ALLAMA IQBAL OPEN UNIVERSITY

USMAN NOTES
www.usmannotes.com

Name:

Registration No.:

Course Code: 841

Program: M.A./M.Ed

Semester: Autumn 2026

ALLAMA IQBAL OPEN UNIVERSITY, ISLAMABAD

Course: Educational Measurement and Evaluation (841) • Level: M.A./M.Ed.

Semester: Autumn 2026

Assignment No. 2 (Units 5–9)

Total Marks: 100 • Pass Marks: 40

SOLVED ASSIGNMENT GUIDE

(For Study & Reference)

USMAN NOTES www.usmannotes.com

Note: This guide is prepared for study and reference purposes. Students are advised to read the relevant units of the course book, understand the concepts, and then write the assignment in their own words.

Question 1: Procedure for Constructing Different Types of Objective Test Items (20 Marks)

Discuss the procedure for constructing different types of objective test items.

Introduction

An objective test item is a question which can be scored objectively, because there is a single correct answer — or a fixed scoring key — on which all competent markers will agree. Objective items dominate modern testing because they allow a wide sampling of the syllabus, quick and consistent scoring, and freedom from the marker's mood, handwriting impressions and

favouritism. Their principal types are the multiple-choice item, the true-false item, the matching item, and the short-answer or completion item. Constructing good items of any type, however, is a disciplined craft, not an improvisation: badly written items measure guessing ability and test-wiseness instead of learning. This answer first sets out the general procedure common to all objective item construction, and then discusses the specific rules for constructing each type.

The General Procedure of Item Construction

The construction of objective items follows a definite sequence of steps, whether the test is a classroom quiz or a public examination.

- Step 1 — Decide the purpose and content of the test: The constructor first states what the test is for (achievement, diagnosis, selection), which course or unit it covers, and at what level of difficulty its results will be used.
- Step 2 — Prepare a table of specifications (blueprint): Content topics are listed down one side and levels of objectives — knowledge, understanding, application, analysis and higher skills — across the top. Each cell is given a weight according to the importance and teaching time of that topic and objective. The blueprint decides how many items of each type will be written on each topic, so that the finished test represents the whole course fairly instead of the item-writer's favourite corners.
- Step 3 — Write the items: Items are drafted according to the rules for each type given below. The writer works from the textbook and teaching notes, phrases items in language slightly below the students' reading level so that the test measures subject mastery rather

than reading difficulty, and prepares roughly one and a half to two times as many items as the test will finally need, to allow for discarding weak ones.

- Step 4 — Prepare the scoring key: The correct answer to every item is fixed in advance and checked independently by another subject expert. An item whose key cannot be agreed upon is rewritten or dropped.
- Step 5 — Review and edit: Each item is reviewed for technical faults — ambiguity, clues to the answer, trivial content, negative phrasing, bias — and for editorial faults of grammar, spelling and punctuation. A useful device is to review the item from the student's seat: is there any way of answering it correctly without knowing the subject, or of answering it wrongly despite knowing the subject?
- Step 6 — Try out (pre-testing): Where the test is important, the items are administered to a sample similar to the real candidates. The time required, the clarity of instructions and the behaviour of every item are observed.
- Step 7 — Item analysis: For each item two indices are computed. The difficulty index is the proportion of candidates who answer correctly; very easy items (above about 0.90) and very hard items (below about 0.20) contribute little and are usually revised or removed from a norm-referenced test. The discrimination index compares success on the item between the highest-scoring and lowest-scoring groups; a good item is answered correctly by far more of the strong group, and an item with zero or negative discrimination is defective by definition. For multiple-choice items, the functioning of each distractor is also examined.

- Step 8 — Assemble, administer and store: Surviving items are arranged in the final test, usually from easy to difficult and grouped by type, with clear directions and examples. After use, good items are filed in an item bank with their statistics, so that future tests are built from proven material.

Constructing Multiple-Choice Items

The multiple-choice item consists of a stem (the question or incomplete statement) and several options, one correct answer and the rest plausible but wrong options called distractors. It is the most versatile and most recommended objective type, capable of measuring understanding and application as well as recall. Its construction rules are: the stem must state a single, complete problem, so that a knowledgeable student can answer before reading the options; every option must be plausible, grammatically consistent with the stem, and of similar length, so that the correct answer does not betray itself by being longest or most detailed; distractors should be built from students' common errors and misconceptions, which is what makes them attract the weak student and not the strong one; options like “all of the above” and “none of the above” are used sparingly because they reward partial knowledge and test-wiseness; negative stems (“which is NOT…”) are avoided or the negative word is underlined; and the position of the correct answer is varied randomly among the options so that no answering pattern emerges. A test usually offers four or five options per item, which reduces the chance of guessing correctly to one in four or five.

Constructing True-False Items

A true-false item presents a single statement which the student judges true or false. Its construction is treated separately in Question 2 of this assignment; in brief, the statement must be wholly true or wholly false without qualification, must test one idea only, must avoid specific determiners such as “always” and “never” which cue the answer, must avoid direct copying from the textbook, and the numbers of true and false statements must be roughly balanced in no predictable pattern.

Constructing Matching Items

A matching exercise presents two columns — premises and responses — and the student pairs each premise with its correct response. Good construction requires that all premises and responses belong to one homogeneous topic, that the responses outnumber the premises (or may be used more than once) so that the last pair cannot be obtained by elimination alone, that the whole exercise fits on one page, that directions state exactly how matching is to be done and recorded, and that premises are kept short with the longer material placed in the response column for easy reading.

Constructing Short-Answer and Completion Items

In the short-answer item the student supplies a word, number, date or brief phrase in response to a direct question; in the completion item he fills a blank in a statement. These items remove guessing almost entirely and are easily constructed, but they must be built so that only one answer is acceptable: the question must be specific, the blank must fall on the key word (preferably near the end of the sentence), blanks must be of uniform length so they give no

clue, and trivial details, names and dates copied verbatim from the book are avoided. The scorer's key must list all acceptable alternative answers in advance, including reasonable spelling and wording variations.

Conclusion

To conclude, constructing objective test items is a systematic procedure: define the purpose, build a blueprint, draft more items than are needed according to the rules of each type, fix and verify the key, review and edit, try out, analyse difficulty and discrimination, and only then assemble the final test. Multiple-choice items demand a clear stem and plausible distractors; true-false items demand unqualified single statements; matching items demand homogeneous columns with surplus responses; and completion items demand blanks with a single acceptable answer. Followed faithfully, this procedure produces tests that are valid, reliable and fair — the qualities on which every educational decision depends.

Question 2: Significance of True-False and Matching Tests (20 Marks)

What is the significance of true-false and matching tests? Discuss.

Introduction

Among objective test items, the true-false and matching types are the oldest, the simplest to construct and among the most frequently used in classroom testing. Neither can do everything: each has a definite territory in which it is efficient and valuable, and definite limitations which the teacher must respect. This answer explains the nature of each type, sets out its significance — the purposes for which it is genuinely useful — and then discusses its limitations and the precautions by which good practice makes the most of each type.

The True-False Test: Nature and Significance

In a true-false item the student reads a declarative statement and marks it true or false (or right or wrong, yes or no). The format is particularly significant, first, for the wide and rapid sampling of factual knowledge. Because each item takes only a few seconds to answer, fifty or more items can be administered in the time an essay question would need, covering the whole unit and therefore yielding a more representative and more reliable picture of the student's factual mastery. For the same reason, true-false tests are invaluable as quick diagnostic instruments at the start or end of a lesson: a teacher can discover within minutes exactly which facts and ideas a class has grasped and which it has misunderstood.

Second, the true-false format is uniquely suited to testing the ability to distinguish fact from misconception, correct principles from common errors, and fact from opinion. Well-written false statements are built from the actual errors students make — the confusions a teacher hears in class — so the test does not merely record ignorance, it exposes specific wrong ideas that teaching can then correct. Third, construction and scoring are economical: a teacher can write many items quickly once the rules are mastered, scoring is perfectly objective and can be done by key, stencil or machine, and results are available immediately for feedback. Fourth, because the language load is small — one short statement at a time — true-false items are fairer than many formats for younger pupils and weaker readers, and they can be used from the primary school upwards. Modifications such as requiring the student to correct the false statement, or to underline the word that makes it false, raise the level of thinking demanded and reduce guessing, extending the usefulness of the basic format.

The limitations must be stated with equal honesty. The fifty per cent chance of guessing correctly is the standing objection: a student who knows nothing still scores about half marks on average, so true-false tests need many items, and should not alone carry high-stakes decisions. They chiefly measure recognition of factual statements, and writing statements that are wholly true or wholly false without being trivial or tricky is harder than it looks — statements lifted from textbooks reward memorising the book's wording rather than understanding. Ambiguous statements penalise the thoughtful student who sees the exception. These limitations are managed by the standard rules of construction: one idea per statement, no specific determiners ("always", "never", "all" usually signal false; "generally", "sometimes" usually signal true), no double negatives, roughly equal true and false keys in irregular order, and statements of similar length.

The Matching Test: Nature and Significance

A matching exercise consists of a list of premises (terms, events, questions, descriptions) and a list of responses (names, dates, definitions, places), and the student matches each premise with its correct response. Its significance lies, first, in its compactness and efficiency: a large number of associations — ten to fifteen — can be tested in a single exercise occupying a few lines and a few minutes. No other format tests so much factual relationship-knowledge in so little space and time, which is why matching is the natural format for vocabulary, terminology, authors and books, events and dates, inventors and inventions, countries and capitals, and terms and definitions.

Second, matching measures a genuine and important kind of learning: the ability to see relationships and to discriminate between closely related alternatives. The student is not merely recalling an isolated fact; he must hold several related facts in mind and distinguish, for example, which definition belongs to which term when all the definitions are visible together. Third, guessing is substantially reduced compared with true-false items: if there are more responses than premises, the chance of a correct guess on later items does not rise by elimination as quickly as students expect, and the probability of matching a whole set correctly by chance is very small. Fourth, matching exercises are easy to construct from textbook material, perfectly objective to score, and attractive to students, who often experience them as a puzzle; they are therefore excellent for revision tests and classroom quizzes. Their diagnostic value is also real: the pattern of a student's wrong matches shows exactly which pairs of ideas he confuses.

Against these strengths stand definite limitations. Matching is confined almost entirely to factual association at the knowledge level; it cannot measure understanding in depth, organisation of ideas, or the ability to apply and create. The exercise must be homogeneous — one topic throughout — or the matching becomes trivially easy, since a date can only match an event and a capital only a country. Poorly built exercises contain clerical and logical clues: if the numbers of premises and responses are equal, the last pair is given free; grammatical agreement between a premise and only one response betrays answers; and long exercises spread over two pages waste time in searching. Good construction therefore provides surplus responses (or permits responses to be used more than once), keeps the exercise on one page, orders responses logically (dates chronologically, names alphabetically), puts the longer text in

the premises and short responses in the answer column, and gives directions with a worked example.

Comparative Note

The two formats complement each other. True-false items test one statement at a time and maximise the number of independent facts sampled: matching items test sets of related facts together and maximise relationship-learning per page. Both are objective, rapidly scored and best at the knowledge level; both mislead when used alone for important decisions; and both earn their place inside a balanced test in which multiple-choice and constructed-response items carry the measurement of higher abilities. A teacher who knows their significance uses true-false items to sweep quickly across the factual ground of a unit and matching exercises to consolidate the connected sets of facts — terms, persons, events — around which a subject is organised.

Conclusion

To conclude, the significance of true-false tests lies in their speed, breadth of sampling, ease of objective scoring and power to expose misconceptions; the significance of matching tests lies in their compactness, their measurement of association and discrimination among related facts, and their reduced guessing. Both have real limitations — guessing in the one, restriction to factual association in the other — which careful construction and sensible combination with other item types keep within bounds. Within those bounds they remain indispensable, economical tools of classroom evaluation.

Question 3: Internal and External Assessment Critically Examined (20 Marks)

Critically examine internal and external assessment.

Introduction

Assessment in education is conducted either by the student's own teachers within his institution — internal assessment — or by an outside agency, such as a board of intermediate and secondary education or a university examinations department, which sets and marks examinations for students it has never taught — external assessment. Every educational system must decide how much weight to give each, and the debate between them is one of the oldest in evaluation, because each possesses strengths the other lacks and weaknesses the other exposes. This answer defines each form, examines its merits and defects critically, and concludes with the balanced arrangement toward which modern systems, including Pakistan's, have been moving.

Internal Assessment: Meaning and Merits

Internal assessment is the evaluation of students by their own institution, based on evidence gathered throughout the course: class tests, assignments, projects, practical work, participation, attendance and the teacher's continuing observation. Its first merit is continuity: it samples performance on many occasions across the whole course, so a single bad day, illness or accident cannot ruin a student's record, as it can in a single external examination. Because the evidence is broad, internal assessment can measure what an external paper cannot —

practical skills, oral work, project work, attitudes, regularity, initiative and the process of learning as well as its product.

Its second merit is its effect on learning. When assessment is continuous and its results return quickly, it becomes formative: students receive feedback while improvement is still possible, develop regular study habits instead of cramming at the end, and are freed from much of the fear and unhealthy pressure of the all-or-nothing final examination. Teachers, thirdly, can align internal assessment with exactly what they taught, at the moment they taught it, and can use the results diagnostically to improve their own teaching. Fourthly, internal assessment respects individual differences and institutional variety: a school may emphasise its own strengths and local needs in ways a common external paper cannot. Finally, it is economical and administratively light, being absorbed into the ordinary work of the school.

Internal Assessment: Defects

The defects of internal assessment are, however, serious, and they concern above all comparability and trust. First, standards differ from school to school and teacher to teacher: eighty marks in one school may represent less learning than sixty in another, so internal marks cannot fairly be compared for admission, scholarships or employment across institutions. Second, the teacher is not a disinterested judge. Favouritism, personal liking, pressure from parents and influential persons, and the school's desire to advertise good results all push marks upward — the well-documented phenomenon of mark inflation — while personal dislike can push an individual student's marks down. The student, knowing his teacher controls his marks, may cultivate favour rather than learning.

Third, teachers vary in assessment competence: not all can construct reliable tests or mark consistently, and heavy teaching loads make careful continuous assessment burdensome, so in practice it often degenerates into record-keeping and routine tests of recall. Fourth, internal assessment by itself lacks public credibility; employers and universities, unable to verify the conditions under which the marks were earned, discount them. Fifth, because internal marks are awarded piecemeal by many hands, moderation (statistical adjustment between schools) is technically difficult, and attempts at it regularly provoke disputes.

External Assessment: Meaning and Merits

External assessment is conducted by an agency independent of the teaching institution, which sets common examination papers, administers them under uniform conditions, and marks them centrally, often anonymously. Its first merit follows directly: uniformity and comparability. Every candidate answers the same questions under the same rules, marked to the same standard, so results can legitimately be compared across schools, districts and years, and used for selection far beyond the school. Its second merit is impartiality and credibility: the examiner does not know the candidate, cannot be approached by his parents, and has no interest in the school's reputation, so the certificate it issues commands public confidence at home and abroad.

Third, external examinations exert a powerful, if double-edged, influence on standards: they guarantee a common minimum curriculum coverage in every school, however remote, and give the system data by which weak schools and neglected topics can be identified. Fourth, they relieve the classroom teacher of the invidious role of final judge, preserving the teacher-student

relationship as one of helper rather than examiner. Fifth, centralised examining concentrates expertise: paper-setting, moderation, marking standardisation and result processing are done by specialists with resources no single school possesses.

External Assessment: Defects

Critically examined, external assessment has defects as grave as its merits. It is a one-occasion, high-stakes snapshot: three hours on one day decide a year or more of work, so chance — the questions that happen to be set, the candidate's health, family circumstances, examination hall conditions — weighs far too heavily. Its backwash on teaching is notorious: teachers teach to the examination, the curriculum narrows to what is examinable, predictable questions encourage memorisation of prepared notes and guidebooks, and unexamined skills — practical work, oral expression, creativity, character — are neglected by teachers and students alike. Objective validity is therefore compromised: the examination claims to certify education but samples a thin and distorted slice of it.

Further, external examinations provide no useful feedback for learning: results arrive months later, usually as bare marks, when the teaching is over. They impose heavy stress with documented effects on students' health and honesty — the use of unfair means in examination halls is largely a product of the stakes attached to a single event. They are expensive to run at national scale, vulnerable to paper leakage and administrative failure, and insensitive to local conditions and to the actual course of study followed in a particular school. Finally, by making a distant agency the arbiter, they can depress teacher professionalism, reducing the teacher to a coach for someone else's test.

Table 1: Internal and external assessment compared

Aspect	Internal Assessment	External Assessment
Who assesses	Own teachers/institution	Independent board or agency
Timing	Continuous, throughout the course	Once, at the end of the course
Comparability	Low — standards vary by school	High — common paper and marking
Impartiality	At risk of favouritism and inflation	High — anonymous, disinterested
What can be measured	Broad: skills, projects, attitudes, process	Narrow: mainly written, examinable content
Feedback for learning	Immediate and formative	Delayed; usually marks only
Effect on teaching	Supports regular study and variety	Backwash: teaching to the test, cramming
Public credibility	Limited outside the institution	High — trusted certificates

USMAN NOTES www.usmannotes.com

Towards a Balanced System

The critical conclusion of the comparison is that neither form is dispensable and neither is sufficient. The defects of each are roughly the mirror image of the merits of the other: internal assessment is valid in scope but unreliable across institutions; external assessment is reliable and comparable but narrow in scope and harsh in its effects. Modern reform therefore combines them. A moderate, moderated internal component — commonly twenty to thirty per cent of the total, or school-based assessment of practicals, projects and coursework — is added to the external examination, with safeguards: common criteria and rubrics, samples of internally marked work re-marked externally, statistical moderation against the school's external results, and training and monitoring of teacher-assessors. In this arrangement the external examination guarantees fairness and standards, while internal assessment restores breadth, continuity and feedback to the system.

Conclusion

To conclude, internal assessment offers continuity, breadth, feedback and freedom from one-day chance, but suffers from incomparable standards, bias and inflation; external assessment offers uniformity, impartiality and public trust, but suffers from narrowness, harmful backwash, stress and absence of feedback. A critically informed system does not choose between them: it assigns the external examination the certification and selection functions where comparability is paramount, assigns internal assessment the developmental and breadth functions where only the teacher can judge, and binds the two together with moderation so that the weaknesses of each are controlled by the strengths of the other.

Question 4: Main Components of Internal Assessment and External Assessment (20 Marks)

Discuss the main components of internal assessment and external assessment.

Introduction

Question 3 of this assignment compared the merits and defects of internal and external assessment; the present question asks what each system is actually made of — its components. A component here means a distinct element of the assessment system: who performs it, what evidence it uses, when it occurs, how its results are recorded and combined, and what safeguards maintain its quality. Describing the components matters practically, because an administrator who must design or reform an assessment scheme has to assemble these elements deliberately; a scheme fails when a component — say, record keeping in internal assessment, or moderation in external assessment — is missing. This answer therefore

describes the main components of internal assessment and then of external assessment, element by element.

Components of Internal Assessment

Internal assessment is not a single act but a system with the following interrelated components.

- Assessors: The primary assessors are the student's own teachers, supervised by the head of the institution. In richer schemes, peer assessment and student self-assessment are added as supplementary components, both for their educational value and because they train students in the criteria of good work.
- Evidence base (instruments): Internal assessment draws on many instruments spread over the course — periodic class tests and term tests, homework and written assignments, projects and term papers, practical and laboratory work, oral tests and presentations, classroom participation, and structured observation of attitudes and skills. The breadth of this evidence base is precisely what distinguishes internal assessment: no single instrument carries the whole judgement.
- Schedule and continuity: Assessment events are distributed across the term or year according to a published assessment calendar, so that evidence accumulates continuously and students know in advance what will be assessed and when. Weighting rules specify how much each kind of evidence contributes to the term or final internal score.
- Criteria, rubrics and marking schemes: For each task, the expected standard is expressed in criteria or rubrics — what counts as excellent, good, average or weak work — and in

marking schemes for tests. Shared criteria across the teachers of a subject are essential if internal marks within one school are to mean the same thing in different classrooms.

- Record keeping: A formal record system — mark registers, cumulative record cards and, increasingly, a computerised result system — stores every score with its date and instrument. Records are the memory of the system: they allow progress to be traced, parents to be informed, and the final internal grade to be computed transparently rather than from impression.
- Feedback and reporting: Internal assessment includes returning work promptly with comments, discussing results with students, issuing progress reports to parents, and holding parent-teacher meetings. This reporting component converts scores into guidance and is the point at which assessment improves learning.
- Aggregation and internal moderation: At the end of the period, component scores are combined by the announced weights into the internal grade. Before grades are finalised, the head or a moderation committee reviews them — comparing distributions across teachers and subjects, checking borderline cases — to correct obvious leniency or severity within the institution.
- Remedial follow-up: A complete internal system includes action on the results: remedial teaching for identified weaknesses, enrichment for the advanced, and counselling where the record shows personal or study problems. Assessment that changes nothing is an incomplete system.

Components of External Assessment

External assessment, conducted by a board or university, has a different and more industrial set of components, each designed to secure uniformity and security at large scale.

- The examining authority: A statutory board or examinations department with legal authority to set examinations, register candidates, issue rules and award certificates. Its committees — board of studies, paper-setters' and examiners' panels — are its working organs.
- Syllabus and examination regulations: The published syllabus, scheme of studies, paper pattern, marks distribution and rules of eligibility define what will be examined and under what conditions candidates may enter. These documents bind setters, schools and candidates alike, and are the first guarantee of fairness.
- Paper setting and moderation: Trained paper-setters draft question papers with marking schemes against the syllabus and a table of specifications; moderation committees review the papers for coverage, level, clarity and balance, and check the marking scheme against model answers. Several sets may be prepared as security against leakage.
- Registration and logistics: Candidate registration, issuance of roll numbers and date sheets, establishment of examination centres, appointment of superintendents and invigilators, and the secure printing, storage and distribution of question papers form the administrative component — invisible when it works, catastrophic when it fails.
- Conduct of the examination: Uniform instructions, fixed timings, identity checks, seating plans and invigilation enforce the same conditions for every candidate. Rules and penalties

for unfair means, with flying squads and inspection teams, protect the integrity of the event.

- Marking system: Scripts are marked centrally or by appointed examiners working to the approved marking scheme, with head-examiners standardising the marking of their teams, sample re-checking, and coding of scripts to preserve anonymity. This component carries the whole burden of comparability.
- Result processing and moderation of results: Marks are totalled, verified and processed — nowadays by computer — with checks for totalling errors, grace-mark rules applied uniformly, and results declared with mark sheets showing subject-wise performance.
- Certification, verification and grievance handling: The authority issues certificates and detailed mark sheets, provides verification to employers and institutions, and operates re-checking, re-totalling and appeals procedures through which a candidate can challenge an error. Post-examination analysis of results — item statistics, school-wise comparisons — feeds back into future paper setting and into the system's quality control.

Comparison of the Two Component Structures

Comparing the two sets of components shows the logic of each system. Internal assessment is organised around the teacher and the learning process: its characteristic components are continuous evidence, criteria shared within the school, records, feedback and remediation — components that serve improvement. External assessment is organised around the authority and the examination event: its characteristic components are regulation, secure logistics, standardised conduct, central marking and certification — components that serve uniformity and trust. Notice that both systems nevertheless require the same underlying skeleton: defined

content, instruments, marking criteria, records, quality control and a route of appeal. Where a school's internal assessment lacks components such as rubrics or moderation it degenerates into impression; where an external system lacks components such as paper moderation or result verification it degenerates into scandal. In combined systems the interface between the two is itself a component: internal marks are submitted on prescribed proformas, verified, moderated statistically and merged with external marks by published rules.

Table 2: Components at a glance

Component Area	Internal Assessment	External Assessment
Responsible agent	Teachers and head of institution	Board/university examination authority
Content basis	Taught course and class programme	Published syllabus and paper pattern
Main instruments	Class tests, assignments, projects, practicals, observation	Common written (and practical) examination papers
Timing	Continuous, many occasions	Fixed examination session at course end
Quality control	Shared criteria, head's review, internal moderation	Paper moderation, standardised marking, re-checking
Output	Progress reports and internal grade	Mark sheet and certificate
Appeal route	Teacher and head of institution	Formal re-checking/appeal procedures of the board

Conclusion

To conclude, internal assessment is composed of teacher-assessors, a broad continuous evidence base, a published schedule with weights, shared criteria and marking schemes, systematic records, prompt feedback and reporting, internal moderation and remedial follow-up. External assessment is composed of an examining authority, syllabus and regulations, moderated paper setting, secure registration and logistics, standardised conduct, scheme-

based central marking, verified result processing, and certification with verification and appeals. Understanding these components is the practical key to running either system well — and to combining them, as modern systems do, without importing the weaknesses of one into the other.

Question 5: Important Methods of Grading in Educational Evaluation (20 Marks)

Write important methods of grading in connection with educational evaluation.

Introduction

After students have been measured, their results must be reported. The traditional way is the mark: 63 out of 100, an apparently precise number. Yet marks carry well-known defects — they pretend to a precision that marking cannot deliver, they differ in meaning between subjects and examiners, and a difference of one mark can separate pass from fail although it is well within the error of measurement. Grading is the alternative: instead of reporting a fine numerical score, performance is placed in one of a small number of broad categories — letter grades such as A, B, C, D and F, or descriptive grades such as Excellent, Good, Average — each standing for a range of achievement. This answer explains the main methods by which grades are determined: absolute grading, relative (norm-referenced) grading on the curve, criterion-based grading, and the supporting practices of grade points, cumulative averages and descriptive reporting, and it evaluates the strengths and limits of grading as a whole.

USMAN NOTES www.usmannotes.com

Absolute Grading

In absolute grading, fixed mark ranges are decided in advance and every student who reaches a range receives its grade, regardless of how many others do: for example, 80 per cent and above = A (Excellent), 70–79 = B (Very Good), 60–69 = C (Good), 50–59 = D (Satisfactory), 40–49 = E (Pass) and below 40 = F (Fail). The method is simple to explain, transparent to students and parents, and stable across years as long as the papers are of comparable difficulty: a student knows his target before he sits the examination and competes only with the standard, not with his class-fellows. Its weakness is the mirror image of that stability: if one year's paper is harder than another's, the same grade no longer means the same attainment, and in practice either very many or very few students crowd into the top grade, bringing the standard itself into question. Absolute grading therefore works best where papers are carefully standardised from year to year or where grades are awarded subject by subject by experienced examiners who defend the boundaries.

Relative Grading (Grading on the Curve)

Relative grading — also called norm-referenced grading or grading on the curve — awards grades according to the student's position in his group, assuming that achievement in a large group is distributed approximately according to the normal probability curve. A common scheme assigns the top 7 per cent of students grade A, the next 24 per cent B, the middle 38 per cent C, the next 24 per cent D and the lowest 7 per cent E or F; other schemes use proportions derived from standard-deviation units, each grade band being a slice of the normal curve. The figure below illustrates the principle: the curve of scores is cut into bands, and a student's grade depends on where in the distribution his score falls.

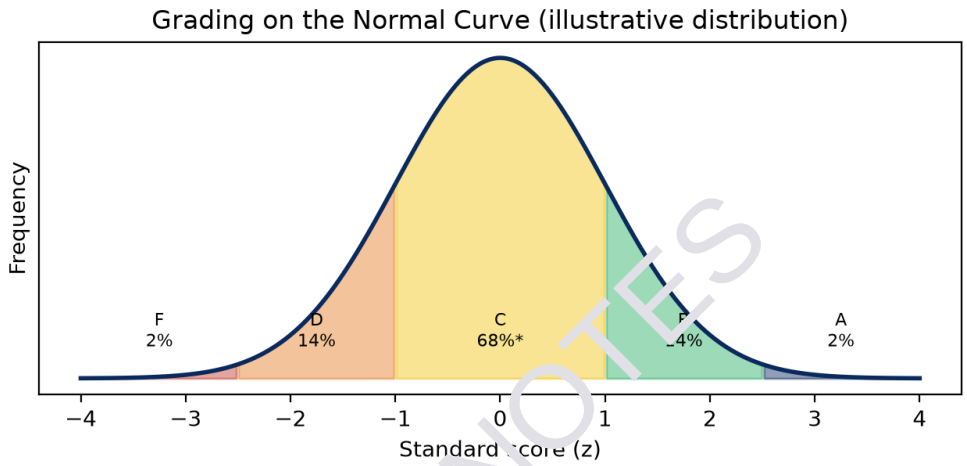


Figure 1: Relative grading — the normal curve is divided into grade bands, and grades are awarded by a student's position in the group (proportions illustrative; the middle band shown combines the central bands of a typical scheme).

The strengths of relative grading are real. It automatically adjusts for a paper that turns out too easy or too hard, it spreads students in a predictable and defensible way, and it suits selection situations where a fixed proportion must be chosen — for example, awarding a fixed number of scholarships. Its defects are equally real. It is unjust in small or unrepresentative groups: the best student of a weak class receives A while a stronger student in an able class may receive C, and if a whole class is excellent, some must still be graded low. It makes students compete against each other rather than against a standard, discouraging cooperation, and it reports nothing about what a student actually knows — an A in a year when the cohort was weak certifies position, not mastery. For these reasons pure curve grading is now generally reserved for large, stable cohorts, and is often softened by fixing minimum absolute standards beneath which no grade is awarded however the distribution falls.

Criterion-Referenced and Descriptive Grading

Criterion-referenced grading awards the grade by comparing the student's performance with a described standard of competence rather than with either fixed percentages or other students. Grade descriptors state what the holder can do: for instance, A = "can solve unfamiliar problems and explain the method clearly", C = "can solve routine problems with minor errors". This method is standard in competency-based and vocational education, in practical subjects and in the assessment of skills, where the social question is not "who is best?" but "who can do the job safely and well?" Its strength is meaning: the grade tells an employer or a parent exactly what has been mastered. Its cost is the labour of writing good descriptors and training assessors to apply them identically.

Closely allied is descriptive or qualitative grading, widely used in primary education: instead of letters or marks, the report uses graded descriptions — Outstanding, Very Good, Good, Satisfactory, Needs Improvement — often separately for each subject, skill and even for conduct, neatness, participation and other traits. Because young children's development is uneven and because fine mark differences among them are educationally meaningless and emotionally harmful, descriptive grading reports growth without branding a child with a rank. Many systems combine it with a brief narrative comment by the teacher.

Grade Points, GPA and Other Reporting Practices

To combine grades across subjects, grades are converted into grade points — for example A = 4, B = 3, C = 2, D = 1, F = 0 — weighted where necessary by the credit hours of each course, and averaged to give the Grade Point Average (GPA) for a term and the Cumulative GPA (CGPA) for

the whole programme. This is the standard practice of universities, including the semester systems in Pakistan, and it allows a single summary figure for standing, honours and probation decisions while the transcript retains the individual subject grades beneath it. Other methods in use include the nine-point stanine scale, which divides the normal distribution into nine bands with fixed percentages (4, 7, 12, 17, 20, 17, 12, 7, 4), percentile reporting, and pass/fail (satisfactory/unsatisfactory) grading used for courses where fine discrimination is unnecessary. Some institutions report both marks and grades side by side, taking from each what it does best.

Table 1. Principal methods of grading compared

Method	Basis of Award	Best Suited To	Chief Limitation
Absolute grading	Fixed mark ranges set in advance	School and board examinations with stable standards	Sensitive to variation in paper difficulty
Relative grading (on the curve)	Student's position in the group distribution	Large cohorts; selection with fixed quotas	Unfair in small or unusual groups; says nothing about mastery
Criterion-referenced grading	Described standard of competence	Skills, practical and vocational subjects	Needs careful descriptors and trained assessors
Descriptive grading	Qualitative levels with comments	Primary classes; traits and co-curricular areas	Less precise for selection purposes
Grade points / GPA	Grades converted to points, averaged (weighted by credits)	University semester systems, transcripts	Averages can hide weakness in a single subject

Merits and Limitations of Grading in General

Whichever method is chosen, grading as such has common merits. It is more honest than marks, because a band openly admits the margin of error that a figure like 61 conceals. It reduces unhealthy competition, stress and the stigma of small differences; a student graded B

is not tormented by the classmate one mark above him. It improves comparability between subjects whose marking scales differ — raw marks in mathematics and in essay subjects notoriously do not mean the same thing, while grades awarded judiciously come closer to doing so. And it simplifies reporting and decision-making for promotion, honours and admission. The general limitations must also be faced: a grade band hides real differences within the band, so information is lost for fine selection decisions; students tend to aim at the lower edge of a desired band; and if grade boundaries and methods are not published, grading can appear arbitrary. The practical remedy adopted by most systems is a hybrid: grades for reporting and certification, supported by published criteria, with subject marks retained in the record for those decisions that genuinely need finer detail.

Conclusion

To conclude, the important methods of grading are absolute grading by fixed mark ranges, relative grading on the normal curve, criterion-referenced grading against described standards, descriptive grading for young learners and special traits, and the conversion of grades into grade points and grade point averages for cumulative reporting. Each method answers a different need — stability, comparability, meaning, kindness to the learner, or convenient summary — and each has limits that the others relieve. A well-designed evaluation system therefore chooses its grading method deliberately, publishes it openly, and uses it consistently, so that a grade becomes what it should be: a fair, intelligible and trustworthy summary of a student's achievement.