Reliability and replicability

Lesson 2.4.4 | AS research methodology

Core ideas

Reliability is about consistency in procedures, measures or findings.

A reliable measure gives similar results when used in the same way under similar conditions.

Inter-observer reliability checks whether different observers record behaviour in similar ways.

Test-retest reliability checks whether a measure gives similar results across time when the target should be stable.

Replicability means a study can be repeated to see whether similar findings are found again.

High reliability supports confidence in evidence, but a reliable measure still needs to be valid.

Key terms

Reliability: Consistency of a measure, procedure or finding.

Replicability: The ability to repeat a study and check whether similar findings occur.

Inter-observer reliability: Agreement between observers recording the same behaviour.

Reliability and replicability

Lesson 2.4.4 | AS research methodology

Continued

Test-retest reliability: Consistency of scores when the same measure is used again later.

Standardised procedure: A procedure kept the same across participants or repeats.

Consistency: Similarity across measurements, observers or occasions.

Measurement error: Inaccuracy or variation caused by the measuring process.

Validity: How well a study or measure captures what it claims to study.

Further detail and evidence

Inter-rater reliability is agreement between people judging or scoring the same material; inter-observer reliability applies to observers recording behaviour. Describe how agreement would be checked and improve unclear categories if agreement is poor.

Inter-rater reliability is agreement between people scoring the same material; inter-observer reliability concerns behaviour recording. Test-retest reliability assesses consistency across time when the target should be stable. Replicability requires enough procedural detail for another researcher to repeat the study.

Reliability and replicability

Lesson 2.4.4 | AS research methodology

Applying the ideas

Train independent observers using the same category definitions and compare coding of the same events. If agreement is poor, identify ambiguous categories and recheck after refinement. For a stable questionnaire construct, repeat the same measure after a suitable interval while considering memory effects.

Worked example: A playground observation uses two observers. How could they check reliability? The observers could independently record the same playground behaviour using the same behavioural categories, then compare how often their records match. High agreement would suggest good inter-observer reliability. If agreement is low, the categories may need clearer definitions or the observers may need more training.

Worked example: Why is a replicable study more useful than a one-off finding? A replicable study is more useful because other researchers can repeat the procedure and check whether similar results occur again. If findings appear across repeated studies, psychologists can have more confidence that the result was not just chance, bias or a one-time mistake. Replicability strengthens evidence, although the original measure still needs to be valid.

Evaluation and connections

Agreement does not establish validity: two observers may share the same flawed interpretation. Change across time can be genuine rather than unreliability. Standardised instructions and scoring support consistency; reporting only "it is reliable because it is a lab" does not explain a check or its result.