

Face, Content & Construct Validity

- Kinds of attributes we measure
- Face Validity
- Content Validity
- Construct Validity
 - Discriminant Validity → Convergent & Divergent evidence
- Summary of Reliability & Validity types and how they are demonstrated

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...

- Social Skills -- achievement or personality ??
- Aptitude -- “how well some will perform after then are trained and experiences” but measures before the training & experience”
 - some combo of achievement, personality and “likes”
- IQ -- is it achievement (things learned) or is it “aptitude for academics, career and life” ??



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?”

The statistical assessment of Construct Validity ...

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....

	New Dep	Dep1	Dep2	Anx	Happy	PhyHlth	FakBad
New Dep							
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	

Tell the elements of discriminant validity tested and the “conclusion”

Evaluate this measure of depression....

	New Dep	Dep1	Dep2	Anx	Happy	PhyHlth	FakBad
New Dep							
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	

convergent validity (but bit lower than $r(\text{dep1}, \text{dep2})$)

more correlated with anx than dep1 or dep2

corr w/ happy about same as Dep1-2

“too” r with PhyHlth

“too” r with FakBad

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”
- Item-analysis -- looking for items not “positive monotonic”
- Chronbach’s α -- 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