Validity: Estimating and Evaluating Convergent and Discriminant Validity Evidence

I. Estimating and Evaluating Convergent Validity

II. Validity generalization

III. Discriminant validity

IV. Multitrait multimethod matrices

V. Predictive validity

VI. Validity Miscellaneous

I. Estimating and Evaluating Convergent Validity

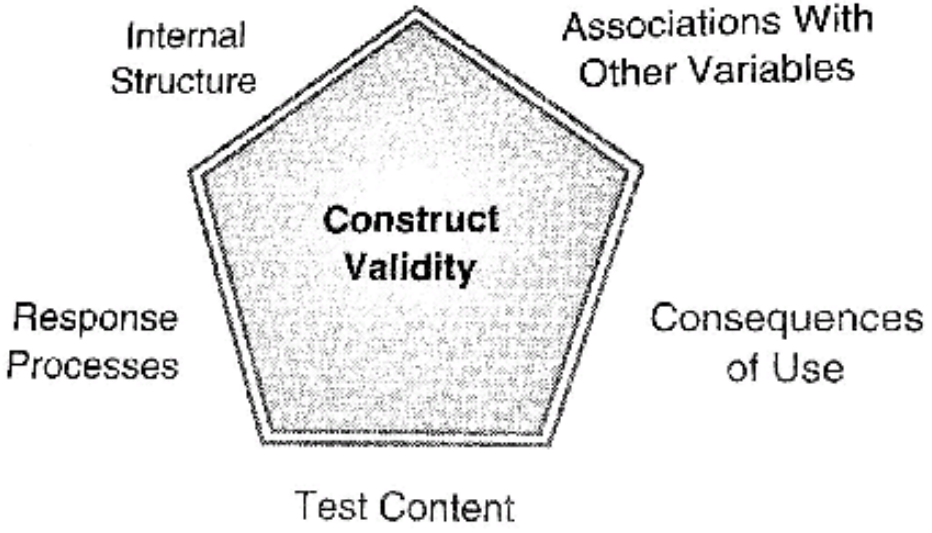


Figure 8.1 A Contemporary Perspective of Types of Information Relevant to Test Validity

I. Estimating and Evaluating Convergent Validity

Validity coefficients

- Common validation strategy is to estimate correlation of the new measure (*test*) with similar measures (*criterion* measures)
- Provides information about convergent validity
- Choose criterion measures based on theory
  - Example: a new measure of self-esteem should be expected to be related to the Rosenberg self-esteem scale (high correlation expected)
  - Example: self-esteem may be related physical self-image but should not be considered the same thing (smaller correlation expected)
- Validity of criterion measures also needs to be considered

I. Estimating and Evaluating Convergent Validity

Validity coefficients

- Correlation (Pearson r) with the other measure is the *validity coefficient*
- Squaring correlation (r^2) gives the proportion shared variance (*coefficient of determination*)
- No conventional cutoff for acceptable validity coefficient
- Equivalent correlation coefficients (point-biserial, if criterion is binary, or phi, if both are binary) or regression coefficients (slopes) may also be reported by authors

I. Estimating and Evaluating Convergent Validity

Factors Affecting Validity Coefficients

- True relationships
- Measurement error attenuates correlations
- Restricted range (including floor or ceiling type effects)
- Outliers
- Method variance (different methods, such as self-report and observation will tend to be more weakly related)
- Time
- Single events

II. Validity Generalization

Validity Generalization

- Establishment of validity takes place across large set of studies
  - Sample size
  - Representativeness (and diversity) of samples
  - Variation in procedures and settings
  - Cultural or group comparisons
  - Meta-analysis (quantitative summary of many studies) may be used to assess state of evidence of measures validity after much research

II. Discriminant Validity

Discriminant validity – new measure should not be related to measures of unrelated constructs

- Also recommended to include measures of other constructs not expected to be related to the new measure's construct
- Correlations should be close to zero and/or nonsignificant
- Also, no real conventional cutoff for acceptable values
- Convergent and discriminant evidence usually gathered at the same time

IV. Multitrait Multimethod Matrices

Multitrait Multimethod Matrices (MTMMM; Campbell & Fiske, 1959)

- Method variance – the method used in the measure will account for some of the variance
- *Monomethod* correlations should be higher than *heteromethod* correlations (e.g., self-reported anger vs. observations of anger)
- *Monotrait* correlations should be higher than *heterotrait* correlations
- So, expect monotrait-monomethod correlations to be the highest and heterotraitmethod-heteromethod correlatios to be the lowest
- Provides information about how much method affects measure and what real convergent validity might be

IV. Multitrait Multimethod Matrices

Table 9.3 Example of MTMMM Correlations

Methods	Traits	Self-Report: Social Skill	Self-Report: Impulsivity	Self-Report: Conscientiousness	Self-Report: Emotional Stability	Acquaintance Report: Social Skill	Acquaintance Report: Impulsivity	Acquaintance Report: Conscientiousness	Acquaintance Report: Emotional Stability	Interviewer Report: Social Skill	Interviewer Report: Impulsivity	Interviewer Report: Conscientiousness	Interviewer Report: Emotional Stability
Self report	Social skill	(.85)											
	Impulsivity	.14	(.81)										
	Conscientiousness	.20	.22	(.75)									
	Emotional stability	.35	.24	.19	(.82)								
Acquaintance	Social skill	.40	.14	.10	.22	(.76)							
	Impulsivity	.13	.32	.13	.19	.18	(.80)						
	Conscientiousness	.09	.17	.36	.14	.14	.26	(.68)					
	Emotional stability	.20	.23	.11	.41	.30	.28	.18	(.78)				
Interviewer report	Social skill	.34	.11	.19	.20	.23	.01	.11	.19	(.81)			
	Impulsivity	.03	.25	.12	.19	.06	.24	.10	.14	.22	(.77)		
	Conscientiousness	.09	.09	.30	.14	.09	.08	.20	.06	.24	.30	(.86)	
	Emotional stability	.14	.16	.08	.33	.13	.12	.06	.19	.44	.38	.29	(.78)

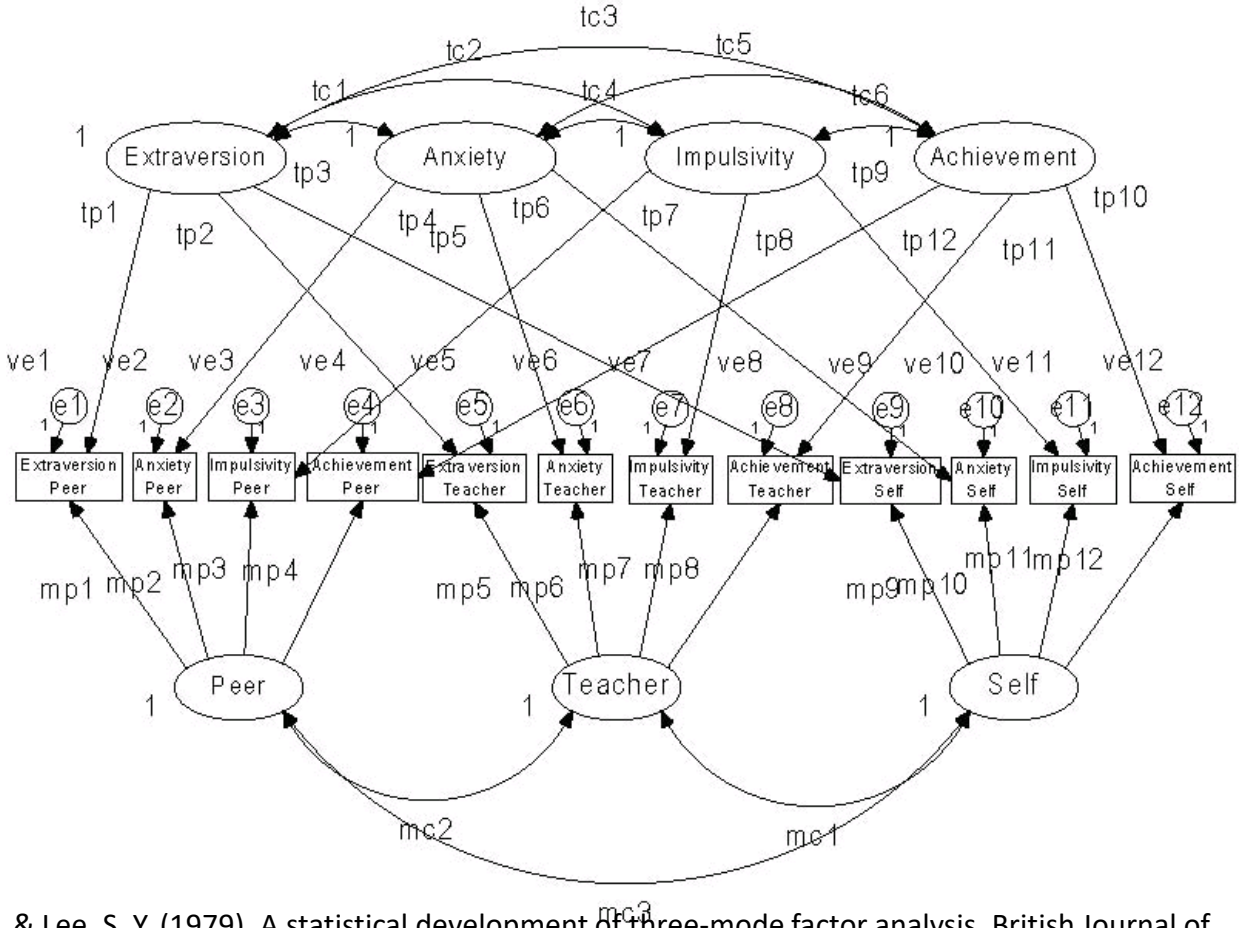
IV. Multitrait Multimethod Matrices

Estimates of associations can be attenuated because of measurement error

May lead to incorrect conclusions of relative size of correlations (Reichardt & Coleman, 1995)

Structural equation modeling (confirmatory factor analysis) approach has advantages, because measurement error can be estimated and removed

IV. Multitrait Multimethod Matrices



Bentler, P. M., & Lee, S. Y. (1979). A statistical development of three-mode factor analysis. British Journal of Mathematical and Statistical Psychology, 32(1), 87-104.

V. Predictive Validity

Predictive and criterion validity closely related

Predictive validity usually used when predicting future state, whereas criterion usually used for association with concurrent state

Either can be assessed in study of convergent validity

V. Predictive Validity

Validity of a measure can be assessed by predicting classification, often in comparison with a "gold standard"

Examples

- Self-reported depression scale and diagnosis by clinician
- Educational test and admission to college
- Personnel test and job success

V. Predictive Validity

Cutoff of measure established theoretically or empirically to compare to classification

Correct classification rate provides information about predictive validity of the measure

V. Predictive Validity

Four ways of predicting category membership in the text:

- Binomial effect size display (BESD: Rosenthal & Rubin, 1982)
- Taylor-Russell tables (Taylor & Russell, 1939)
- Utility analysis (Brogden & Taylor, 1950)
- Sensitivity and specificity

V. Predictive Validity

Binomial effect size display (BESD)

2 × 2 table constructed for below average vs. above average on test and below average vs. above average performance (e.g., SAT and college GPA)

Table 9.6 Example of the Binomial Effect Size Display

	College GPA	
(a) For a correlation of $r = .00$		
Test Score	*Below Average*	*Above Average*
Below average	50	50
Above average	50	50
(b) For a correlation of $r = .48$		
Test Score	*Below Average*	*Above Average*
Below average	A	B
	74	26
Above average	C	D
	26	74

NOTE: GPA = grade point average.

V. Predictive Validity

Taylor-Russell tables (Taylor & Russell, 1939)

- Used to evaluate personnel tests for hiring decisions – test vs. later success in position
- Similar to BESD except other criteria than average can be used (e.g., hiring only 4% of applicants

V. Predictive Validity

Utility analysis (Brogden & Taylor, 1950)

Cost vs. benefit (utility) of testing procedure – does the test predict sufficiently beyond not using a test to be worth the monetary cost?

V. Predictive Validity

Sensitivity and specificity

Widely used in medical and clinical settings to evaluate the predictive accuracy of a particular test

V. Predictive Validity

Sensitivity represents the probability that a test indicates a client has a condition (e.g., depression) when the client truly does have the condition

Specificity then is when the test indicates the client does not have the condition when the client truly does not have it.

V. Predictive Validity

Positive predictive value (PPV) represents the proportion of those classified as depressed who really are depressed

Negative predictive value (NPV) represents the proportion of those who are classified as not depressed who really are not depressed

V. Predictive Validity

		True (Diagnosis)	
		Depressed	Not
Measure (Above Cutoff)	Depressed	*A*	*B*
	Not	*C*	*D*

$$Sensitivity = \frac{A}{(A+C)}$$

$$Specificity = \frac{D}{(B+D)}$$

V. Predictive Validity

		True (Diagnosis)	
		Depressed	Not
Measure (Above Cutoff)	Depressed	*A*	*B*
	Not	*C*	*D*

$$PPV = \frac{A}{(A+B)}$$

$$NPV = \frac{D}{(D+C)}$$

VI. Validity Miscellaneous

Furr and Bacharach discuss *alerting* ($r_{alerting\text{-}CV}$) and *contrast* ($r_{contrast\text{-}CV}$) correlations (Rosenthal, Rosnow, & Rubin, 2000)

Special computations of construct validity (the "CV" here) coefficients that compare predicted correlations to obtained correlations

VI. Validity Miscellaneous

Higher $r_{alerting\text{-}CV}$ values if close match between convergent and divergent correlations obtained and predicted by a set of expert judges

Higher ($r_{contrast\text{-}CV}$) also reflect good match between obtained and predicted correlations but adjusts for absolute magnitude of the correlations

VI. Validity Miscellaneous

Relies on judges' accurate predictions about convergent and discriminant validity coefficients, which may be difficult

May result in concluding high estimates of validity even if judges do not agree on prediction values

VI. Validity Miscellaneous

Statistical significance is used to infer whether or not the validity is different from 0 in the population

Statistical test of Pearson r correlation

Null hypothesis that $\rho = 0$ more likely to be rejected if

- Sample value (*effect size*) is large
- Sample size is large
- Sampling variability is small

VI. Validity Miscellaneous

Need to distinguish between

- Statistical significance
- Large effect size

VI. Validity Miscellaneous

Cohen's (1988) effect size standards for correlation

- Small $r \leq .10$ ($r^2 = .01$)
- Medium $r > .1$ and $\leq .30$ ($r^2 = .09$)
- Large $r \geq .30$ and $\leq .50$ ($r^2 = .25$)

Researchers commonly refer to these effect size standards, but they are arbitrary descriptors

VI. Validity Miscellaneous

The term *face validity* refers to whether the measure or item seems to be a good reflection of the hypothetical construct on its surface

There are no tests for face validity—just a judgement of whether it makes sense

Could use of multiple raters, expert judges, cognitive testing, focus groups could all be used to support face validity

In some instances face validity might be undesirable, because overly obvious questions could lead to social desirable responses, lying, or concealment/faking

VI. Validity Miscellaneous

Some scales that are “empirically derived” may lack face validity, but may predict well

In such measures, items are selected based on their criterion validity without regard to face validity

Subtle items as opposed to obvious items are thought to be less subject to faking, but evidence is not always in support of this assertion (e.g., Thornton & Giersach, 1980)

VI. Validity Miscellaneous

Minnesota Multiphasic Personality Inventory (MMPI) is an example of a test with many items that are not obvious

"I prefer a shower to a bath" is predictive of higher empathy (Hogan Empathy Scale; Hogan, 1969)

VI. Validity Miscellaneous

Revision of a scale to improve reliability or validity should be confirmed in another sample

- Example: in one sample, a particular item may have a lower correlation with the total scale just by random chance
- Example: a cutoff for clinical anxiety which works well for one sample may not be optimal for another sample