

Educational Measurement & Evaluation

Practice Test 1

60 original multiple-choice questions for teacher test preparation (FPSC / FGEI / PPSC style). Choose one answer for each question. The full answer key is on the last page — worked explanations for every question are free on the test page at QuaidianPoint.com.

Name: _____ Class: _____ Score: _____ / 60

Measurement, Assessment and Evaluation

Q1. A teacher records that Ali scored 42 out of 50 on a class quiz. This act of assigning a number to performance is best described as:

- (A) evaluation
- (B) measurement
- (C) diagnosis
- (D) grading policy

Q2. After reviewing quiz scores, homework and class participation, a teacher judges that a student has made satisfactory progress. This overall judgement is an example of:

- (A) scoring
- (B) scaling
- (C) ranking
- (D) evaluation

Q3. Which activity is primarily assessment rather than measurement alone?

- (A) Collecting portfolio work, observations and test scores to understand learning
- (B) Counting the number of correct answers on a test
- (C) Converting raw marks into percentages
- (D) Recording attendance in a register

Q4. A school uses end-of-term results mainly to decide whether a new teaching method should continue. This purpose is best called:

- (A) formative assessment
- (B) diagnostic testing
- (C) evaluation of a programme
- (D) item banking

Q5. Which statement correctly distinguishes measurement from evaluation?

- (A) Measurement judges worth, while evaluation only counts correct answers
- (B) Measurement is quantitative description, while evaluation adds a judgement of value
- (C) Measurement is always informal, while evaluation is always formal
- (D) Measurement needs no instrument, while evaluation needs a test

Q6. A teacher gives a short quiz during a lesson to find gaps and adjust teaching the same day. This is an example of:

- (A) summative evaluation
- (B) placement testing
- (C) certification testing
- (D) formative assessment

Q7. Which of the following is a quantitative measure?

- (A) A written comment that a student writes fluently
- (B) A score of 18 out of 20 on a spelling test
- (C) A teacher impression that a class is active
- (D) A peer remark that a presentation was clear

Q8. Before starting a new unit, a teacher tests what students already know in order to plan the starting point. This is best described as:

- (A) diagnostic assessment
- (B) summative assessment
- (C) norm referencing only
- (D) final grading

Q9. A final examination used to certify that students have completed a course is mainly an example of:

- (A) formative feedback
- (B) classroom observation
- (C) peer assessment
- (D) summative assessment

Q10. Which example shows criterion-referenced interpretation?

- (A) Sana scored 80 percent, meeting the set pass standard of 70 percent
- (B) Sana ranked third in a class of forty students
- (C) Sana scored higher than most of her classmates
- (D) Sana improved her rank from tenth to fifth

Q11. A teacher compares a student's score with the class average and percentile rank. This interpretation is:

- (A) diagnostic only
- (B) formative only
- (C) norm-referenced
- (D) criterion-referenced

Q12. Which purpose fits continuous assessment best?

- (A) A single final judgement with no feedback during the term
- (B) Regular evidence over time to track progress and support learning
- (C) Ranking schools for public league tables only
- (D) Replacing all teaching with testing

Scales of Measurement

Q13. Labelling students by house - Red, Blue, Green - without any order uses which scale?

- (A) Ordinal scale
- (B) Interval scale
- (C) Nominal scale
- (D) Ratio scale

Q14. Ranking three essays as first, second and third, without knowing the gap between them, uses which scale?

- (A) Ordinal scale
- (B) Nominal scale
- (C) Ratio scale
- (D) Interval scale

Q15. Temperature in degrees Celsius, where zero does not mean absence of heat, is an example of which scale?

- (A) Nominal scale
- (B) Ordinal scale
- (C) Ratio scale
- (D) Interval scale

Q16. Measuring the weight of science equipment in kilograms, where zero means no weight, uses which scale?

- (A) Nominal scale
- (B) Ratio scale
- (C) Ordinal scale
- (D) Interval scale

Q17. Which statistic is appropriate for nominal data such as favourite subject categories?

- (A) Mean
- (B) Standard deviation
- (C) Pearson correlation
- (D) Mode

Q18. For ordinal data, such as class positions, the most suitable central measure is the:

- (A) Mean
- (B) Geometric mean
- (C) Median
- (D) Standard deviation

Q19. On an interval scale, which statement is correct?

- (A) Equal differences between scores represent equal differences in the trait
- (B) A score of 40 is twice as much ability as a score of 20
- (C) Scores can only be counted, not ordered
- (D) Zero always means complete absence of the trait

Q20. A teacher says a student who scored 20 marks has twice the knowledge of one who scored 10 marks. This claim is unsafe for most test scores because they are at best:

- (A) Nominal
- (B) Interval, not ratio
- (C) Ordinal only
- (D) Categorical only

Q21. Which of the following is a nominal variable in a school record?

- (A) Medium of instruction recorded as Urdu or English
- (B) Height of students in centimetres
- (C) Time taken to finish a test
- (D) Examination marks out of 100

Q22. Height, weight and response time share which property that Celsius temperature lacks?

- (A) They can be ordered
- (B) They can be counted
- (C) They can be grouped
- (D) They have a true zero point

Q23. A satisfaction rating of poor, fair, good and excellent is best treated as:

- (A) Nominal data
- (B) Ordinal data
- (C) Ratio data
- (D) Interval data

Q24. Which scale allows the mean and standard deviation to be used meaningfully?

- (A) Nominal scale only
- (B) Ordinal scale only
- (C) Interval scale
- (D) Nominal and ordinal scales

Validity

Q25. A test gives highly consistent scores every time it is used, but it measures memorisation rather than the problem-solving skill it claims to test. The test is best described as:

- (A) Valid but not reliable
- (B) Neither valid nor reliable
- (C) Valid for all purposes
- (D) Reliable but not valid for its claimed purpose

Q26. A mathematics test that looks, to students and teachers, as if it tests mathematics has good:

- (A) Predictive validity
- (B) Face validity
- (C) Construct validity
- (D) Split-half reliability

Q27. A panel of subject experts checks that a science test covers all syllabus topics in fair proportion. This establishes mainly:

- (A) Content validity
- (B) Concurrent validity
- (C) Test-retest reliability
- (D) Internal consistency

Q28. A new reasoning test is checked by seeing whether its scores relate to an accepted theory of reasoning ability. This is evidence for:

- (A) Face validity
- (B) Predictive validity
- (C) Construct validity
- (D) Scoring reliability

Q29. A school compares a new entrance test with students current examination results collected at the same time. This checks:

- (A) Test-retest reliability
- (B) Concurrent validity
- (C) Content validity
- (D) Face validity

Q30. A teacher uses admission test scores to predict final examination grades one year later. The relevant validity is:

- (A) Content validity
- (B) Face validity
- (C) Construct validity
- (D) Predictive validity

Q31. Which situation most threatens content validity?

- (A) The test is marked twice with the same result
- (B) The test instructions are clear
- (C) Important syllabus topics are missing from the test
- (D) Students find the test format familiar

Q32. If a reading test requires advanced general knowledge unrelated to reading skill, its scores may suffer from:

- (A) Construct-irrelevant variance
- (B) High internal consistency
- (C) Strong predictive power
- (D) Clear content coverage

Q33. Criterion-related validity is best described as:

- (A) Expert judgement about topic coverage
- (B) Appearance of the test to test takers
- (C) The relationship between test scores and an external criterion
- (D) Consistency between two halves of a test

Q34. Which method would best support the claim that a test measures anxiety rather than general stress?

- (A) Checking that items look relevant
- (B) Showing expected links with anxiety measures and weak links with unrelated measures
- (C) Making the test longer
- (D) Using only multiple-choice items

Q35. A valid test for one purpose, such as placement, is:

- (A) Automatically valid for certification
- (B) Automatically reliable for all groups
- (C) Free from all measurement error
- (D) Not automatically valid for a different purpose or group

Q36. The validity coefficient is usually reported as a:

- (A) Correlation between test scores and a criterion
- (B) Percentage of students who pass
- (C) Count of items in the test
- (D) Difference between two group means

Reliability

Q37. Reliability refers mainly to the:

- (A) Consistency of measurement
- (B) Coverage of the syllabus
- (C) Difficulty of the items
- (D) Appearance of the test

Q38. Giving the same test to the same group after two weeks and correlating the scores estimates:

- (A) Split-half reliability
- (B) Inter-rater reliability
- (C) Test-retest reliability
- (D) Content validity

Q39. Two equivalent forms of a test are given to one group and the scores are correlated. This estimates:

- (A) Face validity
- (B) Construct validity
- (C) Item difficulty
- (D) Parallel-forms reliability

Q40. A test is divided into odd and even items and the two half-scores are correlated. This method is called:

- (A) Test-retest method
- (B) Split-half method
- (C) Parallel-forms method
- (D) Inter-rater method

Q41. KR-20 is most appropriate for a test in which items are:

- (A) Open-ended essays
- (B) Rated on a five-point scale
- (C) Scored right or wrong
- (D) Marked by several judges

Q42. If a test's items vary greatly in difficulty, which reliability estimate is more accurate?

- (A) KR-20, because it uses each item's actual difficulty
- (B) KR-21, because it assumes equal difficulty
- (C) Split-half, because it needs two test administrations
- (D) Test-retest, because it needs two separate testings

Q43. Two teachers independently mark the same essays and their marks agree closely. This shows good:

- (A) Test-retest reliability
- (B) Inter-rater reliability
- (C) Split-half reliability
- (D) Predictive validity

Q44. Increasing the number of good-quality items in a test usually:

- (A) Lowers reliability
- (B) Leaves reliability unchanged
- (C) Makes scores harder to interpret
- (D) Increases reliability

Q45. A reliability coefficient of 0.88, compared with 0.52, indicates:

- (A) Lower consistency
- (B) Higher consistency
- (C) Higher validity automatically
- (D) Easier items

Q46. The standard error of measurement is used to:

- (A) Count the items in a test
- (B) Judge content coverage
- (C) Estimate the range within which a true score probably lies
- (D) Compare two different tests directly

Q47. If a test has a standard deviation of 10 and reliability of 0.84, the standard error of measurement is:

- (A) 4
- (B) 6
- (C) 8
- (D) 10

Q48. Which factor is most likely to lower reliability?

- (A) Clear instructions
- (B) A wide range of ability in the group
- (C) Standard testing conditions
- (D) Ambiguous items and guessing

Item Analysis

Q49. In item analysis, the difficulty index of an item is usually denoted by:

- (A) D
- (B) p
- (C) r
- (D) z

Q50. An item answered correctly by 80 out of 100 students has a difficulty index of:

- (A) 0.80
- (B) 0.20
- (C) 0.50
- (D) 0.08

Q51. A difficulty index of 0.95 for a multiple-choice item usually suggests that the item is:

- (A) Too difficult
- (B) Moderately difficult
- (C) Very easy
- (D) Highly discriminating

Q52. The discrimination index compares performance on an item between:

- (A) Boys and girls only
- (B) First and last items
- (C) Two different tests
- (D) High-scoring and low-scoring groups

Q53. In the upper group 24 of 30 students answer correctly and in the lower group 6 of 30 do so. The discrimination index is:

- (A) 0.20
- (B) 0.40
- (C) 0.80
- (D) 0.60

Q54. A discrimination index of -0.25 means that:

- (A) High scorers outperform low scorers
- (B) The item is of medium difficulty
- (C) Low scorers answer the item correctly more often than high scorers
- (D) The item has no distractors

Q55. An item with difficulty 0.50 and discrimination 0.45 is generally:

- (A) A good item to retain
- (B) Too easy to use
- (C) In need of deletion at once
- (D) Suitable only for essays

Q56. Distractor analysis is used to check whether:

- (A) The key is the longest option
- (B) The test is long enough
- (C) Each wrong option attracts some low scorers but few high scorers
- (D) All students finish on time

Q57. A distractor that no student selects is best described as:

- (A) A non-functioning distractor
- (B) An effective distractor
- (C) A discriminating item
- (D) A valid criterion

Q58. An item with difficulty 0.10 and discrimination 0.05 should usually be:

- (A) Retained without change
- (B) Used as a model item
- (C) Made slightly easier only
- (D) Revised or removed from the test

Q59. Which difficulty range is generally preferred for most items in a classroom test?

- (A) 0.05 to 0.15
- (B) 0.85 to 0.95
- (C) 0.30 to 0.70
- (D) 0.90 to 1.00

Q60. If a distractor attracts more high scorers than low scorers, the most likely problem is that the distractor is:

- (A) Too obviously wrong
- (B) Misleading or possibly also defensible
- (C) Too short
- (D) Placed first in the list

Answer Key

Q1: B	Q11: C	Q21: A	Q31: C	Q41: C	Q51: C
Q2: D	Q12: B	Q22: D	Q32: A	Q42: A	Q52: D
Q3: A	Q13: C	Q23: B	Q33: C	Q43: B	Q53: D
Q4: C	Q14: A	Q24: C	Q34: B	Q44: D	Q54: C
Q5: B	Q15: D	Q25: D	Q35: D	Q45: B	Q55: A
Q6: D	Q16: B	Q26: B	Q36: A	Q46: C	Q56: C
Q7: B	Q17: D	Q27: A	Q37: A	Q47: A	Q57: A
Q8: A	Q18: C	Q28: C	Q38: C	Q48: D	Q58: D
Q9: D	Q19: A	Q29: B	Q39: D	Q49: B	Q59: C
Q10: A	Q20: B	Q30: D	Q40: B	Q50: A	Q60: B

Worked explanations for all 60 questions:

quaidianpoint.com/pedagogy-educational-measurement-evaluation-practice-test-1/