Psychometrics and test construction

Supported GRE teaching and original transfer practice

Psychometrics and test construction

- Name the kind of reliability from the design that produced it.
- Use the Spearman–Brown formula and the **standard error of measurement** 测量标准误.
- Separate content, criterion and **construct validity** 构想效度.
- Convert between z-scores, T-scores and deviation IQ, and read item statistics.

Reliability

- Test–retest: the same test given twice.
- Alternate forms: two similar versions of the test.
- Internal consistency: agreement among items in one sitting (split-half, Cronbach's alpha).

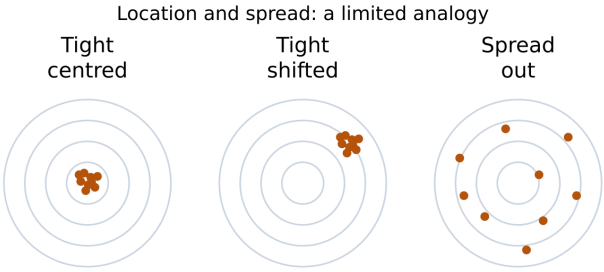
Reliability (2)

- Inter-rater: agreement between scorers. A high correlation shows that raters rank people the same way, not that they give the same scores.
- Spearman–Brown: a half-test correlation r predicts a full-test reliability $\dfrac{2r}{1+r}$.

Validity

- Content validity: the items cover the domain fairly.
- Criterion validity: scores predict an outcome, now (concurrent) or later (predictive).
- **Construct validity**: convergent evidence (correlates with related measures) and discriminant evidence (does not correlate with unrelated ones).
- Reliability limits validity, but never establishes it.

Figure: Validity



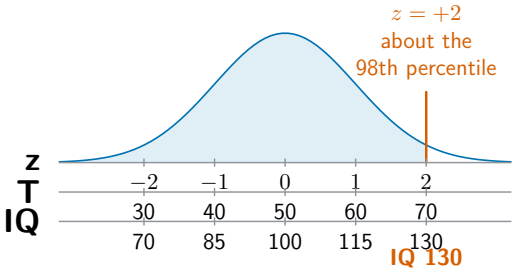
- Tight grouping is like reliability, being on target like validity: a test can be reliable but not valid
- Match each part of the diagram to the text.

Standard scores

- $z = +2$ gives $T = 70$ and IQ 130, about the 98th percentile.
- About 68% of scores lie within 1 SD of the mean, and 95% within 2 SD.
- Standard scores describe relative position. They do not show that a test is fair across groups.

$$z = \frac{X - \overline{X}}{SD}, \qquad T = 50 + 10z, \qquad IQ = 100 + 15z$$

Figure: Standard scores



- One curve, three scales: $T = 50 + 10z$ and $IQ = 100 + 15z$.
- Read the axes and units first.

Measurement error

- The **standard error of measurement** describes how far observed scores spread around a person's true score.
- A band of ± 1 SEM around a score covers the true score about 68% of the time; ± 2 SEM about 95%.
- It is not the standard error of a sample mean.

$$SEM = SD\sqrt{1 - r}$$

Item analysis

- Item difficulty p: the proportion who answer correctly. Items with p near 0 or 1 tell you little.
- Item discrimination: the correlation between the item and the total score. A negative value suggests a wrong key or a misleading item.

Worked example: reliability

- Two similar vocabulary tests a week apart correlate at 0.90: alternate-forms reliability, because the form changed.
- A split-half correlation of 0.80 predicts a full-length reliability of $2(0.80)/1.80 = 0.89$.
- With SD $= 15$ and $r = 0.91$: $SEM = 15\sqrt{0.09} = 4.5$ points.

Worked example: a score

- $z = (65 - 50)/10 = 1.5$; $T = 50 + 10(1.5) = 65$.
- $SEM = 10\sqrt{0.16} = 4$ points, so a 95% band is about 57 to 73.
- The band describes measurement uncertainty, not validity.

Common errors

- Calling a two-form correlation test–retest reliability.
- Citing high reliability as validity evidence.
- Reading a high inter-rater correlation as exact agreement.

Common errors (2)

- Confusing the SEM with the standard error of a mean.
- Keeping items with negative discrimination.

Key terms

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

Check yourself

- What IQ corresponds to $z = -1$?
- A test has SD 20 and reliability 0.75. What is its SEM?
- A new anxiety scale correlates 0.7 with an older anxiety scale and 0.05 with height. What kind of evidence is this?