

Cambridge AS & A Level · Psychology 9990 · Notes

Validity, reliability and replicability

AS

Read online at pastopic.com/cambridge/psychology-9990/notes/validity-and-reliability

Last checked 6 October 2026

Read first: [Research methods](#), [Aims, hypotheses and variables](#)

© 2026 Pastopic. Free for personal study. Please share the link, not the file.

What you need to know

Validity and reliability are tested on almost every Paper 2, in short questions about the core studies and about new studies, and they are the most common evaluation points in Paper 1. By the end of this topic you should be able to:

1. **Describe validity**, including **ecological validity**: what it means for a study or a measure to be valid, and what makes findings apply to everyday life.
2. **Evaluate studies by their validity**, using **subjectivity and objectivity**, **demand characteristics** and **generalisability**.
3. **Apply validity to a new study**: say why a measure, category or procedure is or isn't valid, and suggest how to make it more valid.
4. **Describe the types of reliability**: **inter-rater**, **inter-observer** and **test-retest** reliability.
5. **Evaluate studies by their reliability**, and **apply reliability to a new study**: how to check it and how to improve it.
6. **Understand replicability** and use it when planning a study.

Understand it

Validity: measuring what you claim to measure

A study is **valid** when it really tests what it says it tests, and a measure is valid when it really measures the thing it is meant to. A "kindness" score that mostly picks up how chatty people are is not valid, however carefully it is used.

There are several kinds of validity. You need to be able to say which one you mean:

Kind	The question it asks	How to raise it
Validity of a measure (does it measure the right thing?)	Does this test, question or category capture the behaviour we want?	Clear operational definitions; questions that only fit one meaning; categories that can't be confused with other behaviours

Kind	The question it asks	How to raise it
Internal validity (did the IV cause the change?)	Is the difference in the DV due to the IV, or to something else?	Control extraneous variables; standardise the procedure; random allocation or matched pairs; reduce demand characteristics
Ecological validity	Do the findings apply outside the study, in other settings and in everyday life?	Realistic tasks, materials and settings (field rather than laboratory)
Generalisability (population validity)	Do the findings apply to people beyond this sample?	A larger, more varied sample (ages, cultures, jobs, genders)

Ecological validity is about how far findings from the setting where the study was done would apply to other settings, especially real life. A memory test in a quiet laboratory using lists of random letters may not tell us how people remember a shopping list in a noisy supermarket. Saying a study is "realistic" or "natural" is not a definition: the point is whether the **findings carry over** to other situations.

Three threats to validity named in the syllabus

Subjectivity and objectivity. Data is **objective** when it doesn't depend on who records it: a reaction time, a count of correct answers, a brain scan. It is **subjective** when it depends on someone's judgement or opinion: deciding whether a facial expression is "friendly", or interpreting an open answer. Subjective data can be biased by what the researcher expects to find, so it is less valid. Clear categories and closed questions make data more objective, but they can also miss what people really meant (open answers let people explain themselves).

Demand characteristics. Participants who work out the aim may change their behaviour to fit it (or to spoil it), so the DV shows what they think is wanted rather than their real behaviour. Researchers reduce this with a cover story, filler questions that hide the real items, an independent measures design so nobody sees both conditions, a single-blind procedure, or by observing people who don't know they are being studied.

Generalisability. Findings from a sample can only be applied to people who are like that sample. A study of students from one university, or one person in a case study, may not apply to other ages, cultures or jobs. When you discuss it, say **who** the findings might not apply to and **why** that group could behave differently.

Reliability: consistency

Reliability is the **consistency** of a procedure, task or measure: a reliable measure gives the same result each time it is used in the same way on the same thing. A bathroom scale that shows 60 kg, then 63 kg, then 58 kg for the same person standing still is unreliable.

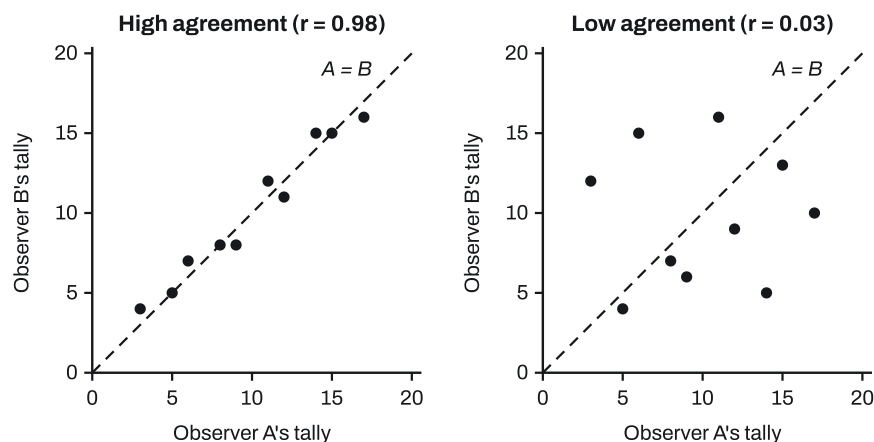
Reliability is about the **measure or procedure**, not about running the whole study again (that is replicability, below). It also matters for validity: if a measure is inconsistent, a difference between conditions could be caused by the measuring rather than by the IV, so the study can't show cause and effect. A measure can be reliable without being valid (it consistently measures the wrong thing), but an unreliable measure can't be valid.

Types of reliability

Inter-rater reliability is how far two or more researchers give the **same scores** when they score the **same data** independently, for example rating the same interview answers or essays on a scale. **Inter-observer reliability** is the same idea for observations: whether two or more observers watching the **same behaviour** at the same time (or on the same video) make the **same records**.

The key word is **agreement**. Having two observers doesn't make a study reliable on its own; it only lets you **check** reliability by comparing what they recorded. To check it, each observer records independently, then their scores for each participant are **correlated**. A strong positive correlation (researchers usually want about +0.8 or more) means high agreement.

Two observers' tallies for the same 10 children



In Bandura et al., the children's aggression was observed through a one-way mirror, with a record every 5 seconds over 20 minutes (240 response units). The male model scored every child and a second observer independently scored half of them; their scores agreed closely. That matters because the group differences in aggression can then be put down to the children's behaviour, not to how each observer judged it.

Test-retest reliability checks whether a measure gives the **same results with the same people on two occasions**. Give the same test, in the same way, to the same participants twice (with a gap, so they don't simply remember their answers, but not so long that the trait could really change), then correlate the two sets of scores. Similar scores (a strong positive correlation) mean the measure is reliable.

How to improve reliability

- **Operational definitions and behavioural categories**: define each category exactly (what counts and what doesn't), so everyone records the same thing.
- **Training and a pilot**: observers or raters practise together on the same video or transcripts, discuss differences, and repeat until their records match.
- **A standardised procedure**: same instructions, materials, timing, setting and questions for every participant.
- **Time sampling**: record at fixed intervals, as in Bandura et al.'s 5-second units.

- **Closed questions and rating scales:** answers need no interpretation, so different researchers score them the same way.

Checking reliability (correlating two raters, or retesting) shows **whether** a measure is consistent; it doesn't make it more consistent. To improve it you must change the procedure, the definitions or the training.

Replicability

A study is **replicable** if another researcher can **repeat it exactly**: the same procedure, materials, instructions and measures. Replicability comes from a **standardised** procedure that is **fully reported**. If a replication gives the same results, we can be more confident that the findings are real, not a one-off or the result of a mistake.

When planning a study, build replicability in: write down the exact instructions (a script), the operational definitions of variables, the materials, the timings and how participants are chosen and allocated. Case studies and unstructured interviews are hard to replicate, because the questions follow the person's answers and the person is unique.

Don't mix the two ideas up. "Some participants did different activities at home" is a problem for **reliability** (the procedure wasn't the same for everyone), not for replicability, because the study could still be run again.

Validity and reliability in the core studies

Core study	Feature	What it does
Milgram	Real-looking shock generator, a sample 45-volt shock given to each participant, the rigged draw, the experimenter's grey technician's coat, Yale University	Participants believed the shocks and the authority were real, so they behaved as they really would (internal validity, no guessing the aim)
Milgram	The same verbal prods, in the same order, for everyone	Standardised, so differences can't come from the experimenter (reliability, replicability)
Milgram	40 men from a range of jobs and ages	Wider generalisability than one group, though all were men from one area of the USA
Bandura et al.	Two observers; 5-second intervals; operationalised categories	Inter-observer reliability; objective, quantitative records
Dement and Kleitman	Told to avoid alcohol and caffeine, eat normally and arrive just before bedtime	Controls: these change sleep and dreaming, so the sleep was typical and changes in recall came from the stage of sleep
Fagen et al.	Performance test: pass at 80% correct	A clear standard (reliable), but the other 20% could each be slightly wrong or very wrong, and judging a response is partly subjective (validity)

Core study	Feature	What it does
Baron-Cohen et al.	Revised eyes test: 36 still black-and-white photos of the eye region, four words each, complex mental states	More valid than the original (fewer lucky guesses, no ceiling effect), but low ecological validity: real eyes move, and we normally see the whole face, body and context

Key facts and formulas

There is no formula list in the 9990 exam. "Learn it" means you must remember it.

Fact	Status
Validity: whether a study or measure tests what it claims to test	Learn it
Ecological validity: whether findings from the setting of a study apply to other settings, especially everyday life	Learn it
Internal validity: whether the change in the DV was caused by the IV, not by an extraneous variable	Learn it
Generalisability: whether findings from the sample apply to the wider population	Learn it
Objective data does not depend on who records it; subjective data depends on personal judgement	Learn it
Demand characteristics: participants guess the aim and change their behaviour to fit it	Learn it
Reliability: the consistency of a procedure, task or measure	Learn it
Inter-rater reliability: agreement between researchers scoring the same data independently	Learn it
Inter-observer reliability: agreement between observers recording the same behaviour independently	Learn it
Test-retest reliability: the same measure gives similar results with the same people on two occasions	Learn it
Agreement is checked by correlating the two sets of scores; about +0.8 or above counts as high	Learn it
Replicability: a study can be repeated exactly by others, because its procedure is standardised and fully reported	Learn it

How to do it

We have tagged 31 questions on this topic, all from Paper 2, from 2021 to March 2026. 8 of them mix more than one type (one has three), so the counts add up to 40.

Question type	Questions
Explain why a measure, category or task is (or is not) valid	8
Suggest how to improve validity (incl. ecological validity)	8
Define reliability, validity or ecological validity (with an example)	6
Explain how a control or procedure feature protects validity (extraneous variables, demand characteristics)	5
Describe, check or improve inter-rater / inter-observer / test-retest reliability	5
Explain why findings can or cannot be generalised from the sample	4
Explain why reliability matters, or how a feature makes a study reliable	3
Decide whether a stated problem is about reliability or validity	1

1. Explain why a measure, category or task is (or is not) valid (8 questions)

Usually 1 mark for a reason, or 2 for a reason plus a link to the study.

1. **Look at the measure itself:** what exactly is recorded, and how.
2. **Valid because** it clearly captures the behaviour: easy to see, hard to confuse with anything else, objective, or it lets people answer as they really feel.
3. **Not valid because** it could be recorded when the behaviour isn't happening, or missed when it is: the same action can mean something else, it is hard to see from where the observer is, the scoring is subjective, or the answer options force people into a choice that isn't true for them.

Variations you will meet:

- **A behavioural category in an observation.** Give a reason that fits **that** category, not "because it is non-verbal" (which repeats the question). Clapping is valid because it is loud and obvious; it may not be valid because people clap to keep warm or to get a dog's attention.
- **Open versus closed questions.** Open data can be **more** valid: people can explain why and say what they really mean, instead of choosing "yes" when they mean "sometimes". It can be **less** valid: answers must be interpreted, so they are subjective and open to researcher bias; closed answers need no interpretation.
- **A test with a pass mark** (Fagen et al.): question the **test**, not the animals. Passing at 80% correct means up to 1 response in 5 could be wrong, and the test doesn't show whether each of those was slightly wrong or very wrong; judging whether a response is correct is also partly subjective.
- **A rating scale or physiological measure:** a 10-point rating of "accuracy" is subjective (two raters might score the same answer 6 and 8); a measure such as heart rate is objective but can rise for other reasons (exercise, caffeine).

- **A problem other than generalisability**, for example asking adults about childhood: they may not remember accurately, so they may leave out real experiences or report ones that never happened.

2. Suggest how to improve validity (incl. ecological validity) (8 questions)

1. **Say the change** (1): what the researcher would do differently.
2. **Say how it helps** (1, sometimes 2): which problem it removes and why the data is then more valid.

Variations you will meet:

- **Differences between people in two groups** (participant variables): allocate participants to groups at random, so each kind of person is spread across both; or measure the variable first (a questionnaire, a pre-test) and use matched pairs, one of each pair in each condition.
- **A vague measure**: tighten the operational definition (say exactly what counts), or raise the pass standard (for example 90% or 100% correct instead of 80%).
- **Low ecological validity**: make the stimuli or setting more like real life. For a test using still photos of eyes: use video clips of moving eyes, or photos of the whole face and body. For a staged event: make the actor's appearance and behaviour more convincing.
- **A case study** (often 3 marks: the way + 2 details): triangulation (several sources, such as interviews with the person and family, observation and questionnaires, to see whether they agree); pre-validated, standardised tests; both qualitative and quantitative data; keeping a professional distance so the researcher stays objective.
- **A weakness of your own design**: name the problem in your procedure (an observer can't see every child at once), then the change (two observers, or film it) and why it helps. Ethics and sampling score 0 if the question excludes them.
- **Removing an extraneous variable** can create a new problem: excluding some participants biases the sample; stopping people's normal activities can upset them.

3. Define reliability, validity or ecological validity (with an example) (6 questions)

1. **Definition in your own words** (1). Reliability: the consistency of a procedure, task or measure. Validity: whether it tests what it claims to. Ecological validity: whether findings apply to other settings, especially everyday life.
2. **Example or link** (1): a study, real or made-up, where it is high or low. Piliavin et al. was carried out on real subway trains, so its findings are more likely to apply to everyday helping; Milgram's laboratory with a shock generator is far from everyday life.

6 marks: "Describe reliability and validity, using any examples." Up to 2 marks for each term's definition and detail (types of reliability, kinds of validity), up to 2 for each term's examples. Define each term, describe a type of each, then give an example of each.

Avoid: "reliability means getting the same results when the study is repeated" (that is replication), "trustworthy", "accurate", "the realism of a study".

4. Explain how a control or procedure feature protects validity (extraneous variables, demand characteristics) (5 questions)

1. **Name the feature** (1), with the detail from the study: Dement and Kleitman's participants avoided alcohol and caffeine.
2. **Explain how it protects validity** (1): what would otherwise have changed the DV (alcohol and caffeine affect sleep and dreaming, so their sleep would not be typical), or why the result can now be put down to the IV.

Variations you will meet:

- **A feature that made the situation believable** (Milgram): the shock generator looked real and each participant felt a real 45-volt sample shock, so they believed they were giving shocks; they were unlikely to guess the aim, so their obedience was real, not a response to demand characteristics. Saying "it looked real" on its own, or "authority figure" without a detail, isn't enough.
- **Participant differences between groups**: say what the difference is and how it would change the DV in one group (someone who loves films may lose track of time while watching one).
- **How demand characteristics could change the results** (1 for where the clue comes from, 1 for the effect), then **how to reduce them** without a false aim: filler questions, extra tasks that hide the real measure, an independent measures design.
- **A control that causes a new problem**: say the problem and link it to the study.

5. Describe, check or improve inter-rater / inter-observer / test-retest reliability (5 questions)

Describe (often 6 marks, "using any example(s)"): for each type, what it is (1) and a detail (1), plus examples (up to 2 per type). Without any example, or about one type only, the most is 4. Model points: **inter-rater**: how closely the scores of two or more researchers match when each scores the same data separately (1), for example the same interview transcripts rated on a scale (1). **Inter-observer**: how closely two or more observers' records match when they watch the same behaviour at the same time (1), each tallying the same behavioural categories in a structured observation (1). **Either**: if reliable, the two sets of scores correlate strongly and positively (1). **Example**: Bandura et al. had a second observer score half the children independently, and the two observers' scores agreed closely (1).

Check: both researchers score the **same** responses (or the same video) **independently**, then **correlate** their scores or compare them; strong agreement means high inter-rater reliability.

Test-retest (often 3 marks): the same measure is given to the same people twice, in the same way (1 + 1 for the detail); matching scores on the two occasions show that the measure is reliable (1).

Improve: give the raters precise operational definitions or categories, and have them practise together on the same recordings until their records match. "Add another observer" or "measure inter-rater reliability" earns nothing as an improvement.

6. Explain why findings can or cannot be generalised from the sample (4 questions)

1. **Give a reason linked to the behaviour studied** (1): why people unlike the sample might act differently.
2. **Give an example** (1): a specific group and how it would differ.

For a core study, link the spread of the sample to what was measured. In Milgram's study, **occupation** matters because some jobs involve following orders (soldiers and police officers must obey superiors), while others involve giving them (managers), so obedience could differ by job. **Age** matters because people of different ages are expected to obey to different degrees (children are expected to obey parents and teachers). Writing "to improve generalisability" just repeats the question and scores 0.

For a single case or an unusual sample: one person is unique, or lives in an unusual setting, so others in the same position might feel or behave differently. For animals: one species (fish) may not behave like others, or like humans.

7. Explain why reliability matters, or how a feature makes a study reliable (3 questions)

- **Why it matters in experiments** (2): an inconsistent measure adds its own random ups and downs to the DV, so a gap between conditions might be produced by the measuring rather than by the IV, and the study can't show that the IV caused it (1); for example, if Bandura et al.'s observers had scored inconsistently, the gap between the aggressive and non-aggressive model groups could have been due to who did the scoring (1).
- **A feature that makes a study reliable** (2): a standardised part of the procedure (1), with the detail from the study (1): every participant had the same instructions, the same materials or the same time limit.
- **Improve reliability** (1): any change that makes the procedure, task or measure more consistent (the same apparatus, room and timing for every participant; participants with the same experience of the task).

8. Decide whether a stated problem is about reliability or validity (1 question)

The mark is for the **reason**, not for picking the problem.

- **Reliability** if the problem is about **consistency**: the same answer can be scored in different ways, or the same person gets different scores each time.
- **Validity** if the problem is about **measuring the wrong thing**: an item meant to measure one thing really measures another.

Write "this is reliability **because** ..." every time.

Worked examples

Example 1

A researcher studies how quickly drivers react to a child running into the road. Participants sit at a desk and press a key when a cartoon child appears on a computer screen.

(a) Using this study, explain why its ecological validity may be low. [2]

(b) Explain the difference between reliability and replicability. [2]

(a) Ecological validity is how far the findings of a study carried out in one setting would also apply to other settings, especially everyday life **(1)**. Pressing a key at a desk when a cartoon appears is very different from braking a real car while watching traffic, feeling the danger and handling the pedals, so reaction times in the study may not show how fast people react on a real road **(1)**.

(b) Reliability is the consistency of a procedure, task or measure, for example whether a questionnaire gives the same score to the same person on two occasions **(1)**. Replicability is whether the whole study could be repeated exactly by another researcher, because the procedure is standardised and fully written down **(1)**.

Example 2

Tomas is observing how children aged 6 share play equipment in a school playground at break time. He records how often each child takes turns, grabs equipment from another child, and offers equipment to another child. His colleague Amira observes at the same time.

(a) For the category "grabs equipment":

(i) Give one reason why "grabs equipment" is a good measure of poor sharing. [1]

(ii) Give one reason why Tomas might record "grabs equipment" when no poor sharing has happened. [1]

(b) Explain how Tomas and Amira could check their inter-observer reliability. [2]

(c) Suggest how Tomas and Amira could improve their inter-observer reliability. [2]

(a)(i) Pulling something out of another child's hands is the opposite of sharing, and it is an obvious movement that an observer is unlikely to miss **(1)**.

(a)(ii) Two friends might pass a ball quickly between them as part of a game, which could look like grabbing from a distance even though nobody is unhappy **(1)**.

(b) They should each tally the categories for the same children at the same time, without talking to each other **(1)**. They then correlate their tallies for each child; a strong positive correlation (about +0.8 or more) means high inter-observer reliability **(1)**.

(c) They could write an exact operational definition of each category, for example "grabs: takes an object from another child's hands without being offered it" (1), and practise together on a video of a playground first, comparing tallies until they record the same behaviours (1).

Example 3

Dr Osei is testing whether chewing gum improves concentration. Twenty students chew gum while doing a 15-minute proofreading task, and a different twenty students do the same task without gum. The DV is the number of spelling errors each student finds in the text.

- (a) The students in the gum group may differ from those in the no-gum group. Describe one such difference and explain how it could produce a false result. [2]
- (b) Describe a change to the way students are put into the two groups that would deal with the problem in part (a). [2]
- (c) Explain one way that demand characteristics could affect the results, and suggest how Dr Osei could reduce them without giving a false aim. [3]
- (d) Explain why the use of a standardised procedure makes this study replicable. [2]
- (a) Some students may be much better at spelling than others, for example those studying languages (1). If more of them happened to be in the gum group, that group would find more errors because of their spelling skill, not the gum (1).
- (b) She could give every student a short spelling test first and use matched pairs, putting one student of each similar-scoring pair in each group (1), so any difference in errors found is due to the gum, not to spelling ability (1).
- (c) Students given gum may guess that the study is about gum helping concentration (1) and try extra hard on the task, so they find more errors than they normally would (1). She could hide the real aim by also asking them to rate the flavour of the gum and the readability of the text, so the gum seems to be part of a taste or design study (1).
- (d) Because the same 15-minute task, text, instructions and scoring would be used for every student (1), another researcher could run the study again in exactly the same way and check whether gum has the same effect (1).

Common mistakes

- **Confusing reliability with replication.** Reliability is the consistency of a procedure, task or measure, not "repeating the study and getting the same results" (June 2021, November 2022 and June 2023 reports). Repeating counts only for test-retest: the same measure with the same people.
- **Talking about the results instead of the measure.** Reliability is about the procedure and measures, not just the findings; "accurate" or "trustworthy" is not a definition (June 2021 and June 2022 reports).

- **Thinking two observers make a study reliable.** Inter-rater and inter-observer reliability are about the **agreement** between them; a second observer only lets you check it (June 2022, November 2022, June 2023 and March 2025 reports).
- **Naming inter-rater reliability as a way to improve reliability.** It is a measure of reliability; to improve it, change the definitions, training or procedure (November 2022 report).
- **Answering a different question about inter-rater reliability:** describing how to test it, or defining reliability in general, when asked what inter-rater reliability is (November 2021 report).
- **Giving the textbook definition of validity as the judgement.** "It doesn't measure what it is supposed to measure" earns nothing on its own; look at the details in the question and judge that measure (March 2025 report).
- **Repeating the question.** "The category is valid because it is a non-verbal behaviour" scores 0, and "you can smile for many reasons" is too vague: give a specific other reason for the behaviour (March 2026 report).
- **Blaming the participants when asked about a test's validity.** For Fagen et al.'s performance test, the elephants' health or personality, or the test being "too hard", scores nothing (March 2025 report).
- **Writing the term without saying what it means or why.** "Ecological validity" alone earns nothing; you must say what the weakness is, such as "it lacked ecological validity" (November 2021 report). Simply saying something is "not reliable or valid" doesn't show understanding (March 2026 report).
- **Mixing up reliability and replicability.** Participants doing different things at home is a problem for reliability, not replicability (March 2025 report).
- **Vague answers such as "to make it reliable"** with no detail of how (November 2022 report).

How to get the marks

- **"State" or "outline what is meant by":** a definition in your own words that doesn't just reuse the term: consistency for reliability, measuring what it claims for validity, findings applying to other settings for ecological validity.
- **"Explain ..., using this study as an example" or "Use an example":** the definition (1) plus a specific detail from the study (1). A study name alone is not an example.
- **"Explain one reason it would (or may not) be valid":** a reason about this exact measure or category, never "because it measures what it should".
- **"Suggest how to improve validity / reliability":** the change (1) plus how it helps or a linked detail (1). The change must actually change the study.
- **"Describe X and Y, using any example(s)" (6 marks):** both terms, each with a definition and a detail, plus at least one example; without an example, or about one term only, the most is 4.
- **"Explain which problem relates to reliability":** the mark is for "because ...", not for choosing the problem.
- **Generalisability questions:** if the question already says "to improve generalisability", saying so again scores 0; give a reason linked to the behaviour and an example group.

- **Use the right term:** internal validity, ecological validity, generalisability, inter-rater or inter-observer reliability, test-retest reliability. Then explain it.

Quick check

1. State what is meant by "reliability".
2. A psychologist gives the same 20-item anxiety questionnaire to the same 40 people twice, three weeks apart. Their two sets of scores correlate at $+0.35$. Name the type of reliability being checked, and say what the result suggests.
3. A researcher measuring general knowledge has two problems: (1) a participant who sits the same quiz on Monday and again on Friday gets very different scores, although nothing has changed; (2) the quiz is so easy that almost everyone gets full marks, so it can't show who knows more. Explain which problem is about reliability and which is about validity.
4. A study tests memory by asking people to learn lists of 15 random three-letter combinations, such as "KEP" and "ZOR", on a computer in a quiet room. Explain one problem with its ecological validity, and suggest one way to improve it.
5. A study of mobile phone use takes its sample from 18-year-old students at one college in one city. Explain why the findings might not apply to all adults, using an example.

Answers

1. Reliability is the consistency of a procedure, task or measure: whether it gives the same result each time it is used in the same way.
2. Test-retest reliability. A correlation of $+0.35$ is well below about $+0.8$, so people's scores changed a lot between the two occasions: the questionnaire is not reliable.
3. Problem 1 is about reliability, because the quiz does not give consistent scores for the same person on two occasions (low test-retest reliability). Problem 2 is about validity, because a quiz that everyone aces does not measure differences in general knowledge, which is what it claims to measure.
4. People rarely learn meaningless letter strings in a quiet room in everyday life, so the findings may not show how they remember real information, such as a list of things to buy in a busy shop. To improve it, use meaningful, everyday material (a real shopping list or directions) and test people in a more natural setting with normal background noise.
5. Older adults may use their phones very differently, for example a 70-year-old may use a phone mostly for calls rather than social media, so their patterns of use would not match the students'. The sample is one age group from one place, so its findings may not apply to other ages or cultures.

Past questions to try

- [9990/22/F/M/26 Q7](#): why an observation category would and would not be valid.
- [9990/23/O/N/25 Q8](#): how differences between groups lower validity, and how to fix it.
- [9990/23/O/N/23 Q3](#): how test-retest checks reliability, and improving inter-rater reliability.
- [9990/23/M/J/25 Q7](#): deciding which problem is about reliability and which about validity.