

PSYC30013
Research Methods for
Human Inquiry
Week 12: Psychological
Assessment:
Validity Criteria in
Psychological Assessment

Week 12: Day 1 & 2



Agenda for this week

(11.5.) Making tests more reliable: the Spearman-Brown Prophecy formula

12.1 Introduction to validity

12.2 The classic tripartite distinction: Criterion validity, Content validity, Construct validity

12.3 Construct validity as one validity to rule them all

12.4 Evidence for validity: Sensitivity, Specificity, Positive Predictive Power, Negative Predictive Power

12.5 Evidence for validity: Nomological networks and Multitrait-Multimethod Matrices

12.6 Conclusion – tying validity back to reliability and Classical Test Theory

11.5. Making tests more reliable: the Spearman-Brown Prophecy formula

Correlations between measures and correlations between constructs

Classical Test Theory (CTT) holds that the observed correlation between two measures x and y is lower than the true correlation between the underlying constructs, because it is attenuated by measurement error.

If CTT assumptions hold and we have a large sample size, the maximum observed correlation between the constructs represented by x and y can be estimated by

$$r_{xy} = r_{x_t y_t} \sqrt{r_{xx} r_{yy}}$$

Where

- r_{xx} is the reliability of measure x
- r_{yy} is the reliability of measure y
- r_{xy} is the observed correlation between x and y (i.e. the correlation in your data)
- $r_{x_t y_t}$ is the correlation between the constructs represented by x and y (i.e. the correlation between the true scores of x and y)

Maximum observed correlation between constructs: Worked Example

You're studying the relationship between the constructs "extroversion" and "public speaking ability"

- r_{xx} is the reliability of your test of extroversion (0.7)
- r_{yy} is the reliability of your test of public speaking ability (0.5)
- r_{xy} is the observed correlation between extroversion and public speaking ability
- $r_{x_t y_t}$ is the correlation between the constructs extroversion and public speaking ability (1)

$$\begin{aligned} r_{xy} &= r_{x_t y_t} \sqrt{r_{xx} r_{yy}} \\ r_{xy} &= 1 \times \sqrt{(0.7)(0.5)} \\ r_{xy} &= \sqrt{0.35} \\ r_{xy} &= 0.59 \end{aligned}$$

The disattenuation formula

$$r_{x_t y_t} = \frac{r_{xy}}{\sqrt{r_{xx} r_{yy}}}$$

This formula corrects for the fact that measurement error attenuates (makes smaller) the correlation between the two constructs measured

It estimates the disattenuated correlation between the two constructs, i.e. the correlation if the constructs were measured without error

Note: The formula is not often used in research, as its assumptions tend to not hold. But it's implicit in techniques you may learn in later courses.

The disattenuation formula: Worked Example

You're studying the relationship between extroversion and skill at public speaking.

- r_{xx} is the reliability of your test of extroversion (0.7)
- r_{yy} is the reliability of your test of public speaking ability (0.5)
- r_{xy} is the observed correlation between extroversion and public speaking ability (0.4)
- $r_{x_ty_t}$ is an estimate of what the correlation between extroversion and public speaking ability would be if there was no measurement error (i.e. the correlation between the constructs)

$$r_{x_ty_t} = \frac{r_{xy}}{\sqrt{r_{xx}r_{yy}}}$$

$$r_{x_ty_t} = \frac{0.4}{\sqrt{(0.7)(0.5)}} = \frac{0.4}{\sqrt{0.35}} = \frac{0.4}{0.59} = 0.68$$

How to increase the correlation between test scores and constructs?

1. Increase the relationship between the psychological construct and the test
2. Remove sources of inconsistency in test administration and interpretation
3. Increase the number of items on the test

Spearman-Brown prophecy formula

$$r'_{xx} = \frac{nr_{xx}}{1 + (n - 1)r_{xx}}$$

Where

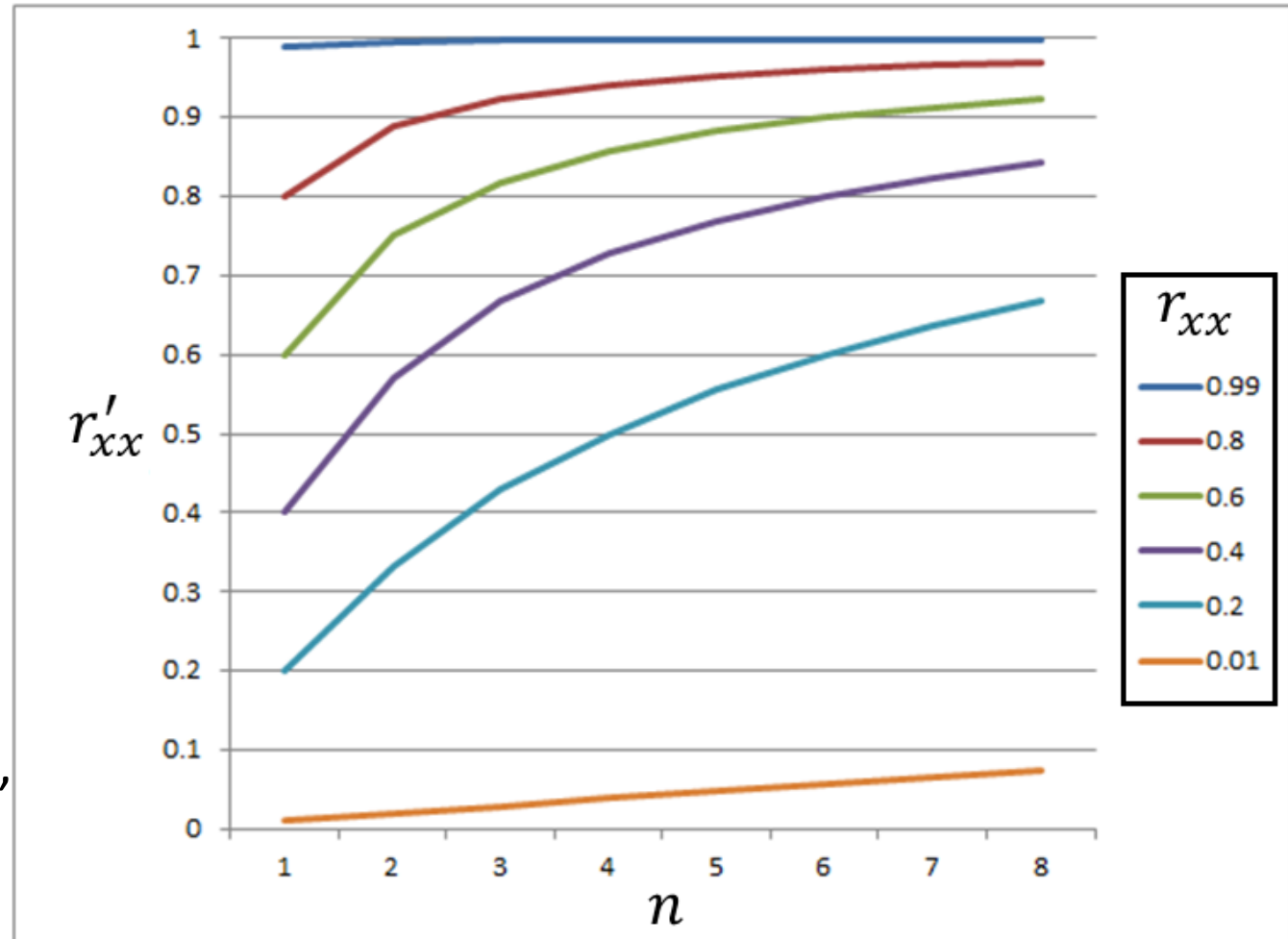
- r'_{xx} is the reliability of the expanded test
- r_{xx} is the reliability of the original test
- n is the expansion factor (so if $n=2$ we double the number of items, $n=\frac{1}{2}$ we halve the number of items, $n=1$ we leave the number of items the same)

Spearman-Brown prophecy formula

$$r'_{xx} = \frac{nr_{xx}}{1 + (n - 1)r_{xx}}$$

Where

- r'_{xx} is the reliability of the expanded test
- r_{xx} is the reliability of the original test
- n is the expansion factor (so if $n=2$ we double the number of items, $n=\frac{1}{2}$ we halve the number of items, $n=1$ we leave the number of items the same)

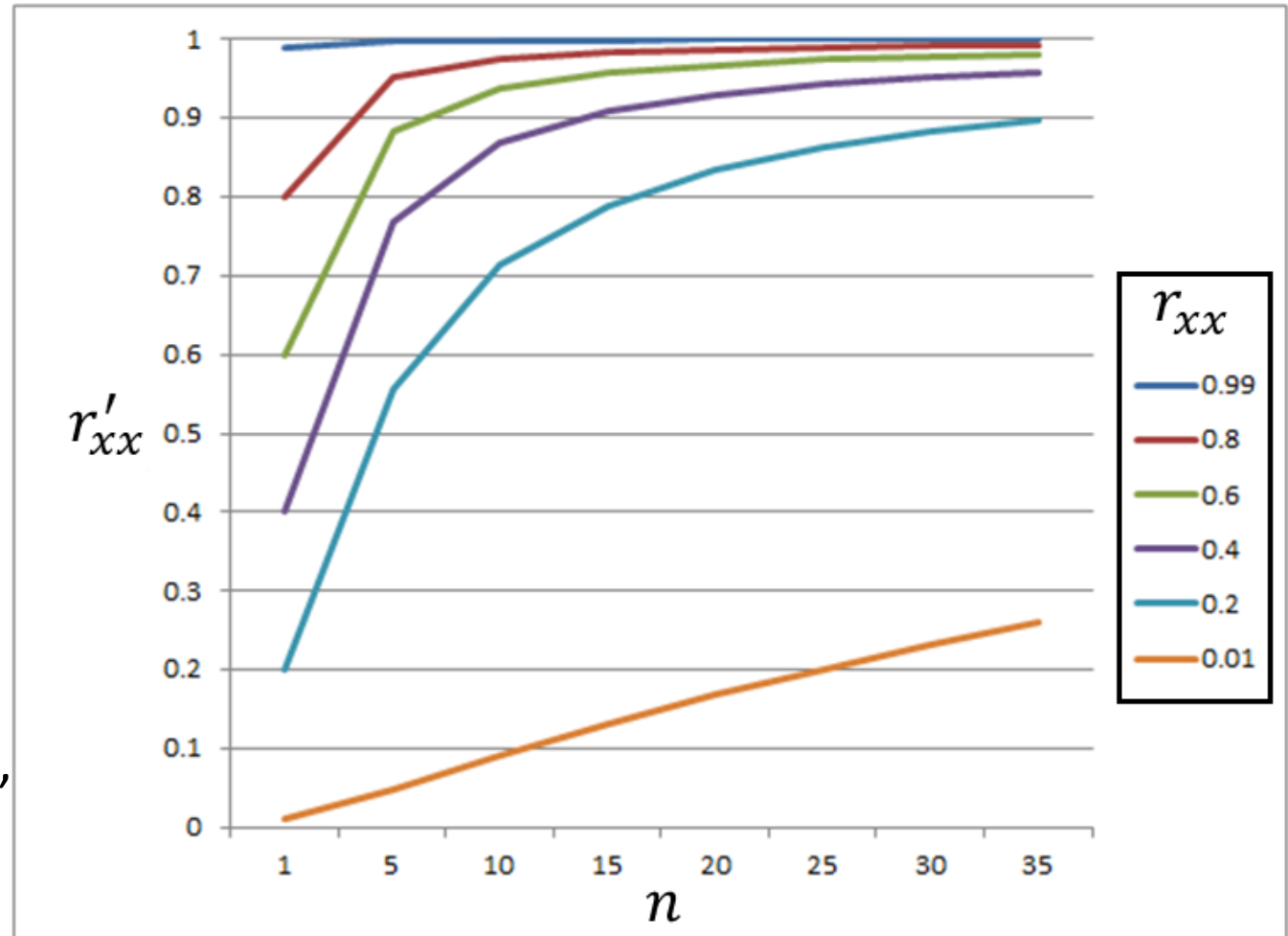


Spearman-Brown prophecy formula

$$r'_{xx} = \frac{nr_{xx}}{1 + (n - 1)r_{xx}}$$

Where

- r'_{xx} is the reliability of the expanded test
- r_{xx} is the reliability of the original test
- n is the expansion factor (so if $n=2$ we double the number of items, $n=\frac{1}{2}$ we halve the number of items, $n=1$ we leave the number of items the same)



Solving for the expansion factor

$$n = \frac{r'_{xx}(1 - r_{xx})}{r_{xx}(1 - r'_{xx})}$$

Where

- r'_{xx} is the reliability of the expanded test
- r_{xx} is the reliability of the original test
- n is the expansion factor (so if $n=2$ we double the number of items, $n=\frac{1}{2}$ we halve the number of items, $n=1$ we leave the number of items the same)

Solving for the expansion factor: Worked Example

The reliability of our test (r_{xx}) is .8.

We want the reliability of the expanded test (r'_{xx}) to be .9 as per the Nunnally and Bernstein (1994) guidelines from last week's lecture.

$$n = \frac{r'_{xx}(1 - r_{xx})}{r_{xx}(1 - r'_{xx})}$$

$$n = \frac{0.9(1 - 0.8)}{0.8(1 - 0.9)} = \frac{0.9(0.2)}{0.8(0.1)} = \frac{0.18}{0.08} = 2.25$$

We need to make our revised test 2.25 times as long as the previous one.

12.1. Introduction to validity

Validity defined

- Last lecture we loosely defined a valid test as one that “accurately measures what it purports to measure”.
- More precisely, validity is "the degree to which evidence and theory support the interpretations of test scores entailed by the proposed uses of tests".
 - Source: American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014) p. 11.
- Key points:
 - Validity concerns interpretations and uses of tests, not merely the test items themselves.
 - Test interpretations and uses must be grounded in relevant evidence and theory
 - Validity is a continuum and not a binary feature
 - Validity tends to require reference to some external criterion

What to think when someone says “This test is valid” or “This test is invalid”.

1. Perhaps they don't understand validity

They may mistakenly think it's a binary attribute of the test rather than a continuous attribute stemming from *interpretations* of the test

2. Perhaps they're expressing themselves in shorthand form.

Example: by “The WAIS-IV IQ test is valid” they mean “Scores on the WAIS-IV IQ test can be validly interpreted as measuring the psychological construct intelligence in adults and older adolescents, as such an interpretation is solidly based on relevant evidence and theory...”

- Common Analogy: Tests as tools



An excess of validities in the literature...

Abstract	Concrete	Content-related	Cross-sectional	Divergent	Face	Indirect	Job component	Particular	Rational	Site	Theory-based
Administrative	Concurrent	Content sampling	Cultural	Domain	Factorial	Inferential	Judgmental	Performance	Raw	Situational	Trait
Aetiological	Concurrent Criterion	Context	Curricular	Domain-selection	Faith	Instructional	Known-groups	Postdictive	Relational	Specific	Translation
Artifactual	Concurrent	Contextual	Decision	Edumetric	Fiat	Internal	Linguistic	Practical	Relevant	Statistical	Translational
Behavior domain	Criterion-related	Convergent	Definitional	Elaborative	Forecast true	Internal test	Local	Predictive	Representational	Status	Treatment
Cash	Concurrent true	Correlational	Derived	Elemental	Formative	Interpretive	Logical	Predictive criterion	Response	Structural	True
Circumstantial	Congruent	Criteria	Descriptive	Empirical	Functional	Intervention	Longitudinal	Predictor	Retrospective	Substantive	User
Cluster domain	Consensual	Criterion	Design	Empirical-judgemental	General	Intrinsic	Lower-order	Prima Facie	Sampling	Summative	Washback
Cognitive	Consequential	Criterion-oriented	Diagnostic	Essential	Generalized	Intrinsic content	Manifest	Procedural	Scientific	Symptom	
Common sense	Construct	Criterion-related	Differential	Etiological	Generic	Intrinsic correlational	Natural	Prospective	Scoring	Synthetic	
Communication	Constructor	Criterion-relevant	Direct	External	Higher-order	Intrinsic rational	Nomological	Psychological & logical	Self-defining	System	
Concept	Construct-related	Cross-age	Discriminant	External test	In situ	Item	Occupational	Psychometric	Semantic	Systemic	
Conceptual	Content	Cross-cultural	Discriminative	Extratest	Incremental	Job analytic	Operational	Quantitative face	Single-group	Theoretical	17

Psychological Assessment Standards

From the Australian Psychological Society (2018):

Ethical guidelines for psychological assessment and the use of psychological tests

From Section 2, “Competence”:

2.3. Psychologists ensure that any test used as part of a formal psychological assessment:

- a) has clear directions for administration and scoring, and adequate information about the properties of scores derived from the test – including the purpose of the test, **the relevant standard errors, and validity and reliability data**;
- b) is **valid for the purpose for which the test is being used, and is also valid for any sub-population of the total population to be included in the particular testing program** (e.g., sub-populations defined according to age, gender, ethnicity, language background or socioeconomic status); and
- c) has **appropriate normative or reference group data to allow for the interpretation of scores** in relation to a clearly defined population.

Very brief history of testing

- Civil service examinations in China trace back to at least the Sui Dynasty (605 CE), and continued until around the fall of the Qing Dynasty (1905 CE)
- Standardized testing in education began to make inroads from around the 1850s
- Binet-Simon Intelligence test introduced in 1905



12.2 The classic tripartite distinction: Criterion validity, Content validity, Construct validity

Criterion Validity

- A test has criterion validity to the extent that it can predict scores on relevant criterion variables
- Sometimes divided into two forms:
 - Concurrent validity: Test scores evaluated against a criterion measured at the same time as the test
 - Predictive validity: Test scores evaluated against a criterion measured later

Examples of criterion validity

Test in green, criterion in red

- Patients are given a newly-developed test for the disease lupus. Results are compared with those of a gold standard test for lupus, which patients are given at the same time.
- An Artificial Intelligence program interviews job applicants and gives them each a score out of 100. Skilled human interviewers listen to the same interviews, and also give applicants a score out of 100.
- Students complete Secondary School and receive an Australian Tertiary Admission Rank (ATAR). This is compared against their final Weighted Average Mark (WAM) at University.
- A test of risk factors for late-onset Alzheimer's disease is administered to 30-year-olds. This is compared against the persons eventually diagnosed with late-onset Alzheimer's disease.

Criterion validity in practice

- Typical practice: calculate the correlation between test scores and scores on the criterion
 - Note that this coefficient will be attenuated
- Why bother making a new test if we already have a criterion?
 - Gold standard test may be more expensive or difficult to apply
 - Some criterion tests can only be done in the future, and we may want to develop a new test that can be done earlier (e.g. the Alzheimer's disease example from the previous slide)

Content Validity

- A test has content validity to the extent its content reflects the full domain of the construct it is supposedly assessing.
- Related concept: Face validity
 - A test has face validity to the extent it *seems* to non-experts (e.g. the test-takers) to have content validity.
 - Generally not taken seriously by test-makers, but can be practically important.

Examples of content validity

- On a VCE Maths Methods exam, all subtopics within the subject are represented approximately in proportion to the amount of teaching time they received.
- In developing a test of what makes a successful Uber driver, we surveyed highly-rated Uber drivers and asked them what skills they considered most important.

Threats to content validity

Two major threats generally identified in the literature:

1. Inclusion of construct-irrelevant content

Example: You take a course with a stated aim of giving students a solid understanding of the foundations of statistics. The final exam contains an item which asks you to “Describe the last 10 minutes of the film *The Godfather*”.

2. Construct underrepresentation

Example: You take a 12-week course with a stated aim of giving students a broad knowledge of Shakespeare’s plays. The course covers one play per week. The assessment is a single exam which covers *Romeo and Juliet* only.

Content validity in practice

Common approach from Lawshe (1975). Ask experts whether individual items are “essential”, “useful but not essential”, or “not necessary”, then calculate the Content Validity Ratio (CVR).

$$CVR = \frac{n_e - (\frac{N}{2})}{(\frac{N}{2})}$$

Where

- n_e is the number of experts responding ‘essential’
- N is the total number of experts

Construct Validity

- A test has construct validity to the extent it reflects the construct it is meant to be reflecting.
- Heavily tied up with two issues:
 - Convergent validity: Test scores should be (strongly) correlated with measures of related constructs
 - Discriminant validity: Test scores should not be (strongly) correlated with measures of unrelated constructs
- Example: test of *anxiety* should correlate strongly with a test of *stress*, but should not correlate strongly with a test of *visual memory*

Summing up the classic validity types

Validity type	What it asks	Typical way of assessing it
Criterion	Do test scores predict scores on relevant criterion variables?	Do test scores (strongly) correlate with scores on relevant criterion variables?
Content	Does the test's content reflect the full domain of the construct it's supposedly assessing?	Do experts say the content appropriately reflects the full domain of the construct?
Construct	Does the test reflect the construct it's meant to be reflecting?	Do test scores (strongly) correlate with measures of related constructs, and not (strongly) correlate with measures of unrelated constructs?



12.3. Construct validity as one validity to rule them all

Criticisms of the tripartite (criterion, content, construct) view of validity

- Messick (1989): “[E]ven for purposes of applied decision making, reliance on criterion validity or content coverage is not enough. The meaning of the measure, and hence its construct validity, must always be pursued-not only to support test interpretation but also to justify test use.” (p. 17)
- Implication: content and criterion validity should be subsumed within construct validity



Content

A Venn diagram consisting of two overlapping circles. The left circle is light green and labeled 'Content'. The right circle is a darker shade of green and labeled 'Typical Evidence'. The circles overlap in the center-right area.

Does test content
match the content
that *should* be
included?

Typical Evidence

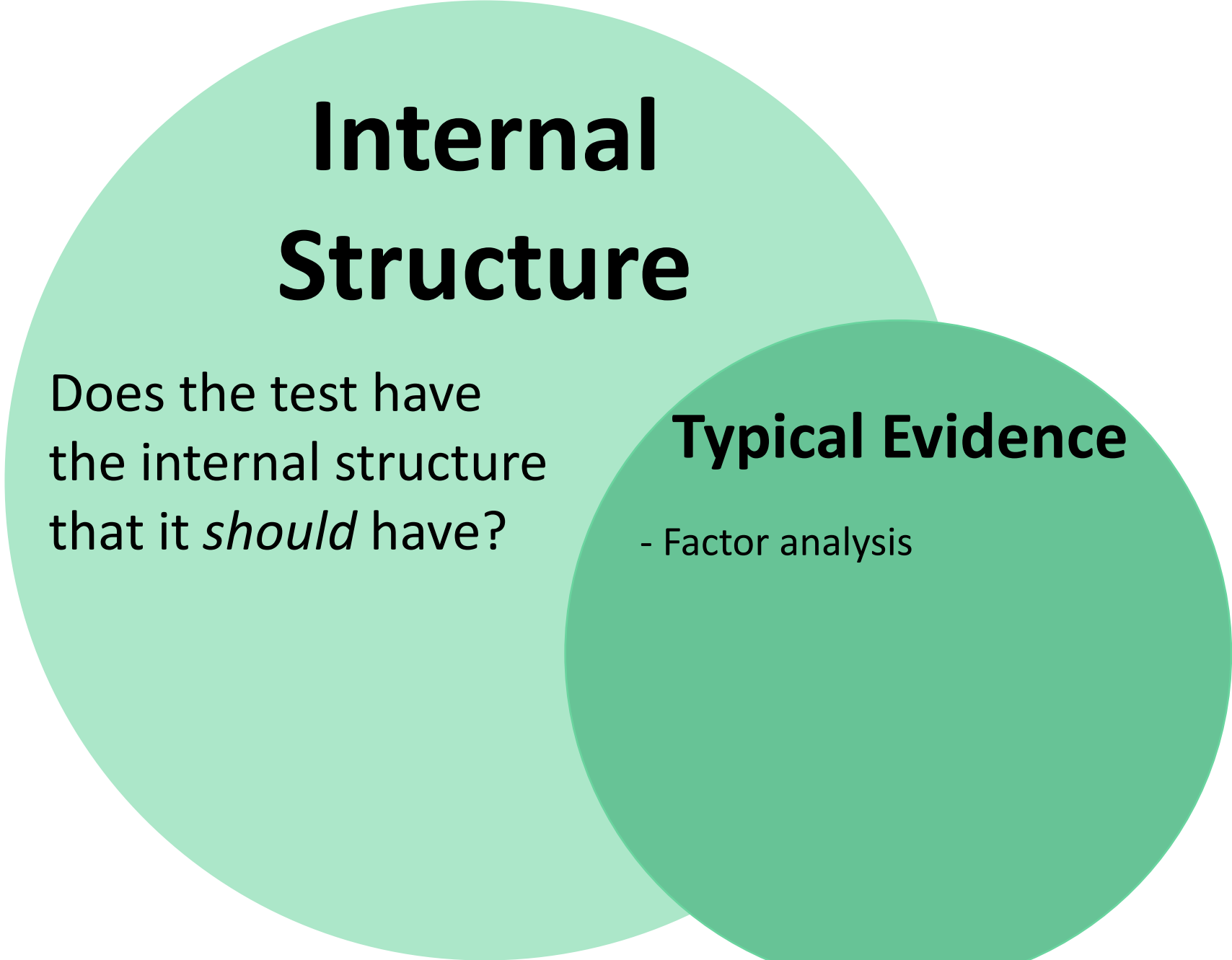
- Expert ratings (e.g. Lawshe's Content Validity Ratio)

Response Processes

Do test-takers use cognitive processes that they *should* be using?

Typical Evidence

- Direct evidence, e.g. interviews in which test-takers are asked how they completed the test
- Indirect evidence, e.g. eye-tracking, response times



Internal Structure

Does the test have
the internal structure
that it *should* have?

Typical Evidence

- Factor analysis

Relations to Other Variables

Does the test have
the relationships
with other variables
that it *should* have?

Typical Evidence

- Relationships with other tests measuring the same construct
- Relationships with criterion variables (assessed through e.g. sensitivity/specificity)
- Appropriate convergent and discriminant relationships (assessed through e.g. Multitrait-Multimethod Matrices)

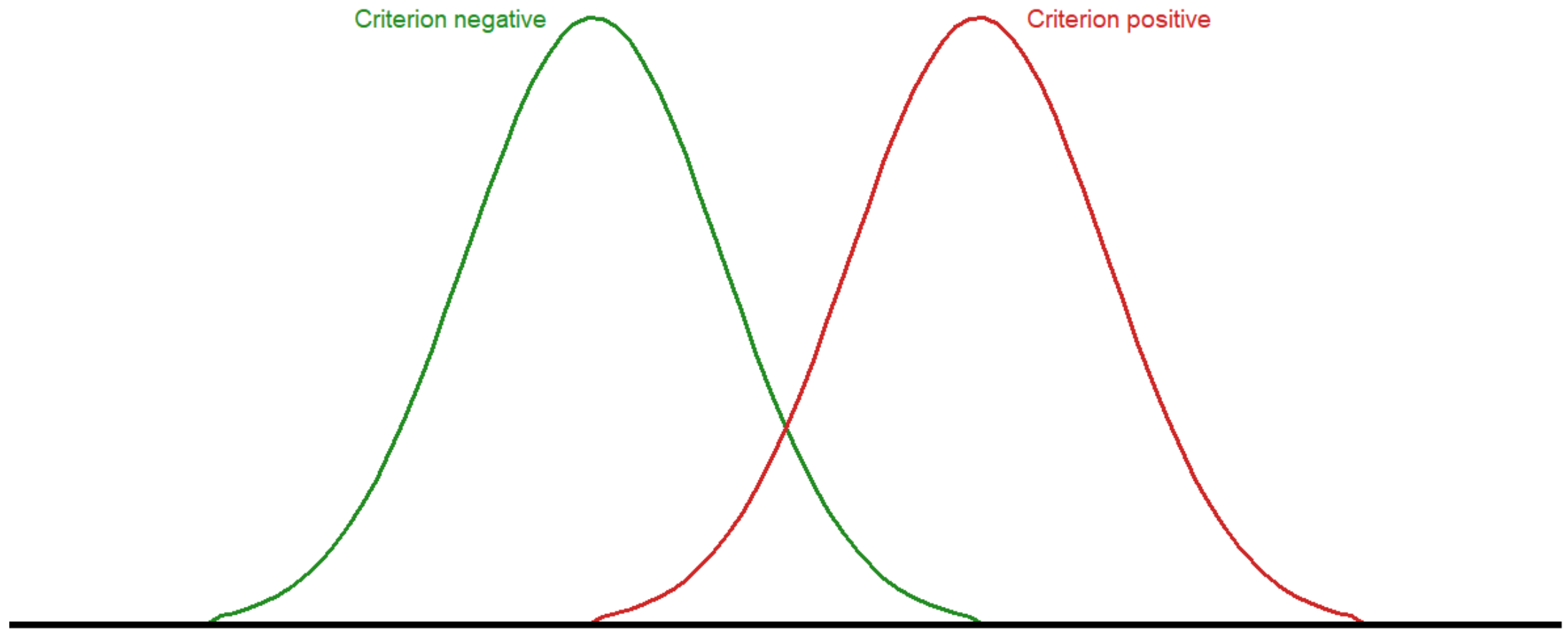
Consequences of Testing

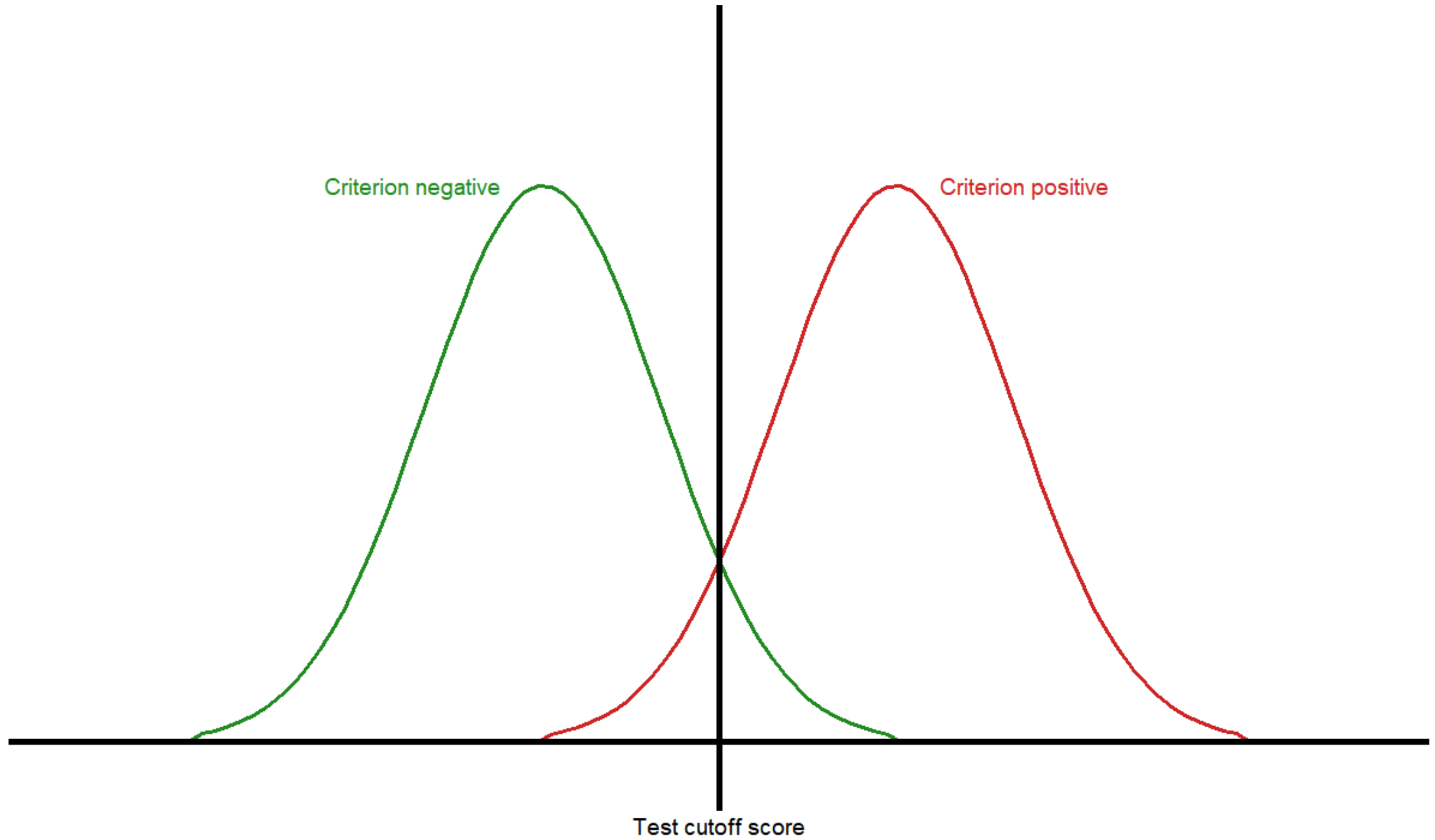
Does testing bring
about the
consequences that it
should bring about?

Typical Evidence

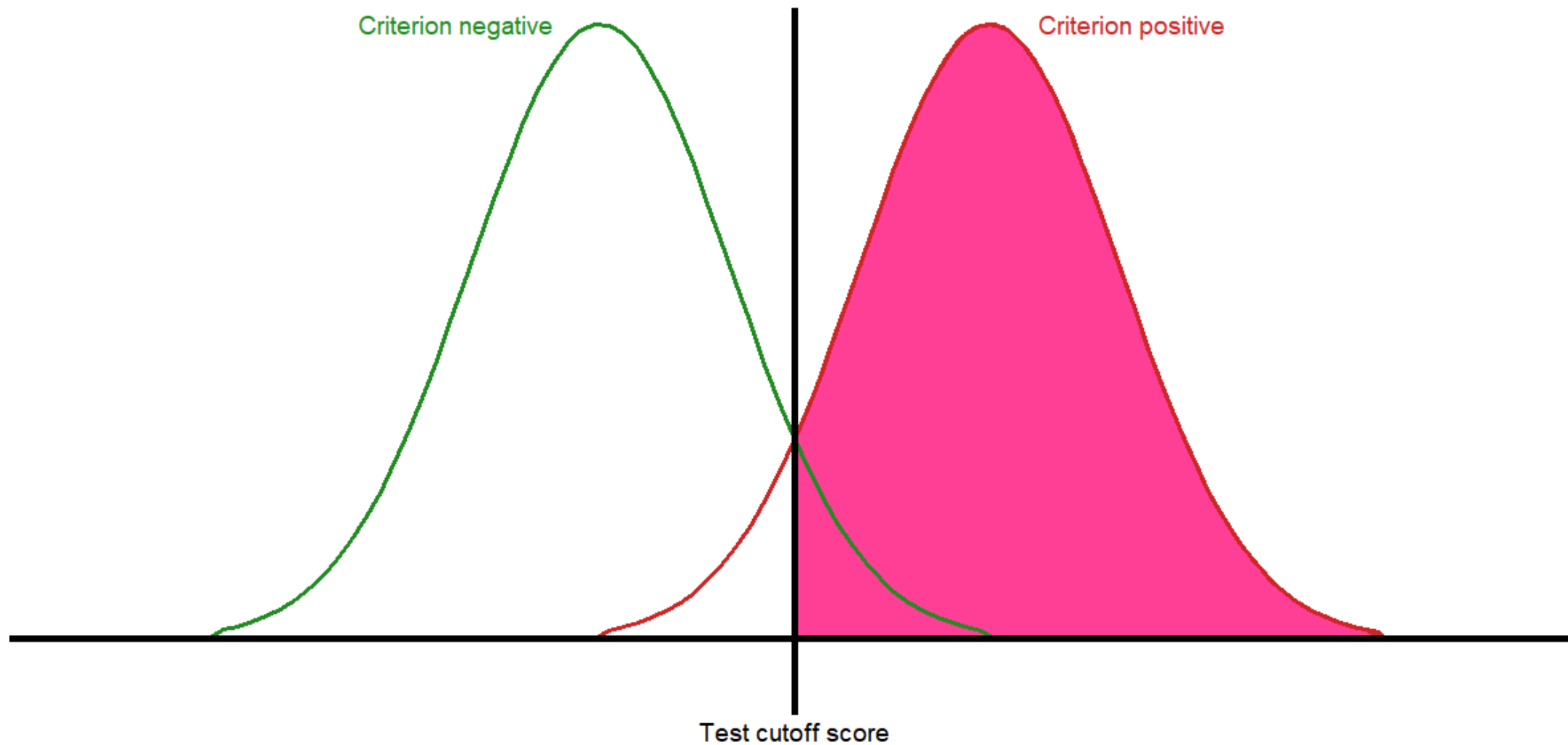
- Investigation of whether test score interpretations cause beneficial outcomes
- Investigation of whether testing itself causes beneficial outcomes
- Investigation of unintended consequences of testing

12.4 Evidence for validity: Sensitivity, Specificity, Positive Predictive Power, Negative Predictive Power





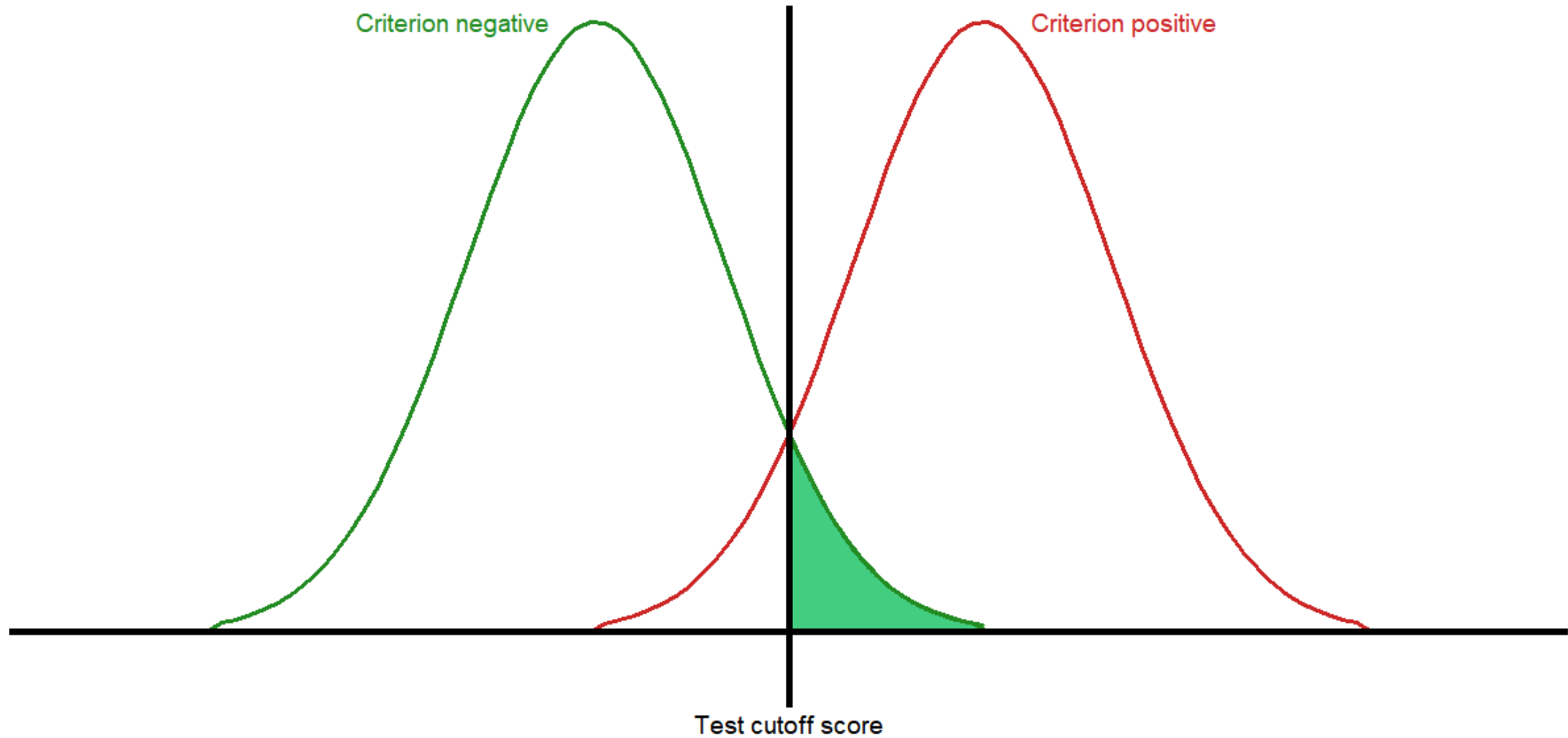
True Positives



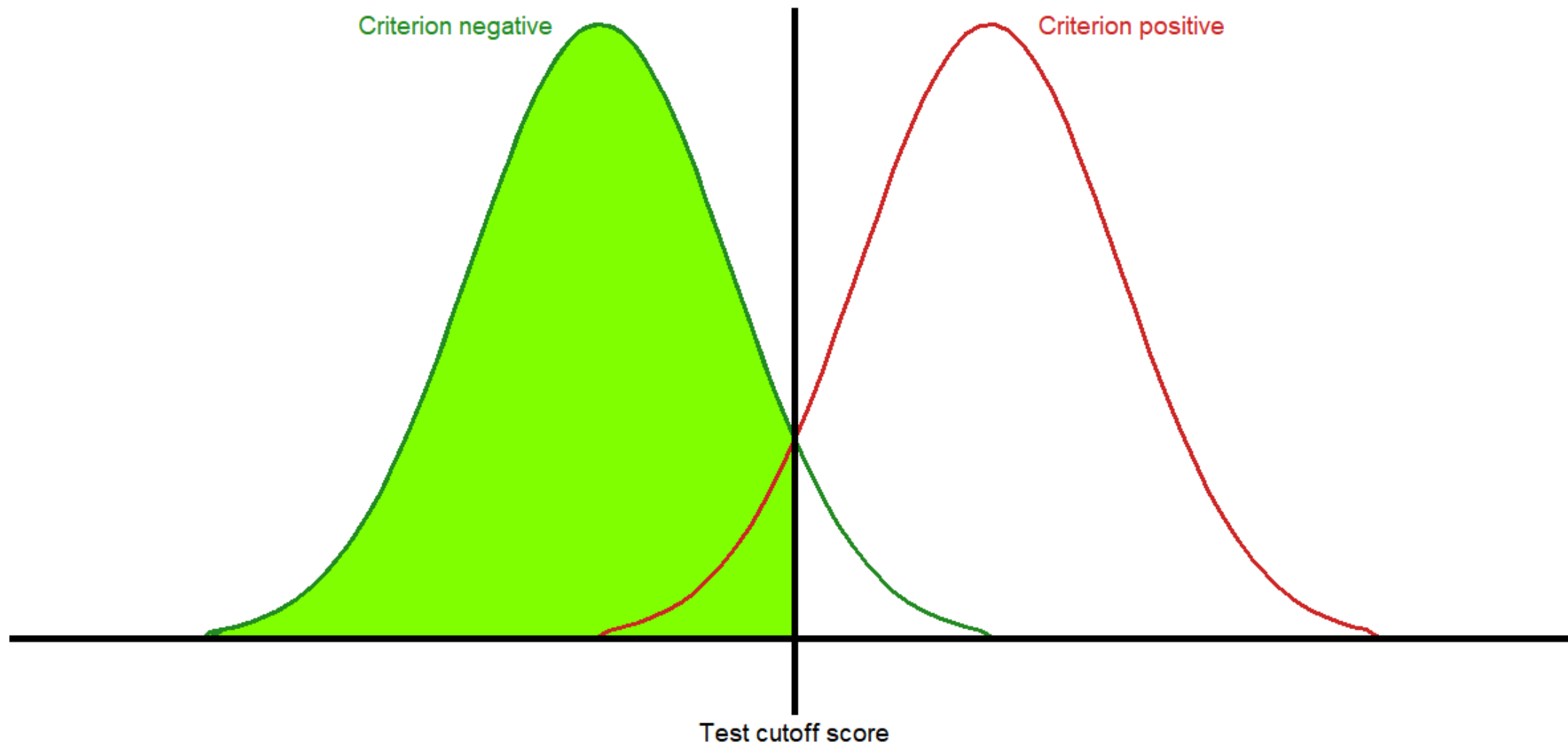
False Negatives



False Positives



True Negatives



Sensitivity

	Criterion Positive	Criterion Negative
Test Positive	True Positives	False Positives
Test Negative	False Negatives	True Negatives

Sensitivity: Test's ability to correctly detect positive cases

$$\text{Sensitivity} = \frac{\text{True Positives}}{\text{True Positives} + \text{False Negatives}}$$

Specificity

	Criterion Positive	Criterion Negative
Test Positive	True Positives	False Positives
Test Negative	False Negatives	True Negatives

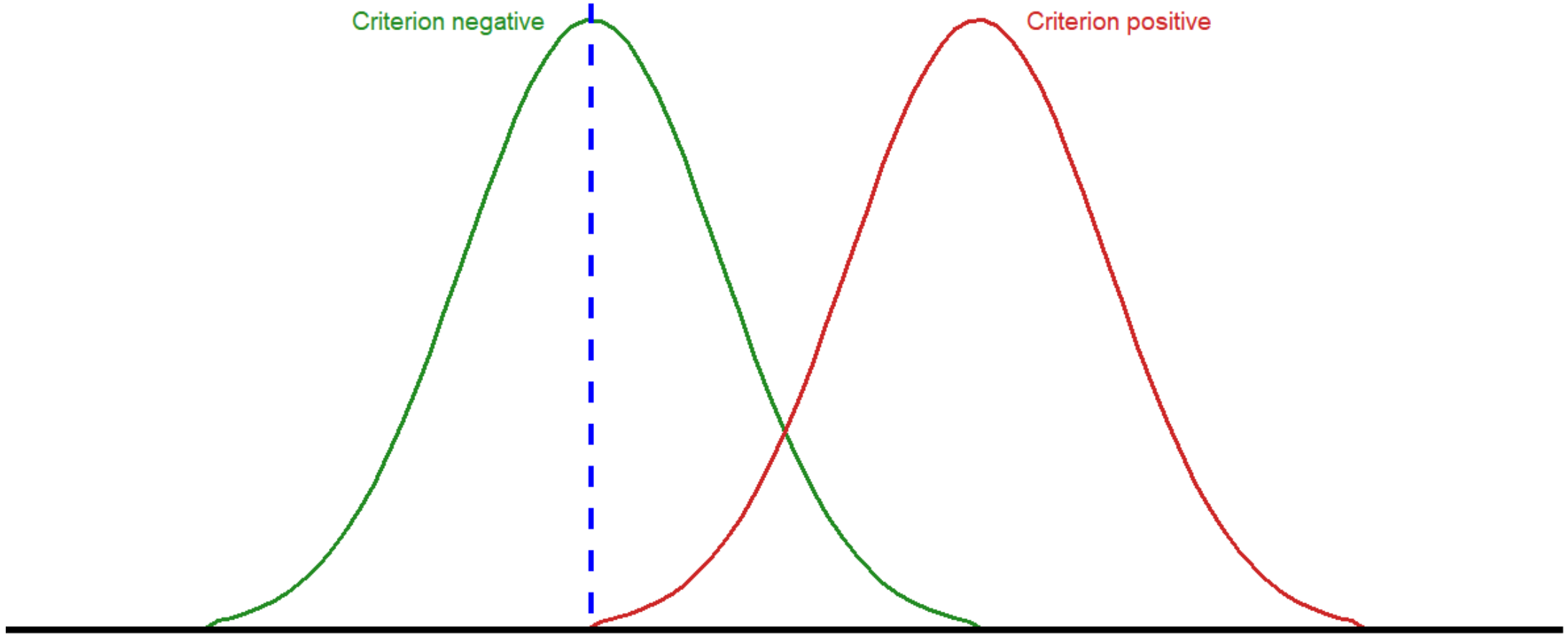
Specificity: Test's ability to correctly detect negative cases

$$\text{Specificity} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Positives}}$$

Cutoff with 100% Sensitivity, 50% Specificity

Criterion negative

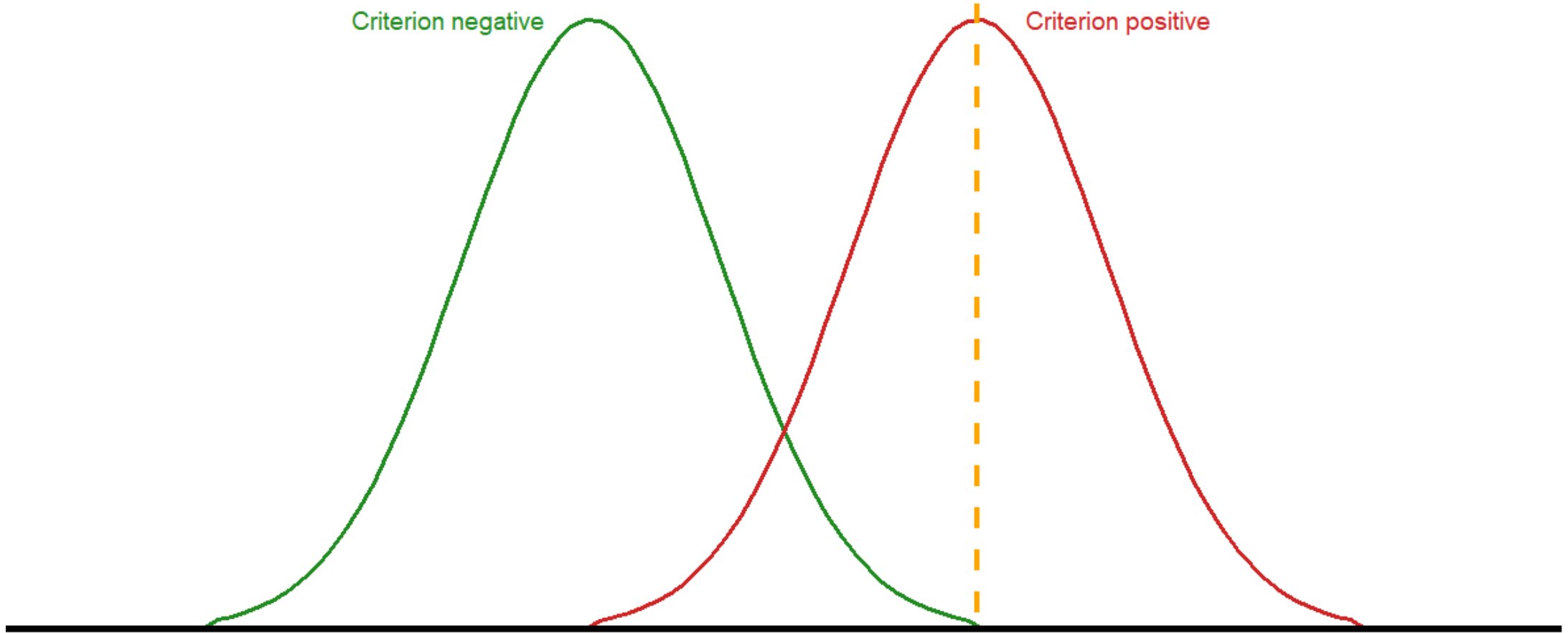
Criterion positive



Cutoff with 50% Sensitivity, 100% Specificity

Criterion negative

Criterion positive



Positive Predictive Power

	Criterion Positive	Criterion Negative
Test Positive	True Positives	False Positives
Test Negative	False Negatives	True Negatives

Positive Predictive Power (PPP): Probability a positive test result indicates a positive case

$$PPP = \frac{\text{True Positives}}{\text{True Positives} + \text{False Positives}}$$

Negative Predictive Power

	Criterion Positive	Criterion Negative
Test Positive	True Positives	False Positives
Test Negative	False Negatives	True Negatives

Negative Predictive Power (NPP): Probability a negative test result indicates a negative case

$$\text{NPP} = \frac{\text{True Negatives}}{\text{True Negatives} + \text{False Negatives}}$$

Prevalence

	Criterion Positive	Criterion Negative
Test Positive	True Positives	False Positives
Test Negative	False Negatives	True Negatives

Prevalence: Probability a random case is criterion positive

$$\text{Prevalence} = \frac{\text{True Positives} + \text{False Negatives}}{\text{True Positives} + \text{False Positives} + \text{False Negatives} + \text{True Negatives}}$$

Useful mnemonics, and a caution

Famous Mnemonics introduced by the medical textbook Sackett et al (2000)

- Sp-P-In: In relation to a highly Specific test, a Positive result tends to rule In the diagnosis.
- Sn-N-Out: In relation to a highly Sensitive test, a Negative result tends to rule Out the diagnosis.

Useful mnemonics, but be cautious

- Perfect diagnosis will only happen if a test has 100% Sensitivity and 100% Specificity, which in practice is never the case
- For Sp-P-In to work well, decent Sensitivity is also needed
- For Sn-N-Out to work well, decent Specificity is also needed

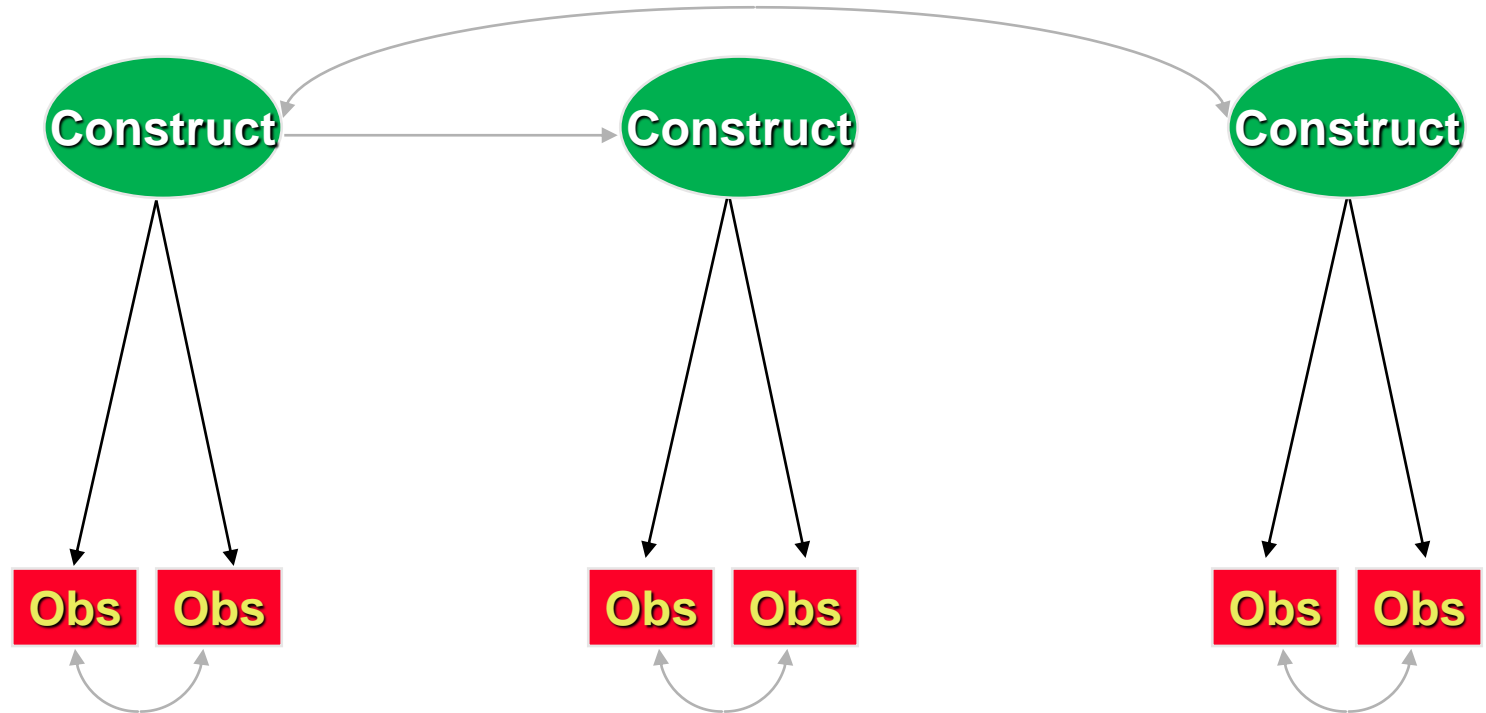
Common issue in interpreting NPP and PPP

- The reported prevalence will relate to a population under consideration for a particular study. However, in most clinical/applied settings the prevalence of diagnosis will be higher than in the aforementioned population.
- Implication: if we have a test with less than perfect specificity/sensitivity, be wary of ruling in/out a condition on the basis of test results.

12.5. Evidence for validity: Nomological networks and Multitrait-Multimethod Matrices

General philosophical approach to construct validity: Nomological networks

- Nomological networks (Cronbach & Meehl, 1955): the interlocking system of laws which constitute a theory and that relate
 - Observable properties to each other
 - Theoretical constructs to observables
 - Theoretical constructs to each other



Evaluating convergent/discriminant validity: Multitrait-Multimethod Matrices

- Technique proposed by Campbell & Fiske (1959), building on the conceptual foundation for construct validity established by Cronbach & Meehl (1955).
- Obtain multiple measures of traits, and also multiple applications of the same methods applied to different traits
- Attempts to get at convergent/discriminant validity by evaluating two sources of variance: trait variance and method variance

Multitrait-Multimethod Matrices: Example

- We investigate four traits:
 - Social skill
 - Impulsivity
 - Conscientiousness
 - Emotional stability
- We use three methods:
 - Self-report
 - Acquaintance ratings
 - Interviewer ratings

Multitrait-Multimethod Matrices: Two flavours of variance



Trait variance

- Measures tend to correlate if they're based on the same (or similar) traits



Method variance

- Measures tend to correlate if they're based on the same (or similar) methods
- Problem: Correlations between measures could be explained by trait variance, or method variance, or both

Multitrait-Multimethod Matrices: Four flavours of correlation



Heterotrait-heteromethod correlations

- Different constructs measured using different methods



Heterotrait-monomethod correlations

- Different constructs measured using the same method



Monotrait-heteromethod correlations

- Same constructs measured using different methods



Monotrait-monomethod correlations

- Same constructs measured using the same method

Relationship between the two constructs		Method Used to Measure the Two Constructs	
		Different Methods (e.g. self-report for one construct, acquaintance report for other)	Same Method (e.g. self-report for both constructs)
Different constructs	<i>Label</i>	Heterotrait-heteromethod correlations	Heterotrait-monomethod correlations
	<i>Example</i>	Self-report measure of social skill correlated with acquaintance report measure of emotional stability	Self-report measure of social skill correlated with self-report measure of emotional stability
	<i>Desired correlation</i>	Very weak	Ideally weaker than Monotrait-heteromethod
Same (or similar) constructs	<i>Label</i>	Monotrait-heteromethod correlations	Monotrait-monomethod correlations
	<i>Example</i>	Self-report measure of social skill correlated with acquaintance report measure of social skill	Self-report measure of social skill correlated with self-report measure of social skill (i.e. reliability)
	<i>Desired correlation</i>	Stronger than Heterotrait-heteromethod, and ideally stronger than Heterotrait-monomethod	Very strong (see Week 11 Day 2 for why we want high reliability coefficients)

Multitrait-Multimethod Matrix

		Method 1			Method 2			Method 3		
Traits		A ₁	B ₁	C ₁	A ₂	B ₂	C ₂	A ₃	B ₃	C ₃
Method 1	A ₁	(.89)								
	B ₁	.51	(.89)							
	C ₁	.38	.37	(.76)						
Method 2	A ₂	.57	.22	.09	(.93)					
	B ₂	.22	.57	.10	.68	(.94)				
	C ₂	.11	.11	.46	.59	.58	(.84)			
Method 3	A ₃	.56	.22	.11	.67	.42	.33	(.94)		
	B ₃	.23	.58	.12	.43	.66	.34	.67	(.92)	
	C ₃	.11	.11	.45	.34	.32	.58	.58	.60	(.85)


Multitrait-Multimethod Matrix

		Method 1			Method 2			Method 3		
Traits		A ₁	B ₁	C ₁	A ₂	B ₂	C ₂	A ₃	B ₃	C ₃
Method 1	A ₁	(.89)								
	B ₁	.51	(.89)							
	C ₁	.38	.37	(.76)						
Method 2	A ₂	.57	.22	.09	(.93)					
	B ₂	.22	.57	.10	.68	(.94)				
	C ₂	.11	.11	.46	.59	.58	(.84)			
Method 3	A ₃	.56	.22	.11	.67	.42	.33	(.94)		
	B ₃	.23	.58	.12	.43	.66	.34	.67	(.92)	
	C ₃	.11	.11	.45	.34	.32	.58	.58	.60	(.85)


Monotrait-monomethod correlation
Same constructs measured using the same method

Multitrait-Multimethod Matrix

		Method 1			Method 2			Method 3		
Traits		A ₁	B ₁	C ₁	A ₂	B ₂	C ₂	A ₃	B ₃	C ₃
Method 1	A ₁	(.89)								
	B ₁	.51	(.89)							
	C ₁	.38	.37	(.76)						
Method 2	A ₂	.57	.22	.09	(.93)					
	B ₂	.22	.57	.10	.68	(.94)				
	C ₂	.11	.11	.46	.59	.58	(.84)			
Method 3	A ₃	.56	.22	.11	.67	.42	.33	(.94)		
	B ₃	.23	.58	.12	.43	.66	.34	.67	(.92)	
	C ₃	.11	.11	.45	.34	.32	.58	.58	.60	(.85)

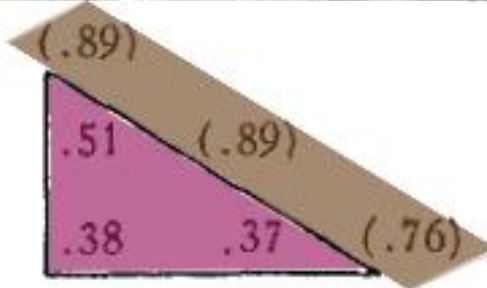


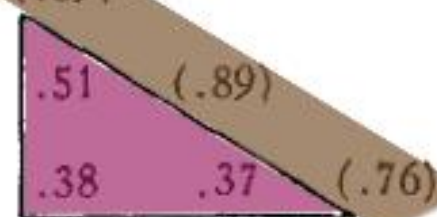
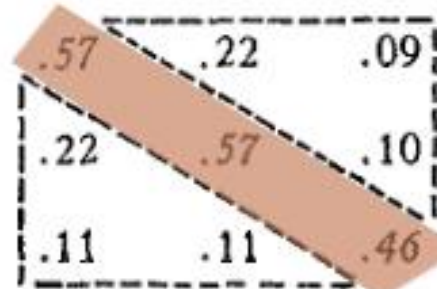
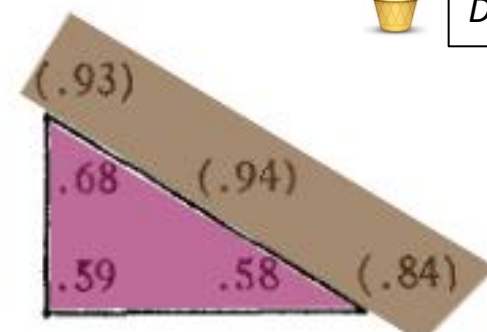

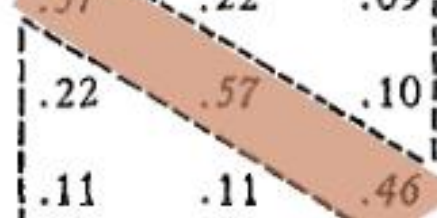
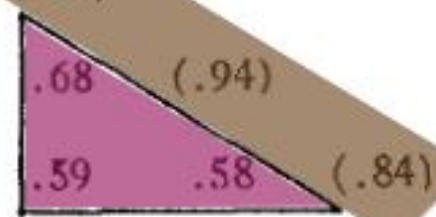
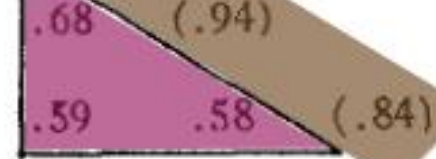
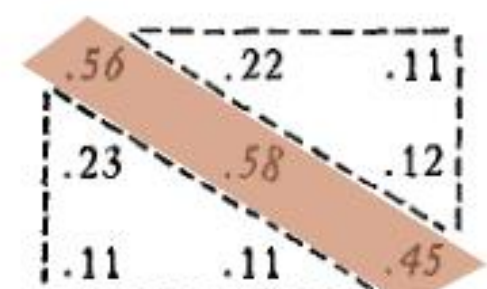
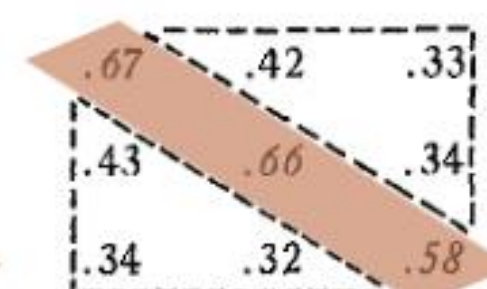
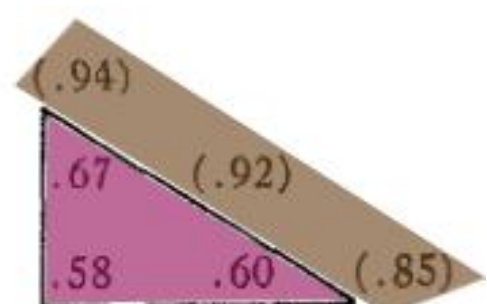
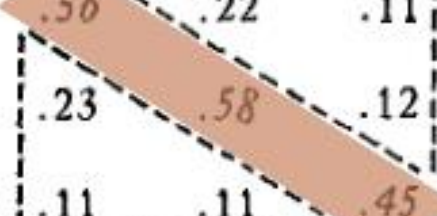
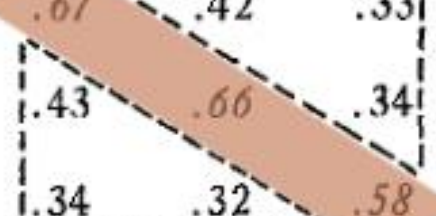
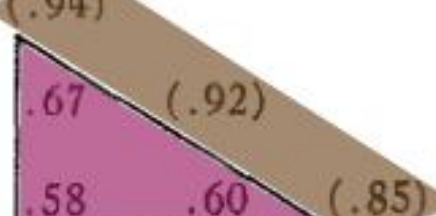
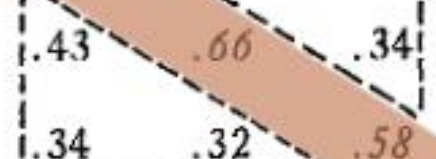


Monotrait-monomethod correlation
Same constructs measured using the same method



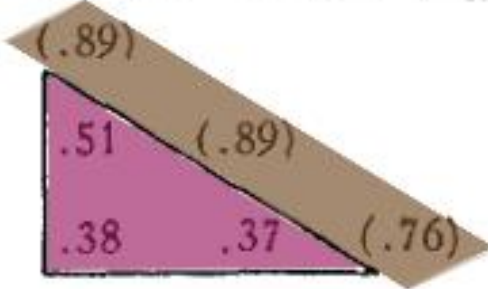

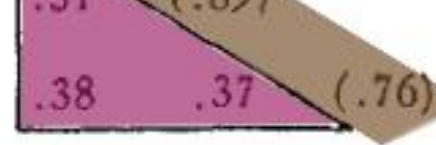

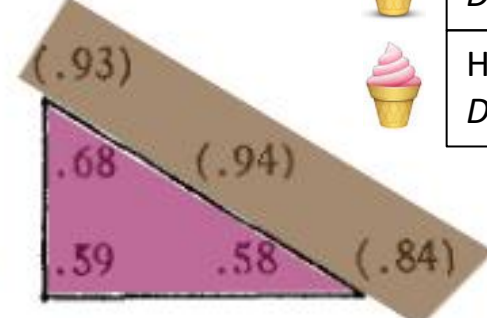




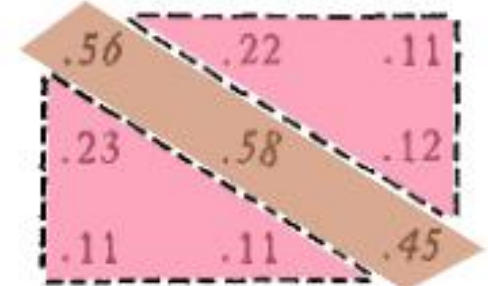
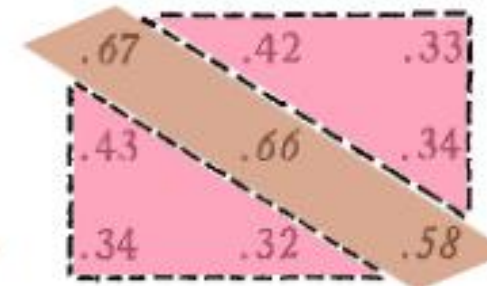
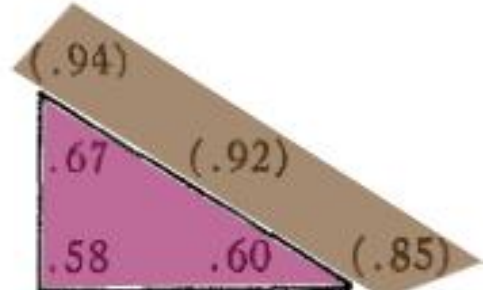




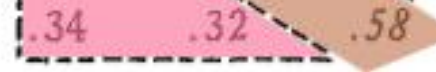

Monotrait-heteromethod correlations
Same constructs measured using different methods

Multitrait-Multimethod Matrix

		Method 1			Method 2			Method 3								
Traits		A ₁	B ₁	C ₁	A ₂	B ₂	C ₂	A ₃	B ₃	C ₃						
Method 1	A ₁				 Monotrait-monomethod correlation <i>Same constructs measured using the same method</i>			 Monotrait-heteromethod correlations <i>Same constructs measured using different methods</i>								
	B ₁															
	C ₁															
Method 2	A ₂							 Heterotrait-monomethod correlations <i>Different constructs measured using the same method</i>								
	B ₂															
	C ₂															
Method 3	A ₃															
	B ₃															
	C ₃															

64

Multitrait-Multimethod Matrix

		Method 1			Method 2			Method 3		
Traits		A ₁	B ₁	C ₁	A ₂	B ₂	C ₂	A ₃	B ₃	C ₃
Method 1	A ₁	 (.89)								
	B ₁	 .51 (.89)								
	C ₁	 .38 .37 (.76)								
Method 2	A ₂	 .57 .22 .09			 (.93)					
	B ₂	 .22 .57 .10			 .68 (.94)					
	C ₂	 .11 .11 .46			 .59 .58 (.84)					
Method 3	A ₃	 .56 .22 .11			 .67 .42 .33			 (.94)		
	B ₃	 .23 .58 .12			 .43 .66 .34			 .67 (.92)		
	C ₃	 .11 .11 .45			 .34 .32 .58			 .58 .60 (.85)		

Multitrait-Multimethod Matrix

Methods	Traits	Self-Report				Acquaintance Report				Interviewer Report			
		Social Skill	Impulsivity	Conscientiousness	Emotional Stability	Social Skill	Impulsivity	Conscientiousness	Emotional Stability	Social Skill	Impulsivity	Conscientiousness	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)

Traps to watch out for when investigating Multitrait-Multimethod correlations

- Remember that they can be deflated by various factors, including
 - Measurement error
 - Restricted range
 - Time lags
- This applies not just to Multitrait-Multimethod correlations, but validity correlations of all sorts

12.6. Conclusion – tying validity back to reliability and Classical Test Theory

A metaphor for reliability and validity



High reliability, low validity



High reliability, high validity



Low reliability, low validity

Tying last week's content to validity

- Reliability as a precondition for believing a test to be valid
- Limitations of Classical Test Theory
 - Struggles to generalize results to other assessments
 - See Generalizability Theory
 - Struggles to calibrate different items to measure a common attribute
 - See Modern Test Theory approaches
 - Tends to adhere to a purely criterion-oriented view of validity
 - See both Generalizability Theory and Modern Test Theory

Future directions

- “Great Validity Debate” still ongoing in psychology (and education)
- Two major textbooks explore the current controversies
 - Markus & Borsboom (2013)
 - Newton & Shaw (2014)
- Some current topics of debate
 - Are tests the proper subject of validity, or interpretations of tests? (Markus & Borsboom, 2013)
 - Are nomological networks necessary? (Lissitz & Samuelson, 2007)

Example: The implicit-association test (IAT)

- Uses categorization task reaction times to measure individual differences in associations of concepts
- Greenwald, McGhee, & Schwartz (1998), which introduced the technique, cited 12,500+ times as of 2020



Example: The implicit-association test (IAT)

Imagine you work at a centre devoted to fostering Australia-USA relations. Your boss has staff complete an IAT. Staff with “high anti-Australia/anti-USA implicit bias” get training to reduce their implicit bias.

- What’s the reliability?
 - Not properly studied, but the IAT’s architects estimate .55 across IAT subtypes (and ~.42 for the race subtype)
- What about the validity of the test for the proposed application?
 - Construct validity?
 - Criterion validity?
 - Meta-analysis by Forscher et al. (2019) found that
 - Implicit measures can be changed, but the changes are weak
 - Such changes don’t mediate changes in explicit measures or behaviour

References

Optional reading:

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. Washington, DC: American Educational Research Association.

Australian Psychological Society. (2018). Ethical guidelines for psychological assessment and the use of psychological tests. Retrieved 15 May 2019 from <https://www.psychology.org.au/getattachment/Membership/APS-Ethical-Guidelines/Ethical-guideline-psych-ax.pdf>

Brown, W. (1910). Some experimental results in the correlation of mental abilities. *British Journal of Psychology*, 3, 296-322.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin*, 56, 81-105.

Cohen, R. J., & Swerdlik, M. E. (2005). *Psychological testing and assessment: An introduction to tests and measurement* (6th ed.). Boston, MA: McGraw-Hill. (Chapter 6).

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological Bulletin*, 52, 281-302.

Forscher, P. S., Lai, C. K., Axt, J. R., Ebersole, C. R., Herman, M., Devine, P. G., & Nosek, B.A. (2019). A meta-analysis of procedures to change implicit measures. *Journal of Personality and Social Psychology*, 117, 522-559.

Furr, R. M. (2021). *Psychometrics: An introduction*. Los Angeles, CA: Sage.

Greenwald, A. G., McGhee, D. E., & Schwartz, J. L. (1998). Measuring individual differences in implicit cognition: the implicit association test. *Journal of Personality and Social Psychology*, 74, 1464-1480.

Lawshe, C. H. (1975). A quantitative approach to content validity. *Personnel Psychology*, 28, 563-575.

Lissitz, R. W., & Samuels, K. (2007). A suggested change in terminology and emphasis regarding validity and education. *Educational Researcher*, 36, 437-448.

Markus, K. A., & Borsboom, D. (2013). *Frontiers of test validity theory: Measurement, causation, and meaning*. New York, NY: Routledge.

Messick, S. (1989). Validity. In Linn, R. L. *Educational Measurement* (3rd ed.). New York, NY: Macmillan.

Newton, P., & Shaw, S. (2014). *Validity in educational and psychological assessment*. London, UK: Sage.

Nunnally, J. C., & Bernstein, I. H. (1994). *Psychometric theory* (Chapter 7). New York, NY: McGraw-Hill.

Sackett, D. L., Straus, S. E., Richardson, W. S., Rosenberg, W., & Haynes, R. B. (2000). *Evidence-based Medicine: How to Practice and Teach EBM*. Edinburgh, UK: Churchill Livingstone.

Spearman, C. (1910). Correlation calculated from faulty data. *British Journal of Psychology*, 3, 271-295.