

Psychometrics and test construction

Supported GRE teaching and original transfer practice

Identify the evidence

Identify the evidence

Distinguish reliability types by the design that produced them

Use the overview measurement lesson VI for research inference.

Concepts 1 (1)

Concepts 1 (1)

Reliability is consistency, and its type is named by the design that produced it: the same test twice gives test-retest reliability; two similar forms give alternate-forms reliability; consistency within one sitting (split-half, Cronbach alpha) is internal consistency; agreement between scorers is inter-rater reliability. The Spearman-Brown formula $r_{\text{full}} = 2r / (1 + r)$ rescales a half-test correlation to full test length: $r = .80$ becomes about .89.

Concepts 1 (2)

Concepts 1 (2)

Validity is evidence for the proposed interpretation, not of the test itself: content validity asks whether items cover the domain; criterion validity relates scores to an outcome, concurrently or predictively; construct validity assembles convergent and discriminant evidence that the score tracks the intended trait. Reliability caps validity but never establishes it: a consistent scale can systematically measure something else.

Concepts 2

Concepts 2

Standard scores re-express raw scores against a norm group: $z = (X - \text{mean}) / \text{SD}$, $T = 50 + 10z$, and Wechsler deviation IQ $= 100 + 15z$. A z of $+2$ gives IQ 130 and about the 98th percentile. The standard error of measurement $\text{SEM} = \text{SD} \times \sqrt{1 - r}$ converts reliability into a score band: $\text{SD} = 15$ with $r = .91$ gives $\text{SEM} = 15 \times \sqrt{.09} = 4.5$.

Item analysis checks each item: difficulty p is the proportion answering correctly, and discrimination is the item-total correlation. Items that nearly everyone or no one answers correctly carry little information; negative discrimination flags a likely miskeyed or misleading item.

Reason through the worked case

Reason through the worked case

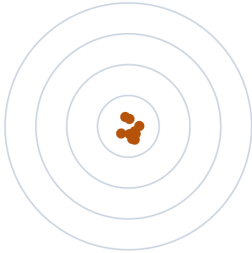
A fictional scale has mean 50 points, SD 10 points and reliability .84. A score is 65. Compute z and SEM, showing the formulas.

$z = (X - \text{mean}) / \text{SD}$; $z = (65 \text{ points} - 50 \text{ points}) / (10 \text{ points}) = 1.5$, dimensionless.

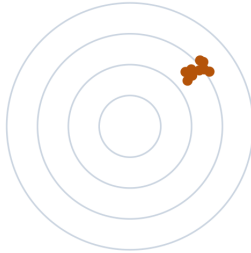
$\text{SEM} = \text{SD} \sqrt{1 - r}$; $\text{SEM} = (10 \text{ points}) \sqrt{1 - .84} = 4 \text{ points}$. SEM is a model-based uncertainty measure, not proof of validity. $T = 50 + 10z$; $T = 50 + 10(1.5) = 65$ on the T scale.

Read the reviewed representation

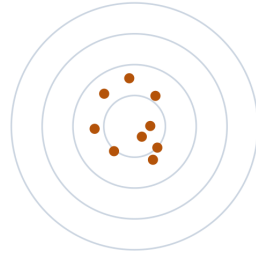
Reliable + valid



Reliable, not valid



Unreliable



Original schematic for Psychometrics and test construction. Illustrative relationships and numbers are not research results.

Attempt an unfamiliar case

Attempt an unfamiliar case

Two raters repeatedly disagree by five points but rank scripts identically. Does a high correlation establish agreement? Suggest a check.

Pause and write the explanation before continuing.

Check the transfer argument

Check the transfer argument

No. Rank association can remain perfect with systematic score differences. Inspect paired differences and an agreement measure suited to the use, with a clear rubric and rater training. Correlation alone does not establish interchangeable scores. Distinguish consistency from absolute agreement.

Accept a different explanation when its evidence and limits are justified.

Limit the conclusion

Limit the conclusion

Calling a two-different-forms correlation test-retest reliability, or citing high reliability as if it were validity evidence.

Original classroom cases are not official questions or personal clinical advice.

Check yourself

Independent retrieval

A highly reliable score predicts poorly in a new population. Explain why reliability did not guarantee this validity claim.

Write your reasoning before comparing it with the methods and limits taught in this lesson.