Testing & Measurement in Psychology

Validity I

Chapter 8:
Content and Construct-oriented Validation Strategies

What is Validity?

- Validity Defined
  - Traditional- Test measures what it is intended to/ claims to/designed to/purports to measure.
  - Current- "the degree to which evidence and theory support the interpretations of test scores entailed by proposed uses of tests"

What is Validity?

- It is not the test itself is validated, but rather the inferences and conclusions that you reach on the basis of test scores.
- Validity refers to the "interpretation of test scores."
  - So, the goal is to understand the meaning and the implications of test scores.
- For example, "Is the interpretation of performance on the WISC-IV as reflecting intelligence valid?"

What is Validity?

- E.g. ALES
  - If an undergraduate student receives a higher scores on ALES, a graduate admissions committee might use that score to predict or infer that s/he will do well in graduate school.
  - If that student do well in grad school and if students with lower scores do less well, they drew valid conclusions from scores on this test.

2 Major Types of Validity

- Validity of Measurement: Does the test measure what it is supposed to measure and how well?
  - Content validity
  - Construct validity
- Validity of Decisions: How accurate are the decisions based on the test?
  - Predictive validity
  - Concurent validity

2 Major Types of Validity

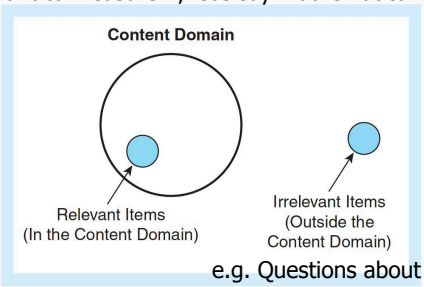
- A company use an inventory called Leadership Skills Profile to select managers.
  - 1st: does this inventory tell you anything about a person's leadership skills?
  - 2nd: do people who got higher scores on this test become a good manager?

Content Validity

- Extent to which the items on a test are representative of the construct the test measures.
- Is established by examining the test itself.
- If your test is supposed to be on chapters 1-5, this is representative of content validity.

Content Validity

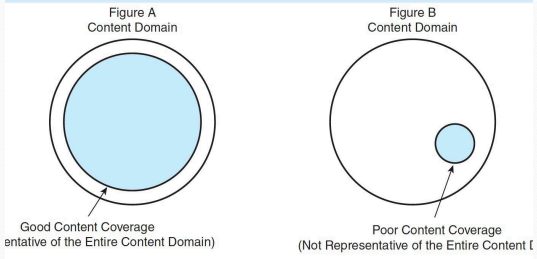
I want to measure X, let's say mathematical ability.



e.g. Questions about sports.

A content domain has boundaries.

Content Validity



Content Validity

- No exact statistical method to assess content validity
- The process consists of 3 steps:
  - Describe the content domain.
  - Determine the domain areas that are measured by each test item.
  - Compare the structure of the test with the structure of the domain.
  - **Tests with high content validity should cover all parts of the domain and most important aspects should be covered with longest number of items.**

Construct Validity

- How well the test measures the theoretical construct or trait.
- Defined as the extent to which the test measures a theoretical construct.
- Deals with the assumed relationships between and among hypothetical constructs.
- We test construct validity by looking at the patterns of relationships of measures of constructs (i.e., are the tests scores correlated the way we think they would be?)

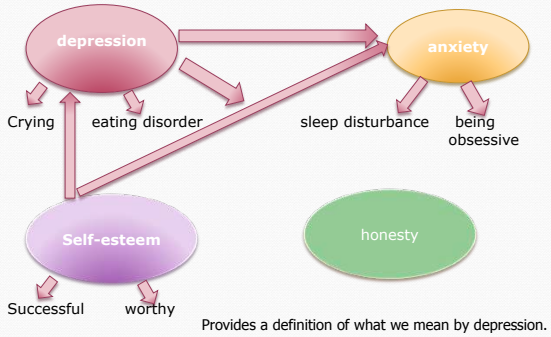
What is a Construct?

- Construct is abstract, hypothetical trait that summarizes some regularity in nature.
- related to Concrete-behaviors (activities that are observable and measurable) In psychology, a sample of behaviors.
  - leadership ability, overcontrolled hostility, depression, and intelligence, beauty, love, self-esteem, honesty, etc.
- Think of gravity, we can not see gravity, but we can see an apple fall from a tree.

Construct Explication

- Translating the abstract construct into concrete, behavioral terms.
  - Identify the behaviors that relate to the construct.
  - Identify other constructs that may be related to the construct being explained.
  - Identify behaviors related to similar constructs and determine whether these behaviors are related to the original construct.

Construct Explication



Provides a definition of what we mean by depression.

Construct Explication

Table 8-3 STEPS IN DESCRIBING THE CONSTRUCT "AGGRESSIVENESS IN SCHOOLCHILDREN"

1. Identify behaviors related to aggressiveness.

Construct	*Behavior*
Aggressiveness	Assaults other students
	Pushes to head of line
	Dominates games

2. Identify other constructs and decide whether they are related to aggressiveness.

Construct	*Behavior*
Need for power	
Aggressiveness	Assaults other students
	Pushes to head of line
	Dominates games
Honesty	

3. Identify behaviors that are related to each construct, and determine their relation to the construct to be measured.

Construct	*Behavior*
Need for power	Makes decisions in groups
Aggressiveness	Dominates games
	Assaults other students
	Pushes to head of line
Honesty	Refrains from cheating
	Tells truth to the teacher

Note: Constructs and behaviors that are related to one another are connected with a solid line. Unrelated constructs or behaviors are not connected.

Construct Explication

Table 8-4 EXPECTED CORRELATIONS BETWEEN A GOOD MEASURE OF AGGRESSIVENESS AND MEASURES OF SPECIFIC BEHAVIORS

Behaviors	*Relationship with aggressiveness*	*Expected correlation*
Assaulting others	Direct	Strong Positive
Pushing in line	Direct	Strong Positive
Dominating games	Direct	Strong Positive
Making decisions	Indirect—related to need for power	Weak Positive
Refraining from cheating	None	None
Telling truth to teacher	None	None

Construct Explication

Table 8-5 CORRELATIONS BETWEEN TEST SCORES AND BEHAVIOR MEASURES FOR TWO TESTS

Behaviors	Expected correlations	Actual correlations: Test A	Actual correlations: Test B
Assaulting others	Strong Positive	.59	−.22
Pushing in line	Strong Positive	.70	.14
Dominating games	Strong Positive	.65	.02
Making decisions in groups	Weak Positive	.30	−.40
Refraining from cheating	None	.09	.56
Telling truth to teacher	None	−.04	.39
		(Test with *high* level of construct validity)	(Test with *low* level of construct validity)

Gathering Psychometric Evidence for Construct Validity

- Reliability
  - Correlations to other relevant tests.
- Multitrait-multimethod matrix (MTMM)
  - Convergent Validity
  - Discriminant Validity
- Factor Analysis
  - Items correlations to the factor (construct)

MTMM

- Combines evidence of reliability, convergent validity and discriminant validity into one study.
- Investigators choose 2 or more constructs that are unrelated in theory and 2 types of different methods (self-report, observation, peer rating, etc)
- An example could be the 5 different personality traits assessed by self, peer, and observer.
- The honesty, the aggressiveness and intelligence of a group of schoolchildren can be measured using 3 methods: teacher ratings, paper-pencil tests, ratings from outside observers.

Table 8-7 A MULTITRAIT-MULTIMETHOD MATRIX

Method	Trait	Teacher ratings: Honesty	Teacher ratings: Aggressiveness	Teacher ratings: Intelligence	Tests: Honesty	Tests: Aggressiveness	Tests: Intelligence	Observers' ratings: Honesty	Observers' ratings: Aggressiveness	Observers' ratings: Intelligence
Teacher ratings	Honesty									
	Aggressiveness	.43								
	Intelligence	.36	.32							
Tests	Honesty	.62	.03	.20						
	Aggressiveness	.22	.70	.13	.40					
	Intelligence	.10	.13	.64	.22	.30				
Observers' ratings	Honesty	.59	.11	.02	.60	.20	.21			
	Aggressiveness	.14	.82	.16	.13	.61	.23	.30		
	Intelligence	.21	.10	.72	.06	.19	.52	.49	.36	

Note: Convergent validity coefficients are underlined. Correlations between different constructs using the same method (e.g., teacher ratings) are enclosed in solid triangles. Correlations between measures of different constructs using different methods are enclosed in broken triangles.

Method bias

Correlations might be high in a particular way of measuring.

Convergent validity

- Establishes construct by comparing it to same construct collected by another method.
- Convergent validity is demonstrated when a test correlates highly with other variables or tests with which it shares an overlap of constructs.
- Example: compare a new measure of anxiety with an old, established measure of anxiety.
  - Correlate new and old measure
  - If correlation is strong, positive = good convergent validity

Discriminant Validity

- Establishes construct by differentiating it from separate constructs.
- Discriminant validity is demonstrated when a test does not correlate with variables or tests from which it should differ.
- Compare construct to unrelated constructs
  - Example: compare extraversion to neuroticism
  - To analyze: correlate construct with other unrelated constructs.
  - Expect negative correlations. Pearson's r should be low or 0.

Convergent and Discriminant validity

- _____: Different measures of the same construct should correlate highly.
- _____:theoretically independent constructs should not be correlated.

Characteristics of a good test

- Consistent scores of the same construct from different tests.
- Test scores do not correlate with unrelated constructs (when the same method is used)
- Method of measurement used by the test show little evidence of bias.

Factor Analysis

- Factor analysis is a specialized statistical technique that is particularly useful for investigating construct validity.
- The goal is to find a smaller set of factors that can account for the observed array of intercorrelations among individual tests.

Factor Analysis

Table 4-5 CORRELATION AMONG TWO DEPTH-PERCEPTION MEASURES, A READING COMPREHENSION TEST, AND A VOCABULARY TEST

	RC	*VOCAB*	*FR*	*EF*
Reading comprehension (RC)	1.0			
Vocabulary (VOCAB)	.62	1.0		
Figure rotation (FR)	.12	.09	1.0	
Exploded figures (EF)	.04	.11	.76	1.0

Spatial ability and verbal ability.
These 2 abilities are independent.

RC and VOCAB are highly correlated because they measure the same factor.

Factor Analysis

Table 4-6 RESULTS OF A FACTOR ANALYSIS OF THE CORRELATIONS SHOWN IN TABLE 4-5

	Factor loadings	
Variables	*Factor I*	*Factor II*
Reading comprehension	.88	.09
Vocabulary	.76	.15
Figure rotation	.04	.72
Exploded figures	.20	.78

So, there are obviously 2 distinct factors, factor loadings.

Face Validity

- Not technically a form of validity.
- A test has face validity if it looks valid to test users, examiners, and especially the examinees.
- How a test taker perceives that appropriateness of a test.
- Important because it can influence how test takers approach the test.

Differences between content and construct validity

1. Content validity is assessed by checking if the test provides representative sample of the content domain.

Construct validity is showed when the pattern of correlations is the same as hypothesized.

Differences between content and construct validity

2. Content validity is established if a test looks like a valid measure.

Construct validity is established if a test acts like a valid measure.

Testing & Measurement in Psychology

Validity II

Chapter 9:
Validity for Decisions:
Criterion-related Validity

Are our decisions valid?

- Validity of Decisions: How accurate are the decisions based on the test?
- Criterion-related validity: Do the test scores correlate with criterion, in other words, predict outcome?
  - Can test scores predict performance on a criterion? (e.g., SAT predict college GPA)
  - Validity coefficients: correlation coefficient between test scores and criterion.

Are our decisions valid?

Table 9-1 MECHANICAL COMPREHENSION SCORES OF FIVE APPLICANTS

Applicant	Scores on mechanical comprehension test (100 = perfect score)
A	98
B	82
C	81
D	43
E	29

On the basis of person's score on measurement,
you try to estimate which applicants
would perform well on the job.
A good test allows you to make reasonably accurate predictions.

Are our decisions valid?

- The variable of primary interest is the outcome measure, called a criterion.
- Criterion is a measure that could be used to determine the accuracy of decision. (e.g. Work performance on the job)
- Criteria must be relevant, uncontaminated and representative of the domain to be predicted.

Are our decisions valid?

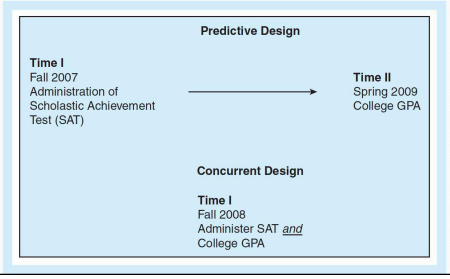
Evidence of criterion-related validity, when the test demonstrates that its scores are systematically related to relevant criterion.

Two types;

Predictive: Test scores are used to estimate outcomes to be measured at a later date.

Concurrent: Test scores and criterion information are obtained simultaneously.

- Predictive studies involve a time interval between test and criterion.
- In concurrent studies, the test and criterion are measured at the same time.



Predictive validation

- Test scores are obtained before making decisions.

1. Obtain test scores, but do not use the test, either directly or indirectly, in making hiring decisions.
2. At some later time, obtain performance measures for those people hired, and correlate these measures with test scores to obtain the predictive validity coefficient.

Predictive validation

- Obtain criterion (e.g. performance)
  - Measure and correlate test scores
  - Needs a random sample.
  - Theoretically the best strategy, but has many practical and ethical problems.
    - Not a realistic one.
    - It is impractical to hire people, admit them to school on a random basis.
    - Decisions are made about applicants without test scores.
    - Failure on the job is a very negative experience. Have substantial losses in terms of training costs and lost productivity.

Concurrent validation (the practical alternative)

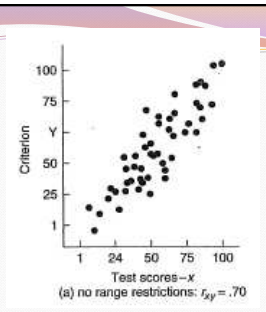
- Test and criterion scores are obtained in the same time in a preselected sample.
- The most fundamental difference between predictive and concurrent validity is not time interval. Concurrent validity coefficient is obtained in a preselected sample (e.g. Present employees, students already accepted)
  - The sample is preselected, not a random sample.
  - Correlation between test and criteria.
  - (e.g. Correlation between test scores and school grades)

Concurrent validation (the practical alternative)

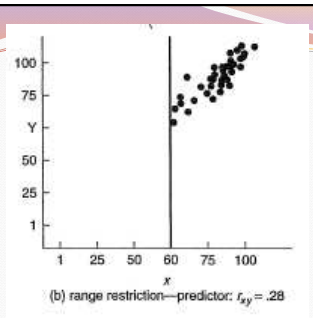
- Adv: practical, quick (test and criterion scores obtained simultaneously, no time interval), easy (no random sampling)
- Disadv: range restriction. Range is smaller.
  - Caused by selection because people are selected according to their test scores (e.g. Bad performers drop out) So, only those with high test scores are selected.
  - In a restricted sample, test measures the difference between moderate and good workers. The worst end of the distribution is missing.

Validity Coefficient

- The relationship between test scores and criterion measures
  - r_{xy}



- In the population of applicants, the correlation between test and criterion.



-
-

those with test scores above 60

- When people are selected on the basis of their test scores, the range of the predictor is directly restricted.

Validity Coefficient

- Usually quite small (r=.30-.50)
  - The lower reliability of the tests, the lower validity coefficients.
- Correlation between test scores and criterion does not give the full picture.
- Another problem is about population. For example, in a work setting, there is a number of workers who have extensive experience on the job as well as new workers. They might have completely different abilities. So, a test that predicts the performance of experienced ones may not be useful in predicting performance of new ones.

Decision Theory

- False positive-True positive
- False negative-True negative

	Reject	Accept
Success	FN	TP
Failure	TN	FP

- Sensitivity & Specificity
  - Sensitivity: detect the presence of a condition.
  - Specificity: detect the absence of a condition.

Decision Theory

- Base ratio=proportion of the population who meet the criterion.
- (e.g. if 50% pass a training course, base ratio is .50)
- High base ratio means many true positive and some false negatives.
- Low base ratio means many true negative and some false positive decisions.
- Base ratio of .50 is the best for test use in decisions.

Decision Theory

- Selection ratio= rating of positions to applicants.
- (e.g. 12 applicants for 10 positions. 0.83)
- The lower the selection ratio is, more a test's validity influences the proportion of success.
- 100 applicants but only 1 position. In this case, the test should be valid to select the perfect 1 applicant!

- The selection rate, base rate and the validity influence the outcomes of decisions.

Construct and Criterion-Related Validity

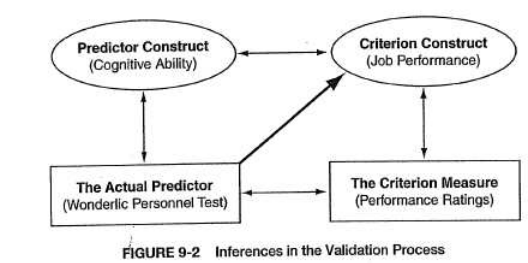
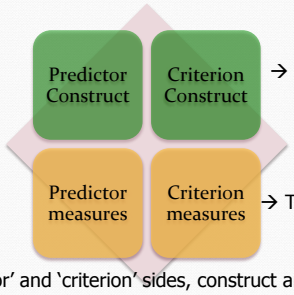


FIGURE 9-2 Inferences in the Validation Process

Construct and Criterion-Related Validity



In both 'predictor' and 'criterion' sides, construct and measures are needed.

Validity Evidence from Meta-Analyses

- There might be several studies which are similar to the one you are designing.
- Meta-analysis refers to methods for combining research results from a large number of studies.
  - E.g.50 studies on validity of the Wonderlic personnel test as a predictor of job performance.
  - Results can be evaluated the criterion-related validity
  - Because large number of studies rarely use precisely the same tests and criterion measures. And also they might use different performance measures, they might have different sample size...

Tests and Decisions

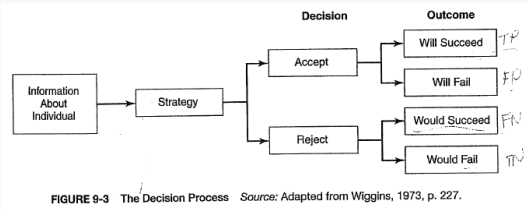
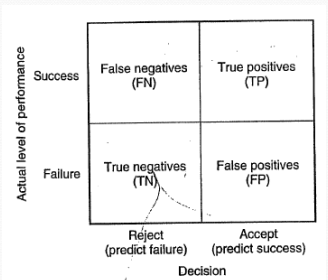


FIGURE 9-3 The Decision Process *Source:* Adapted from Wiggins, 1973, p. 227.

Tests and Decisions

- We need to evaluate the accuracy of decisions. So compare predictions with the outcome of decisions.



Tests and Decisions

- True positive. True negative → represent accurate decisions.
- False positive. False negative→ decision error.

	Reject	Accept
Success	FN	TP
Failure	TN	FP

Tests and Decisions

- To fully evaluate the effect of a test, we should consider these:
- Base rate=proportion of the population who meet the criterion. Proportion of an applicant pool who would succeed on the job.
- (e.g. if 50% pass a training course, base ratio is .50)
- High base ratio means many true positive and some false negatives.
- Low base ratio means many true negative and some false positive decisions.
- Base ratio of .50 is the best for test use in decisions.

Tests and Decisions

- Selection ratio= rating of positions to applicants.
- (e.g. 12 applicants for 10 positions. %83. 0.83)
- (e.g. 30 people apply for 3 jobs, %10 selection ratio)
- The lower the selection ratio is, the more a test's validity influences the proportion of success.
- 100 applicants but only 1 position. In this case, the test should be valid to select the perfect 1 applicant!
- When selection ratio is low, a test with very modest validity can work.

Tests and Decisions

The selection rate, base rate and the validity influence the outcomes of decisions.

Table 9-4 TAYLOR-RUSSELL TABLE SHOWING THE EXPECTED PROPORTION OF SUCCESSES WITH A BASE RATE OF .50

	Selection ratio									
Validity	*.05*	*.10*	*.20*	*.30*	*.40*	*.50*	*.60*	*.70*	*.80*	*.90*
.00	.50	.50	.50	.50	.50	.50	.50	.50	.50	.50
.10	.54	.54	.53	.52	.52	.51	.51	.51	.51	.50
.20	.67	.64	.61	.59	.58	.56	.55	.53	.53	.52
.30	.74	.71	.67	.64	.62	.60	.58	.56	.54	.52
.40	.82	.78	.73	.69	.66	.63	.61	.58	.56	.53
.50	.88	.84	.78	.74	.70	.67	.63	.60	.57	.54
.60	.94	.90	.84	.79	.75	.70	.66	.62	.59	.45
.70	.98	.95	.90	.85	.80	.75	.70	.65	.60	.55
.80	1.00	.99	.95	.90	.85	.80	.73	.67	.61	.55
.90	1.00	1.00	.99	.97	.92	.86	.78	.70	.62	.56
1.00	1.00	1.00	1.00	1.00	1.00	1.00	.83	.71	.63	.56

Tests and Decisions

When decisions are made on a random basis…

$$P(TP) \text{ decision when decisions are made at random} = BR \times SR$$

A base rate of .60 and a selection ratio of .50, 30% of the decisions made at random will be true positives.

	Reject	Accept	
P(FN) = BR – P(TP)	False negative P(FN) = .60 – .30 = .30	True positive P(TP) = .60 × .50 = .30	P(TP) = BR × SR
P(TN) = 1 – P(TP) – P(FP) – P(FN)	True negative P(TN) = 1 – .30 – .30 – .20 = .20	False positive P(FP) = .50 – .30 = .20	P(FP) = SR – P(TP)

Tests and Decisions

When decisions are made on a valid test…

$$P(TP) = BR \times SR + r_{xy}\sqrt{BR(1 - BR)SR(1 - SR)}$$

When the validity of the test is equal to 0.0, the probability of TP is exactly same as when decisions are made at random. When validity coefficient gets higher, we observe an increment in the likelihood of TP decisions.

Random decisions		Using a test with a validity of .40		Using a test with a validity of .70	
FN .30	TP .20	FN .21	TP .29	FN .19	TP .31
TN .30	FP .20	TN .39	FP .11	TN .41	FP .09

11/2/15

Tests and Decisions

Tests are used for making BETTER decisions than without the tests. But, how much better?

Utility Theory suggests 2 things to consider a test's impact on decisions:

- It's ability to increase the number of correct decisions. (who can be a good pilot?)
- The value of correct decision. (which is more important and hard to decide: choosing a good pilot or choosing a good psychologist)

11/2/15 59

Utility Theory

- When base rate, selection ratio and coefficient are known, the effect of test can be determined easily.
- What about determining the value?
- Productivity gain: 'the amount money gained if a test is used.'

Utility theory provides a method for estimating, in dollar terms, the gain (per year) in productivity that will result if valid tests are used in personnel selection. This gain is estimated by

$$\text{Productivity gain} = Kr_{xy}SD_y\overline{Z}_s \quad [9\text{-}4]$$

where

K = number of persons selected
r_{xy} = validity coefficient
SD_y = standard deviation of the criterion
$\overline{Z}_s$ = average (standard) test score among those selected

Class Activity

- Decide on a human behavior; either a measure of intelligence, emotional intelligence or self-esteem
- The in-class assignment is to outline a plan to conduct a validation study of the measure Include any statistical analyses in the plans.
- Once finished, share with your partner. Volunteers will present their own plan to the class.