

Educational Measurement & Evaluation

Practice Test 4

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 counts spelling errors in each essay and reports the counts. This step is:

- (A) Evaluation
- (B) Measurement
- (C) Diagnosis
- (D) Certification

Q2. A school board judges a literacy programme as successful after comparing results with its targets. This judgement is:

- (A) Scoring
- (B) Scaling
- (C) Item writing
- (D) Evaluation

Q3. Which tool best supports assessment of speaking skills?

- (A) Attendance register only
- (B) Textbook page count only
- (C) Observation of oral presentations with criteria
- (D) Counting library books only

Q4. A quiz used only to give students feedback and guide revision, with no grade recorded, is:

- (A) Formative
- (B) Summative
- (C) Certifying
- (D) Selective

Q5. End-of-year board examinations that decide promotion are mainly:

- (A) Diagnostic
- (B) Formative
- (C) Informal
- (D) Summative

Q6. A teacher finds through a quick check that most students can add fractions but not subtract them. The next step of re-teaching subtraction shows:

- (A) Diagnostic use of assessment
- (B) Final certification
- (C) Norm referencing
- (D) External ranking

Q7. A student score is reported as above the 80th percentile. This interpretation is:

- (A) Criterion-referenced
- (B) Mastery-based only
- (C) Norm-referenced
- (D) Diagnostic only

Q8. A practical skills test requires students to perform each step correctly to pass. This pass decision is:

- (A) Norm-referenced only
- (B) Criterion-referenced
- (C) Peer-referenced only
- (D) Random

Q9. Which assessment practice gives the fullest picture of learning?

- (A) One test on one day
- (B) One oral question
- (C) Multiple methods over time, including tests, projects and observation
- (D) One homework task

Q10. Students review each other drafts using a checklist. This is an example of:

- (A) Standardised testing
- (B) Peer assessment
- (C) Final grading
- (D) Placement testing

Q11. Measurement without evaluation can tell a teacher:

- (A) What score was obtained, but not whether it is good enough
- (B) Whether a programme should continue
- (C) Whether a student should be promoted
- (D) Whether a course is excellent

Q12. A teacher uses weekly short tests to monitor progress and adjust support for weak students. This is best called:

- (A) One-off certification
- (B) External examination
- (C) Final summative grading
- (D) Continuous formative assessment

Scales of Measurement

Q13. Sorting students into science, arts and commerce streams for records uses which scale?

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

Q14. A music competition ranks performers gold, silver and bronze. This ranking is:

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

Q15. Test scores expressed as standard scores with equal units but no true zero are treated as:

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

Q16. The mass of chemicals used in an experiment, measured in grams, is on a:

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

Q17. For data in categories such as blood groups, the appropriate average is the:

- (A) Mode
- (B) Mean
- (C) Median of codes
- (D) Standard deviation of codes

Q18. A claim that 20 degrees Celsius is twice as hot as 10 degrees is wrong because Celsius is:

- (A) Nominal
- (B) Ordinal
- (C) Ratio
- (D) Interval, with an arbitrary zero

Q19. Which data can be ordered but should not be averaged as if gaps are equal?

- (A) Heights in centimetres
- (B) Weights in kilograms
- (C) Class ranks
- (D) Times in seconds

Q20. The difference between first and second place may be tiny, while the gap between second and third may be large. This illustrates a limit of:

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

Q21. A survey asks teachers to rate workload as low, medium or high. These ratings are:

- (A) Ratio
- (B) Ordinal
- (C) Nominal
- (D) Interval with a true zero

Q22. Which operation is valid for interval data?

- (A) Saying one score is twice another
- (B) Counting only
- (C) Comparing differences between scores
- (D) Ordering categories only

Q23. A researcher wants to report the coefficient of variation, which divides the standard deviation by the mean. This statistic is defensible only for data on which scale?

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

Q24. A teacher reports the median class position rather than the mean position because positions are:

- (A) Ordinal, so the median is safer
- (B) Ratio, so the mean is unsafe
- (C) Nominal, so no average works
- (D) Interval, so the median is wrong

Validity

Q25. An Urdu test that uses only poetry questions, although the syllabus also covers prose and grammar, lacks:

- (A) Reliability
- (B) Scoring speed
- (C) Content validity
- (D) Face appeal only

Q26. A test that looks like a proper maths test to parents and students, with no further evidence, has:

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

Q27. A test of leadership is studied to see if its scores match leadership theory, including differences between trained and untrained groups. This is evidence for:

- (A) Construct validity
- (B) Face validity only
- (C) Inter-rater reliability
- (D) Split-half reliability

Q28. A college compares new admission scores with current first-semester grades. This checks:

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

Q29. An entry test is used to forecast training success after six months. The relevant validity is:

- (A) Content validity
- (B) Predictive validity
- (C) Face validity
- (D) Internal consistency

Q30. A history test includes many questions on one small topic and none on others. The main threat is to:

- (A) Scoring reliability
- (B) Test length
- (C) Content validity
- (D) Item difficulty alone

Q31. A test intended for Grade 10 is given to Grade 6 students and fails to predict their later grades. The likely issue is that validity depends on:

- (A) The population and use, not the test alone
- (B) The number of items in the test alone
- (C) The time limit used for the test alone
- (D) The difficulty of one item alone

Q32. If a test measures test anxiety as well as maths skill, its maths scores contain:

- (A) Pure true variance
- (B) No error at all
- (C) Only content error
- (D) Construct-irrelevant variance

Q33. The best first step to support content validity when building a new test is to:

- (A) Prepare a table of specifications from objectives
- (B) Print the test in colour
- (C) Add more easy items
- (D) Shorten instructions

Q34. A validity coefficient near zero between a test and its intended criterion suggests the test:

- (A) Is highly reliable
- (B) Covers content well
- (C) Looks professional
- (D) Does not predict or reflect that criterion

Q35. Which evidence belongs to criterion-related validity?

- (A) Expert ratings of coverage
- (B) Student opinions on appearance
- (C) Correlation with job performance ratings
- (D) Agreement between two markers

Q36. A test developer finds that a reasoning test correlates 0.70 with another reasoning test and 0.10 with a clerical speed test. This supports:

- (A) Face validity only
- (B) Construct validity through convergent and discriminant patterns
- (C) Split-half reliability only
- (D) Content validity only

Reliability

Q37. A test that gives nearly the same score when repeated under similar conditions is:

- (A) Reliable
- (B) Valid for all uses
- (C) Easy
- (D) Criterion-referenced

Q38. Administering the same test twice with a suitable gap and correlating scores estimates:

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

Q39. Two parallel forms given close together mainly control for:

- (A) Inter-rater disagreement
- (B) Memory of specific items
- (C) Content coverage
- (D) Face appeal

Q40. A split-half correlation of 0.60 for half a test, corrected by Spearman-Brown for full length, will usually become:

- (A) Lower than 0.60
- (B) Exactly 0.60
- (C) Higher than 0.60
- (D) Zero

Q41. KR-21 is preferred over KR-20 when the test developer:

- (A) Has essay items only
- (B) Needs judge agreement
- (C) Has two administrations
- (D) Wants a quick estimate assuming equal item difficulty

Q42. A detailed answer key and marking guide for short-answer items mainly improves:

- (A) Content validity
- (B) Inter-rater reliability
- (C) Predictive validity
- (D) Item difficulty

Q43. If SD is 12 and reliability is 0.91, the SEM is closest to:

- (A) 3.6
- (B) 6.0
- (C) 9.0
- (D) 12.0

Q44. Two scores 70 and 74 on a test with SEM 5 should be interpreted as:

- (A) Clearly different abilities
- (B) Proof of cheating
- (C) Possibly the same true score, given measurement error
- (D) Proof of different tests

Q45. Which practice most harms reliability?

- (A) Clear timing
- (B) Ambiguous stems and unclear directions
- (C) Standard scoring
- (D) Enough testing time

Q46. A test with items that all measure different, unrelated topics will likely show:

- (A) Low internal consistency
- (B) High internal consistency
- (C) Perfect predictive validity
- (D) No measurement error

Q47. Reliability coefficients range, in practice, from:

- (A) 0 to 10
- (B) -1 to 0 only
- (C) -1.00 to +1.00, like a correlation
- (D) 0.00 to 1.00

Q48. A longer testing time, up to a reasonable limit, usually helps reliability by:

- (A) Adding fatigue only
- (B) Reducing the number of items answered with care
- (C) Reducing guessing pressure
- (D) Making items harder

Item Analysis

Q49. The difficulty index is also called:

- (A) Discrimination power
- (B) p-value of the item
- (C) Reliability coefficient
- (D) Validity coefficient

Q50. An item answered by 120 of 300 students has p equal to:

- (A) 0.12
- (B) 0.25
- (C) 0.40
- (D) 0.60

Q51. An item with $p = 0.92$ is most useful as:

- (A) A strong discriminator
- (B) A mastery check only
- (C) A difficult challenge item
- (D) An easy warm-up or mastery item, not a discriminator

Q52. Discrimination index D is calculated as:

- (A) (Upper correct minus lower correct) divided by group size
- (B) Total correct divided by total students
- (C) Upper correct plus lower correct
- (D) Number of distractors divided by items

Q53. Upper 18 of 20 correct, lower 8 of 20 correct. D equals:

- (A) 0.20
- (B) 0.40
- (C) 0.50
- (D) 0.90

Q54. A D value of -0.30 should lead a developer to:

- (A) Keep the item unchanged
- (B) Use it for top students only
- (C) Increase its marks
- (D) Investigate and likely revise or delete it

Q55. An item with D between 0.20 and 0.29 is generally:

- (A) Excellent
- (B) Acceptable, but could be improved
- (C) Completely useless
- (D) Negative

Q56. A distractor that attracts mainly strong students is a warning that the option may be:

- (A) Ambiguous or partly correct
- (B) Too obviously wrong
- (C) Too short
- (D) Placed last

Q57. Which distractor is non-functioning?

- (A) One chosen by 30 percent of low scorers
- (B) One chosen by 20 percent of all students
- (C) One chosen mainly by low scorers
- (D) One chosen by almost nobody

Q58. After analysis, an item has $p = 0.35$, $D = 0.38$ and all options used. The best decision is to:

- (A) Delete it at once
- (B) Retain it
- (C) Make it much easier
- (D) Change the key at random

Q59. Item analysis helps a test developer to:

- (A) Improve items and build a better item bank
- (B) Avoid all future testing
- (C) Rank schools directly
- (D) Set the pass mark without judgement

Q60. A positive point-biserial for an item means students who do well overall tend to:

- (A) Skip the item
- (B) Choose distractors
- (C) Answer the item correctly
- (D) Take longer on the item

Answer Key

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

Worked explanations for all 60 questions:

quaidianpoint.com/pedagogy-educational-measurement-evaluation-practice-test-4/