

Exercise walk-through

10.1

Psychometrics and test construction: evidence and transfer

GRE · GRE Subject Test · GRE Psychology

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.

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.

Exercises 2.1·2.2

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

Exercises 2.1·2.2

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

2.1 Answer: C. $r_{\text{full}}=2r/(1+r)=1.6/1.8\approx.89$. The other options do not satisfy the stated defining conditions.

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

2.2 Answer: C. Construct validity rests on convergent and discriminant evidence; the others are reliability or sampling properties. Compare the exact process, design or evidence, rather than an associated keyword.

Exercises 2.3-3.1

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

3.1 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]

Exercises 2.3-3.1

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

2.3 Use the worked case evidence to support a bounded claim and explain why a causal, universal or stronger claim needs additional evidence. The claim must fit the specific design and observed process.

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]

3.1 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. Award B1 each (authored) for the interpretation, its reasoning, a relevant limitation/alternative and an appropriate check, up to four points. Accept a different justified answer satisfying these demands.