AQA A-Level Psychology: Reliability and Validity

Complete Guide to Research Methods Specification Points

This comprehensive guide covers all required content for the AQA A-level Psychology specification on Reliability and Validity across all methods of investigation. Each section includes detailed AO1 knowledge and understanding, three-level AO3 evaluation with clear "SO" statements showing the significance of each point.

4.2.3.1 Reliability Across All Methods of Investigation

4.2.3.1.1 AO1: Knowledge and Understanding

Reliability refers to the consistency of measurements provided by a research technique or test. A reliable measure should produce consistent results when repeated under the same conditions.

Key Characteristics of Reliability:

- **Consistency**: The measure produces similar results on different occasions
- **Stability**: Results remain stable over time and across different researchers
- **Repeatability**: The same findings emerge when the study is replicated

Reliability Across Different Research Methods:

Experiments: Reliability concerns whether the same results would be obtained if the experiment were repeated with the same participants under identical conditions.

Observations: Reliability involves whether different observers would record the same behaviors when watching the same events (inter-observer reliability) and whether the same observer would record consistently across different time periods (intra-observer reliability).

Self-Report Methods: Reliability refers to whether questionnaires or interviews produce consistent responses from the same participants on different occasions.

Correlational Studies: Reliability involves whether the same relationship between variables would be found if the study were repeated with different samples.

4.2.3.1.2 Ways of Assessing Reliability

Test-Retest Reliability

Definition: A method of assessing reliability by administering the same test to the same participants on two separate occasions and measuring the correlation between the two sets of scores.

Procedure:

1. Administer the test/measure to a group of participants
2. Wait for an appropriate time interval (typically 2-4 weeks)
3. Re-administer the same test to the same participants under identical conditions
4. Calculate the correlation coefficient between the two sets of scores
5. A correlation of +0.8 or higher is generally considered to indicate good reliability

Example: A researcher develops a new anxiety questionnaire. They give it to 100 students, wait 3 weeks, then give the same questionnaire to the same students again. If the correlation between the two sets of scores is +0.85, this indicates good test-retest reliability.

Inter-Observer Reliability

Definition: A method of assessing reliability in observational studies by having two or more observers independently record the same behaviors and measuring the agreement between their observations.

Procedure:

1. Two or more observers use the same behavioral categories/coding system
2. Observers receive training to ensure they understand the coding system consistently
3. Observers independently record behaviors during the same observation period
4. The level of agreement between observers is calculated (often using correlation)
5. A correlation of +0.8 or higher indicates good inter-observer reliability

Example: In a study of children's aggressive behavior in playgrounds, two observers independently record instances of physical aggression using the same behavioral checklist. Their observations are then compared to calculate inter-observer reliability.

4.2.3.1.3 Improving Reliability

Improving Test-Retest Reliability:

- Use appropriate time intervals (not too short to avoid practice effects, not too long to avoid genuine change)
- Ensure identical testing conditions on both occasions
- Use standardized procedures and instructions
- Control for external factors that might influence performance

Improving Inter-Observer Reliability:

- Provide comprehensive training for all observers

- Use clear, operational definitions for all behavioral categories
- Conduct pilot observations to identify and resolve disagreements
- Regular reliability checks throughout the study
- Use video recordings to allow repeated analysis

General Methods to Improve Reliability:

- **Standardization**: Use identical procedures, instructions, and conditions
- **Clear operationalization**: Define variables precisely and unambiguously
- **Pilot studies**: Test procedures beforehand to identify potential problems
- **Training**: Ensure all researchers understand procedures consistently
- **Environmental control**: Minimize external factors that could affect results

---

4.2.3.1.4 AO3 Evaluation: Reliability

Level 1 – Assessment Methods Critique

Strength: Test-retest reliability provides an objective, quantifiable measure of consistency that can be statistically analyzed. **SO**: This allows researchers to make evidence-based decisions about whether their measures are sufficiently reliable for research purposes, with the +0.8 threshold providing clear guidance for acceptability.

Limitation: Test-retest reliability may be affected by practice effects, where participants perform differently on the second administration due to familiarity with the test. **SO**: This means that improvements in performance between test occasions might be mistakenly attributed to poor reliability rather than genuine learning, potentially leading to rejection of actually reliable measures.

Strength: Inter-observer reliability ensures that behavioral observations are not dependent on the subjective interpretation of a single observer. **SO**: This increases confidence that observed behaviors reflect genuine phenomena rather than observer bias, making findings more credible and scientifically valid.

Limitation: High inter-observer reliability can sometimes be achieved by using overly simplistic behavioral categories that miss important nuances. **SO**: This creates a tension between reliability and validity, as the most reliable measures may not capture the complexity of human behavior that researchers actually want to study.

Level 2 – Methodological and Conceptual Issues

Strength: Reliability is essential for the replication of psychological research, which is fundamental to the scientific process. **SO**: This ensures that psychological findings can be verified independently by other researchers, building a cumulative body of knowledge and increasing confidence in psychological theories.

Limitation: The emphasis on reliability can lead to a preference for easily measurable behaviors over more complex psychological constructs. **SO**: This may result in psychology focusing on superficial aspects of human behavior rather than deeper psychological processes, potentially limiting the theoretical advancement of the field.

Strength: Standardized reliability assessment procedures allow for comparison between different studies and measures within psychology. **SO**: This enables researchers to make informed choices about which measures to use and allows for meta-analyses that combine results across studies with known reliability levels.

Limitation: Reliability does not guarantee validity – a measure can be consistently wrong. **SO**: This means that researchers must balance the pursuit of reliable measures with ensuring those measures actually assess what they claim to measure, requiring careful consideration of both concepts together.

Level 3 – Real-World Applications and Theoretical Implications

Strength: Reliable psychological assessments are crucial for clinical diagnosis and treatment planning in mental health services. **SO**: This ensures that patients receive consistent diagnoses and appropriate treatments, with unreliable assessments potentially leading to misdiagnosis and inappropriate interventions that could harm patient wellbeing.

Limitation: Over-emphasis on reliability in applied settings may lead to rigid adherence to standardized procedures that don't account for individual differences. **SO**: This could result in important individual variations being overlooked in clinical, educational, or occupational settings, potentially reducing the effectiveness of interventions tailored to specific needs.

Theoretical Implication: The concept of reliability reflects psychology's attempt to apply scientific principles to human behavior study. **SO**: This demonstrates psychology's commitment to scientific rigor, but also highlights the challenge of studying complex, dynamic human behavior using methods designed for more predictable physical phenomena.

4.2.3.2 Types of Validity Across All Methods of Investigation

4.2.3.2.1 AO1: Knowledge and Understanding

Validity refers to whether a research method measures what it claims to measure or whether the conclusions drawn from research are legitimate and well-founded.

Face Validity

Definition: The extent to which a test or measure appears, on the surface, to assess what it claims to measure.

Characteristics:

- Based on subjective judgment rather than statistical analysis
- Evaluated by examining whether the test content seems relevant to the construct
- Often assessed by experts in the field or target population

Example: A questionnaire measuring depression that includes items about sadness, hopelessness, and sleep problems would have good face validity because these symptoms are commonly associated with depression.

Concurrent Validity

Definition: The extent to which a new test or measure correlates with an established measure of the same construct when both are administered at approximately the same time.

Characteristics:

- Involves comparison with a "gold standard" or well-established measure
- Assessed through correlation analysis
- Demonstrates that the new measure produces similar results to existing valid measures

Example: A new intelligence test would have good concurrent validity if scores correlate highly (+0.8 or above) with scores on an established IQ test like the WAIS-IV when both tests are given to the same participants.

Ecological Validity

Definition: The extent to which research findings can be generalized to real-world, everyday situations and environments.

Characteristics:

- Concerns the realism of research settings and tasks
- High when research conditions closely resemble natural environments
- Important for determining practical applications of research findings

Example: A memory study conducted in a realistic classroom setting would have higher ecological validity than one conducted in a sterile laboratory environment, making findings more applicable to educational contexts.

Temporal Validity

Definition: The extent to which research findings remain valid and applicable across different time periods.

Characteristics:

- Concerns whether findings are historically bound or generalizable across time
- High when research findings remain consistent despite social, cultural, or technological changes
- Important for determining the lasting value of psychological research

Example: Early studies of conformity from the 1950s may have lower temporal validity today due to changes in social attitudes, individualism, and cultural values that affect how people respond to group pressure.

4.2.3.2.2 Assessment of Validity

Assessing Face Validity

- Expert judgment: Specialists in the field examine the measure
- Target population feedback: Ask intended users whether items seem relevant
- Content analysis: Systematic review of test items for relevance
- Peer review: Independent researchers evaluate the measure's apparent validity

Assessing Concurrent Validity

- Correlation with established measures: Calculate correlation coefficient between new and established tests
- Criterion comparison: Compare performance on new measure with known outcomes
- Statistical analysis: Use correlation values (+0.8 or higher typically considered good)
- Cross-validation: Test concurrent validity across different samples

Assessing Ecological Validity

- Environmental realism: Compare research setting to natural environments
- Task relevance: Evaluate whether research tasks resemble real-world activities
- Population representativeness: Assess whether sample reflects target population
- Outcome generalizability: Test whether findings apply in natural settings

Assessing Temporal Validity

- Longitudinal studies: Repeat research across different time periods
- Cross-temporal meta-analysis: Compare findings from studies conducted at different times

- Replication studies: Conduct identical studies in contemporary settings
- Historical analysis: Examine whether findings remain consistent over time

---

4.2.3.2.3 Improving Validity

Improving Face Validity

- Use clear, relevant item wording that obviously relates to the construct
- Consult experts and target populations during measure development
- Avoid ambiguous or confusing language
- Ensure comprehensive coverage of the construct being measured

Improving Concurrent Validity

- Compare new measures with multiple established "gold standard" tests
- Use appropriate statistical procedures for correlation analysis
- Ensure both measures are administered under similar conditions
- Test concurrent validity across diverse samples

Improving Ecological Validity

- Conduct research in naturalistic settings when possible
- Use realistic tasks and materials
- Include diverse, representative samples
- Consider cultural and contextual factors

Improving Temporal Validity

- Regularly update and re-validate measures
- Account for historical and cultural changes in research design
- Use contemporary samples and settings
- Consider cohort effects in longitudinal research

4.2.3.2.4 AO3 Evaluation: Types of Validity

Level 1 – Individual Validity Types Critique

Strength (Face Validity): Face validity provides an initial, accessible check of whether a measure seems appropriate and relevant. **SO**: This enables researchers to quickly identify obviously inappropriate measures and helps ensure participant cooperation by using measures that appear meaningful and relevant to the research purpose.

Limitation (Face Validity): Face validity is subjective and may not reflect the actual validity of a measure – something that looks valid may not actually be valid. **SO**: This means researchers cannot rely solely on face validity when selecting measures, as superficial appearance may mask fundamental problems with what the measure actually assesses.

Strength (Concurrent Validity): Concurrent validity provides objective, statistical evidence that a measure assesses the same construct as established tests. **SO**: This gives researchers confidence that new measures are scientifically sound and allows for meaningful comparison of results across studies using different but concurrently valid measures.

Limitation (Concurrent Validity): Concurrent validity assumes that existing "gold standard" measures are themselves valid, which may not always be the case. **SO**: This creates circular reasoning where new measures are validated against potentially flawed existing measures, perpetuating methodological problems rather than improving measurement quality.

Strength (Ecological Validity): Research with high ecological validity provides findings that are more likely to apply in real-world situations. **SO**: This ensures that psychological research can inform practical applications and interventions, making psychology more relevant to addressing real-world problems and improving people's lives.

Limitation (Ecological Validity): Pursuing high ecological validity often requires sacrificing experimental control, potentially reducing internal validity. **SO**: This creates a methodological dilemma where researchers must balance realism against precision, potentially compromising either the applicability or the scientific rigor of their findings.

Strength (Temporal Validity): Research with high temporal validity provides findings that remain useful across different historical periods. **SO**: This ensures that psychological theories and interventions maintain their relevance over time, providing lasting value for the field and preventing the need for constant replacement of outdated findings.

Limitation (Temporal Validity): Assessing temporal validity requires long-term studies that are expensive, time-consuming, and difficult to conduct. **SO**: This means that many psychological findings may have unknown temporal validity, limiting confidence in their long-term applicability and potentially leading to inappropriate application of outdated research.

Level 2 – Methodological and Conceptual Issues

Strength: The different types of validity address different aspects of measurement quality, providing a comprehensive framework for evaluation. **SO**: This allows researchers to systematically assess measures from multiple perspectives, ensuring robust evaluation and helping identify specific areas where measures might be improved or strengthened.

Limitation: The various types of validity can sometimes conflict with each other, creating methodological tensions in research design. **SO**: This forces researchers to make difficult trade-offs between different aspects of validity, potentially compromising overall research quality and making it difficult to achieve optimal measurement across all validity dimensions.

Strength: Validity assessment provides standardized criteria for evaluating research quality across different areas of psychology. **SO**: This enables meaningful comparison of research quality across studies and helps establish standards for publication, funding, and practical application of psychological research.

Limitation: The concept of validity itself is complex and contested, with different researchers emphasizing different aspects of validity. **SO**: This creates inconsistency in how validity is assessed and reported, making it difficult to compare studies and potentially leading to disagreements about research quality within the psychological community.

Level 3 – Real-World Applications and Theoretical Implications

Strength: Valid measures are essential for evidence-based practice in clinical, educational, and occupational psychology. **SO**: This ensures that psychological interventions and assessments are based on sound measurement principles, protecting the public from ineffective or harmful practices and maintaining professional credibility.

Limitation: The emphasis on validity may lead to over-reliance on quantitative measures that can demonstrate statistical validity but miss important qualitative aspects of human experience. **SO**: This could result in a narrow understanding of psychological phenomena, potentially overlooking valuable insights from more interpretive approaches and limiting the richness of psychological theory and practice.

Theoretical Implication: The concept of validity reflects fundamental questions about the nature of psychological constructs and whether they can be objectively measured. **SO**: This highlights ongoing philosophical debates within psychology about realism versus constructivism, influencing how the field approaches research methodology and theoretical development.

Practice Questions with Real AQA Mark Schemes

Question 1 (4 marks): *From AQA June 2022 Paper 2*

Question: Explain how the reliability of the controlled observation could be assessed through inter-observer reliability.

Mark Scheme Answer: Award 1 mark for each of the following points:

- two observers would use same behavioural categories/discuss and agree on an interpretation of each of the social behaviours in the category system

- two observers would make independent observations/tallies (of the same child at the same time/the 5-minute sessions are filmed and each observer watches and records the data for each film)
- the two observers' tally charts would be compared to check for agreement/calculate the correlation between the recordings of the two observers to determine the level of inter-observer reliability
- researchers generally accept +0.8 correlation as a reasonable degree of reliability.

Note - For responses with no explicit application a maximum of 3 marks can be awarded.

Question 2 (4 marks): *From AQA June 2018 Paper 2*

Question: The psychologist wanted to assess the reliability of the content analysis. Explain how the reliability of the content analysis could be assessed.

Mark Scheme Answer:

Test-retest reliability - Award 1 mark for each of the following points:

- content analysis repeated on a second occasion using the same interview data
- compare the results of the two separate analysis (number of occurrences of each)
- researchers could calculate the correlation between the two ratings
- researchers generally accept 0.8 correlation (accept 0.7-0.9) between the test and the re-test.

Inter-rater reliability - Award 1 mark for each of the following points (up to 4 marks):

- use a second person to work with the original researcher
- they could read the interviews (separately) and devise a set of categories (and agree operational definitions)
- they could tally the occurrences of each of the categories of the interviews (separately)
- they could compare their tally charts looking for agreement
- researchers could calculate the correlation between the two ratings
- researchers generally accept 0.8 correlation (accept 0.7-0.9) between the test and the re-test.

Question 3 (2 marks): *From AQA June 2019 Paper 2*

Question: Explain what it means for a test to have high concurrent validity.

Mark Scheme Answer: 2 marks for a clear and appropriate definition such as:

- the test produces similar results/correlates highly with another established/accepted test that measures the same thing
- when both tests are taken at the same/similar time

1 mark for a limited definition that shows some understanding.

Question 4 (2 marks): *From AQA June 2019 Paper 2*

Question: Briefly explain one method the psychologist could use to check the validity of the data she collected in this study.

Mark Scheme Answer: 2 marks for a clear and detailed explanation applied to this study. 1 mark for a partial or muddled explanation or one that is only loosely applied to the study.

Credit answers based on any type of validity. Most answers will refer to either face or concurrent as follows:

- asking other people if verbal errors are a good measure of verbal fluency (face validity)
- giving participants an alternative/established verbal fluency test and checking to see that the two sets of data are positively correlated (concurrent validity).

Question 5 (16 marks): Essay Question

Question: Discuss the importance of reliability and validity in psychological research. Refer to ways of assessing and improving both reliability and validity in your answer.

Suggested Essay Structure:

Introduction: Define reliability and validity, explaining their fundamental importance in psychological research.

AO1 Content (8 marks):

- Definition and explanation of reliability (consistency)
- Ways of assessing reliability: test-retest and inter-observer
- Definition and explanation of validity (measuring what you claim to measure)
- Types of validity: face, concurrent, ecological, temporal
- Methods for improving both reliability and validity

AO3 Evaluation (8 marks):

- Strengths and limitations of different assessment methods
- The relationship between reliability and validity

- Real-world applications and importance for evidence-based practice
- Methodological tensions and trade-offs between different types of validity
- Examples from psychological research demonstrating importance

Conclusion: Synthesize the importance of both concepts for scientific psychology.

Question 6 (8 marks):

Question: Evaluate the use of inter-observer reliability as a way of assessing the reliability of observational research.

Mark Scheme Guidance:

- **AO3 = 8 marks**
- **Level 4 (7-8 marks)**: Evaluation is thorough and effective. The answer demonstrates sound analysis and understanding. The answer is well focused and shows coherent elaboration. Specialist terminology is used effectively.
- **Level 3 (5-6 marks)**: Evaluation is mostly effective. The answer demonstrates reasonable analysis and understanding. The answer is mostly focused and shows reasonable elaboration. Specialist terminology is mostly used appropriately.
- **Level 2 (3-4 marks)**: Evaluation is partially effective. The answer demonstrates basic, superficial understanding. The answer is sometimes focused and shows some evidence of elaboration. Specialist terminology is used inappropriately on occasions.
- **Level 1 (1-2 marks)**: Evaluation is limited. The answer demonstrates very little understanding. The answer occasionally lacks focus and shows little evidence of elaboration. Specialist terminology is either absent or inappropriately used.

Potential evaluation points:

- Objectivity vs. subjectivity of behavioral categories
- Practical challenges in training observers
- Statistical thresholds for acceptable reliability
- Relationship between reliability and validity in observations
- Cost and time implications
- Alternative approaches to assessing observational reliability