

10.1 Psychometrics and test construction: evidence and transfer

Name: _____ Class: _____ Date: _____

Total: 8 marks

KEY WORDS standard error of measurement • construct validity

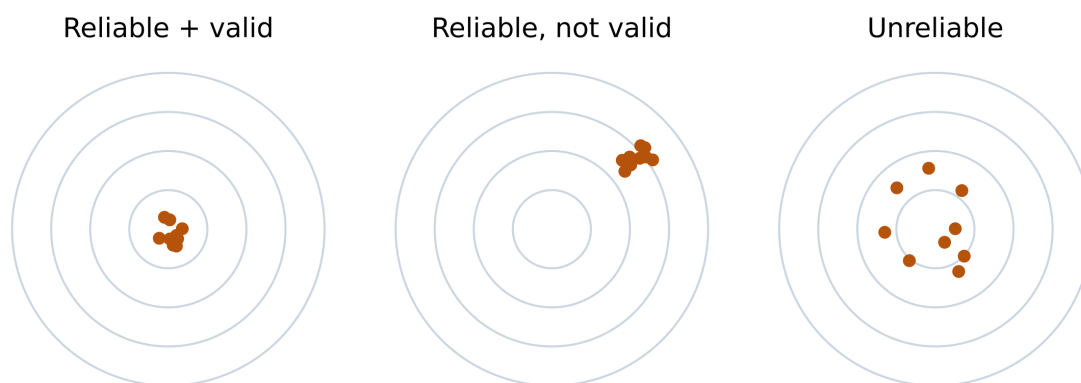
Objective

Distinguish reliability types by the design that produced them; Separate content, criterion and construct validity evidence; Compute standardised scores and the standard error of measurement. Original practice and authored marking guidance.

1 Key knowledge and worked examples

standard error of measurement : The expected spread of observed scores around the true score, $SD \times \sqrt{1-r}$. **construct validity** : Convergent and discriminant evidence that a score reflects the intended trait.

Choose the reliability coefficient by the source of variation being studied. A standard error of measurement concerns measurement uncertainty under a model; it is not the standard error of a sample mean. The Spearman–Brown estimate assumes sufficiently comparable halves; a formula cannot repair unrelated item content. Standard scores describe relative position and do not show that a test measures the intended construct fairly across groups.



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.

$z = (X - \text{mean}) / SD$; $z = (65 \text{ points} - 50 \text{ points}) / (10 \text{ points}) = 1.5$, dimensionless.
 $SEM = SD \sqrt{1-r}$; $SEM = (10 \text{ points}) \sqrt{1-.84} = 4 \text{ points}$. SEM is a model-based un-

certainty measure, not proof of validity. $T=50+10z$; $T=50+10(1.5)=65$ on the T scale.

2 Practice

2.1 Odd and even halves of a test correlate $r=.80$. After Spearman-Brown correction, the estimated full-test reliability is about: [1]

A. 0.8 B. 0.85 C. 0.89 D. 0.91 E. 0.96

2.2 Evidence that a score tracks the intended trait rather than a different one is primarily: [1]

A. high test-retest reliability B. high internal consistency C. convergent and discriminant correlations D. a large normative sample E. high mean item difficulty

2.3 Revisit the worked case. State one conclusion its evidence supports and one stronger conclusion it does not support. Explain the difference. [2]

3 Exam-style questions

3.1 Two raters repeatedly disagree by five points but rank scripts identically. Does a high correlation establish agreement? Suggest a check. Give the interpretation, supporting reasoning, a relevant alternative or limitation, and an appropriate check. [4]
