

# CTT Approaches to Validity

- Today's topics:
  - Levels of Measurement
  - Properties of Scales
  - CTT Framework for Validity Evidence
  - How to Mess up a Construct Validity Study
  - Wrapping Up...

# Levels of Measurement

- We may think about psychometric instruments AS IF the rules of physical measurement apply...
  - 10 feet is twice as long as 5 feet
- In reality, it may not work so neatly...
  - e.g., Likert scale of
    - 1 = strongly disagree
    - 2 = disagree
    - 3 = neither disagree nor agree
    - 4 = agree
    - 5 = strongly agree

Is 'agree' twice as much as 'disagree'?

Is the difference in the latent trait between 2 and 3 the same as between 4 and 5?

# 3 Criteria for Levels of Measurement

1. Whether greater quantities indicate an ordering
2. Whether equal intervals represent same distance
3. Whether measure has a true zero point

	1: Ordering	2: Distance	3: Zero
Ratio (Length)	Yes	Yes	Yes
Interval (Temperature)	Yes	Yes	
Ordinal (Likert items)	Yes		

# Properties of Scales

- Measurement: Process of assigning numbers of an otherwise unobserved attribute via an instrument (scale)
  - Can be bidirectional: 5 = strongly agree, the end.
  - Can be unidirectional: Different ways to get same total correct
- All instruments have possible *ceilings* and *floors* (even if not observed within a given sample)
- Simple 'total correct' fulfills a basic requirement – it assigns a number to an ability, which can then be located (relatively speaking) on the construct map

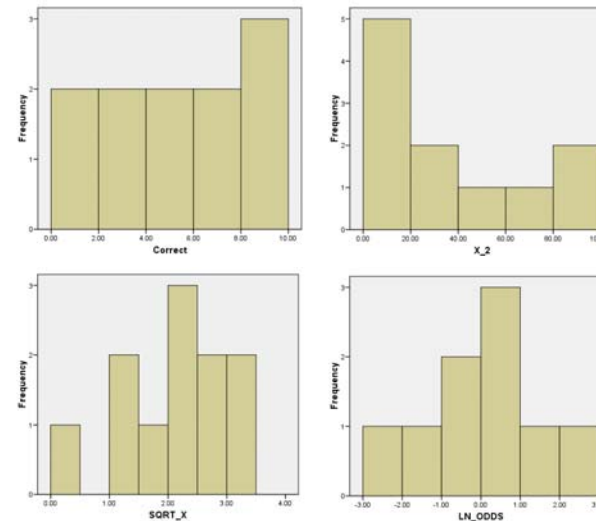
# Choices of Scale Metrics

- Metric: the choice of a set of numbers to assign to the observations that indicate performance level
- Need to choose origin and unit of measurement

TABLE 4.1  
Alternative Scales

	# Correct	X <sup>2</sup>	√X	LN Odds
Passed all items	10	100	3.16	(?)
Passed 9 items	9	81	3	2.18
Passed 8 items	8	64	2.83	1.39
Passed 7 items	7	49	2.65	0.85
Passed 6 items	6	36	2.45	0.41
Passed 5 items	5	25	2.24	0
Passed 4 items	4	16	2	-0.41
Passed 3 items	3	9	1.73	-0.85
Passed 2 items	2	4	1.14	-1.39
Passed 1 item	1	1	1	-2.18
Failed all items	0	0	0	(-?)

# Sample Distributions by Metric



*Nonlinear* transformations can be helpful for improving normality (which is assumed by many statistical models).

*Nonlinear* transformations will change the relative distances (intervals) between persons, but not their rank order.

# Other Transformations

- Standardizing via Z-scores:  $Z = \frac{X - \mu_X}{\sigma_X}$
- Z-scores might be further transformed onto a numerically specific metric:
  - General formula: NewScore = cZ + a → IQ = 15\*Z + 100
    - c = new standard deviation
    - a = new mean
  - Note that this is a LINEAR transformation – and thus the relative distances between persons will stay the same, as well as their rank ordering (i.e., will be interval still if it started out that way)
  - Also note this will not ‘fix’ ceilings or floors...
  - Nor will it make ‘ordinal’ into ‘interval’

# Scale Score Interpretation in CTT

- Criterion-referenced measurement:
  - Scores have absolute meaning relative to items/test
  - Metric is interpreted without reference to distribution
  - e.g., need 8 out of 10 correct for ‘mastery’
  - e.g., mean of > 4 out of 5 indicates ‘depression’
  - Raw scores may be most directly useful to do so
- Norm-referenced measurement:
  - Scores have only relative meaning
  - Metric is chosen based on distribution
  - Transformed scores may be most directly useful

## 2 Big Concerns about Scale Scores

- **Reliability:**
  - “Extent to which the instrument does what it is supposed to *with sufficient consistency* for its intended usage”
  - “Extent to which same results would be obtained from the instrument after repeated trials”
  - Operationalized in several ways, which we’ll get to next time...
- **Validity:**
  - “Extent to which the instrument measures *what it is supposed to* (i.e., it does what it is intended to do)” or “Validity for WHAT?”
  - Is measure of degree, and depends on USAGE or INFERENCES
    - Scales are not “valid” or “invalid” – validity is NOT a scale property
    - e.g., Test of intelligence: Measure IQ? Predict future income?

## Another Way to Think About Reliability and Validity

$$\text{Observed score} = \text{true score} + \text{error} \quad (Y = T + e)$$

- Error can be ‘random’
  - Random error can be due to many sources (internal, external, instrument-specific issues, rater issues)
  - **Random error compromises reliability**
- Error can also be ‘non-random’
  - Non-random error is due to constant source of variation that get measured consistently along with the construct (e.g., acquiescence)
  - **Non-random error compromises validity**
- In other words... reliability concerns how well you can hit the bulls-eye of the target... Validity concerns whether you hit the right target!



## More about Validity

- The process of ‘establishing’ validity should be seen as building an argument:
  - To what extent can we use this instrument for its intended purpose (i.e., as a measure of construct X in this context)?
- Validity evidence can be gathered in two main ways:
  - Internal evidence
    - From construct map – does the empirical order of the items along the construct map match your expectations of their order?
    - From ‘explanatory’ item response models... stay tuned
  - External evidence
    - Most of CTT is focused on this kind of evidence
    - This will be our focus for now...

## Historical Classification of Types of Validity

- In 1954, the American Psychological Association (APA) issued a set of standards for validity, defining 4 types
  1. Predictive Validity
  2. Concurrent Validity
  3. Content Validity
  4. Construct Validity
- Cronbach and Meehl (1955) then expanded (admittedly unofficially) on the logic of construct validity

## Predictive and Concurrent Validity

- Predictive and concurrent validity are often categorized under 'criterion-related validity' (which makes it 3 kinds)
  - Predictive validity/utility: New scale relates to future criterion
  - Concurrent validity: New scale relates to simultaneous criterion
- Criterion-related validity implies that there is some known comparison (e.g., scale, performance, behavior, group membership) that is immediately and undeniably relevant
  - e.g., Does newer, shorter test 'work as well' as older, longer test?
  - e.g., Do SAT scores predict college success?
  - This requirement limits the usefulness of this kind of validity evidence, however...

## Content Validity

- Content validity concerns how well a scale covers the plausible universe of the construct...
  - e.g., Construct: Spelling ability of 4<sup>th</sup> graders – Is the sample of words on this test representative of all the words they should know how to spell?
- 'Face validity' is sometimes mentioned in this context
  - Does the scale 'look like' it measures what it is supposed to?
- What might be some potential problems with 'establishing' these kinds of validity evidence?

## The Big One: Construct Validity

- Extent to which scale can be interpreted as a measure of the latent construct (and for that context, too)
  - Involved whenever construct is not easily operationally defined...
  - Required whenever a ready comparison criterion is lacking...
- Depends on having a 'theoretical framework' from which to derive expectations...
  - The more elaborate the theoretical framework around your construct, the pickier you need to be...

## Construct Validity: 3 Steps for Inference

- 1. Predict** relationships with related constructs
  - Convergent validity
    - Shows expected relationship (+/-) with other related constructs
    - Indicates "what it IS" (i.e., similar to, the opposite of...)
  - Divergent validity
    - Shows expected lack of relationship (0) with other constructs
    - Indicates "what it is NOT" (unrelated to...)
- 2. Find** those relationships in your sample
  - No small task...
- 3. Explain** why finding that relationship means you have shown something useful
  - Must argue based on 'theoretical framework'

## 3 Ways to Mess Up a Construct Validity Study...

So your evidence didn't turn out like you thought it would?

1. Do you have a crappy measure or a crappy inferential test of the validity scales?
  - Validity issues → Not really a measure of that construct
  - Poor reliability (perhaps specifically for different population)
  - Lack of statistical power or improper statistical analysis
    - Watch out for discrepant EFA-based studies...
2. Wrong 'theoretical framework'
  - Relationships really wouldn't be there in a perfect world

Psyc 948 Class 1c  
17 of 20

## 3 Ways to Mess Up a Construct Validity Study...

- If the framework is ok and the measures of the related constructs and statistical tests thereof are acceptable...
  - ... then, it's not them; it's you
    - Does your measure have problems with reliability?
      - Reliability precedes validity (or at least examination of it does)
    - Did you do your homework before starting out?
      - Pilot testing/feedback from relevant people
      - Clear (as possible, anyway) definition of construct and of the construct map (i.e., an ordering of persons and items)

Psyc 948 Class 1c  
18 of 20

## The 3<sup>rd</sup> Way to Mess Up a Construct Validity Study...

3. Did you fool yourself into thinking that once the study (or studies) are over, that your scale 'has validity'?
  - Can you accept that the development and evaluation of your new measure may never really be "finished"?
  - Are the items still temporally or culturally relevant?
  - It is being used in the way that's intended, and is it working like it was supposed to in those cases?
  - Has the theory of your construct evolved, such that you need to reconsider the dimensionality of your construct?
  - Do the response anchors still apply?
  - Can you make it shorter or adaptive to improve efficiency?

Psyc 948 Class 1c  
19 of 20

## Wrapping Up...

- The numerical properties of your scale will be important:
  - For choosing which kinds of statistics will be appropriate
  - For what kind of inferences can be made
- Reliability is a precursor to validity...
  - ... and we'll get there next week
- CTT approaches to validity are largely external...
  - Depend on relationships with other constructs, which can be found or not for many reasons besides validity
  - This kind of externally-oriented validity is called 'nomological net'
  - We'll get to an alternative, more internal approach later...

Psyc 948 Class 1c  
20 of 20