid), unless it is consistent ??
• internal consistency -- all items reflect the same construct
• test-retest consistency -- scale yields repeatable scores
• statistically -- limited reliability means that some of the variability in the measure is systematic, but part is unsystematic (unreliable)
• low reliability will “attenuate” the validity correlation
• much like range restriction -- but this is a restriction of the “systematic variance”, not the overall variance

It is possible to statistically “correct” for this attenuation
-- like all “statistical correction”, this must be carefully applied!
Various “correction for attenuation” formulas...

Note:  \( y = \text{criterion} \ \ x = \text{measure being assessed} \)

- estimates what would be the validity coefficient if both the criterion and the measure were perfectly reliable \((\alpha = 1.00)\)
- estimates what would be the validity if the criterion were perfectly reliable
- a more useful formula estimates the validity coefficient if each measure’s reliability “improved” to a specific value

\[
\frac{\hat{r}_{YX}}{\sqrt{\alpha_Y \cdot \sqrt{\alpha_X}}} = \frac{\hat{r}_{YX}}{\sqrt{\alpha_Y}} \cdot \sqrt{\alpha_X}
\]

You can also use this last formula to modify the reliability of just one of the involved measures.

Example...

You have constructed an interview which is designed to predict employee performance
- scores on this interview \((X)\) correlate \(0.40\) with supervisor’s ratings \((Y)\)
- the interview has an \(a_Y = 0.50\)
- the supervisor rating scale (the criterion) has \(a_Y = 0.70\)

Correcting both the interview and criterion to perfect reliability...

\[
\frac{\hat{r}_{YX}}{\sqrt{\alpha_Y \cdot \sqrt{\alpha_X}}} = \frac{0.40}{\sqrt{0.50} \cdot \sqrt{0.70}} = 0.68
\]

Correcting just the interview to perfect reliability ...

\[
\frac{\hat{r}_{YX}}{\sqrt{\alpha_Y}} = \frac{0.40}{\sqrt{0.70}} = 0.48
\]

Correcting the interview to \(a\approx 7\) to and criterion to \(a=9\)...

\[
\frac{\hat{r}_{YX}}{\sqrt{\alpha_Y \cdot \sqrt{\alpha_X}}} = \frac{0.40}{\sqrt{0.90} \cdot \sqrt{0.70}} = 0.53
\]

So, What’s our best estimate of the “true” criterion-related validity of our instrument -- \(0.40\) ?? \(0.48\) ?? \(0.53\) ?? \(0.68\) ??

Hmmm... One must use these “correction” formulas with caution!

“Poorer uses...”
- using only the corrected values to evaluate the measure’s validity (remember, “best case” seldom represents “best guess”)

Good uses...
- ask how the validity would be expected to change if the reliability of the new measure were increased to a certain value, as a prelude to working to increase the reliability of the new measures to that reliability (adding more good items)
- ask how the validity would be expected to change if the reliability of the criterion were increased to a certain value, as a prelude to finding a criterion with this increased reliability

Another good use -- a sort of reality check...
- you can modify the formula to ask “to what reliability must I improve my measure to reach the desired validity?”
  \[
  \alpha_X' = \frac{\hat{r}_{YX}}{\sqrt{\alpha_Y \cdot \sqrt{\alpha_X}}}^2
  \]
  \[
  k = \frac{\alpha_X' \cdot (1 - \alpha_X')}{\alpha_X' \cdot (1 - \alpha_X')}
  \]
- then check how many items you’ll need to add to your measure to reach this reliability
- now you have an estimate of by how much you must increase the number of items, to increase the reliability of the measure, to increase the validity of the measure, but...
- sometimes it isn’t possible to raise the reliability “enough” (max \(\alpha = 1.00\))
- need that many “as good” items
Face Validity
• Does the test “look like” a measure of the construct of interest?
  • “looks like” a measure of the desired construct to a member of the target population
  • will someone recognize the type of information they are responding to?
• Possible advantage of face validity ...
  • If the respondent knows what information we are looking for, they can use that “context” to help interpret the questions and provide more useful, accurate answers
• Possible limitation of face validity ...
  • if the respondent knows what information we are looking for, they might try to “bend & shape” their answers to what they think we want -- “fake good” or “fake bad”

Content Validity
• Does the test contain items from the desired “content domain”?
  • Based on assessment by experts in that content domain
  • Is especially important when a test is designed to have low face validity
  • e.g., tests of “honesty” used for hiring decisions
  • Is generally simpler for “achievement tests” than for “psychological constructs” (or other “less concrete” ideas)
  • e.g., it is a lot easier for “math experts” to agree whether or not an item should be on an algebra test than it is for “psychological experts” to agree whether or not an item should be on a measure of depression.
  • Content validity is not “tested for”. Rather it is “assured” by the informed item selections made by experts in the domain.

Construct Validity
• Does the test interrelate with other tests as a measure of this construct should?
• We use the term construct to remind ourselves that many of the terms we use do not have an objective, concrete reality.
  • Rather they are “made up” or “constructed” by us in our attempts to organize and make sense of behavior and other psychological processes
• attention to construct validity reminds us that our defense of the constructs we create is really based on the “whole package” of how the measures of different constructs relate to each other
  • So, construct validity “begins” with content validity (are these the right types of items) and then adds the question, “does this test relate as it should to other tests of similar and different constructs?”

Discriminant Validity
• Does the test show the “right” pattern of interrelationships with other variables? -- has two parts
  • Convergent Validity -- test correlates with other measures of similar constructs
  • Divergent Validity -- test isn’t correlated with measures of “other, different constructs”
  • e.g., a new measure of depression should ...
    • have “strong” correlations with other measures of “depression”
    • have negative correlations with measures of “happiness”
    • have “substantial” correlation with measures of “anxiety”
    • have “minimal” correlations with tests of “physical health”, “faking bad”, “self-evaluation”, etc.
Evaluate this measure of depression:

<table>
<thead>
<tr>
<th>New Dep</th>
<th>Dep1</th>
<th>Dep2</th>
<th>Anx</th>
<th>Happy</th>
<th>PhyHlth</th>
<th>FakBad</th>
</tr>
</thead>
<tbody>
<tr>
<td>New Dep</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old Dep1</td>
<td>.61</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Old Dep2</td>
<td>.49</td>
<td>.76</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Anx</td>
<td>.43</td>
<td>.30</td>
<td>.28</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Happy</td>
<td>-.59</td>
<td>-.61</td>
<td>-.56</td>
<td>-.75</td>
<td></td>
<td></td>
</tr>
<tr>
<td>PhyHlth</td>
<td>.60</td>
<td>.18</td>
<td>.22</td>
<td>.45</td>
<td>-.35</td>
<td></td>
</tr>
<tr>
<td>FakBad</td>
<td>.55</td>
<td>.14</td>
<td>.26</td>
<td>.10</td>
<td>-.21</td>
<td>.31</td>
</tr>
</tbody>
</table>

Tell the elements of discriminant validity tested and the "conclusion".

Convergent validity (but bit lower than r(dep1, dep2)

Old Dep1 .61
Old Dep2 .49 .76
Anx .43 .30 .28
Happy -.59 -.61 -.56 -.75
PhyHlth .60 .18 .22 .45 -.35
FakBad .55 .14 .26 .10 -.21 .31

This pattern of results does not show strong discriminant validity!:

Summary
Based on the things we've discussed, what are the analyses we should do to "validate" a measure, what order do we do them (consider the flow chart next page) and why do we do each?

- Inter-rater reliability -- if test is not "objective"
- Chronbach's $\alpha$ -- internal reliability
- Test-Retest Analysis ($r$ & wg-t) -- temporal reliability
- Alternate Forms (if there are two forms of the test)
- Content Validity -- inspection of items for "proper domain"
- Construct Validity -- correlation and factor analyses to check on discriminant validity of the measure
- Criterion-related Validity -- predictive, concurrent and/or postdictive

What are the different types of "things we measure"??
The most commonly discussed types are...

- Achievement -- "performance" broadly defined (judgements)
  - e.g., scholastic skills, job-related skills, research DVs, etc.
- Attitude/Opinion -- "how things should be" (sentiments)
  - polls, product evaluations, etc.
- Personality -- "characterological attributes" (keyed sentiments)
  - anxiety, psychoses, assertiveness, etc.

There are other types of measures that are often used...

- IQ -- is it achievement (things learned) or is it "aptitude for academics, career and life"??