

# Educational Measurement & Evaluation

## Practice Test 2

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 uses a checklist to observe group work and records how often each skill appears. The checklist is best seen as a:**

- (A) Final grade
- (B) Value judgement
- (C) Measurement instrument
- (D) Teaching method

**Q2. Which question is answered by evaluation rather than by measurement alone?**

- (A) How good is this programme compared with its goals?
- (B) How many items did the student answer correctly?
- (C) What percentage did the student obtain?
- (D) How long did the student take to finish?

**Q3. A portfolio containing drafts, revisions and reflections over a term is mainly a tool for:**

- (A) Single-score ranking
- (B) Assessment of progress over time
- (C) Recording attendance
- (D) Checking textbook coverage only

**Q4. After a unit test, a teacher changes the next lesson plan because many students missed the same concept. This use of results is:**

- (A) Certification
- (B) Selection for a prize
- (C) Placement in a new school
- (D) Formative use of assessment

**Q5. A district compares schools by average test scores to decide funding priorities. This is best described as:**

- (A) Classroom diagnosis
- (B) Peer assessment
- (C) Self-assessment
- (D) Evaluation for decision making

**Q6. Which activity involves measurement but not yet evaluation?**

- (A) Declaring a student successful after meeting a standard
- (B) Counting correct responses and reporting a raw score of 34
- (C) Recommending a student for promotion
- (D) Judging a course as effective

**Q7. A pre-test given to place students into remedial or advanced groups serves mainly a:**

- (A) Placement function
- (B) Summative function
- (C) Certification function
- (D) Ranking function only

**Q8. When a teacher asks students to judge their own drafts against a rubric before submission, this is:**

- (A) External examination
- (B) Norm-referenced grading
- (C) Self-assessment
- (D) Standardised testing

**Q9. A teacher gives marks for classwork, homework and a project, then combines them into a term grade. This combined judgement is closest to:**

- (A) Evaluation based on multiple measures
- (B) A single measurement
- (C) A diagnostic label
- (D) An item analysis

**Q10. Which statement about assessment is most accurate?**

- (A) Assessment is only a final examination
- (B) Assessment always produces a single number
- (C) Assessment must use a standardised test
- (D) Assessment can be formal or informal and ongoing

**Q11. A student is awarded a certificate because she met all course outcomes. The certificate decision depends on:**

- (A) Her rank in class
- (B) Her improvement over peers
- (C) Criterion-based evaluation
- (D) Her attendance rank only

**Q12. A teacher notices, through questioning, that students confuse two science terms and re-teaches them at once. This is an example of:**

- (A) Summative testing
- (B) Informal formative assessment
- (C) Final certification
- (D) External moderation

## **Scales of Measurement**

---

**Q13. Assigning roll numbers to students for identification uses which scale?**

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

**Q14. Which example is on a ratio scale?**

- (A) Letter grades A, B and C
- (B) School names
- (C) Years of the calendar
- (D) Number of books read in a month

**Q15. Calendar years such as 2020 and 2024 are best classified as:**

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

**Q16. Which measure is suitable for ordinal data?**

- (A) Median
- (B) Mean of raw counts only
- (C) Standard deviation of ranks
- (D) Product-moment correlation of categories

**Q17. A teacher ranks debate teams from best to worst. The main limitation of this ranking is that it does not show:**

- (A) How large the gaps between teams are
- (B) The order of teams
- (C) Which team is first
- (D) The names of teams

**Q18. Which operation is NOT meaningful for nominal data?**

- (A) Counting cases in each category
- (B) Finding the most frequent category
- (C) Calculating a mean category value
- (D) Reporting percentages per category

**Q19. If a runner takes 20 seconds and another takes 40 seconds, the first runner took half the time. This ratio statement works because time is measured on a:**

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

**Q20. A test score of zero on a classroom quiz does NOT prove zero knowledge, because classroom scores are usually treated as:**

- (A) Nominal measures
- (B) Interval measures
- (C) Ratio measures
- (D) Pure counts

**Q21. Which data set allows all four operations of ordering, equal intervals and ratios?**

- (A) Favourite colours
- (B) Letter grades
- (C) Celsius temperatures
- (D) Distances in metres

**Q22. The most appropriate summary for the favourite games chosen by a class is the:**

- (A) Mode
- (B) Mean
- (C) Standard deviation
- (D) Range

**Q23. Converting raw ranks into percentile ranks changes ordinal data into a form that is often treated as:**

- (A) Nominal
- (B) Purely categorical
- (C) Approximately interval for practical use
- (D) Ratio with a true zero

**Q24. Which scale is the strongest, carrying the most measurement information?**

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

## Validity

---

**Q25. A test developer writes a table of specifications linking each item to a learning objective. This step mainly supports:**

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

**Q26. A driving test that only checks road signs but never tests actual driving lacks:**

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

**Q27. Two tests of the same skill, given in the same week, correlate at 0.78. This result best supports:**

- (A) Scoring speed
- (B) Concurrent validity
- (C) Face validity
- (D) Item difficulty

**Q28. A university predicts first-year grades from entry test scores. If the prediction is accurate, the entry test has good:**

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

- Q29. A test of creativity correlates strongly with a test of artistic skill but weakly with a test of memory. This pattern supports:**
- (A) Scoring reliability
  - (B) Test length
  - (C) Construct validity
  - (D) Face validity only
- Q30. Which evidence is weakest for validity?**
- (A) A test that merely looks suitable to non-experts
  - (B) Correlation with a relevant criterion
  - (C) Expert review of content
  - (D) Theory-based studies of the trait
- Q31. If a maths test is written in language far above the students reading level, scores will mainly reflect reading ability. This is a threat called:**
- (A) Low reliability
  - (B) Poor item difficulty
  - (C) Short test length
  - (D) Construct-irrelevant variance
- Q32. A test claims to measure problem solving but only asks recall of facts. The main problem is:**
- (A) Low face validity
  - (B) Construct under-representation
  - (C) High reliability
  - (D) Long testing time
- Q33. Which pair belongs under criterion-related validity?**
- (A) Face and content validity
  - (B) Concurrent and predictive validity
  - (C) Split-half and test-retest reliability
  - (D) Internal consistency and stability
- Q34. A test with high reliability but poor validity is best described as:**
- (A) Useful for all purposes
  - (B) Free of error
  - (C) Easy to score
  - (D) Consistently measuring the wrong thing
- Q35. To judge content validity, the most suitable judges are:**
- (A) Subject-matter experts
  - (B) Students who took the test
  - (C) Parents of students
  - (D) Test printers
- Q36. Restriction of range in the criterion group will usually:**
- (A) Increase item difficulty
  - (B) Increase test length
  - (C) Lower the observed validity coefficient
  - (D) Improve face validity

## Reliability

- 
- Q37. A bathroom scale that shows a different weight each time a person steps on it lacks:**
- (A) Content coverage
  - (B) Face appeal
  - (C) Predictive power
  - (D) Reliability
- Q38. Which reliability method needs only one test administration?**
- (A) Test-retest method
  - (B) Parallel-forms method
  - (C) Split-half method
  - (D) Stability method over time
- Q39. The Spearman-Brown formula is used in split-half reliability to:**
- (A) Correct for the shorter length of each half
  - (B) Find item difficulty
  - (C) Judge content coverage
  - (D) Score essay answers
- Q40. A test-retest interval that is too short may inflate reliability because students:**
- (A) Forget the items completely
  - (B) Remember items and answers from the first sitting
  - (C) Change in ability greatly
  - (D) Receive new teaching between sittings
- Q41. Cronbach alpha is most similar to which approach?**
- (A) Test-retest over months
  - (B) Internal consistency across items
  - (C) Agreement between judges
  - (D) Prediction of future grades
- Q42. If two markers give very different marks for the same scripts, the marking suffers from low:**
- (A) Content validity
  - (B) Predictive validity
  - (C) Item difficulty
  - (D) Inter-rater reliability
- Q43. A clear marking scheme for essays is most likely to improve:**
- (A) Test length
  - (B) Syllabus coverage
  - (C) Inter-rater reliability
  - (D) Item difficulty
- Q44. A test with SD 8 and reliability 0.75 has a standard error of measurement of:**
- (A) 4
  - (B) 6
  - (C) 2
  - (D) 8

**Q45. A smaller standard error of measurement means scores are:**

- (A) More difficult
- (B) More valid automatically
- (C) More precise
- (D) Less consistent

**Q46. Which condition tends to increase reliability?**

- (A) Vague directions
- (B) Homogeneous items measuring one trait
- (C) Very few items
- (D) Noisy testing room

**Q47. A reliability coefficient is best interpreted as:**

- (A) The percentage of students who pass
- (B) The average item difficulty
- (C) The number of items in the test
- (D) The proportion of score variance due to true differences

**Q48. A test given to a group with a very wide range of ability will usually show:**

- (A) Higher reliability, because true score variance is larger
- (B) Lower reliability, because scores vary too much
- (C) No change in reliability
- (D) Lower validity in all cases

## Item Analysis

---

**Q49. In item analysis, the upper and lower groups are usually formed from the top and bottom:**

- (A) 10 percent each
- (B) 50 percent each
- (C) 27 percent each
- (D) 5 percent each

**Q50. An item answered correctly by 135 of 150 students has difficulty index:**

- (A) 0.15
- (B) 0.90
- (C) 0.45
- (D) 0.10

**Q51. An item with  $p = 0.20$  is best described as:**

- (A) Very easy
- (B) Moderate
- (C) Too easy to discriminate
- (D) Difficult

**Q52. A discrimination index of 0.40 or above is generally regarded as:**

- (A) Very good discrimination
- (B) Poor discrimination
- (C) Negative discrimination
- (D) No discrimination

**Q53. Upper group 20 of 25 correct, lower group 10 of 25 correct. The discrimination index is:**

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

**Q54. An item with discrimination near zero should be:**

- (A) Retained as a model item
- (B) Revised or discarded, since it does not separate strong and weak students
- (C) Made compulsory for all
- (D) Used only in easy tests

**Q55. Which distractor pattern shows a working distractor?**

- (A) Chosen by no one
- (B) Chosen mainly by high scorers
- (C) Chosen mainly by low scorers, with few high scorers
- (D) Chosen equally by all groups

**Q56. If the key is chosen by fewer low scorers than a distractor, the item most likely has:**

- (A) Ideal difficulty
- (B) Perfect discrimination
- (C) A very easy key
- (D) A faulty key or misleading distractor

**Q57. An item analysis shows  $p = 0.55$  and  $D = 0.35$ . The best action is to:**

- (A) Delete the item at once
- (B) Make the item much harder
- (C) Make the item much easier
- (D) Retain the item, with minor review if needed

**Q58. The point-biserial correlation for an item is used to show:**

- (A) The time taken per item
- (B) How many students chose each option
- (C) How well the item relates to total test performance
- (D) The number of items in the test

**Q59. A test developer wants maximum discrimination for a norm-referenced test. The best average difficulty to aim for is about:**

- (A) 0.10
- (B) 0.50
- (C) 0.90
- (D) 0.95

**Q60. After item analysis, an item with  $p = 0.88$ ,  $D = 0.08$  and one unused distractor is best handled by:**

- (A) Revising the item and its distractors, or removing it
- (B) Keeping it unchanged as a core item
- (C) Using it as the first item in every test
- (D) Doubling its marks

---

## Answer Key

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

Worked explanations for all 60 questions:

[quaidianpoint.com/pedagogy-educational-measurement-evaluation-practice-test-2/](https://quaidianpoint.com/pedagogy-educational-measurement-evaluation-practice-test-2/)