

# Educational Measurement & Evaluation

## Practice Test 5

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](https://QuaidianPoint.com).

Name: \_\_\_\_\_ Class: \_\_\_\_\_ Score: \_\_\_\_\_ / 60

---

### Measurement, Assessment and Evaluation

---

**Q1. A teacher records reading fluency as words read correctly per minute. This number is a:**

- (A) Final judgement
- (B) Programme decision
- (C) Measure of fluency
- (D) Promotion rule

**Q2. Deciding that a reading programme is worth expanding, based on fluency gains and costs, is:**

- (A) Evaluation
- (B) Scoring only
- (C) Ranking only
- (D) Item writing

**Q3. Which situation is assessment rather than a single measurement?**

- (A) Counting correct answers on one quiz
- (B) Recording one test score
- (C) Timing one task
- (D) Combining quizzes, projects and observations to judge progress

**Q4. A mid-course survey asks students what they find difficult so the teacher can adjust support. This is:**

- (A) Summative grading
- (B) Formative feedback use
- (C) Final certification
- (D) External ranking

**Q5. A final portfolio grade awarded at the end of a course is mainly:**

- (A) Summative
- (B) Diagnostic only
- (C) Informal only
- (D) Placement only

**Q6. Before teaching fractions, a teacher checks which students already understand equal parts. This check is:**

- (A) Summative
- (B) Certifying
- (C) Ranking for prizes
- (D) Diagnostic and placement-related

**Q7. A student passes because she scored at least 50 percent, the set standard for the unit. This is:**

- (A) Norm referencing
- (B) Peer referencing
- (C) Criterion referencing
- (D) Random grading

**Q8. A teacher reports that Ahmed scored higher than 90 percent of test takers nationally. This report uses:**

- (A) Criterion referencing
- (B) Norm referencing
- (C) Diagnostic labelling
- (D) Formative scoring only

**Q9. Which assessment is most likely to improve learning while it is happening?**

- (A) A single final examination with no feedback
- (B) A ranking published once a year
- (C) A certificate issued at course end
- (D) Frequent low-stakes checks with feedback

**Q10. A school uses test results, attendance and teacher reports together to decide support for a student. This decision is best described as:**

- (A) A single measurement
- (B) A raw score
- (C) An evaluative judgement from multiple sources
- (D) An item analysis

**Q11. Self-assessment helps learning mainly because students:**

- (A) Avoid all teacher feedback
- (B) Compare their work with criteria and spot their own gaps
- (C) Receive a final grade sooner
- (D) Skip practice tasks

**Q12. The key difference between measurement and evaluation is that evaluation:**

- (A) Adds a judgement of value against a standard
- (B) Never uses numbers
- (C) Never uses tests
- (D) Only counts errors

## **Scales of Measurement**

---

**Q13. Recording students home districts as categories for a survey uses which scale?**

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

**Q14. Ranking essays as excellent, good, satisfactory and weak gives data on which scale?**

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

**Q15. A score scale where a difference of 10 points means the same amount of change anywhere on the scale is:**

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

**Q16. The length of a classroom in metres is measured on a:**

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

**Q17. Which statistic is misleading if applied to nominal codes such as 1 for boys and 2 for girls?**

- (A) Mode
- (B) Frequency count
- (C) Percentage in each group
- (D) Mean of the codes

**Q18. A teacher orders students by test score and reports the middle student score. That middle value is the:**

- (A) Mean
- (B) Median
- (C) Mode only
- (D) Range

**Q19. On a ratio scale, which statement is valid?**

- (A) Categories can only be counted
- (B) Order is absent
- (C) One quantity can be twice another
- (D) Zero is arbitrary

**Q20. A Likert response from strongly agree to strongly disagree is best treated, strictly, as:**

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

**Q21. Which variable is ratio?**

- (A) Number of correct answers
- (B) Favourite subject
- (C) Calendar year
- (D) Letter grade

**Q22. The mean and standard deviation are most defensible for data that are at least:**

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

**Q23. Converting heights into ranks loses information about:**

- (A) Order
- (B) Exact differences between heights
- (C) The tallest person
- (D) The shortest person

**Q24. Zero marks on a test do not prove zero learning because test scores, in general, lack:**

- (A) Nominal
- (B) Ordinal only
- (C) Interval with equal units
- (D) Ratio with a true zero for ability

## Validity

---

**Q25. A test of computer skill that only asks about keyboard shortcuts, ignoring files, safety and problem solving, suffers from:**

- (A) High reliability
- (B) Good face validity only
- (C) Clear scoring
- (D) Poor content validity through under-representation

**Q26. A candidate feels a test looks fair and relevant. This impression is:**

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

**Q27. A new test of teaching aptitude is compared with classroom observation ratings made the same term. This gives evidence of:**

- (A) Predictive validity
- (B) Concurrent validity
- (C) Content validity only
- (D) Inter-rater reliability only

**Q28. A screening test used to predict which trainees will complete a course is judged mainly by:**

- (A) Predictive validity
- (B) Face validity only
- (C) Split-half reliability only
- (D) Item difficulty only

**Q29. A measure of motivation should rise after a motivational programme and stay stable without one, if it has good:**

- (A) Face validity only
- (B) Scoring reliability only
- (C) Construct validity
- (D) Test length

**Q30. Expert review of a new test blueprint mainly establishes:**

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

**Q31. A test is valid for selecting trainees but is then used to certify experienced teachers without new evidence. The problem is that:**

- (A) Reliability is automatic
- (B) Face validity is enough
- (C) Content coverage is automatic
- (D) Validity does not transfer automatically to a new use

**Q32. An English test written in very complex language for young learners will mainly threaten:**

- (A) Test length
- (B) Validity, through construct-irrelevant difficulty
- (C) Scoring speed
- (D) Item count

**Q33. A validity study with a very narrow, high-ability sample will probably show:**

- (A) A higher validity coefficient
- (B) A lower validity coefficient due to restricted range
- (C) No effect on validity
- (D) Perfect prediction

**Q34. Which combination is criterion-related?**

- (A) Face plus content
- (B) Content plus construct
- (C) Split-half plus test-retest
- (D) Concurrent plus predictive

**Q35. A reliable test that measures the wrong skill for a job is:**

- (A) Consistent but invalid for that job decision
- (B) Automatically valid
- (C) Free from error
- (D) Useful without further evidence

**Q36. The strongest validity evidence for a new theory-based trait measure comes from:**

- (A) Appearance alone
- (B) Length alone
- (C) Multiple theory-based studies, including group differences and correlations
- (D) One easy item

## Reliability

---

**Q37. A stopwatch that gives the same time for the same race on repeated checks is:**

- (A) Valid for all races
- (B) Reliable
- (C) Easy to read only
- (D) Criterion-referenced

**Q38. Giving the same spelling test in September and again in October, then correlating scores, estimates:**

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

**Q39. Splitting a test into first-half and second-half scores and correlating them is a form of:**

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

**Q40. A test with many items all scored right or wrong can have internal consistency estimated by:**

- (A) Face judgement only
- (B) Timing only
- (C) Item count only
- (D) KR-20

**Q41. KR-21 will usually give a slightly lower estimate than KR-20 when item difficulties:**

- (A) Vary across items
- (B) Are all equal
- (C) Are all zero
- (D) Are unknown

**Q42. Two judges rate dance performances and agree closely after training on a rubric. This shows improved:**

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

**Q43. If a test has SD 6 and reliability 0.64, its SEM is:**

- (A) 2.4
- (B) 3.0
- (C) 4.8
- (D) 3.6

**Q44. A test with a large SEM should be used with:**

- (A) Complete confidence in single scores
- (B) No reporting at all
- (C) Caution, reporting a band around each score
- (D) Only rank orders

**Q45. Which factor usually raises reliability?**

- (A) Fewer, ambiguous items
- (B) Noisy conditions
- (C) More good items and clear directions
- (D) Rushed guessing

**Q46. A reliability of 0.60 for a high-stakes selection test is generally:**

- (A) Too low for high-stakes decisions alone
- (B) Perfect
- (C) Better than 0.90
- (D) Proof of validity

**Q47. A test-retest gap of six months for a rapidly changing skill may underestimate reliability because:**

- (A) Memory effects dominate
- (B) Items are forgotten
- (C) Scoring changes
- (D) True ability has changed in between

**Q48. Standardising test administration helps reliability by:**

- (A) Making items easier
- (B) Reducing irrelevant differences between sittings
- (C) Shortening the test
- (D) Removing all true differences

## Item Analysis

---

**Q49. An item answered correctly by 90 of 180 students has difficulty:**

- (A) 0.50
- (B) 0.90
- (C) 0.18
- (D) 0.25

**Q50. An item with  $p = 0.05$  is best described as:**

- (A) Moderate
- (B) Easy
- (C) Ideal for all tests
- (D) Extremely difficult

**Q51. Upper group 15 of 30 correct, lower group 3 of 30 correct. The discrimination index is:**

- (A) 0.20
- (B) 0.40
- (C) 0.50
- (D) 0.60

**Q52. A discrimination index of 0.00 means the item:**

- (A) Is very difficult
- (B) Is very easy
- (C) Does not separate high and low scorers at all
- (D) Has a wrong key for sure

**Q53. Which item should be revised first?**

- (A)  $p = 0.55$ ,  $D = 0.45$
- (B)  $p = 0.20$ ,  $D = -0.20$
- (C)  $p = 0.50$ ,  $D = 0.40$
- (D)  $p = 0.60$ ,  $D = 0.35$

**Q54. In distractor analysis, the key should ideally be chosen by:**

- (A) Fewer high scorers than low scorers
- (B) Nobody at all
- (C) More high scorers than low scorers
- (D) Only the lowest group

**Q55. A distractor chosen by 25 percent of students, mostly low scorers, is:**

- (A) Working well
- (B) Dead
- (C) Mis-keyed for sure
- (D) Too easy

**Q56. The best general difficulty range for items in a norm-referenced test is around:**

- (A) 0.05 to 0.10
- (B) 0.90 to 0.95
- (C) 0.10 to 0.20 only
- (D) 0.30 to 0.70

**Q57. An item with  $p = 0.82$ ,  $D = 0.12$  and two unused distractors is best handled by:**

- (A) Retaining it unchanged
- (B) Doubling its marks
- (C) Using it as a model
- (D) Revising the item and distractors, or dropping it

**Q58. Item analysis is most useful for:**

- (A) Selecting and improving items for future tests
- (B) Setting the pass mark for a single class
- (C) Ranking schools by their average scores
- (D) Replacing classroom teaching with testing

**Q59. A negative point-biserial for an item suggests that:**

- (A) The item is too easy
- (B) The item has no distractors
- (C) Low scorers do better on the item than high scorers
- (D) High scorers tend to answer it correctly

**Q60. A test developer keeps an item with  $p = 0.45$ ,  $D = 0.50$  and all distractors attracting some low scorers. This decision is sound because the item is:**

- (A) Too difficult to use
- (B) Moderate, discriminating and well-distracted
- (C) Valid for essays only
- (D) Reliable without a trial

---

## Answer Key

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

Worked explanations for all 60 questions:

[quaidianpoint.com/pedagogy-educational-measurement-evaluation-practice-test-5/](http://quaidianpoint.com/pedagogy-educational-measurement-evaluation-practice-test-5/)