South Valley University
Qena Faculty of Education
Department of Educational Psychology

Lectures on

SCALES and TESTS

For 3rd year students

Prepared by
Mahsoub Abdelkader
Professor of Educational Psychology

Academic Year
2022-2023

Chapter I
Common Terms in Scales and Tests

Measurement

According to Norman E. Ground (1985): Measurement is the process of obtaining a numerical description of the degree to which an individual possesses a particular characteristic.

According to Ebel & Frisbie (1991): Measurement is the process of assigning numbers to individuals or their characteristics according to specified rules.

Also, measurement defined as:

- Measurement- Process of collecting data on attribute of interest.
- Measurement is an act or process that involves the assignment of numerical values to whatever is being tested. So it involves the quantity of something.
- Measurement is the term used to describe the assignment of number to a given assessment. The number can be a raw score or a score based on a normal distribution curve. The process of quantifying this number is separate from using this information to evaluate student outcomes and achievement.
- The process of gathering information to monitor progress and make educational decisions if necessary.

Some of the basic principles of measurement in educational evaluations would be raw scores, percentile ranks, derived scores, standard scores, etc.

Measurement Process Involves Four Steps

1. Define the characteristics that you want to measure.
2. Select the appropriate test. This may also mean to select the appropriate testing instrument.
3. Administer the test. If an instrument is involved in the testing, this also means to use the instrument correctly.
4. Collect and record the measurement from the test.

Considerations When Taking Measurements

1. Remember that you are measuring a characteristic of the person- You are not measuring the person themselves. Thus, make no judgments about the person.
2. Make no comical remarks regarding the collected data.
3. Have high ethical standards when collecting the data.
4. Be professional.

Types of Measurements

- Objective measurements

An objective measurement is one that cannot be interpreted differently because of numerical values.

- Subjective measurements

A subjective measurement is one that can possibly be interpreted differently.

Purposes of Test & Measurement

- For getting knowledge about the progress.

- For preparation of effective planning.
- For knowing the abilities and capacities.
- For giving motivation.
- For knowing the achievements in future.
- For research and experimentations.

Assessment

According to Murry Print (1993): Assessment involves the interpretation of measurement data. It makes sense of the data collected on student performance.

According to Dictionary of Education (1989): Assessment is the process whereby one attempts to measure the quality and quantity of learning and teaching using various assessment techniques.

Also, assessment defined as:

- Assessment can focus on the individual learner, the learning community (class, workshop, or other organized group of learners), the institution, or the educational system.
- The process of gathering quantitative and qualitative data of what a student can do. And how much a student possesses.
- An assessment may include a test, but also includes methods such as observations, interviews, behavior monitoring, etc.
- Assessment is the process of documenting, usually in measurable terms, knowledge, skill, attitudes and beliefs.

Assessment in education is the process of gathering, interpreting, recording & using information about pupils' responses to an educational Task.

In education, the term assessment refers to the wide variety of methods or tools that educators use to evaluate, measure, and document the academic readiness, learning progress, skill acquisition, or educational needs of students.

Purpose of Assessment is:

- To diagnoses students' strengths and weaknesses.
- To assign grads,
- To determine the teachers' effectiveness.
- To monitor students' progress.
- To help evaluate teachers.

Also, we can enlist the Purposes of Assessment as:

- motivating and directing learning,
- providing feedback to student on their performance,
- providing feedback on instruction and/or the curriculum,
- Ensuring standards of progression are met.

Assessment is used to…

(1) Measure students learning and progress.
(2) Improve teaching techniques.
(3) Improve the quality of teaching progress.
(4) Raise student achievement.
(5) Report to parents on student progress.

Classroom assessment can help teachers answer the following specific questions:

- To what extent are my students achieving the stated goals?
- How should I allocate class time for the current topic?

- Can I teach this topic in a more efficient or effective way?
- What parts of this course/unit are my students finding most valuable?
- How will I change this course/unit the next time I teach it?
- Which grades do I assign my students?

Three forms of assessment

- ❖ Diagnostic- pre testing students to see what they know before teaching the unit.
- ❖ Formative- assessing students' strengths and weaknesses, and providing feedback during the unit.
- ❖ Summative- Testing the student's knowledge at the end of teaching a unit.

Diagnostic Assessment

You can diagnose the student's strengths and weaknesses. This will help you identify what you need to do to help the student become stronger.

Formative Assessment

Are in-process evaluations of student learning that are typically administered multiple times during a unit, course, or academic program. The general purpose of formative assessment is to give educators in-process feedback about what students are learning or not learning so that instructional approaches, teaching materials, and academic support can be modified accordingly. Formative assessments are usually not scored or graded, and they may take a variety of forms, from more formal quizzes and

assignments to informal questioning techniques and in-class discussions with students.

Formative assessment is the ongoing analysis of a student's needs - we must recognize when it needs to be learned, motivated, and provided with positive reinforcement in order for it to grow. Most formative assessment is informal. The feedback and response involves both teacher and student.

Formative assessment has the greatest impact on learning and achievement. Formative assessment works best when the students understand what the learning objectives are.

Summative Assessment

Summative assessment are used to evaluate student learning at the conclusion of a specific instructional period—typically at the end of a unit, course, semester, program, or school year. Summative assessments are typically scored and graded tests, assignments, or projects that are used to determine whether students have learned what they were expected to learn during the defined instructional period.

Summative assessment of the measuring of the student's growth at a point. The measurement tells us how much the students have grown. It does not affect the growth of the students. Measures the progress a student has made. Is often used at the end of a unit of work.

Formative assessments are commonly said to be *for* learning because educators use the results to modify and improve teaching techniques during an instructional period, while

summative assessments are said to be *of* learning because they evaluate academic achievement at the conclusion of an instructional period. Or as assessment expert Paul Black put it, "When the cook tastes the soup, that's formative assessment. When the customer tastes the soup, that's summative assessment."

Measurement error in education generally refers to either (1) the difference between what a test score indicates and a student's actual knowledge and abilities or (2) errors that are introduced when collecting and calculating data-based reports, figures, and statistics related to schools and students.

Because some degree of measurement error is inevitable in testing and data reporting, education researchers, statisticians, data professionals, and test developers often publicly acknowledge that performance data, such as high school graduation rates or college-enrollment rates, are not perfectly reliable (they may even report the "margin of error" for a given statistic or finding) or that test scores don't always accurately reflect what students know or can do—i.e., that there is no such thing as a perfectly reliable test of student knowledge and skill acquisition.

Measurement errors in testing may result from a wide variety of factors, such as a student's mental and emotional state during the test period or the conditions under which the test was administered. For example, students may have been unusually tired, hungry, or emotionally distressed, or distractions such as loud noises, disruptive peers, or technical problems could have adversely affected test performance. Test scores for young

children are often considered to be especially susceptible to measurement error, given that young children tend to have shorter attention spans and they may not be able to fully comprehend the importance of the test and take it seriously. In addition, young children of the same chronological age or grade level may be at very different stages of social, cognitive, and emotional development, and if a young child experiences a rapid developmental growth spurt, test results could quickly become outdated and therefore misrepresentative.

The following is a representative list of a few additional factors and problems that may give rise to measurement error in testing:

- Ambiguously phrased questions or inaccurate answers.
- Test items, questions, and problems may not address the material students were actually taught.
- Performance levels and cutoff scores, such as those considered to be "passing" or "proficient" on a particular test, may be flawed, poorly calibrated, or misrepresentative.
- The scoring process may be poorly designed, and both human scorers and computer-scoring systems may make mistakes.
- Test administrators could give students incorrect directions, help students cheat, or fail to create calm and conducive test-taking conditions.
- Test-result data may be inaccurately recorded and reported.

Measurement errors in the reporting of education data and statistics are common and, to a greater or lesser extent, both expected and unavoidable. While human error may lead to inaccurate reporting, data systems and processes are intrinsically

limited—i.e., it is simply not possible to create perfect data systems or collect data flawlessly, particularly as systems grow in scale and scope. National or statewide data systems—e.g., systems administered by government agencies to track important educational data such as high school graduation rates—are especially prone to measurement error, given the massive complexities entailed in collecting data from thousands of schools on the performance of hundreds of thousands or millions of students. For this reason, most large-scale education data are openly qualified as estimates.

The following is a representative list of a few additional factors and problems that may give rise to measurement error in educational data:

- Flawed, imprecise, or mismanaged data-collection processes resulting in incorrect reports, records, figures, and statistics.
- An absence of clear and understandable rules, guidelines, and standards for data collection and reporting processes, or ambiguous guidelines that give rise to misinterpretation and error.
- Small sample sizes—such as in rural schools that may have small student populations and few minority students—that may distort the perception of performance for certain time periods, graduating classes, or student groups.
- Divergent data-collection and data-reporting processes—such as the unique data-collection systems and requirements developed by states—that can lead to misrepresentative comparisons or systems incompatibilities that produce errors.

- High rates of transfer in and out of school systems—e.g., by the children of transient workers—that make it more difficult to accurately track the enrollment status of students.
- Lack of adequate training, experience, or technical expertise in proper data-collection and -reporting procedures among those responsible for collecting and reporting data at the school, district, and state levels.
- Intentional misrepresentations of student performance and enrollment, such as those that may accompany high-stakes testing.

Reform

While some degree of measurement error is—and perhaps always will be—unavoidable, many educators, schools, districts, government agencies, and test developers are taking steps to mitigate measurement error in both testing and data reporting.

In testing, measurement error is generally considered a relatively minor issue for low-stakes testing—i.e., when test results are not used to make important decisions about students, teachers, or schools. As the stakes attached to test performance rise, however, measurement error becomes a more serious issue, since test results may trigger a variety of consequences.

Measurement error is one reason that many test developers and testing experts recommend against using a single test result to make important educational decisions. For example, the Standards for Educational and Psychological Testing—a set of proposed guidelines jointly developed by the American Educational Research Association, American Psychological Association, and

the National Council on Measurement in Education—recommends that "in elementary or secondary education, a decision or characterization that will have a major impact on a test taker should not automatically be made on the basis of a single test score."

The following are a few representative strategies that educators and test developers may employ to reduce measurement error in testing:

- Test developers can carefully review questions for test bias and fairness, and remove or revise items that may adversely affect the performance of students of different races, cultural groups, or genders.
- Test developers can conduct pilot tests to get feedback on difficulty levels, phrasing clarity, and bias, and then revise tests before they are administered.
- To reduce errors in the human scoring of questions that cannot be scored by computer, such as open-response and essay questions, two or more scorers can score each item or essay. If they disagree, the item can be passed on to additional scorers.
- Schools can tighten security practices to combat and prevent cheating by those administering and taking the tests.
- Policy makers can lower or eliminate the consequences resulting from test results to minimize score inflation and reduce the motivation to manipulate results.
- Instead of relying on one potentially inaccurate measure, schools can get more comprehensive information by using multiple methods to assess student achievement and learning growth.

In educational data collection and reporting, measurement error can also become a significant issue, particularly when school-funding levels, penalties, or the perception of performance are influenced by publicly reported data, such as dropout rates or graduation rates, for example. For these and other reasons, improving the quality and accuracy of data systems, collection processes, and reporting requirements has become a growing priority for schools, policy makers, and government agencies, and a variety of organizations and initiatives, such as the Data Quality Campaign and the Common Education Data Standards, are working to improve quality, consistency, and reliability of education data.

The following are a few representative strategies that educators and data experts may employ to reduce measurement error in data reporting:

- "Unique student identifiers," such as state-assigned codes or social-security numbers, can be used to facilitate the tracking of individual students and increase data reliability as they move from grade to grade or school to school.
- Common data-collection and -reporting standards can be developed to improve the reliability of data and allow for performance comparisons across schools and states.
- Redundant processes—multiple systems and people checking for errors—can be used to improve reporting accuracy.
- Clearer guidelines and better training can be provided to those compiling and calculating data.
- Improved technology and the use of compatible or interoperable systems can facilitate data quality and the

exchange of data among different schools, organizations, and states.

Evaluation

Evaluation is the process of determining the worth or value of something. This involves assigning values to the thing or person being evaluated.

Evaluation is a systematic process of determining what the actual outcomes are but it also involves judgment of desirability of whatever outcomes are demonstrated (Travers, 1955).

Evaluation is the process of delineating, obtaining and providing useful information for judging decision alternatives (Stufflebeam et al 1971).

Educational Evaluation means examining student performance and comparing and judging its quality. Determining whether or not the learner has met the course objectives and how well.

Evaluation is concerned with a whole range of issues in and beyond education, lessons, programs, and skills can be evaluated. It produce a global view of achievements usually based on many different types of information such as observation of lessons, test scores, assessment reports, course documents or interviews with students and teachers.

Evaluation is the process of making overall judgment about one’s work or a whole school work.

Evaluation is a process of determining to what extend the educational objectives are being realized.

Functions of Evaluation

Evaluation answers the question:

- "How *well* did we do?"
- "How *much* did we do?

The Purposes of Evaluation

- To determine the relative effectiveness of the program in terms of students' behavioral outputs.
- To make reliable decisions about educational planning.
- To identify students' growth or lack of growth in acquiring desirable knowledge, skills, attitudes and societal values.
- To help teachers determine the effectiveness of their teaching techniques and learning materials.
- To help motivate students.
- To identify problems.
- To predict the general trend in the development of the teaching-learning process.

We can also enlist some purpose of evaluation as the follow:

- Clarify and define objects.
- Facilitate the improvement program.
- Motivate participants.
- Establish and maintain standards to, meet legal, professional and academic credentials.
- Test the efficiency of teachers.

Steps involved in making an evaluation

- Define the objective or the purpose of the test.
- Measure the performance or administer the test.
- Find or develop a standard.
- Compare a person's performance on the test to a standard.
- Make the evaluation then discuss and distribute the results in the most appropriate manner.

Forms (Types) of Evaluation

- Process Evaluation: it refers to evaluation taking place during the program or learning activity. It is conducted while the event to be evaluated is occurring & focuses on identifying the progress towards purposes, objectives, or outcomes to improve the activities, courses, curriculum, program or teaching and student. It is also known as formative evaluation.
- Product Evaluation: examines the effect of outcomes of some object. It conducted at the end of the course. It is also known as summative evaluation. It evaluates the progress towards an established outcomes..

Formative Evaluation

Formative evaluation provides diagnostic feedback to students and instructors at short-term intervals (during a class or on a weekly basis).

of all the evaluation strategies, formative evaluation is the most valuable strategy for supporting students' learning and for promoting students' independence and responsibility as learners.

Formative evaluation is a diagnostic use of evaluation to provide feedback to teachers and students over the course of instruction.

Means also, gathering of data during a time program is being develop. This likewise provide feedback for the improvement of an instruction or for the improvement of the program.

Formative evaluation is administrated at the end of the day's lesson. The goal of formative evaluation is to monitor students learning to provide ongoing feedback that can be used by instructors to improve their teaching and by students to improve their learning, and determines whether the teacher delivers quality instruction in a particular day-base on the particular result.

Advantages of Formative Evaluation

Formative evaluation enables the teacher to;

- Draw more reliable inference about his students than an external assessor, although he may not be as objective as the latter.
- Identify the levels of cognitive process of his students.
- Choose the most suitable teaching techniques and materials.
- Determine the feasibility of a program within the classroom setting.
- Determine areas needing modifications or improvement in the teaching-learning process.
- Determine to a great extent the outcome of summative evaluation.

Summative Evaluation

Summative Evaluation provides a description of students' level of attainment upon completion of an activity, module, or course.

A summative evaluation is one attempt to evaluate student learning for a specific time period. For example, a unit test would be a summative evaluation. Use to determine the mastery & achievement of the student. Done usually at the end of a chapter or unit. Accountability of success or failure. Use primary in a evaluating grades.

A summative evaluation is designed to determine the extent to which the instructional objectives have been achieved. The goal of summative evaluation is to evaluate student learning at the end of an instructional unit by comparing it against some standard or benchmark.

<u>Advantages of Summative Evaluation:</u>

1. Its existence (learners will need to be aware of it from the start). Provides motivation and helps create an appropriate learning environment.
2. Positive results give the trainees a boost in confidence and can act as a springboard into subsequent behavior change back in the workplace.
3. Trainers can identify those areas where results are consistently lower and can then consider alternative delivery methods- helping to develop the training for future events.
4. The results provide a measurable way of determining the success of the training program, directly comparable from one intake to the next.

Test

Test, appropriately, should be an instrument based on competence, i.e. it reflects the ability to present a correct answer to a specific item/question.

Test is an instrument or activity and systematic procedure for measuring a sample of behavior “How well does the individual perform”.

Test an instrument or activity used to accumulate data on a person’s ability to perform a specified task. It is an assessment intended to measure a test-taker’s knowledge, skill, aptitude, performance, or classification in many other topics.

Test is a method to determine a student's ability to complete certain tasks or demonstrate mastery of a skill or knowledge of content. Some types would be multiple choice tests, or a weekly spelling test.

Scale

Scale is commonly used with respect to attitudes or constructs for which there is not a correct response, but demands endorsing of an alternative among the presented.

Scale is an instrument, usually a questionnaire, designed to measure a person's position on some dimension. Examples: Intelligence scale; conscientiousness scale; self-esteem scale, etc.

Can be responded to by oneself or by others – yielding what is called self-report or other-report. Can assess ability, or attitudes, or personality

Inventory

Inventory is more often used to measure personality traits and vocational tendencies, for instance. They do not demand a correct answer, but the indication of which option is more adequate to characterizing the individual with regarding to a specific statement.

A self-report inventory is a type of psychological test in which a person fills out a survey or questionnaire with or without the help of an investigator. Self-report inventories often ask direct questions about personal interests, values, symptoms, behaviors, and traits or personality types. Inventories are different from tests in that there is no objectively correct answer; responses are based on opinions and subjective perceptions.

Questionnaire

Questionnaire is a broad concept. It can include a set of different instrument types (e.g., scales, inventories).

The questionnaire is the main instrument for collecting data in survey research. Basically, it is a set of standardized questions, often called items, which follow a fixed scheme in order to collect individual data about one or more specific topics. Sometimes questionnaires are confused with interviews. In fact, the questionnaire involves a particular kind of interview—a formal contact, in which the conversation is governed by the wording and order of questions in the instrument. The questionnaire often is

administered in a standardized fashion, that is, in the same way to all the respondents of the survey. The logic behind the standardization of questions and answers is that only if a stimulus is the same for all the respondents of a survey.

What's the difference between a test and examination?

A test and an exam both tests the knowledge of a student. So, in most cases tests and exams are synonyms. Both test the knowledge of the students with a series of questions and will grade the questions to get a result. They differ in one aspect: an exam is more formal then a test.

Examples of Tests

- International Olympiad of Science (IOS).
- Brilliant International Olympiad of Science (BIOS).

International Olympiad of Science (IOS): It is a test of competence and proficiency in Science and is held annually at the National and International levels, based on syllabus prescribed by CBSE/ ICSE and the other State Boards.

Brilliant International Olympiad of Science endeavors to assess the competency and proficiency of the students in the field of Science at national and international level every year. Its syllabus comprises of the syllabus of CBSE/ICSE and various State Boards. The BIOS examination is purely based on Science, which provides ample opportunity to the students not only to test their skills but also sharpen them accordingly.

The Program for International Student Assessment (PISA) is a worldwide study by the Organization for Economic Co-

operation and Development (OECD) in member and non-member nations intended to evaluate educational systems by measuring 15-year-old school pupils' scholastic performance on mathematics, science, and reading. It was first performed in 2000 and then repeated every three years. Its aim is to provide comparable data with a view to enabling countries to improve their education policies and outcomes. It measures problem solving and cognition.

Definition of Score

Score is a particular subject's value on a variable. For example Ahmed's score on the Physics exam, and Nada's score on the intelligence test.

Descriptive Statistics

Procedures for summarizing a set of scores or otherwise making them more comprehensible.

Descriptive statistics techniques that are used to describe or characterize the obtained sample data.

Measures of Central Tendency

A measure of central tendency is a descriptive statistic that describes the average, or typical value of a set of scores. There are three common measures of central tendency: the mode, the median, and the mean:

Median is the middle score of a sequence of all the scores in a distribution arranged from highest to lowest.

Mode is the value with the greatest frequency in a distribution.

Mean (the arithmetic average) is the arithmetic average of a group of scores; the sum of the scores divided by the number of scores. Is defined as the some of the scores divided by the number of scores in equation form:

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{X_1 + X_2 + ... + X_n}{n}$$

Where:
x_1, ……, x_n = raw scores
n= number of scores

The mean is:

- the arithmetic average of all the scores (ΣX)/N
- the number, m, that makes Σ(X - m) equal to 0
- the number, m, that makes $\Sigma(X - m)^2$ a minimum

Example: Calculate the mean of the following data:

1 5 4 3 2

- Sum the scores (ΣX): 1 + 5 + 4 + 3 + 2 = 15
- Divide the sum (ΣX = 15) by the number of scores (N = 5):

$$\bar{X} = \frac{15}{5} = 3$$

Practice problem

Calculate the mean for each of the following sample sets of scores:

a. x: 3, 5, 6, 8, 14

b. 20, 22, 28, 30, 37, 38

c. 2.2, 2.4, 3.1, 3.1

The Range

Is defined as the difference between the highest and lowest scores in the distribution. In equation form,

range = highest-lowest

For example, the range of the set of data: 2, 2, 5, 8, 10 is range = 10-2 = 8.

Deviation Score

Deviations from the average of the group. A deviation score tells how far away the raw score is from the mean of its distribution or a score minus the mean of all scores:

$$\sum(X - \bar{X})$$

Calculate deviation scores for the following distributions:

2, 4, 6, 8, 10

X	$x - \bar{x}$
2	2-6= -4
4	4-6 = -2
6	6-6= 0
8	8-6= +2
10	10-6 = +4

$$\bar{x} = \frac{2+4+6+8+10}{5} = \frac{30}{5} = 6$$

Practical Problem

Use the deviation scores to judge the individual differences according to the following data: 1, 5, 4, 3, 2

scores	average	deviations	Differences description
1	3	1-3 = -2	Below average
5		5-3 = +2	Above average
4		4-3 = +1	Above average
3		3-3 = 0	Average
2		2-3 = -1	Below average
Sum		zero	

Practical Problem

- Find the values of the expressions listed here:
- Find $\sum_{i=1}^{6} X_i$ for the scores $X_1=2$, $X_2=3$, $X_3=4$, $X_4=6$, $X_5=9$, $X_6=11$, $X_7=14$.
- Find $\sum_{i=3}^{N-1} X_i$ for the scores $X_1=22$, $X_2=24$, $X_3=28$, $X_4=35$, $X_5=38$, $X_6=40$.

Practical Problem

A biological psychologist records the number of cells in a particular brain region of cats that respond to a tactile stimulus. Nine cats are used. The following cell counts/animal are recorded: 15, 28, 33, 19, 24, 17, 21, 34, 12. Compute the following:

a. Mean

b. Range

Practical Problem

A cognitive psychologist measures the reaction times of 6 subjects to emotionally laden words. The following scores in milliseconds are recorded: 250, 310, 360, 470, 425, 270. Compute the following:

a. Mean

b. Range

Linking Assessment and Instruction

Assessment is an integral part of teaching and learning. Quality instruction and assessment are not necessarily different activities, and in fact, should become nearly indistinguishable.

When students experience difficulties and receive no useful feedback, they are likely to attribute their problems to a lack of ability, and give up. But when they receive specific information about ways in which they can improve and are given opportunities to revise their work, they receive a clear message from the teacher that gives them confidence and enables them to improve.

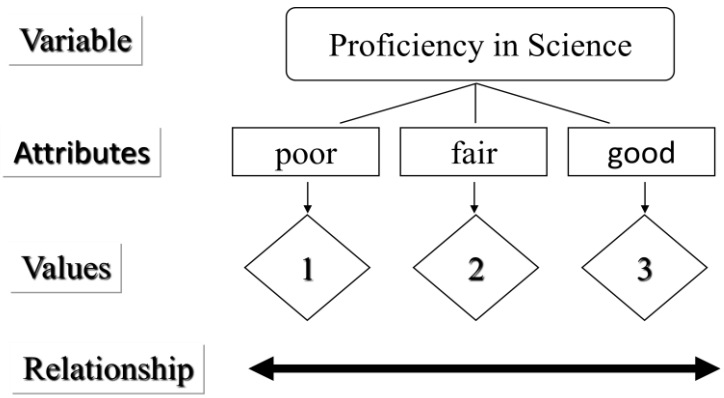
Chapter II
Levels of Measurement

Measurement is the process of assigning numbers to quantities. The process is so familiar that perhaps we often overlook its fundamental characteristics.

What Is Level of Measurement?

Attributes have properties that are similar to numbers. When we assign numbers to attributes, we can do so poorly, in which case the properties of the numbers to not correspond to the properties of the attributes. In such a case, we achieve only a "low level of measurement. On the other hand, if the properties of our assigned numbers correspond properly to those of the assigned attributes, we achieve a high level of measurement.

Level of measurement represents the relationship of the values that are assigned to the attributes for a variable:



There are typically four levels of measurement that are defined:

(1) Nominal
(2) Ordinal
(3) Interval
(4) Ratio

Why is level of measurement important?

- Helps you decide what statistical analysis is appropriate on the values that were assigned.
- Helps you decide how to interpret the data from that variable.

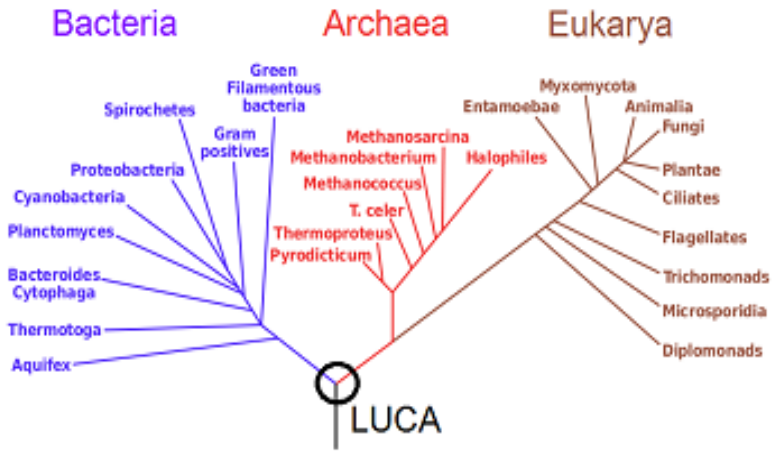
Nominal Measurement

A nominal scale is the lowest level of measurement and is most often used with variables that are qualitative in nature rather than quantitative.

Examples of qualitative variables are gender, nationality, religious affiliation, language, marital status, jersey numbers in football, type of car, ID numbers, and eye color.

In grammar, the parts of speech: noun, verb, preposition, article, pronoun, etc.

In biology, the taxonomic ranks below domains: Archaea, Bacteria, and Eukarya:



The values "name" the attribute uniquely. The name does not imply any ordering of the cases. Units of the scale are categories. Objects are measured by determining the category to which they belong. There is no magnitude relationship between the categories.

Ordinal Measurement

An ordinal scale needs no equal intervals between its points. Instead, only rank order is requisite. For instance, if a principal ranked five teachers in order of their effectiveness, this would be an example of data representing an ordinal scale (Popham, 1990: 89).

Ordinal measurement possesses the property of magnitude. Can rank-order the objects according to whether they possess more, less, or the same amount of the variable being measured. Thus can determine whether A > B, A= B, A < B. cannot determine how much greater or less A is than B in the attribute being measured.

In the case of ordinal measurement the distances between attributes do not have any meaning. For example, code gender as 0=female; 1= male. Is the distance from 0 to 1 have any meaning?

The ordinal type allows for rank order (1^{st}, 2^{nd}, 3^{rd}, etc.) by which data can be sorted, but still does not allow for relative degree of difference between them. Examples include, on one hand, dichotomous data with dichotomous (or dichotomized) values such as 'sick' vs. 'healthy' when measuring health, 'guilty' vs. 'not-guilty' when making judgments in courts, 'wrong/false' vs. 'right/true' when measuring truth value, and, on the other hand, non-dichotomous data consisting of a spectrum of values, such as 'completely agree', 'mostly agree', 'mostly disagree', 'completely disagree' when measuring opinion.

For ordinal data, any monotonic functional transform:

$$Y = F(X)$$

Interval Measurement

An interval scale is one which allows us to believe that there are actually equal intervals between equidistant points on the scale. For example, the ten-point difference between scores of 48 and 58 on a 100-item scale would be considered identical to the ten-point difference between scores of 88 and 98. If we are convinced that such is the case, then we have an interval scale (Popham, 1990: 89).

The interval type allows for the degree of difference between items, but not the ratio between them. Examples include

temperature with the Celsius scale, which has two defined points (the freezing and boiling point of water at specific conditions) and then separated into 100 intervals, date when measured from an arbitrary epoch (such as AD), location in Cartesian coordinates, and direction measured in degrees from true or magnetic north. Ratios are not meaningful since 20 °C cannot be said to be "twice as hot" as 10 °C (unlike temperature in Kelvins), nor can multiplication/division be carried out between any two dates directly. However, ratios of differences can be expressed; for example, one difference can be twice another. Interval type variables are sometimes also called "scaled variables", but the formal mathematical term is an affine space (in this case an affine line).

Interval Measurement possesses the properties of magnitude and equal intervals between adjacent units. Can do same determinations as with an ordinal scale, plus can determine whether A – B = C – D, A – B > C – D. When distance between attributes has meaning, for example, temperature (in Fahrenheit) -- distance from 30-40 is same as distance from 70-80. Note that ratios don't make any sense -- 80 degrees is not twice as hot as 40 degrees (although the attribute values are).

For interval data, any linear transform of the form:

$$Y = aX + b, \ldots\ldots a \succ 0$$

Ratio Measurement

A ratio scale, incidentally, is an interval scale for which a zero point exists, such as a weight scale or a height scale in which

it is possible to have a true zero. We encounter few ratio scales in educational measurement (Popham, 1990: 89).

Ratio measurement possesses the properties of magnitude, equal interval between adjacent units, and absolute zero point. Can do all the mathematical operations usually associated with numbers, including ratios. The data have all the properties of interval data and the ratio of two values is meaningful.

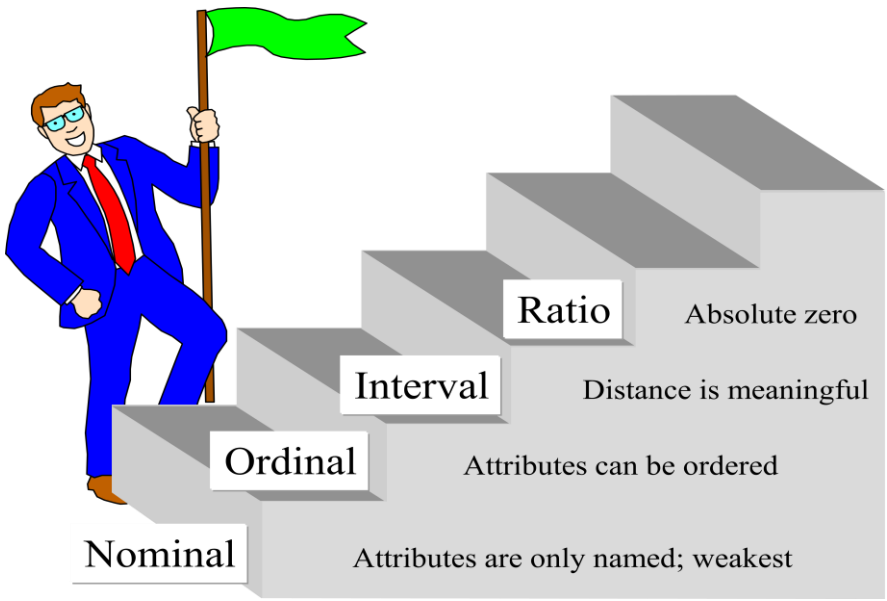
A ratio scale possesses a meaningful (unique and non-arbitrary) zero value. Most measurement in the physical sciences and engineering is done on ratio scales. Examples include mass, length, duration, plane angle, energy and electric charge. In contrast to interval scales, ratios are now meaningful because having a non-arbitrary zero point makes it meaningful to say, for example, that one object has "twice the length". Very informally, many ratio scales can be described as specifying "how much" of something (i.e. an amount or magnitude) or "how many" (a count). The Kelvin temperature scale is a ratio scale because it has a unique, non-arbitrary zero point called absolute zero.

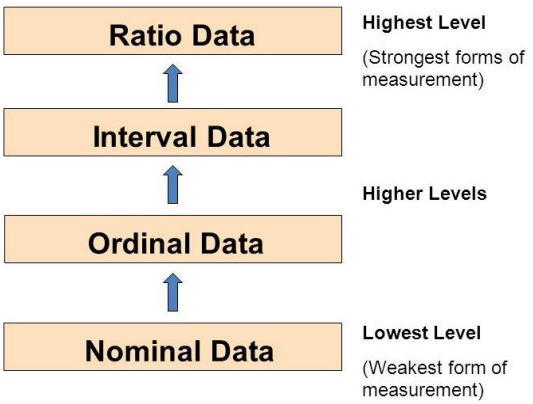
Ratio measurement has an absolute zero that is meaningful. Can construct a meaningful ratio (fraction), for example, number of students in past three years.

For ratio data, any multiplicative transform of the form:

$$Y = aX, \ldots\ldots\ldots a \succ 0$$

The Hierarchy of Levels:





Questions and Problems

(1) Identify the scaling of each of the variables:

a) The IQ of your teachers (assume equal interval scaling).

b) Proficiency in Science graded in the categories of poor, fair, and good.

c) Anxiety over public speaking scored on a scale of 0-100 (assume the difference in anxiety between adjacent units throughout the scale is not the same).

d) The weight of a group of dieters.
e) The time it takes to react to the sound of a tone.
f) Proficiency in mathematics is scored on a scale of 0-100. The scale is well standardized and can be thought of as having equal intervals between adjacent units.
g) Ratings of professors by students on a 50-point scale. There is an insufficient basis for assuming equal intervals between adjacent units.
h) Number of bicycles ridden by students in the freshman class.
i) Types of bicycles ridden by students in the freshman class.

…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………
…………………………………………………………………………

(2) A student is measuring assertiveness with an interval scale. Is it correct to say that a score of 30 on the scale represents half as much assertiveness as a score of 60? Explain.

Indicate which level of measurement is being used in the given scenario: "The teacher of a class of third graders records the height of each student".

Nominal
Ordinal
Interval
Ratio

Indicate which level of measurement is being used in the given scenario: "The teacher of a class of third graders records the eye color of each student".

Nominal
Ordinal
Interval
Ratio

Indicate which level of measurement is being used in the given scenario: "The teacher of a class of third graders records the letter grade for mathematics for each student".

Nominal
Ordinal
Interval
Ratio

Indicate which level of measurement is being used in the given scenario: "The teacher of a class of third graders records the percentage that each student got correct on the last science test".

Nominal
Ordinal
Interval
Ratio

Indicate which level of measurement is being used in the given scenario: "A meteorologist compiles a list of temperatures in degrees Celsius for the month of May".

Nominal
Ordinal
Interval
Ratio

Indicate which level of measurement is being used in the given scenario: "A meteorologist compiles a list of temperatures in degrees Fahrenheit for the month of May".

Nominal
Ordinal
Interval
Ratio

Indicate which level of measurement is being used in the given scenario: "A film critic lists the top 50 greatest movies of all time".

Nominal
Ordinal
Interval
Ratio

Indicate which level of measurement is being used in the given scenario: "A car magazine lists the most expensive cars for 2020".

Nominal
Ordinal
Interval
Ratio

Indicate which level of measurement is being used in the given scenario: "The roster of a basketball team lists the jersey numbers for each of the players".

Nominal
Ordinal
Interval
Ratio

Indicate which level of measurement is being used in the given scenario: "A local animal shelter keeps track of the breeds of dogs that come in".

Nominal
Ordinal
Interval
Ratio

Indicate which level of measurement is being used in the given scenario: "A local animal shelter keeps track of the weights of dogs that come in".

Nominal
Ordinal
Interval
Ratio

Identify the scale of measurement for the following: military title -- Lieutenant, Captain, Major.

Nominal
Ordinal
Interval
Ratio

Which is the corresponding level of measurement for the following data: The Science teacher's name is Sam Wilson

Nominal
Ordinal
Interval
Ratio

Which is the corresponding level of measurement for the following data: The teacher is 58 years old.

Nominal
Ordinal
Interval
Ratio

Which is the corresponding level of measurement: The years the school principal served are 2000, 2006, and 2012.

Nominal
Ordinal
Interval
Ratio

Identify the scale of measurement for the following categorization of clothing: hat, shirt, shoes, and pants.

Nominal
Ordinal
Interval
Ratio

Identify the scale of measurement for the following: heat measured in degrees Celsius.

Nominal
Ordinal
Interval
Ratio

City of birth is an example of a(n)

- Nominal
- Ordinal
- Interval
- Ratio

Choose the appropriate scale of measurement: As part of a test preparation course, students are asked to take a practice version of the Graduate Record Examination (GRE). This is a standardized test. Scores can range from 200 to 800.

- Nominal
- Ordinal
- Interval
- Ratio

Children in elementary school are evaluated and classified as non-readers (0), beginning readers (1), grade level readers (2), or advanced readers (3). The classification is done in order to place them in reading groups.

- Nominal
- Ordinal
- Interval
- Ratio

South Valley University wants to know which dormitories the students prefer. The administration counts the number of applications for each dorm. Administrators assign a rank to each dorm based on the number of applications received.

- Nominal
- Ordinal
- Interval
- Ratio

Chapter III
Achievement Tests

What is an Achievement Test?

Achievement is the accomplishment or proficiency of performance in a given skill or body of knowledge. Therefore, it can be said that achievement implies the overall mastery of a pupil on a particular context.

Achievement is any measuring instrument that measures the attainments or accomplishments of a pupil's achievement must be valid and reliable.

Testing is a systematic procedure for comparing the behavior of two or more persons. This way an achievement test is an examination to reveal the relative standing of an individual in the group with respect to achievement.

As achievement is the competence of a person in relation to a domain of knowledge_An Achievement Test is a test of knowledge or proficiency based on something learned or taught. The purpose of an achievement test is to determine student's knowledge in a particular subject area.

Achievement test is an important tool in school evaluation and has great significance in measuring instructional progress and progress of the students in the subject area.

Accurate achievement data are very important for planning curriculum and instruction and for program evaluation.

Achievement test measures present proficiency, mastery and understanding of general and specific areas of knowledge.

Achievement test is a test designed to measure the knowledge or proficiency of an individual in something that has been learned or taught, as science or social sciences.

It is a test that measures a student's achievement and progression in a specific subject or topic over a set period of time.

Achievement test is a test of a student's knowledge of a subject, which can be compared with the performance of other students taking the same test.

Or, any test that measures the attainments and accomplishments of an individual after a period of training or learning (NM Downie).

Achievement test is the type of ability test that describes what a person has learned to do (Thorndike and Hagen).

Also, defined as a systematic procedure for determining the amount a student has learned through instructions (Groulund).

Functions of Achievement Test

- It Provides basis for promotion to the next grade.
- To find out where each student stands in various academic areas.
- To motivate the students before a new assignment has taken up.

- To expose student's difficulties which the teacher can help them to solve.
- It helps in determination about the placement of the students in a particular section.

Characteristics of Achievement Tests

Measurement tools can be judged on a variety of merits. These include practical issues as well as technical ones. All instruments have strengths and weaknesses no instrument is perfect for every task. Some of the practical issue that need to be considered includes:

- Reliability
- Validity
- Ease in administration.
- Cost.
- Time.
- Acceptability.
- Objectivity.
- Equilibrium.
- Specificity.
- Precise & Clear.

(1) Reliability

Reliability of a test refers to the degree of accuracy with which an exam, test measures what it seeks to measure a given variable. A test good reliability means that the test taker will obtain the same test score over repeated testing as long as no other extraneous factors have affected the score. A good instrument will produce consistent scores. An instrument's reliability is estimated using a correlation coefficient of one type or another.

Reliability of a test refers to its consistency or stability. A test good reliability means that the test taker will obtain the same test score over repeated testing as long as no other extraneous factors have affected the score. Reliability is the extent to which the measurements resulting from a test are the result of characteristics of those being measured. For example, reliability has elsewhere been defined as “the degree to which test scores for a group of test takers are consistent over repeated applications of a measurement procedure and hence are inferred to be dependable and repeatable for an individual test taker” (Berkowitz, Wolkowitz, Fitch, and Kopriva, 2000).

Technically, the theoretical definition of reliability is the proportion of score variance that is caused by systematic variation in the population of test-takers. This definition is population-specific. If there is greater systematic variation in one population than another, such as in all public school students compared with only eighth-graders, the test will have greater reliability for the more varied population. This is a consequence of how reliability is defined. Reliability is a joint characteristic of a test and examinee group, not just a characteristic of a test. Indeed, reliability of any one test varies from group to group

Reliability is the quality of a test which produces scores that are not affected much by chance. Students sometimes randomly miss a question they really knew the answer to or sometimes get an answer correct just by guessing; teachers can sometimes make an error or score inconsistently with subjectively scored tests.

(2) Validity (meaningfulness)

Validity is the quality of a test which measures what it is supposed to measure. It is the degree to which evidence, common

sense, or theory supports any interpretations or conclusions about a student based on his/her test performance. More simply, it is how one knows that a math test measures students' math ability, not their reading ability.

Validity like reliability also depends upon certain factors, they are:

1. Adequacy in sampling
2. Objectivity in scoring
3. Aim

Thus, a valid measurement tool does a good job of measuring the concept that it purports to measure. It is important to remember that the validity of an instrument only applies to a specific purpose with a specific group of people.

A test is valid when it:

- produces consistent scores over time.
- correlates well with a parallel form.
- measures what it purports to measure.
- can be objectively scored.
- has representative norms.

(3) Ease in Administration

A test is good only when the conditions of answering are simple (scientific and logical). Its instruction should be simple and clear.

(4) Cost

A good test should be inexpensive, not only from the view point of money but also from the view point of time and effort

taken in the construction of a test. Fortunately there is no direct relationship between cost and quality.

(5) Time

Generally the time given to students is always in short supply, however the students too do not accept very long tests. Therefore a test should neither be very long nor very short.

(6) Acceptability

A good test should be acceptable to student to whom its being given without regard to any specific situation that is the question given in the test should be neither very difficult nor very easy.

(7) Objectivity

A test is objective when the scorer's personal judgment doesn't affect the scoring.

(8) Equilibrium

Achievement of the correct proportion among questions allotted to each of the objectives & teaching content.

(9) Specificity

The items in a test should be specific to the objectives.

(10) Precise & Clear

Items should be precise, clear so that the students can answer well and score marks.

Definitions

Achievement test is the tool which helps in measures the capacities and capabilities of an individual. It is also helpful in upgrading the standard of education in an energetic way so that the individual can see with their own eyes that what they achieve by their past learning.

Some of the definitions of achievement test are given below:

According to Waters, “Achievement test act as useful aids in diagnosing the student’s specific learning needs for identifying his relative strengths and weaknesses”.

According to Super, “An achievement test or proficiency test is used to ascertain what and how much has been learnt or how well a task can performed”.

According to Free Man, “Achievement test is a test designed to measure knowledge, understanding and skills in a specified subject or a group of subjects”.

According to N.M. Downie, “Any test that measures the attainments or accomplishments of an individual after period of training or learning is called achievement test. It helps to permute the student to next class.”

Achievement test is most probably the very important area of appraisal for a guidance program for the benefit of the individual. Scores on achievement test are excellent means for evaluating educational (academic) attainments and for the individual in the concerned area of the subject covered by the test.

It involves a determination of how quickly, how accurately and at what level an individual can perform the tasks taken to represent accomplishment.

Achievement test measures present proficiency, mastery and understanding of general and specific areas of knowledge. Achievement tests attempt to measure what and how individual has learnt, viz. his present standard of performance. Scores of achievement test indicate the academic status of the individual learner in different subjects as a whole or individually.

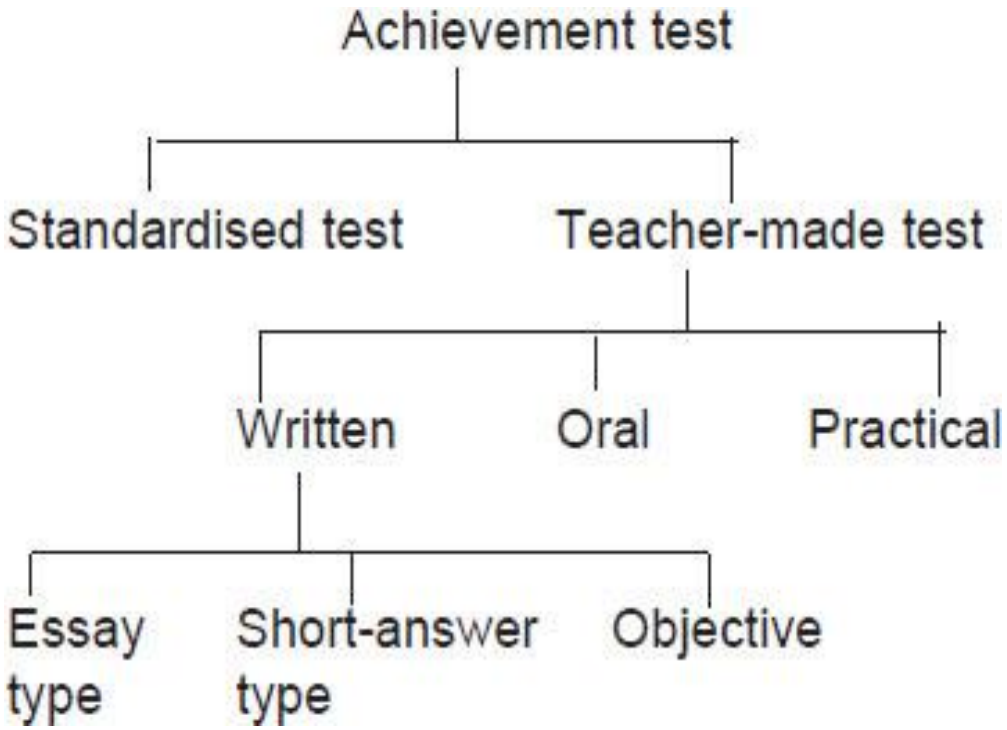
Achievement test scores are quite helpful clues for vocational guidance since these mostly related to aptitudes and interests. In the circumstances the achievement test should be based on systematic testing program of every school that desires to undertake suitable guidance service for the individuals.

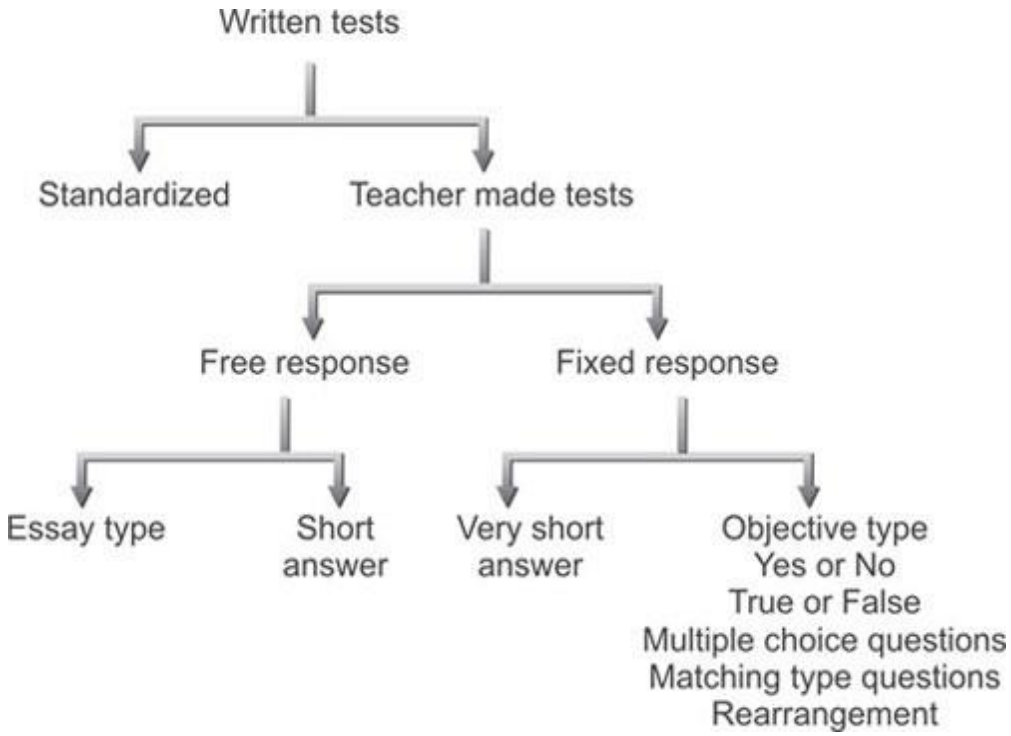
Types of Achievement Tests

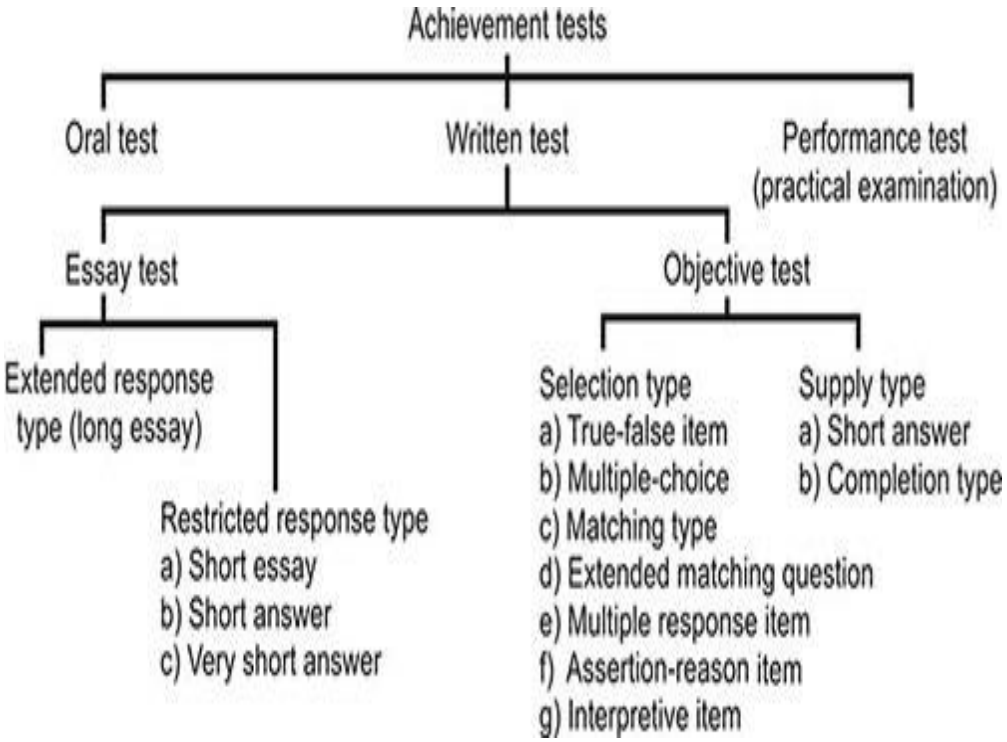
Broadly, based on quality, achievement tests can be grouped into two categories –the standardized test and the classroom tests.

- Classroom Tests - for formative or summative assessment
- Standardized tests – to meet multiple objectives, including evaluation

Examples of achievement tests: The small-scale achievement tests that are usually administered in schools include spelling tests, map quizzes, periodic arithmetic, etc., to measure how well each student can demonstrate their knowledge of a particular academic subject or skill.


Achievement test
Standardised test
Teacher-made test
Written
Oral
Practical
Essay type
Short-answer type
Objective


Written tests
Standardized
Teacher made tests
Free response
Fixed response
Essay type
Short answer
Very short answer
Objective type
Yes or No
True or False
Multiple choice questions
Matching type questions
Rearrangement



Achievement tests can be of various categories basing on form, purpose, time, method, and subject area. Achievement test can be of different forms like oral test, written test and practical test. Items of achievement test can be essay type questions or short answer questions or objective type of questions or combination of all these types.

Achievement test may be of different types on the basis of the purpose for which it is administered. They are diagnostic tests, prognostic test, accuracy test, power test, spit test etc.

Achievement tests can be administered in different period of time. When it is based on time or period factor, the test is summative test, daily test, weekly test, fortnightly test, monthly

test, quarterly test, half yearly test, annual test or final examination at the end of course of study of an academic year.

On the basis of content or subject matter, achievement tests are categorized as language test, reading test, spelling test, history test, geography test, mathematic test, science test etc.

Broadly speaking, all these achievement tests can be divided into two on the basis of quality that is standardized test and teacher-made test. Here let us have a discussion on the objective type of achievement test.

The traditional system of examination or the essay type of examination has come under heavy fire. Students reject it because of its heavy strength or pressure. The parents criticize it because of its injurious effect on the physical and mental health of children.

The teachers complain because of its harmful effect on school work. The practical psychologist speaks ill of it because of its unreliability and invalidity and the educational theorist attacks it because it lacks definiteness in aim and purpose.

To remove some of the evils of the essay type examinations, objective tests seem to be very useful. Modern educationists give much stress on this type of tests to supplement the traditional type of tests. The All India Council for Secondary Education has set up an “evaluation unity.”

Many workshops and seminars have been organized during the past 10 years with a view to preparing new type tests. The

services of Dr. Bloom of America, an expert in evaluation were secured for some time for the purpose of popularizing the new type of tests and given training to a large number of teachers in the use of these tests.

Type I. Teacher-made Tests

A teacher uses different terms of evaluation techniques in a class-room situation. Teacher-made test is one of the most valuable instruments in the hands of the teacher to solve this purpose. It is designed to solve the problems or requirements of the class for which it is prepared.

Teacher-made tests are classroom tests and are developed by the teachers. These tests assess students learning every period of time or after a particular unit of study.

It is prepared to measure the outcomes and content of local curriculum. It is very much flexible so that, it can be adapted to any procedure and material. It does not require any sophisticated technique for preparation. It is easy to construct. As standardized test are prepared to measure the learning objectives in general so it is necessary to prepare teacher made tests to suit the local objectives.

In teacher-made test the test items, time limit, instruction and procedure of scoring vary from test to test. These tests may be written or oral in nature. In teacher-made test both objective type and essay type items can be included.

Features of Teacher-made Tests

1. Assess degree of student's progress with reference to classroom activities.
2. Help the teacher to assess individual student's strengths and weaknesses and needs.
3. Simple to use.
4. Provide feedback for teachers as to assess the effectiveness of teaching methods.

Limitations of Teacher-made Tests

- Tests are ambiguous and unclear.
- Tests are either too short or too lengthy.
- Tests do not cover the entire content.
- Tests serve limited purpose.
- Tests are usually hurriedly conducted.

Type II. Standardized Tests

Developed by testing companies and administered to large populations of students.

Standardized tests are carefully constructed tests which have uniformity of procedure in scoring, administering and interpreting the test results. Generally these tests are "norm-referenced tests that measure the pupils' level of achievement in various contents. And skill areas by comparing their test performance with the performance of other pupils in some general reference group."

In the Dictionary of Education C.V. Good has described a standardized test is that "for which content has been selected and checked empirically for which norms have been established for which uniform methods of administering and scoring have been

developed, and which may be scored with a relatively high degree of objectivity."

A standardized test is one in which the procedure, Apparatus, and scoring have been fixed so that precisely the same test can be given at different times and places. (Lee J. Cronbach).

Standardized tool is one for which norms have been established.

A standardized test is prepared after several trials of a test to a large number of students.

From the above discussion we can determine the following characteristics of a standardized test:

1. The test is developed by experts test specialists so that its items are of high technical quality.
2. A clear instruction for administering and scoring the test is given. So that it maintains uniformity in administering and scoring.
3. For the interpretation of test scores norms are provided in the test manual.
4. Equivalent forms of tests are available.
5. These tests possesses high reliability coefficient, generally between .80 to .95.

Importance of Standardized Tests

- Give impartial information about an individual.
- Provides information in much less time than provided by any other devices.

- Tests measures those aspects of behaviors which otherwise could not be obtained.
- In subjective observation we may overlook shy students but these tests discover such cases also.

Types of standardized tests

Several kinds of standardized tests are used in schools today. There are three broad categories of standardized tests: achievement, diagnostic, and aptitude (including interest). examples of standardized achievement tests:

- International Olympiad of Science (IOS).
- Brilliant International Olympiad of Science (BIOS).

Teacher-made Tests

Oral Test

Oral exams—tests during which teachers ask students to answer exam questions aloud—can be undoubtedly stressful, but there are a number of ways to prepare for nontraditional testing or reporting methods like this. Though oral exams are most common for language learners, they are increasingly prevalent across other subjects because they allow teachers to cater syllabuses to students with a variety of learning styles.

Practical (Performance) Test

Practical exams test students' practical skills and techniques usually in laboratory, clinical or field settings. They can be administered individually, in pairs or small groups.

Written Tests

A test in which you have to write the answer.

Types of Test items

- ✓ Selected-response (SR) test item - the student selects from several alternative. Also called an objective item. Examples: multiple-choice, true-false, matching, fill-in-the-blank (completion).

- ✓ Constructed-response (CR) test item – the student must produce a response from scratch (zero point) but within a context. Also called a free-response item (Examples: short answer, essay, …).

So, There are two general categories of test items: (1) objective items which require students to select the correct response from several alternatives or to supply a word or short phrase to answer a question or complete a statement; and (2) subjective or essay items which permit the student to organize and present an original answer. Objective items include multiple-choice, true-false, matching and completion, while subjective items include short-answer essay, extended-response essay, problem solving and, performance test items. For some instructional purposes one or the other item types may prove more efficient and appropriate.

Teachers need to be knowledgeable about constructing both item types. Teachers need to be aware that each item type has strengths and weaknesses and use this information to evaluate item type usage. Teachers need to guard against gullibly accepting the “common sense” rhetoric about one type of item being better in assessing important mental processes.

What is an essay type?

A test item which requires a response composed by the examinee, usually in the form of one or more sentences, of a nature that no single response or pattern of responses can be listed as correct, and the accuracy and quality of which can be judged subjectively only by one skilled or informed in the subject.

Advantages

- Assess higher-order or critical thinking skills.
- Evaluate student thinking and reasoning.
- Provide authentic experience.

Limitations

- Assess a limited sample of the range of content.
- Are difficult and time consuming to grade.
- Provide practice in poor or unpolished writing.

Misconceptions

- Assess higher-order or critical thinking skills regardless of how the responses are written.
- Essay questions are easy to construct.
- The use of essay questions eliminates the problem of guessing.
- Essay questions benefit all students by placing emphasis on the importance of written communication skills.
- Essay questions encourage students to prepare more thoroughly.

When Should Essay Questions Be Used?

- To assess students' understanding of subject-matter content.

- To assess students' abilities to reason with their knowledge of a subject.

How Should Essay Questions Be Constructed?

- Clearly define the intended learning outcome to be assessed by the item.
- Avoid using essay questions for intended learning outcomes that are better assessed with other kinds of assessment.
- Clearly define and situate the task within a problem situation.
- Present a reasonable task to students.
- The task can be written as a statement or question.
- Specify the relative point value and the approximate time limit in clear directions.
- State the criteria for grading.
- Use several relatively short essay questions rather than one long one.
- Avoid the use of optional questions.
- Improve the essay question through preview and review.

What is an objective type?

An objective type of test item is one which the response will be objective. Objective type test item broadly classified into two:

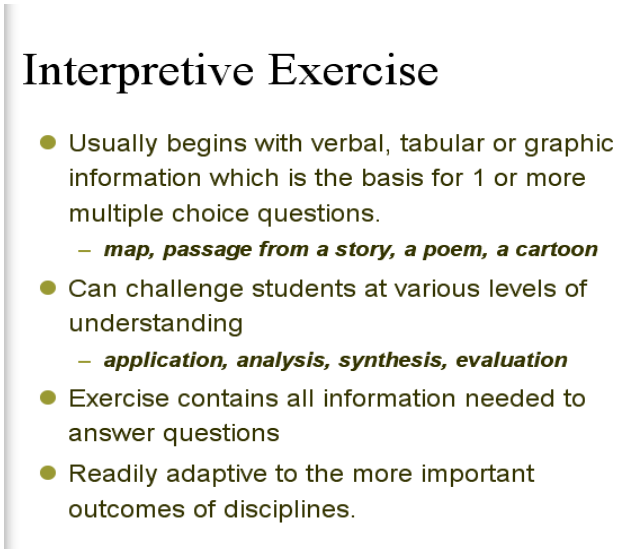
- Supply type (Recall Type): The respondent has to supply the responses.
- Selection type (Recognition Type): The respondent has to select the responses from among the given responses.

Advantages of Objective Type Items

- A large amount of study material can be tested in a very short period time
- Economy of time.
- Objectivity of scoring.
- No bluffing
- It reduces the subjective element of the examiner to the minimum.
- If carefully planned, it can measure the higher mental process of understanding, application, analysis, prediction and interpretation.

Limitations of objective type items

- Difficulty in preparing good items.
- Problem of guessing.
- Problem of cheating.
- Inefficiency in testing complicated skills
- High printing cost.
- Emphasis on testing superficial knowledge.



Interpretive Exercises

Strengths

1. An efficient means of measuring the interpretation of printed information in various forms (e.g., written, charts, graphs, maps, pictures) is provided
2. More meaningful complex learning outcomes can be measured than with a single item format
3. The use of introductory material provides a common basis for responding
4. Scoring is easy, objective and reliable

Limitations

1. It is difficult to construct effective items
2. Written material is highly dependent on reading skill
3. The item type is highly subject to extraneous clues
4. It is ineffective in measuring the ability to originate, organize, and express ideas.

The Interpretive Exercise

Complex learning outcomes can be effectively measured with a paragraph, table, chart, graph, map, or picture

Can be used to measure interpretation of introductory material

In the example (see handout) the use of multiple-choice items

> ***Example 1 - measures the ability to measure unstated assumptions***
>
> ***Example 2 - the ability to identify the meaning of a term***
>
> ***Example 3 - the ability to identify relationships***

Assertion-Reason Questions

The word “assertion” means a confident and forceful statement of fact or belief when we talk about it as a noun. In

these questions, the candidate is provided with a statement. This statement presents an opinion, a fact, or a comment. We call it the assertion. The other statement is the reason. These two statements form a pair of reason and assertion statements. There may or may not be a relation between the two. The most common misconception is that the assertion is a statement and the reason has to be a defending statement or the cause for it. Well, that is what you need to check for.

The questions on Assertion and Reason look something like the following.

Solved Examples

Directions: In the following questions, the Assertions (A) and Reason(s) (R) have been put forward. Read both the statements carefully and choose the correct alternative from the following:
(A) Both the Assertion and the Reason are correct and the Reason is the correct explanation of the Assertion.
(B) The Assertion and the Reason are correct but the Reason is not the correct explanation of the Assertion.
(C) Our Assertion is true but the Reason is false.
(D) The statement of the Assertion is false but the Reason is true.
(E) Both the statements are false.

Example 1: Assertion: All the crow species are entirely black in color.

Reason: The color of the crows is a biological adaptation.

Answer: Well the most rookie mistake that people make is this. They argue that if the assertion and the reason sound similar, the option must be A. In other words, the reason must be the correct explanation of the assertion. Let us first check the statements. The first statement presents a strong opinion or in

other words, a blanket opinion. Such opinions are often not correct. For example, if someone says that all insects are small, the statement is incorrect. rather if you say insects are small, the statement is genuine.

Similarly to say that all crows are black is not correct as there are species of the common crow, that are grey in color. So the assertion is not right. The reason, however, is correct. Therefore the correct option here is D).

More Examples

In the previous section, we saw that we should avoid jumping to conclusions. The first step is to check the truth of the statements themselves and later worry about the relation between the two. Let us see more such examples:

Example 2: Assertion: The Mountains on Moon are way taller than Mount Everest.

Reason: The force of gravity is stronger on Earth than on the moon.

(A) Both the Assertion and the Reason are correct and the Reason is the correct explanation of the Assertion.
(B) The Assertion and the Reason are correct but the Reason is not the correct explanation of the Assertion.

(C) Our Assertion is true but the Reason is false.
(D) The statement of the Assertion is false but the Reason is true.
(E) Both the statements are false.

Answer: The two statements don't seem to have the same topic but that doesn't matter. The reason that the height of Everest is not greater than the mountains on the surface of the moon is gravity. The stronger force of gravity on the surface of earth makes sure that the mountains are shorter. Thus the reason and assertion are not only true but the reason is the correct explanation of the assertion. Therefore the answer is).

Example 3: Assertion: Humans have evolved from an ape-like species of primates.

Reason: There are many fossils that support the theory of evolution in case of humans.

Answer: You might be tempted to select A) here but wait, there is caution here. The answer is not A). Of course, humans are evolved from apes and there are fossils that support the theory of evolution. But the reason has to explain why humans have evolved from apes. A simple trick to see if reason explains the assertion is to change the assertion into a question.

For example, here the assertion becomes, "why have humans evolved from apes?" or "How have humans evolved from apes". The reason doesn't answer that. So both the statements are correct but the reason is the explanation of the assertion. Therefore the correct option here is B).

Direction: In each of the following questions, a statement of Assertion is given followed by a corresponding statement of Reason. Of the statements, mark the correct answer as
(a) Both assertion and reason are true and reason is the correct explanation of assertion.

(b) Both assertion and reason are true but reason is not the correct explanation of assertion.
(c) Assertion is true but reason is false.
(d) Assertion is false but reason is true

Question 1.
Assertion: The number of particles present in one mole of a substance is fixed.
Reason: The mass of one mole of a substance is equal to its relative atomic mass in grams.

Answer

Answer: (a)

Question 2.
Assertion: Atoms always combine to form molecule and ions.
Reason: Atoms of most element are not able to exist independently.

Answer

Answer: (d)

Question 3.
Assertion: Atomicity of ozone is three while that of oxygen is two.
Reason: Atomicity is the number of atoms constituting a molecule.

Answer

Answer: (a)

Question 4.
Assertion: 1 amu equals to 1.66×10^{-24} g.
Reason: 1.66×10^{-24} g equal to $1/12^{th}$ mass of a C-12 atom.

Answer

Answer: (a)

Question 5.
Assertion: On burning magnesium in oxygen, the mass of magnesium oxide formed is equal to the total mass of magnesium and oxygen
Reason: In a chemical substance, the elements are always present in a definite proportion.

Answer

Answer: (b)

Question 6.
Assertion: 1 mole of and H_2 each O_2 occupy 22.4 L at standard temperature and pressure.
Reason: Molar volume for all gases at the standard temperature and pressure has the different values.

Answer

Answer: (c)

Question 7.
Assertion: Molecular weight of oxygen is 16.
Reason: Atomic weight of oxygen is 16.

Answer

Answer: (d)

Question 8.
Assertion: Atomic mass of aluminium is 14.
Reason: An atom of aluminium is 27 times heavier than 1/12th of the mass of carbon-12 atom.

Answer

Answer: (a)

Question 9.
Assertion: The number of moles of He in 52 g of He is 13.
Reason: The number of moles of an atom is the ratio of its given mass to its molar mass.

Answer

Answer: (a)

Question 10.
Assertion: The valency of aluminium is 3 and oxygen is 2.
Reason: The chemical formula of aluminium oxide is Al_3O_2.

Answer

Answer: (c)

Question 11.
Assertion: A molecule is the smallest particle of an element or a compound which is capable of free existence.
Reason: The number of atoms present in one molecule of the substance is called its atomicity.

Answer

Answer: (b)

Question 12.
Assertion: Protons cannot be transferred from one atom to another.
Reason: Protons are present deep inside the atom in its nucleus.

Answer

Answer: (a)

Question 13.
Assertion: Water molecules always contain hydrogen and oxygen in the ratio 1:8.
Reason: Water obeys law of constant proportions irrespective of source and method of preparation.

Answer

Answer: (a)

Question 14.
Assertion: Relative atomic mass of the atom of element is the average masses of the
atom as compared to the mass of one carbon-12 atom.
Reason: Carbon-12 isotope is the standard reference for measuring atomic masses.

Answer

Answer: (c)

Question 15.
Assertion: A sodium ion has positive charge.
Reason: Sodium ion has more protons than a neutral atom.

Answer

Answer: (c)

Question 16.
Assertion : In water, the ratio of mass of hydrogen to the mass of oxygen is always 1:8 whatever the source of water.
Reason : According to law of constant proportion, the elements are always present in definite proportion by mass in a chemical substance.

Answer

Answer: (a)

Question 17.
Assertion : Ions are always positively charged.
Reason : Ions are formed by losing or gaining of electrons.

Answer

Answer: (d)

Question 18.
Assertion : One mole of SO_2 contains double the number of molecules present in one mole of O_2
Reason : Molecular weight of SO_2 is double to that of O_2

Answer

Answer: (d)

Question 19.
Assertion : When 12 g of $CaCO_3$ is decomposed, 4.6 g of residue is left and 4.4 g of escapes. Reason : Law of conservation of mass is followed.

Answer

Answer: (d)

Question 20.
Assertion : Pure water obtained from different sources such as river, well, spring, sea etc. always contains hydrogen and oxygen combined in the ratio of 1 : 8 by mass.
Reason : A chemical compound always contains same elements combined in different fixed proportion by mass.

Answer

Answer: (c)

Question 21.
Assertion : Law of conservation of mass holds good for all the reactions.
Reason : I states that energy can neither be created nor destroyed in a chemical reaction.

Answer

Answer: (a)

Question 22.
Assertion : Atomicity of oxygen is 4.
Reason : 1 mole of an element contains 6.023×10^{23} atoms.

Answer

Answer: (d)

Question 23.
Assertion : Atomicity of O_3 is 2.
Reason : 1 mole of an element contains 6.023×10^{23} atoms.

Answer

Answer: (d)

Question 24.
Assertion : 52g of He contains 13 x 6.023 x 10^{23} atoms.
Reason : 1 mole of an element contains 6.023 x 10^{23} atoms.

Answer

Answer: (a)

Question 25.
Assertion: Atoms can neither be sub-divided, created nor destroyed.
Reason: This postulate of Dalton's theory is the result of law of constant proportion.

Answer

Answer: (c)

Question 26.
Assertion: Carbonates are polyatomic ions.
Reason: The carbonate ion consists of one carbon atom and three oxygen atoms and carries an overall charge of.

Answer

Answer: (a)

Question 27.
Assertion : Number of gram-molecules of SO_2Cl_2 in 13.5 g of sulphur chloride is 0.1.
Reason : Gram molecular mass is equal to two gram molecule.

Answer

Answer: (c)

Question 28.
Assertion : The total number of electrons present in 16 g of methane gas is 6.022 x 10^{23}
Reason : 1 mole of an element contains 6.023 x 10^{23} atoms.

Answer

Answer: (a)

Question 29.
Assertion : All noble gases are monoatomic.
Reason : Noble gases are highly stable and unreactive.

Answer

Answer: (a)

Multiple-choice test items

Many variations, common ones are:

- simple question followed by a number of alternative choices, typically four.
- paragraph(s) followed by the alternatives
- as the two above, but the alternatives are combinations of conditions (I &II, I&III, I&IV, III & IV)
- item sets (i.e., multiple questions based on a single map, chart, table or reading

The multiple-choice item consists of two parts: (a) the stem, which identifies the question or problem and (b) the response alternatives. Students are asked to select the one alternative that best completes the statement or answers the question. For example,

Sample Multiple-Choice Item

(a) Item Stem: Which of the following is a chemical change?

(b) Response Alternatives:

a. Evaporation of alcohol
b. Freezing of water
*c. Burning of oil
d. Melting of wax

*correct response

Advantages In Using Multiple-Choice Items

Multiple-choice items can provide ...

- versatility in measuring all levels of cognitive ability.
- highly reliable test scores.
- scoring efficiency and accuracy.
- objective measurement of student achievement or ability.
- a wide sampling of content or objectives.

- a reduced guessing factor when compared to true-false items.
- different response alternatives which can provide diagnostic feedback.

Limitations In Using Multiple-Choice Items

Multiple-choice items ...

- are difficult and time consuming to construct.
- lead an instructor to favor simple recall of facts.
- place a high degree of dependence on the student's reading ability and instructor's writing ability.

Suggestions for Writing Multiple-Choice Test Items (The Stem)

1. When possible, state the stem as a direct question rather than as an incomplete statement.
   Undesirable: Alloys are ordinarily produced by ...
   Desirable: How are allows ordinarily produced?
2. Present a definite, explicit and singular question or problem in the stem.
   Undesirable: Psychology ...
   Desirable: The science of mind and behavior is called...
3. Eliminate excessive verbiage or irrelevant information from the stem.
   Undesirable: While ironing her formal, Jane burned her hand accidently on the hot iron. This was due to a transfer of heat be ...
   Desirable: Which of the following ways of heat transfer explains why Jane's hand was burned after she touched a hot iron?
4. Include in the stem any word(s) that might otherwise be

repeated in each alternative.

Undesirable: In national elections in the United States the President is officially

a. chosen by the people.
b. chosen by members of Congress.
c. chosen by the House of Representatives.
*d. chosen by the Electoral College.

Desirable: In national elections in the United States the President is officially chosen by

a. the people.
b. members of Congress.
c. the House of Representatives.
*d. the Electoral college.

5. Use negatively stated stems sparingly. When used, underline and/or capitalize the negative word.

Undesirable: Which of the following is not cited as an accomplishment of the Kennedy administration?

Desirable: Which of the following is NOT cited as an accomplishment of the Kennedy administration?

Suggestions *for* Writing Multiple-Choice Test Items (Item Alternatives)

6. Make all alternatives plausible and attractive to the less knowledgeable or skillful student. What process is most nearly the opposite of photosynthesis?

Undesirable	Desirable
a. Digestion	a. Digestion
b. Relaxation	b. Assimilation
*c. Respiration	*c. Respiration

d. Exertion d. Catabolism

7. Make the alternatives grammatically parallel with each other, and consistent with the stem.

Undesirable: What would do most to advance the application of atomic discoveries to medicine?

*a. Standardized techniques for treatment of patients.

b. Train the average doctor to apply radioactive treatments.

c. Remove the restriction on the use of radioactive substances.

d. Establishing hospitals staffed by highly trained radioactive therapy specialists.

Desirable: What would do most to advance the application of atomic discoveries to medicine?

*a. Development of standardized techniques for treatment of patients.

b. Training of the average doctor in application of radioactive treatments.

c. Removal of restriction on the use of radioactive substances.

d. Addition of trained radioactive therapy specialists to hospital staffs.

8. Make the alternatives mutually exclusive.

Undesirable: The daily minimum required amount of milk that a 10 year old child should drink is

a. 1-2 glasses.

*b. 2-3 glasses.

*c. 3-4 glasses.

d. at least 4 glasses.

Desirable: What is the daily minimum required amount of milk a 10 year old child should drink?

a. 1 glass.
b. 2 glasses.
*c. 3 glasses.
d. 4 glasses.

9. When possible, present alternatives in some logical order (e.g., chronological, most to least, alphabetical).
At 7 a.m. two trucks leave a diner and travel north. One truck averages 42 miles per hour and the other truck averages 38 miles per hour. At what time will they be 24 miles apart?

Undesirable		Desirable	
a.	6 p.m.	a.	1 a.m.
b.	9 p.m.	b.	6 a.m.
c.	1 a.m.	c.	9 a.m.
*d.	1 p.m.	*d.	1 p.m.
e.	6 a.m.	e.	6 p.m.

10. Be sure there is only one correct or best response to the item.

Undesirable: The two most desired characteristics in a classroom test are validity and….

a. precision.
*b. reliability.
c. objectivity.
*d. consistency.

Desirable: The two most desired characteristics in a classroom test are validity and

a. precision.
*b. reliability.
c. objectivity.
d. standardization.

11. Make alternatives approximately equal in length.

Undesirable:

The most general cause of low individual incomes in the United States is

*a. lack of valuable productive services to sell.
b. unwillingness to work.
c. automation.
d. inflation.

Desirable: What is the most general cause of low individual incomes in the United States?

*a. A lack of valuable productive services to sell.
b. The population's overall unwillingness to work.
c. The nation's increased reliance on automation.
d. an increasing national level of inflation.

12. Avoid irrelevant clues such as grammatical structure, well known verbal associations or connections between stem and answer.

Undesirable: (grammatical clue)

A chain of islands is called an:

*a. archipelago.
b. peninsula.
c. continent.
d. isthmus.

Undesirable: (verbal association clue)

The reliability of a test can be estimated by a coefficient of:

a. measurement.
*b. correlation.
c. testing.
d. error.

Undesirable: (connection between stem and answer clue)

The height to which a water dam is built depends on

a. the length of the reserve CTE behind the dam.
b. the volume of water behind the dam.
*c. The height of water behind the dam.
d. the strength of the reinforcing wall.

13. Use at least four alternatives for each item to lower the probability of getting the item correct by guessing.

14. Randomly distribute the correct response among the alternative positions throughout the test having approximately the same proportion of alternatives a, b, c, d and e as the correct response.

Use the alternatives "none of the above" and "all of the above" sparingly. When used, such alternatives should occasionally be used as the correct response.

also Multiple Choice Items: Suggestions for Preparing

1. See Haladyna's comprehensive list of rules.
2. Content is King . . . when writing items get the content issue right above all.

3. Don't give clues to the correct answer (either wittingly or unwittingly).
4. Don't be too rigid in rule application. Good items represent a sense of finesse and ingenuity, not just rule application.

A Selection of Multiple Choice Item Writing Guidelines, Sets I. & II.

I. General item-writing (procedure)
    1. Format the item responses vertically, not horizontally.
    2. Keep the vocabulary as simple as possible.
    3. Minimize student reading time while maintaining meaning.
    4. Avoid potentially insensitive content or language.

II. General item-writing (content)
    1. Base all items on your previously stated learning objectives.
    2. Avoid cuing one item with another; keep items independent of one another.
    3. Avoid textbook, verbatim phrasing when developing the item.
    4. Base items on important aspects of the content area; avoid trivial material.

<u>Guidelines, Set III.</u>

III. Stem development
    1. State the stem in either question form or completion form.
    2. When using the completion format, don't place the missing concept at the beginning or middle of the

stem.

3. Include only the material needed to make the problem clear.
4. Don't add extraneous information unless this is part of what you ask the students to identify.
5. Word the stem positively; avoid negative phrasing. If an item must be stated negatively, underline or capitalize the negative word.
6. Include the central idea and most of the phrasing in the stem.

Guidelines, Set IV.

IV. General option development

1. Place options in a logical order, if one exists (e.g., numerical, alphabetical, chronological).
2. Keep the length of options fairly consistent.
3. Avoid, or use sparingly, the phrase “all of the above.”
4. Avoid, or use sparingly, the phrase “none of the above.”
5. Avoid distracters that can clue test-wise examinees; for example, grammatical inconsistencies involving “a" or "an" give clues to the correct answer.

Guidelines, Sets V. & VI.

V. Correct option development

1. Position the correct option so that it appears about the same number of times in each possible position for a set of items.
2. Make sure there is one and only one correct, or clearly best, answer on which experts would agree.

VI. Distracter development (sometimes called foils)

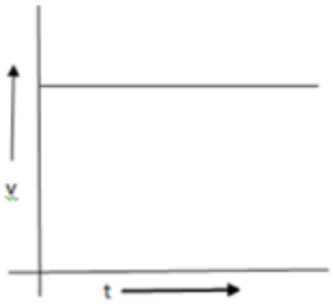
1. Use plausible distracters; avoid illogical distracters.
2. Incorporate common errors of students in distracters.
3. Use familiar yet incorrect phrases as distracters.
4. Use true statements that do not correctly answer the item.
5. Shun/Court humor when developing options.

MCQ Questions for Class 9 Science Chapter 8 Motion

Q.1. If the displacement of an object is proportional to square of time, then the object moves with:
(a) Uniform velocity
(b) Uniform acceleration
(c) Increasing acceleration
(d) Decreasing acceleration

Answer
Answer: (b) Uniform acceleration

Q.2. From the given v-t graph, it can be inferred that the object is
(a) At rest
(b) In uniform motion
(c) Moving with uniform acceleration
(d) In non-uniform motion



Answer
Answer: (b) In uniform motion

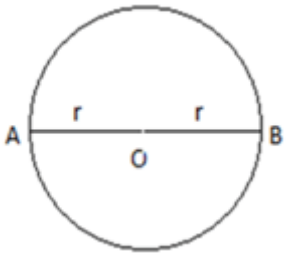
Q.3. Suppose a boy is enjoying a ride on a marry-go-round which is moving with a constant speed of 10 m/s. It implies that the boy is:
(a) At rest

(b) Moving with no acceleration
(c) In accelerated motion
(d) Moving with uniform velocity

Answer

Answer: (c) In accelerated motion

Q.4. A particle is moving in a circular path of radius r.



The displacement after half a circle would be:
(a) Zero
(b) πr
(c) 2r
(d) 2πr

Answer

Answer: (c) 2r

Q.5. Which of the following can sometimes be ‘zero’ for a moving body?
i. Average velocity
ii. Distance travelled
iii. Average speed
iv. Displacement
(a) Only (i)
(b) (i) and (ii)
(c) (i) and (iv)
(d) Only (iv)

Answer

Answer: (c) (i) and (iv)

Q.6. Which of the following statement is correct regarding velocity and speed of a moving body?
(a) Velocity of a moving body is always higher than its speed

(b) Speed of a moving body is always higher than its velocity
(c) Speed of a moving body is its velocity in a given direction
(d) Velocity of a moving body is its speed in a given direction

Answer

Answer:(d) Velocity of a moving body is its speed in a given direction

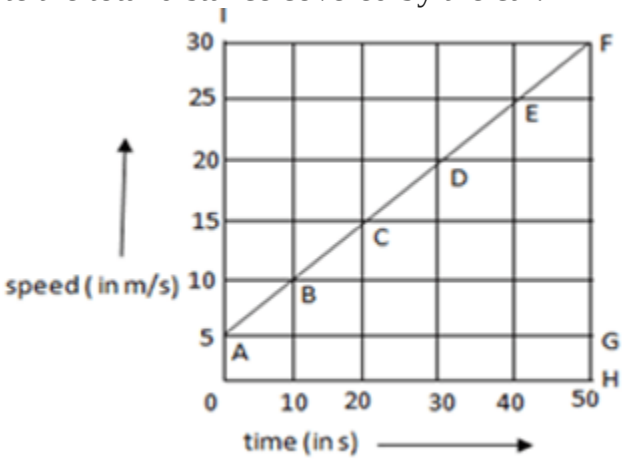
---

Q.7. When a car driver travelling at a speed of 10 m/s applies brakes and brings the car to rest in 20 s, then the retardation will be:
(a) + 2 m/s^2
(b) − 2 m/s^2
(c) − 0.5 m/s^2
(d) + 0.5 m/s^2

Answer

Answer: (d) + 0.5 m/s^2

Q.8. The speed – time graph of a car is given here. Using the data in the graph calculate the total distance covered by the car.



(a) 1250 m
(b) 875 m
(c) 1500 m
(d) 870 m

Answer

Answer:(b) 875 m

Q.9. A car of mass 1000 kg is moving with a velocity of 10 m/s. If the velocity-time graph for this car is a horizontal line parallel to the time axis, then the velocity of the car at the end of 25 s will be:

(a) 40 m/s
(b) 25 m/s
(c) 10 m/s
(d) 250 m/s

Answer

Answer: (c) 10 m/s

Q.10. Which of the following is most likely not a case of uniform circular motion?
(a) Motion of the earth around the sun
(b) Motion of a toy train on a circular track
(c) Motion of a racing car on a circular track
(d) Motion of hours' hand on the dial of a clock

Answer

Answer: (c) Motion of a racing car on a circular track

Q.11. In which of the following cases of motions, the distance moved and the magnitude of the displacement are equal?
i. If the car is moving on a straight road
ii. If the car is moving in circular path
iii. The pendulum is moving to and fro
iv. The earth is moving around the sun
(a) only(ii)
(b) (i) and (iii)
(c) (ii) and (iv)
(d) only (i)

Answer

Answer: (d) only (i)

Q.12. A car is travelling at a speed of 90 km/h. Brakes are applied so as to produce a uniform acceleration of $-0.5\ m/s^2$. Find how far the car will go before it is brought to rest?
(a) 8100 m
(b) 900 m
(c) 625 m
(d) 620 m

Answer

Answer: (c) 625 m

Q.13. In a free fall the velocity of a stone is increasing equally ion equal intervals of time under the effect of gravitational force of the earth. Then what can you say about the motion of this stone? Whether the stone is having:
(a) Uniform acceleration
(b) Non-uniform acceleration
(c) Retardation
(d) Constant speed

Answer

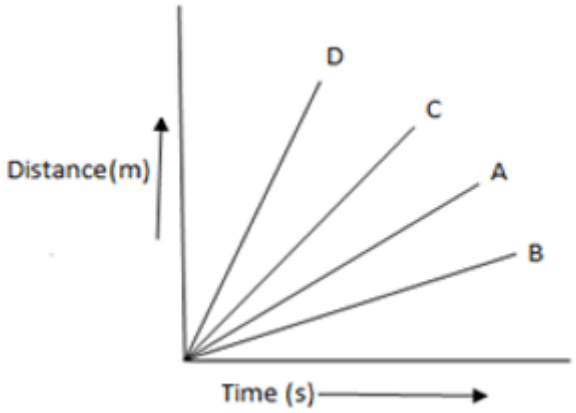
Answer: (a) Uniform acceleration

Q.14. The numerical ratio of displacement to distance for a moving object is:
(a) Always less than 1
(b) Equal to 1 or less than 1
(c) Always more than 1
(d) Equal to 1 or more than one

Answer

Answer: (b) Equal to 1 or less than 1

Q.15. Four cars A, B, C and D are moving on a leveled, straight road. Their distance time graphs are shown in the figure below. Which of the following is the correct statement regarding the motion of these cars?



(a) Car A is faster than car D
(b) Car B is the slowest
(c) Car D is faster than car C
(d) Car C is the slowest

Answer
Answer: (b) Car B is the slowest

True-False Test Items

A true-false item can be written in one of three forms: simple, complex, or compound. Answers can consist of only two choices (simple), more than two choices (complex), or two choices plus a conditional completion response (compound). An example of each type of true-false item follows:

Sample True-False Item: Simple

The acquisition of morality is a developmental process.	True	False

Sample True-False Item: Complex

The acquisition of morality is a developmental process.	True	False	Opinion

Sample True-False Item: Compound

The acquisition of morality is a developmental process.		
If this statement is false, what makes it false?	True	False

Advantages In Using True-False Items

True-False items can provide ...

- the widest sampling of content or objectives per unit of testing time.
- scoring efficiency and accuracy.
- versatility in measuring all levels of cognitive ability.
- highly reliable test scores.
- an objective measurement of student achievement or ability.
- Much of the practical knowledge on which the real world operates reduces to propositions of true or false

- Allows greater coverage of topics during a given testing time.

Limitations In Using True-False Items

True-false items ...

- incorporate an extremely high guessing factor. For simple true-false items, each student has a 50/50 chance of correctly answering the item without any knowledge of the item's content.
- can often lead an instructor to write ambiguous statements due to the difficulty of writing statements which are unequivocally true or false.
- do not discriminate between students of varying ability as well as other item types.
- can often include more irrelevant clues than do other item types.
- can often lead an instructor to favor testing of trivial knowledge.
- Format invites trivial content
- Students love to argue these with counter examples demonstrating the item isn't "always" true or "always" false
- Poor public image – reinforces the public view that tests are just guessing games, a matter of chance (in point of fact it is NOT easy to get a good score on true/false items if there are enough items in the test)

Suggestions For Writing True-False Test Items

1. Base true-false items upon statements that are absolutely true or false, without qualifications or exceptions.

   Undesirable: Nearsightedness is hereditary in origin.

   Desirable: Geneticists and eye specialists believe that the

predisposition to nearsightedness is hereditary.

2. Express the item statement as simply and as clearly as possible.

   Undesirable: When you see a highway with a marker that reads, "Interstate 80" you know that the construction and upkeep of that road is built and maintained by the state and federal government.

   Desirable: The construction and maintenance of interstate highways is provided by both state and federal governments.

3. Express a single idea in each test item.

   Undesirable: Water will boil at a higher temperature if the atmospheric pressure on its surface is increased and more heat is applied to the container.

   Desirable: Water will boil at a higher temperature if the atmospheric pressure on its surface is increased. and/or

   Water will boil at a higher temperature if more heat is applied to the container.

4. Include enough background information and qualifications so that the ability to respond correctly to the item does not depend on some special, uncommon knowledge.

   Undesirable: The second principle of education is that the individual gathers knowledge.

   Desirable: According to John Dewey, the second principle of education is that the individual gathers knowledge.

5. Avoid lifting statements from the text, lecture or other materials so that memory alone will not permit a correct answer.

   Undesirable: For every action there is an opposite and equal reaction.

Desirable: If you were to stand in a canoe and throw a life jacket forward to another canoe, chances are your canoe would jerk backward.

6. Avoid using negatively stated item statements.

Undesirable: The Supreme Court is not composed of nine justices.

Desirable: The Supreme is composed of nine justices.

7. Avoid the use of unfamiliar vocabulary.

Undesirable: According to some politicians, the raison d'etre for capital punishment is retribution.

Desirable: According to some politicians, justification for the existence of capital punishment is retribution.

8. Avoid the use of specific determiners which would permit a test-wise but unprepared examinee to respond correctly. Specific determiners refer to sweeping terms like "all," "always," "none," "never," "impossible," "inevitable," etc. Statements including such terms are likely to be false. On the other hand, statements using qualifying determiners such as "usually," "sometimes," "often," etc., are likely to be true. When statements do require the use of specific determiners, make sure they appear in both true and false items.

Undesirable: All sessions of Congress are called by the President. (F)

The Supreme Court is frequently required to rule on the constitutionality of a law. (T)

An objective test is generally easier to score than an essay test. (T)

Desirable: (When specific determiners are used reverse the expected outcomes.)

The sum of the angles of a triangle is always 1800

. (T)

Each molecule of a given compound is chemically the same as every other molecule of that compound. (T)

The galvanometer is the instrument usually used for the metering of electrical energy used in a home. (F)

9. False items tend to discriminate more highly than true items. Therefore, use more false items than true items (but no more than 15% additional false items).

also consider the following:

T/F: Suggestions for Preparing

Suggestions on writing traditional T/F items

1. Include only one concept in each statement.
2. Phrase items so that a superficial analysis by the student suggests a wrong answer (is this trickery or a mirror of life?).
3. Create approximate equal numbers of T and F.
4. Avoid using negative statements or using the word NOT.
5. Never use double negatives.
6. Use sufficient number of items to counteract guessing.

Suggestion for a not-so-traditional T/F item style

1. If sentence is false, revise the sentence so that it is true.
2. Target the revision by underlining a word or phrase.
3. Decide on how you will award points on these types of items.

Matching Test Items

In general, matching items consist of a column of stimuli presented on the left side of the exam page and a column of

responses placed on the right side of the page. Students are required to match the response associated with a given stimulus. For example,

Sample Matching Test Item

Directions: On the line to the left of each factual statement, write the letter of the principle which bests explains the statement's occurrence. Each principle may be used more than once.

Factual Statements	Principles
1. Fossils of primates first appear in the Cenozoic rock strata, while trilobite remains are found in the Proterozoic rocks.	a. There have been profound changes in the climate on earth.
2. The Arctic and Antarctic regions are sparsely populated.	b. Coordination and integration of action is generally slower in plants than in animals.
3. Plants have no nervous system.	c. There is an increasing complexity of structure and functions from lower to higher forms of life.
4. Large coal beds exist in Alaska.	d. All life comes from life and produces its own kind of living organisms.
	e. Light is a limiting factor to life.

<u>Advantages In Using Matching Items</u>

Matching items ……

- require short periods of reading and response time, allowing you to cover more content.
- provide objective measurement of student achievement or ability.
- provide highly reliable test scores.
- provide scoring efficiency and accuracy.

Limitations In Using Matching Items

Matching items

- have difficulty measuring learning objectives requiring more than simple recall of information.
- are difficult to construct due to the problem of selecting a common set of stimuli and responses.

Suggestions For Writing Matching Test Items

1. Include directions which clearly state the basis for matching the stimuli with the responses. Explain whether or not a response can be used more than once and indicate where to write the answer.

   Undesirable: Directions: Match the following.

   Desirable: Directions: On the line to the left of each identifying location and characteristics in Column I, write the letter of the country in Column II that is best defined. Each country in Column II may be used more than once.

2. Use only homogeneous material in matching items.

   Undesirable: Directions: Match the following.

1.___	Water	A.	NaCl
2.___	Discovered Radium	B.	Fermi
3.___	Salt	C.	NH_3
4.___	Year of the 1st Nuclear Fission by Man	D.	H_2O
5.___	Ammonia	E.	1942
		F.	Curie

Desirable: Directions: On the line to the left of each compound in Column I, write the letter of the compound's formula presented in Column II. Use each formula only once.

Column I	Column II
1. ___ Water	A. H_2SO_4
2. ___ Salt	B. HCl
3. ___ Ammonia	C. NaCl
4. ___ Sulfuric Acid	D. H_2O
	E. H_2HCl

3. Arrange the list of responses in some systematic order if possible (e.g., chronological, alphabetical). Directions: On the line to the left of each definition in Column I, write the letter of the defense mechanism in Column II that is described. Use each defense mechanism only once.

		Undesirable	Desirable
Column I		Column II	
____1.	Hunting for reasons to support one's beliefs.	a. Rationalization	a. Denial of reality
____2.	Accepting the values and norms of others as one's own even when they are contrary to previously held values.	b. Identification	b. Identification
____3.	Attributing to others one's own unacceptable impulses, thoughts and desires.	c. Projection	c. Introjection
____4.	Ignoring disagreeable situations, topics, sights.	d. Introjection	d. Projection
		e. Denial of Reality	e. Rationalization

4. Avoid grammatical or other clues to the correct response.

Undesirable: Directions: Match the following in order to complete

the sentences on the left.

___1.	Igneous rocks are formed	A. a hardness of 7.
___2.	The formation of coal requires	B. with crystalline rock.
___3.	A geode is filled	C. a metamorphic rock.
___4.	Feldspar is classified as	D. heat and pressure.
		E. through the solid-ification of molten lava.

Desirable: Avoid sentence completion due to grammatical clues.

5. Keep matching items brief, limiting the list of stimuli to under 10.

6. Include more responses than stimuli to help prevent answering through the process of elimination.

7. When possible, reduce the amount of reading time by including only short phrases or single words in the response list.

<u>Matching: Suggestions for Preparing (additional)</u>

1. Matching is not appropriate for distinct ideas; all items in a matching section should be related.
2. Use brief lists (not more than 10)
3. Introduce the matching section by describing the relationship you want the student to find.
4. Place the section on a single page.
5. Place the stem list on the left; place the response option list on right. Sometimes, when the stems are lengthy, matching sections are created above (response option list) and below (stem list).
6. Arrange at least one list in a natural order (e.g.,

chronologically, alphabetically).

7. Have at least one more response option than stem items.
8. Indicate to the student when the response options may be used more than once to match the stem list.

Completion Test Items

The completion item requires the student to answer a question or to finish an incomplete statement by filling in a blank with the correct word or phrase. For example,

Sample Completion Item:

According to Freud, personality is made up of three major systems, the _____, the _____ and the _____.

Advantages In Using Completion Items

Completion items ……

- can provide a wide sampling of content.
- can efficiently measure lower levels of cognitive ability.
- can minimize guessing as compared to multiple-choice or true-false items.
- can usually provide an objective measure of student achievement or ability.

Limitations In Using Completion Items

Completion items ……

- are difficult to construct so that the desired response is clearly indicated.
- have difficulty measuring learning objectives requiring more than simple recall of information.
- can often include more irrelevant clues than do other item

types.

- are more time consuming to score when compared to multiple-choice or true-false items.
- are more difficult to score since more than one answer may have to be considered correct if the item was not properly prepared.

Suggestions For Writing Completion Test Items

1. Omit only significant words from the statement.
   Undesirable: Every atom has a central (core) called a nucleus.
   Desirable: Every atom has a central core called a (n) (nucleus).
2. Do not omit so many words from the statement that the intended meaning is lost.
   Undesirable: The _____type of test item is usually more _____ than the _____ type.
   Desirable: The supply type of test item is usually graded more objectively than the _____ type.
3. Avoid grammatical or other clues to the correct response.
   Undesirable: Most of the United States' libraries are organized according to the (Dewey) decimal system.
   Desirable: Which organizational system is used by most of the United States' libraries? (Dewey decimal)
4. Be sure there is only one correct response.
   Undesirable: Trees which shed their leaves annually are seed-bearing, common).
   Desirable: Trees which shed their leaves annually are called(deciduous).
5. Make the blanks of equal length.
   Undesirable: In Greek mythology, Vulcan was the son of

(Jupiter) and (Juno).

Desirable: In Greek mythology, Vulcan was the son of (Jupiter) and (Juno).

6. When possible, delete words at the end of the statement after the student has been presented a clearly defined problem.

Undesirable: (122.5) is the molecular weight of $KClO_3$.

Desirable: The molecular weight of $KClO_3$ is (122.5).

7. Avoid lifting statements directly from the text, lecture or other sources.
8. Limit the required response to a single word or phrase.

Essay Test Items

It is free response test item, help in ensuring a wide coverage of content and variety of objectives and help in evaluating complex skills.

A test item which requires a response composed by the examinee, usually in the form of one or more sentences, of a nature that no single response or pattern of responses can be listed as correct, and the accuracy and quality of which can be judged subjectively only by one skilled or informed in the subject.

The essay test is probably the most popular of all types of teacher-made tests. In general, a classroom essay test consists of a small number of questions to which the student is expected to demonstrate his/her ability to (a) recall factual knowledge, (b) organize this knowledge and (c) present the knowledge in a logical, integrated answer to the question.

An essay test item can be classified as either an extended-

response essay item or a short-answer essay item. The latter calls for a more restricted or limited answer in terms of form or scope.

Short answer type is a question requiring three value points at most may be defined as a short answer question. Value points diminish the subjectivity and help in ensuring wide coverage of content.

Advantages of Short answer Type Items

- Large portion of the content can be covered in a test.
- No opportunity for guessing.
- Easy to construct, because it measures a relatively simple outcomes.
- It can be made quit objective by carefully fixing the value points.
- Useful in evaluating the ability to interpret diagrams, charts, graphs, etc.
- If carefully prepared, deep level objectives understanding, application and problem solving skill can be evaluated.

Limitations of Short answer Type Items

- It is more subjective than the objective type of items.
- It may encourage student to memories fact and develop poor study habits.
- Mechanical scoring is not possible

An example of each type of essay item follows.

<u>Sample Extended-Response Essay Item</u>

Explain the difference between the S-R (Stimulus-Response) and the S-O-R (Stimulus-Organism-Response) theories

of personality. Include in your answer (a) brief descriptions of both theories, (b) supporters of both theories and (c) research methods used to study each of the two theories. (10 pts. 20 minutes)

Sample Short-Answer Essay Item

Identify research methods used to study the S-R (Stimulus-Response) and S-O-R (Stimulus-Organism-Response) theories of personality. (5 pts. 10 minutes)

Advantages In Using Essay Items

Essay items …….

- are easier and less time consuming to construct than are most other item types.
- provide a means for testing student's ability to compose an answer and present it in a logical manner.
- can efficiently measure higher order cognitive objectives (e.g., analysis, synthesis, evaluation).
- Easy to prepare.
- Useful in measuring certain abilities and skills.
- Permit the examinee to write down comprehensively what he knows about something.
- Promote originality and creative thinking.
- Possibility of guess work can be eliminated.
- Reduce chance on the spot copying.
- Low printing cost.

Limitations In Using Essay Items

Essay items ……

- cannot measure a large amount of content or objectives.

- generally provide low test and test scorer reliability.
- require an extensive amount of instructor's time to read and grade.
- generally do not provide an objective measure of student achievement or ability (subject to bias on the part of the grader).
- Minimum validity.
- Lack of reliability.
- No objectivity.
- Rote memory is encouraged.
- It is a time consuming test item.

Suggestions For Writing Essay Test Items

1. Prepare essay items that elicit the type of behavior you want to measure.

   Learning Objective: The student will be able to explain how the normal curve serves as a statistical model.

   Undesirable: Describe a normal curve in terms of: symmetry, modality, kurtosis and skewness.

   Desirable: Briefly explain how the normal curve serves as a statistical model for estimation and hypothesis testing.

2. Phrase each item so that the student's task is clearly indicated.

   Undesirable: Discuss the economic factors which led to the stock market crash of 1929.

   Desirable: Identify the three major economic conditions which led to the stock market crash of 1929. Discuss briefly each condition in correct chronological sequence and in one paragraph indicate how the three factors were inter-related.

3. Indicate for each item a point value or weight and an estimated

time limit for answering.

Undesirable: Compare the writings of Bret Harte and Mark Twain in terms of settings, depth of characterization, and dialogue styles of their main characters.

Desirable: Compare the writings of Bret Harte and Mark Twain in terms of settings, depth of characterization, and dialogue styles of their main characters. (10 points 20 minutes)

4. Ask questions that will elicit responses on which experts could agree that one answer is better than another.
5. Avoid giving the student a choice among optional items as this greatly reduces the reliability of the test.
6. It is generally recommended for classroom examinations to administer several short-answer items rather than only one or two extended-response items.

Suggestions For Scoring Essay Items

1. Choose a scoring model. Two of the more common scoring models are ANALYTICAL SCORING and GLOBAL QUALITY.

ANALYTICAL SCORING: Each answer is compared to an ideal answer and points are assigned for the inclusion of necessary elements. Grades are based on the number of accumulated points either absolutely (i.e., A=10 or more points, B=6-9 pts., etc.) or relatively (A=top 15% scores, B=next 30% of scores, etc.)

GLOBAL QUALITY: Each answer is read and assigned a score (e.g., grade, total points) based either on the

total quality of the response or on the total quality of the response relative to other student answers.

Examples Essay Item and Grading Models

"Americans are a mixed-up people with no sense of ethical values. Everyone knows that baseball is far less necessary than food and steel, yet they pay ball players a lot more than farmers and steelworkers." WHY? Use 3-4 sentences to indicate how an economist would explain the above situation.

Analytical Scoring

Necessary Elements to be Included in Response	Points
Salaries are based on demand relative to supply of such services.	3
Excellent ball players are rare.	2
Ball clubs have a high demand for excellent players.	2
Clarity of Response	2
	9 pts.

Global Quality

Assign scores or grades on the overall quality of the written response as compared to an ideal answer. Or, compare the overall quality of a response to other student responses by sorting the papers into three stacks:

Below Average/ Average/ Above Average

2. Try not to allow factors which are irrelevant to the learning outcomes being measured affect your grading (i.e., handwriting, spelling, neatness).

3. Read and grade all class answers to one item before going on to the next item.
4. Read and grade the answers without looking at the students' names to avoid possible preferential treatment.
5. Occasionally shuffle papers during the reading of answers to help avoid any systematic order effects (i.e., Sally's "B" work always followed Jim's "A: work thus it looked more like "C" work).
6. When possible, ask another instructor to read and grade your students' responses.

Problem Solving Test Items

Another form of a subjective test item is the problem solving or computational exam question. Such items present the student with a problem situation or task and require a demonstration of work procedures and a correct solution, or just a correct solution. This kind of test item is classified as a subjective type of item due to the procedures used to score item responses. Instructors can assign full or partial credit to either correct or incorrect solutions depending on the quality and kind of work procedures presented. An example of a problem solving test item follows.

Example Problem Solving Test Item

It was calculated that 75 men could complete a strip on a new highway in 70 days. When work was scheduled to commence, it was found necessary to send 25 men on another road project. How many days longer will it take to complete the strip? Show your work for full or partial credit.

Advantages In Using Problem Solving Items

Problem solving items ……

- minimize guessing by requiring the students to provide an original response rather than to select from several alternatives.
- are easier to construct than are multiple-choice or matching items.
- can most appropriately measure learning objectives which focus on the ability to apply skills or knowledge in the solution of problems.
- can measure an extensive amount of content or objectives.

Limitations In Using Problem Solving Items

Problem solving items ……

- generally provide low test and test scorer reliability.
- require an extensive amount of instructor time to read and grade.
- generally do not provide an objective measure of student achievement or ability (subject to bias on the part of the grader when partial credit is given).

Suggestions For Writing Problem Solving Test Items

1. Clearly identify and explain the problem.

   Undesirable: During a car crash, the car slows down at the rate of 490 m/sec2. What is the magnitude and direction of the force acting on a 100-kg driver?

   Desirable: During a car crash, the car slows down at the rate of 490 m/sec2. Using the car as a frame of reference, what is the magnitude and direction of the gram force acting on a 100-kg driver?

2. Provide directions which clearly inform the student of the type of response called for.

   Undesirable: An American tourist in Paris finds that he weighs

70 kilograms. When he left the United States he weighed 144 pounds. What was his net change in weight?

Desirable: An American tourist in Paris finds that he weighs 70 kilograms. When he left the United States he weighed 144 pounds. What was his net weight change in pounds?

3. State in the directions whether or not the student must show his/her work procedures for full or partial credit.

   Undesirable: A double concave lens is made of glass with n = 1.50. If the radii of curvature of the two lens surfaces are both 30.0 cm, what is the focal length of the lens?

   Desirable: A double concave lens is made of glass with n = 1.50. If the radii of curvature of the two lens surfaces are both 30.0 cm, what is the focal length of the lens? Show your work to receive full or partial credit.

4. Clearly separate item parts and indicate their point values. A man leaves his home and drives to a convention at an average rate of 50 miles per hour. Upon arrival, he finds a telegram advising him to return at once. He catches a plane that takes him back at an average rate of 300 miles per hour.

   Undesirable: If the total traveling time was 1 3/4 hours, how long did it take him to fly back? How far from his home was the convention?

   Desirable: If the total traveling time was 1 3/4 hours:

   (1) How long did it take him to fly back? (1 pt.)

   (2) How far from his home was the convention? (1 pt.)

Show your work for full or partial credit.

5. Use figures, conditions and situations which create a realistic problem.

   Undesirable: An automobile weighing 2,840 N (about 640 pounds) is traveling at a speed of 300 miles per hour. What is the car's kinetic energy? Show your work. (2 pts.)

   Desirable: An automobile weighing 14,200 N (about 3200 pounds) is traveling at a speed of 12m/sec. What is the car's kinetic energy? Show your work. (2 pts.)

6. Ask questions that elicit responses on which experts could agree that one solution and one or more work procedures are better than others.
7. Work through each problem before classroom administration to double-check accuracy.

Performance Test Items

A performance test item is designed to assess the ability of a student to perform correctly in a simulated situation (i.e., a situation in which the student will be ultimately expected to apply his/her learning). The concept of simulation is central in performance testing; a performance test will simulate to some degree a real life situation to accomplish the assessment. In theory, a performance test could be constructed for any skill and real life situation. In practice, most performance tests have been developed for the assessment of vocational, managerial, administrative, leadership, communication, interpersonal and physical education skills in various simulated situations. An illustrative example of a performance test item is provided below.

Sample Performance Test Item

Assume that some of the instructional objectives of an urban planning course include the development of the student's ability to effectively use the principles covered in the course in various "real life" situations common for an urban planning professional. A performance test item could measure this development by presenting the student with a specific situation which represents a "real life" situation.

For example,

An urban planning board makes a last minute request for the professional to act as consultant and critique a written proposal which is to be considered in a board meeting that very evening. The professional arrives before the meeting and has one hour to analyze the written proposal and prepare his critique. The critique presentation is then made verbally during the board meeting; reactions of members of the board or the audience include requests for explanation of specific points or informed attacks on the positions taken by the professional.

The performance test designed to simulate this situation would require that the student to be tested role play the professional's part, while students or faculty act the other roles in the situation. Various aspects of the "professional's" performance would than be observed and rated by several judges with the necessary background. The ratings could then be used both to provide the student with a diagnosis of his/her strengths and weaknesses and to contribute to an overall summary evaluation of the student's abilities.

Advantages In Using Performance Test Items

Performance test items ……..

- can most appropriately measure learning objectives which focus on the ability of the students to apply skills or knowledge in real life situations.
- usually provide a degree of test validity not possible with standard paper and pencil test items.
- are useful for measuring learning objectives in the psychomotor domain.

Limitations In Using Performance Test Items

Performance test items ……

- are difficulty and time consuming to construct.
- are primarily used for testing students individually and not for testing groups. Consequently, they are relatively costly, time consuming, and inconvenient forms of testing.
- generally provide low test and test scorer reliability.
- generally do not provide an objective measure of student achievement or ability (subject to bias on the part of the observer/grader).

Suggestions for Writing Performance Test Items

- Prepare items that elicit the type of behavior you want to measure.
- Clearly identify and explain the simulated situation to the student.
- Make the simulated situation as "life-like" as possible.
- Provide directions which clearly inform the students of the type of response called for.
- When appropriate, clearly state time and activity limitations in the directions.
- Adequately train the observer(s)/scorer(s) to ensure that

they are fair in scoring the appropriate behaviors.

TWO METHODS FOR ASSESSING TEST ITEM QUALITY

This section presents two methods for collecting feedback on the quality of your test items. The two methods include using self-review checklists and student evaluation of test item quality. You can use the information gathered from either method to identify strengths and weaknesses in your item writing.

CHECKLIST FOR EVALUATING TEST ITEMS

Evaluate your test items by checking the suggestions which you feel you have followed:

Multiple-Choice Test Items

____ When possible, stated the stem as a direct question rather than as an incomplete statement.

____ Presented a definite, explicit and singular question or problem in the stem.

____ Eliminated excessive verbiage or irrelevant information from the stem.

____ Included in the stem any word(s) that might have otherwise been repeated in each alternative.

____ Used negatively stated stems sparingly. When used, underlined and/or capitalized the negative word(s).

____ Made all alternatives plausible and attractive to the less knowledgeable or skillful student.

____ Made the alternatives grammatically parallel with each other, and consistent with the stem.

____ Made the alternatives mutually exclusive.

____ When possible, presented alternatives in some logical order (e.g., chronologically, most to least).

____ Made sure there was only one correct or best response per item.

____ Made alternatives approximately equal in length.

____ Avoided irrelevant clues such as grammatical structure, well known verbal associations or connections between stem and answer.

____ Used at least four alternatives for each item.

____ Randomly distributed the correct response among the alternative positions throughout the test having approximately the same proportion of alternatives a, b, c, d, and e as the correct response.

____ Used the alternatives "none of the above" and "all of the above" sparingly. When used, such alternatives were occasionally the correct response.

<u>True-False Test Items</u>

____ Based true-false items upon statements that are absolutely true or false, without qualifications or exceptions.

____ Expressed the item statement as simply and as clearly as possible.

____ Expressed a single idea in each test item.

____ Included enough background information and qualifications so that the ability to respond correctly did not depend on some special, uncommon knowledge.

____ Avoided lifting statements from the text, lecture or other materials.

____ Avoided using negatively stated item statements.

____ Avoided the use of unfamiliar language.

____ Avoided the use of specific determiners such as "all," "always," "none," "never," etc., and qualifying determiners such as "usually," "sometimes," "often," etc.

____ Used more false items than true items (but not more than 15% additional false items).

Matching Test Items

____ Included directions which clearly stated the basis for matching the stimuli with the response.

____ Explained whether or not a response could be used more than once and indicated where to write the answer.

____ Used only homogeneous material.

____ When possible, arranged the list of responses in some systematic order (e.g., chronologically, alphabetically).

____ Avoided grammatical or other clues to the correct response.

____ Kept items brief (limited the list of stimuli to under 10).

____ Included more responses than stimuli.

____ When possible, reduced the amount of reading time by including only short phrases or single words in the response list.

Completion Test Items

____ Omitted only significant words from the statement.

____ Did not omit so many words from the statement that the intended meaning was lost.

____ Avoided grammatical or other clues to the correct response.

____ Included only one correct response per item.

____ Made the blanks of equal length.

____ When possible, deleted the words at the end of the statement after the student was presented with a clearly defined problem.

____ Avoided lifting statements directly from the text, lecture or other sources.

____ Limited the required response to a single word or phrase.

Essay Test Items

____ Prepared items that elicited the type of behavior you wanted to measure.

____ Phrased each item so that the student's task was clearly indicated.

____ Indicated for each item a point value or weight and an estimated time limit for answering.

____ Asked questions that elicited responses on which experts could agree that one answer is better than others.

____ Avoided giving the student a choice among optional items.

____ Administered several short-answer items rather than 1 or 2 extended-response items.

Grading Essay Test Items

____ Selected an appropriate grading model.

____ Tried not to allow factors which were irrelevant to the learning outcomes being measured to affect your grading (e.g., handwriting, spelling, neatness).

____ Read and graded all class answers to one item before going on to the next item.

____ Read and graded the answers without looking at the student's name to avoid possible preferential treatment.

____ Occasionally shuffled papers during the reading of answers.

____ When possible, asked another instructor to read and grade your students' responses.

Problem Solving Test Items

____ Clearly identified and explained the problem to the student.

____ Provided directions which clearly informed the student of the type of response called for.

____ Stated in the directions whether or not the student must show work procedures for full or partial credit.

____ Clearly separated item parts and indicated their point values.

____ Used figures, conditions and situations which created a realistic problem.

____ Asked questions that elicited responses on which experts could agree that one solution and one or more work procedures are better than others.

____ Worked through each problem before classroom administration.

Performance Test Items

____ Prepared items that elicit the type of behavior you wanted to measure.

____ Clearly identified and explained the simulated situation to the student.

____ Made the simulated situation as "life-like" as possible.

____ Provided directions which clearly inform the students of the type of response called for.

____ When appropriate, clearly stated time and activity limitations in the directions.

____ Adequately trained the observer(s)/scorer(s) to ensure that they were fair in scoring the appropriate behaviors.

STUDENT EVALUATION OF TEST ITEM QUALITY

Using ICES questionnaire items to assess your test item quality:

The following set of ICES (Instructor and Course Evaluation System) questionnaire items can be used to assess the quality of your test items. The items are presented with their original ICES catalogue number. You are encouraged to include one or more of the items on the ICES evaluation form in order to collect student opinion of your item writing quality.

102--How would you rate the	Excellent	Poor

instructor's examination questions?		
103--How well did examination questions reflect content and emphasis of the course?	Well related	Poorly related
109--Were exams, papers, reports returned with errors explained or personal comments?	Almost always	Almost never
114--The exams reflected important points in the reading assignments.	Strongly agree	Strongly disagree
115--Were the instructor's test questions thought provoking?	Yes Definitely	Definitely no
116--Did the exams challenge you to do original thinking?	Yes, very challenging	No, not challenging
117--Examinations mainly tested trivia.	Strongly agree	Strongly disagree
118--Were there "trick" or trite questions on tests?	Lots of them	Few if any
119--Were exam questions worded clearly?	Yes, very clear	No, very unclear
121--How was the length of exams for the time allotted?	Too long	Too short
122--How difficult were the examinations?	Too difficult	Too easy
123--I found I could score reasonably well on exams by just cramming.	Strongly agree	Strongly disagree
125--Were exams adequately	Yes,	No, not

discussed upon return?	adequately	enough

Rules of Thumb

Some rules of thumb exist for how long it takes most students to answer various types of questions according to Linn and Gronlund (2000):

1. A true-false test item takes 15 seconds to answer unless the student is asked to provide the correct answer for false questions. Then the time increases to 30-45 seconds.
2. A seven item matching exercise takes 60-90 seconds.
3. A four response multiple choice test item that asks for an answer regarding a term, fact, definition, rule or principle (knowledge level item) takes 30 seconds. The same type of test item that is at the application level may take 60 seconds.
4. Any test item format that requires solving a problem, analyzing, synthesizing information or evaluating examples adds 30-60 seconds to a question.
5. Short-answer test items take 30-45 seconds.
6. 6. An essay test takes 60 seconds for each point to be compared and contrasted.

An example of multiple choice items in science:

Which statement explains why the Sun appears to rise and set each day?

A. Earth rotates. B. The Sun rotates. C. The Sun revolves around Earth. D. Earth revolves around the Sun.
Which unit is used to measure how warm or cool the air is? A. grams B. kilometers C. degrees Celsius D. cubic centimeters
Which kind of energy is produced when a student beats a drum? A. electrical B. sound C. light D. chemical
Electricity traveling through a wire is an example of … A. a force applied by a simple machine B. energy flowing through the water cycle C. Earth's gravitational pull on an object D. energy being transferred from place to place
We need our _____ to smell. A. Mouth B. Tongue C. Nose

An example of short answer type questions in sciences:

1	When light bends as it enters a different medium the process is known as what?

2	A magnifying glass is what type of lens?
3	Electric resistance is typically measured in what units?
4	A person who studies physics is known as a?
5	Metals expand when heated and do what when cooled?
6	What is the first name of the famous scientist who gave us Newton's three laws of motion?
7	What state of the art computer technology is used to train pilots when wanting to copy the experience of flying an aircraft?
8	Electric power is typically measured in what units?
9	The most recognized model of how the universe begun is known as the?
10	Who is the Hubble Space Telescope named after?
11	The wire inside an electric bulb is known as the what?
12	Theoretical physicist James Maxwell was born in what country?
13	Infrared light has a wavelength that is too long or short to be visible for humans?
14	What kind of eclipse do we have when the moon is between the sun and the earth?
15	True or false? Iron is attracted by magnets.
16	What is the earth's primary source of energy?
17	Conductors have a high or low resistance?
18	Electric current is typically measured in what units?
19	What scientist is well known for his theory of relativity?
20	Earth is located in which galaxy?

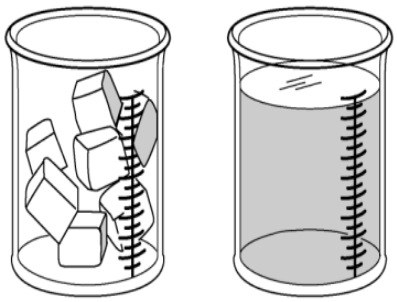
Physics Quiz Answers			
1. Refraction	2. Convex	3. Ohms	4. Physicist
5. Contract	6. Isaac	7. A flight simulator	8. Watts
9. Big bang	10. Edwin Hubble	11. Filament	12. Scotland
13. Long	14. A solar eclipse	15. True	16. The sun
17. Low	18. Amperes	19. Albert Einstein	20. The Milky Way galaxy

An example of true-false type questions in sciences:

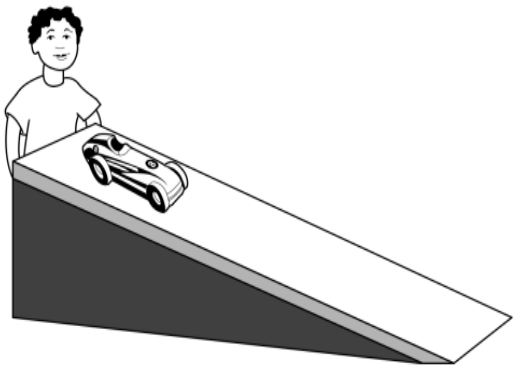
Our eyes send information to the brain with the auditory nerve. A. True

B. false
The eye is an organism for us to see. A. True B. false

An example of short answer type question in sciences:
Victor has two glasses. One glass is filled with ice cubes and the other is filled with water. Give three ways the ice and water are different.



An example of essay type question in sciences:



Jim put a toy car on a ramp. The car slowly moved down the ramp. What was it about the ramp's surface that caused the car

to move slowly? How could Jim change the ramp's surface to get the car to move faster? Write your answer on the lines below.

……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……

Chapter IV
Construction of Achievement Tests

Achievement test is tool for teachers for evaluation of students in school situation. With the help of achievement test we can measure the amount of success of an individual in specific field. In school environment it is used as an instrument to measure success of an individual in particular subject or group of subjects. It gives the knowledge about what an individual acquire by testing his abilities.

Meaning of Teacher Made Test:

Carefully constructed teacher-made tests and standardized tests are similar in many ways. Both are constructed on the basis of carefully planned table of specifications, both have the same type of test items, and both provide clear directions to the students.

Still the two differ. They differ in the quality of test items, the reliability of test measures, the procedures for administering and scoring and the interpretation of scores. No doubt, standardized tests are good and better in quality, more reliable and valid.

But a classroom teacher cannot always depend on standardized tests. These may not suit to his local needs, may not be readily available, may be costly, may have different objectives. In order to fulfill the immediate requirements, the teacher has to prepare his own tests which are usually objective type in nature.

Teacher-made tests are normally prepared and administered for testing classroom achievement of students, evaluating the method of teaching adopted by the teacher and other curricular programs of the school.

Teacher-made test is one of the most valuable instrument in the hands of the teacher to solve his purpose. It is designed to solve the problem or requirements of the class for which it is prepared.

It is prepared to measure the outcomes and content of local curriculum. It is very much flexible so that, it can be adopted to any procedure and material. It does not require any sophisticated technique for preparation.

Taylor has highly recommended for the use of these teacher-made objective type tests, which do not require all the four steps of standardized tests nor need the rigorous processes of standardization. Only the first two steps planning and preparation are sufficient for their construction.

Features of Teacher-Made Tests:

1. The items of the tests are arranged in order of difficulty.
2. These are prepared by the teachers which can be used for prognosis and diagnosis purposes.
3. The test covers the whole content area and includes a large number of items.
4. The preparation of the items conforms to the blueprint.
5. Test construction is not a single man's business, rather it is a co-operative endeavor.
6. A teacher-made test does not cover all the steps of a standardized test.
7. Teacher-made tests may also be employed as a tool for formative evaluation.
8. Preparation and administration of these tests are economical.

9. The test is developed by the teacher to ascertain the student's achievement and proficiency in a given subject.
10. Teacher-made tests are least used for research purposes.
11. They do not have norms whereas providing norms is quite essential for standardized tests.

Construction procedure of an Achievement Test:

Steps/Principles of Construction of Teacher-made Test:

A teacher-made test does not require a well-planned preparation. Even then, to make it more efficient and effective tool of evaluation, careful considerations arc needed to be given while constructing such tests. The following steps may be followed for the preparation of teacher-made test:

Steps in Constructing Teacher-Made Test

1. Planning the Test

In planning the test the following should be observed: the objectives of the subjects, the purpose for which the test is administered, the availability of facilities and equipments, the nature of the testee, the provision for review and the length of the test. Planning of a teacher-made test includes:

- Determining the purpose and objectives of the test, 'as what to measure and why to measure'.
- Deciding the length of the test and portion of the syllabus to be covered.
- Specifying the objectives in behavioral terms. If needed, a table can even be prepared for specifications and weightage given to the objectives to be measured.
- Deciding the number and forms of items (questions) according to blueprint.

- Having a clear knowledge and understanding of the principles of constructing essay type, short answer type and objective type questions.
- Deciding date of testing much in advance in order to give time to teachers for test preparation and administration.
- Seeking the co-operation and suggestion of co-teachers, experienced teachers of other schools and test experts.

Designing is the key step in constructing a test. The exam designers should take care when planning and testing success. They should remember what kind of questions to pose and how they are used. He has to decide the weightage for the various objectives and units of the course.

The following points help the exam setter to design the test methodically.

- Identify objectives
- Measuring the content coverage
- Allocation of time
- Allocation of maximum marks
- Decide to include or not multiple-choice questions
- Focus on each topic and area in the questions that belong

2. Preparing the Test (Preparation of the Test)

The process of writing good test items is not simple – it requires time and effort. It also requires certain skills and proficiencies on the part of the writer. Therefore, a test writer must master the subject matter he/she teaches, must understand his

testee, must be skillful in verbal expression and most of all familiar with various types of tests.

Planning is the philosophical aspect and preparation is the practical aspect of test construction. All the practical aspects to be taken into consideration while one constructs the tests. It is an art, a technique. One is to have it or to acquire it. It requires much thinking, rethinking and reading before constructing test items.

Different types of objective test items viz., multiple choice, short-answer type and matching type can be constructed. After construction, test items should be given lo others for review and for seeking their opinions on it.

The suggestions may be sought even from others on languages, modalities of the items, statements given, correct answers supplied and on other possible errors anticipated. The suggestions and views thus sought will help a test constructor in modifying and verifying his items afresh to make it more acceptable and usable.

After construction of the test, items should be arranged in a simple to complex order. For arranging the items, a teacher can adopt so many methods viz., group-wise, unit-wise, topic wise etc. Scoring key should also be prepared forthwith to avoid further delay in scoring.

Direction is an important part of a test construction. Without giving a proper direction or instruction, there will be a probability of losing the authenticity of the test reliability. It may create a misunderstanding in the students also.

Thus, the direction should be simple and adequate to enable the students to know:

- The time for completion of test,

- The marks allotted to each item,
- Required number of items to be attempted,
- How and where to record the answer? and
- The materials, like graph papers or logarithmic table to be used.

The key elements to consider in design for testing are:
Weightage to objectives -- Proficiency is defined as the standard by which a student acquires grade-appropriate knowledge and skills up to the test stage. Good teaching methods are expected to increase the amount of learning in the school year, thus increasing achievement scores, and producing more "proficient" students than ever before.

Bloom's Taxonomy of Learning Objectives

A group of cognitive psychologists, curriculum theorists and instructional researchers, and testing and assessment specialists published in 2001 a revision of Bloom's Taxonomy with the title *A Taxonomy for Teaching, Learning, and Assessment*. This title draws attention away from the somewhat static notion of "educational objectives" (in Bloom's original title) and points to a more dynamic conception of classification.

In 1956, Benjamin Bloom with collaborators Max Englehart, Edward Furst, Walter Hill, and David Krathwohl published a framework for categorizing educational goals: *Taxonomy of Educational Objectives*. Familiarly known as Bloom's Taxonomy, this framework has been applied by generations of K-12 teachers and college instructors in their teaching.

The framework elaborated by Bloom and his collaborators consisted of six major categories: Knowledge, Comprehension, Application, Analysis, Synthesis, and Evaluation. The categories after Knowledge were presented as "skills and abilities," with the understanding that knowledge was the necessary precondition for putting these skills and abilities into practice.

While each category contained subcategories, all lying along a continuum from simple to complex and concrete to abstract, the taxonomy is popularly remembered according to the six main categories.

The Original Taxonomy (1956)

Here are the authors' brief explanations of these main categories in from the appendix of *Taxonomy of Educational Objectives* (*Handbook One*, pp. 201-207):

- **Knowledge** "involves the recall of specifics and universals, the recall of methods and processes, or the recall of a pattern, structure, or setting."
- **Comprehension** "refers to a type of understanding or apprehension such that the individual knows what is being communicated and can make use of the material or idea being communicated without necessarily relating it to other material or seeing its fullest implications."
- **Application** refers to the "use of abstractions in particular and concrete situations."
- **Analysis** represents the "breakdown of a communication into its constituent elements or parts such that the relative hierarchy of ideas is made clear

and/or the relations between ideas expressed are made explicit."

- **Synthesis** involves the "putting together of elements and parts so as to form a whole."
- **Evaluation** engenders "judgments about the value of material and methods for given purposes."

Revised Bloom's Taxonomy

The Revised Taxonomy (2001)

A group of cognitive psychologists, curriculum theorists and instructional researchers, and testing and assessment specialists published in 2001 a revision of Bloom's Taxonomy with the title A Taxonomy for Teaching, Learning, and Assessment. This title draws attention away from the somewhat static notion of "educational objectives" (in Bloom's original title) and points to a more dynamic conception of classification.

The authors of the revised taxonomy underscore this dynamism, using verbs and gerunds to label their categories and subcategories (rather than the nouns of the original taxonomy). These "action words" describe the cognitive processes by which thinkers encounter and work with knowledge:

A statement of a learning objective contains a verb (an action) and an object (usually a noun).

- The verb generally refers to [actions associated with] the intended cognitive process.
- The object generally describes the knowledge students are expected to acquire or construct. (Anderson and Krathwohl, 2001, pp. 4–5)

The cognitive process dimension represents a continuum of increasing cognitive complexity—from remember to create. Anderson and Krathwohl identify 19 specific cognitive processes that further clarify the bounds of the six categories (Table 1).

Table 1. The Cognitive Process Dimension – categories, cognitive processes (and alternative names)

Remember	recognizing (identifying) recalling (retrieving)
Understand	interpreting (clarifying, paraphrasing, representing, translating) exemplifying (illustrating, instantiating) classifying (categorizing, subsuming) summarizing (abstracting, generalizing) inferring (concluding, extrapolating, interpolating, predicting) comparing (contrasting, mapping, matching) explaining (constructing models)

Apply	executing (carrying out) implementing (using)
Analyze	differentiating (discriminating, distinguishing, focusing, selecting)organizing (finding, coherence, integrating, outlining, parsing, structuring) attributing (deconstructing)
Evaluate	checking (coordinating, detecting, monitoring, testing) critiquing (judging)
Create	generating (hypothesizing) planning (designing) producing (construct)

The **knowledge dimension** represents a range from concrete (factual) to abstract (metacognitive) (Table 2). Representation of the knowledge dimension as a number of discrete steps can be a bit misleading. For example, all procedural knowledge may not be more abstract than all conceptual knowledge. And metacognitive knowledge is a special case. In this model, "*metacognitive knowledge* is knowledge of [one's own] cognition and about oneself in relation to various subject matters . . . " (Anderson and Krathwohl, 2001, p. 44).

Table 2. The Knowledge Dimension

Factual	Conceptual	Procedural	Metacognitive
• knowledge of terminology • knowledge of specific details and elements	• knowledge of classifications and categories • knowledge of principles and generalizations • knowledge of theories, models, and structures	• knowledge of subject-specific skills and algorithms • knowledge of subject-specific techniques and methods • knowledge of criteria for determining when to use appropriate procedures	• strategic knowledge • knowledge about cognitive tasks, including appropriate contextual and conditional knowledge • self-knowledge

Bloom's Revised Taxonomy Model (Responsive)

Note: These are learning objectives – not learning activities. It may be useful to think of preceding each objective with something like, "students will be able to…:

The Knowledge Dimension			
Factual	Conceptual	Procedural	Metacognitive
The basic elements a student must know to be acquainted with a discipline or solve problems in it.	The interrelationships among the basic elements within a larger structure that enable them to function together.	How to do something, methods of inquiry, and criteria for using skills, algorithms, techniques, and methods.	Knowledge of cognition in general as well as awareness and knowledge of one's own cognition

The Cognitive Process Dimension

Remember Retrieve relevant knowledge from long-term memory.	Remember + Factual	Remember + Conceptual	Remember + Procedural	Remember + Metacognitive
	List primary and secondary colors.	Recognize symptoms of exhaustion.	Recall how to perform CPR.	Identify strategies for retaining information.
Understand Construct meaning from instructional messages, including oral, written and graphic communication.	understand + Factual	understand + Conceptual	understand + Procedural	understand + Metacognitive
	Summarize features of a new product.	Classify adhesives by toxicity.	Clarify assembly instructions.	Predict one's response to culture shock.
Apply Carry out or use a procedure in a given	Apply + Factual	Apply + Conceptual	Apply + Procedural	Apply + Metacognitive
	Respond to frequently asked questions.	Provide advice to novices.	Carry out pH tests of water samples.	Use techniques that match one's

situation.				strengths.
Analyze Break material into foundational parts and determine how parts relate to one another and the overall structure or purpose	Analyze + Factual	Analyze + Conceptual	Analyze + Procedural	Analyze + Metacognitive
	Select the most complete list of activities.	Differentiate high and low culture.	Integrate compliance with regulations.	Deconstruct one's biases.
Evaluate Make judgments based on criteria and standards.	Evaluate + Factual	Evaluate + Conceptual	Evaluate + Procedural	Evaluate + Metacognitive
	Check for consistency among sources.	Determine relevance of results.	Judge efficiency of sampling techniques.	Reflect on one's progress.
Create Put elements together to form a coherent whole; reorganize into a new pattern or structure.	Create + Factual	Create + Conceptual	Create + Procedural	Create + Metacognitive
	Generate a log of daily activities.	Assemble a team of experts.	Design efficient project workflow.	Create a learning portfolio.

Why Use Bloom's Taxonomy?

The authors of the revised taxonomy suggest a multi-layered answer to this question, to which the author of this teaching guide has added some clarifying points:

1. Objectives (learning goals) are important to establish in a pedagogical interchange so that teachers and students alike understand the purpose of that interchange.
2. Organizing objectives helps to clarify objectives for themselves and for students.
3. Having an organized set of objectives helps teachers to:

- "plan and deliver appropriate instruction";
- "design valid assessment tasks and strategies"; and
- "ensure that instruction and assessment are aligned with the objectives."

Weightage to content -- Classification of achievement tests based on content or subject, language tests, reading tests, spelling tests etc., and different forms like an oral test.

Weightage to form of questions -- Achievement test items can be essay type questions or short answer questions or objective type questions or a combination of all these types.

Weightage to difficulty level -- The level of rigidity has to be considered while writing the object. It is advisable to sort the questions according to their level of difficulty.

Sample illustrations of Weightage Tables

A. Content Weightage

Sl. No.	Subject	Marks	Percentage (%)
1	Part-1	8	40%
2	Part-2	12	60%
Total		20	100%

B. Objectives Weightage

Sl. No.	Objectives	Marks	Percentage (%)
1	Knowledge based	8	40%
2	Understanding / conceptual	6	30%

3	Application specific	2	10%
4	Skill based	4	20%
Total		20	100%

C. Form of Questions Weightage

Sl. No.	Form of Questions	Marks	Percentage (%)
1	Easy Types	4	20%
2	Short Answer Types	10	50%
3	Very Short Answer Types	4	20%
4	Objective Types	2	10%
Total		20	100%

D. Difficulty Level Weightage

Sl. No.	Difficulty Level	Marks	Percentage (%)
1	Easy Level	6	30%
2	Average Level	10	50%
3	Difficult Level	4	20%
Total		20	100%

3. Reproducing the Test

In reproducing test, the duplicating machine and who will facilitate in typing and mimeographing be considered.

4. Administering the Test

Test should be administered in an environment familiar to the students, sitting arrangements is observed, corrections are

made before the start of the test, distribution and collection of papers are planned, and time should be written on the board. One more important thing to remember is, do not allow every testee to leave the room except for personal necessity.

5. Scoring the Test

The best procedure in scoring objective test is to give one point of credit for each correct answer. In case of a test with only two or three options to each item, the correction formula should be applied. Example: for two option, score equals right minus wrong (S = R-W). For three options, score equals right minus one-half wrong (S = R-1/2 W or S= R-W/2). Correction formula is not applied to four or more options. If correction formula is employed students should be informed beforehand.

Scoring Key - For objective type items where the answers are in the form of some letters or other symbols, a scoring key is created.

Marking Scheme - In the case of short answers and essay type questions, a marking plan is created. When creating a marking scheme, the examiner should list the accumulated value points and determine the points assigned to each value point.

6. Evaluating the Test

The test is evaluated as to the quality of the student's responses and the quality of the test itself. Index difficulty and discrimination index of the test item is considered. Fifty (50) per cent difficulty is better. Item of 100 per cent and zero (0) per cent answered by students are valueless in a test of general achievement.

7. Interpreting Test Results

Standardized achievement tests are interpreted based on norm tables. Table of norm are not applicable to teacher-made test.

Table of Specifications

What is a Table of Specifications?

A Table of Specifications is a two-way chart which describes the topics to be covered by a test and the number of items or points which will be associated with each topic. Sometimes the types of items are described, as well

The purpose of a Table of Specifications is to identify the achievement domains being measured and to ensure that a fair and representative sample of questions appear on the test. Teachers cannot measure every topic or objective and cannot ask every question they might wish to ask. A Table of Specifications allows the teacher to construct a test which focuses on the key areas and weights those different areas based on their importance. A Table of Specifications provides the teacher with evidence that a test has content validity, that it covers what should be covered.

Tables of Specification typically are designed based on the list of course objectives, the topics covered in class, the amount of time spent on those topics, textbook chapter topics, and the emphasis and space provided in the text. In some cases a great weight will be assigned to a concept that is extremely important, even if relatively little class time was spent on the topic. Three steps are involved in creating a Table of Specifications: 1) choosing the measurement goals and domain to be covered, 2) breaking the domain into key or fairly independent parts-

concepts, terms, procedures, applications, and 3) constructing the table. Teachers have already made decisions (or the district has decided for them) about the broad areas that should be taught, so the choice of what broad domains a test should cover has usually already been made. A bit trickier is to outline the subject matter into smaller components, but most teachers have already had to design teaching plans, strategies, and schedules based on an outline of content. Lists of classroom objectives, district curriculum guidelines, and textbook sections, and keywords are other commonly used sources for identifying categories for Tables of Specification. When actually constructing the table, teachers may only wish to use a simple structure, as with the first example above, or they may be interested in greater detail about the types of items, the cognitive levels for items, the best mix of objectively scored items, open-ended and constructed-response items, and so on, with even more guidance than is provided in the second example.

Kubiszyn & Borich, (2003) emphasized the following significance and components of TOS:

1. A Table of Specifications consists of a two-way chart relating instructional objectives to the instructional content. The column of the chart lists the objectives or "levels of skills" (Gredlcr, 1999) to be addressed; the rows list the key concepts or content the test is to measure. According to Bloom, et al. (1971), "We have found it useful to represent the relation of content and behaviors in the form of a two dimensional table with the objectives on one axis, the content on the other".
2. A Table of Specifications identifies not only the content areas covered in class; it identifies the performance

objectives at each level of the cognitive domain of Bloom's Taxonomy. Teachers can be assured that they are measuring students' learning across a wide range of content as well as cognitive processes requiring higher order thinking.

3. A Table of Specifications is developed before the test is written. In fact it should be constructed before the actual teaching begins.
4. The purpose of a Table of Specifications is to identify the achievement domains being measured and to ensure that a fair and representative sample of questions appear on the test. Carey (1988) pointed out that the time available for testing depended not only on the length of the class period but also on students' attention spans.

A simple table might look like this:

TOS for Biology Exam in Circulation Unit

Objectives / Contents	application	analysis	evaluation	weightages
Parts	?	?	?	10 %
Function	?	?	?	40 %
Relation to other systems	?	?	?	50 %
Weightages	50 %	25%	25%	100%

Number of items in each cell = the marginal weight of content × the marginal weight of objective × the total number of Questions needed.

Chapter IV
Conduct the Item Analysis

Introduction

The item analysis is an important phase in the *development of an exam program*. In this phase statistical methods are used to identify any test items that are not working well. If an item is too easy, too difficult, failing to show a difference between skilled and unskilled examinees, or even scored incorrectly, an item analysis will reveal it. The two most common statistics reported in an item analysis are the item difficulty, which is a measure of the proportion of examinees who responded to an item correctly, and the item discrimination, which is a measure of how well the item discriminates between examinees who are knowledgeable in the content area and those who are not. An additional analysis that is often reported is the distractor analysis. The distractor analysis provides a measure of how well each of the incorrect options contributes to the quality of a multiple choice item. Once the item analysis information is available, an item review is often conducted.

Item Analysis Statistics

Item Difficulty Index

The item difficulty index is one of the most useful, and most frequently reported, item analysis statistics. It is a measure of the proportion of examinees who answered the item correctly; for this reason it is frequently called the *p*-value. As the proportion of examinees that got the item right, the *p*-value might more properly be called the item easiness index, rather than the item difficulty. It can range between 0.0 and 1.0, with a higher value indicating that

a greater proportion of examinees responded to the item correctly, and it was thus an easier item. For criterion-referenced tests (CRTs), with their emphasis on mastery-testing, many items on an exam form will have *p*-values of .9 or above. Norm-referenced tests (NRTs), on the other hand, are designed to be harder overall and to spread out the examinees' scores. Thus, many of the items on an NRT will have difficulty indexes between .4 and .6.

Item Discrimination Index

The item discrimination index is a measure of how well an item is able to distinguish between examinees who are knowledgeable and those who are not, or between masters and non-masters. There are actually several ways to compute an item discrimination, but one of the most common is the *point-biserial correlation*. This statistic looks at the relationship between an examinee's performance on the given item (correct or incorrect) and the examinee's score on the overall test. For an item that is highly discriminating, in general the examinees who responded to the item correctly also did well on the test, while in general the examinees who responded to the item incorrectly also tended to do poorly on the overall test.

The possible range of the discrimination index is -1.0 to 1.0; however, if an item has a discrimination below 0.0, it suggests a problem. When an item is discriminating negatively, overall the most knowledgeable examinees are getting the item wrong and the least knowledgeable examinees are getting the item right. A negative discrimination index may indicate that the item is measuring something other than what the rest of the test is measuring. More often, it is a sign that the item has been mis-keyed.

When interpreting the value of a discrimination it is important to be aware that there is a relationship between an item's difficulty index and its discrimination index. If an item has a very high (or very low) *p*-value, the potential value of the discrimination index will be much less than if the item has a mid-range p-value. In other words, if an item is either very easy or very hard, it is not likely to be very discriminating. A typical CRT, with many high item *p*-values, may have most item discriminations in the range of 0.0 to 0.3. A useful approach when reviewing a set of item discrimination indexes is to also view each item's p-value at the same time. For example, if a given item has a discrimination index below .1, but the item's *p*-value is greater than .9, you may interpret the item as being easy for almost the entire set of examinees, and probably for that reason not providing much discrimination between high ability and low ability examinees.

Item Analysis Procedure:

Item analysis procedure gives special emphasis on item difficulty level and item discriminating power. The item analysis procedure follows the following steps:

1. The test papers should be ranked from highest to lowest.
2. Select 27% test papers from highest and 27% from lowest end. For example if the test is administered on 60 students then select 16 test papers from highest end and 16 test papers from lowest end.
3. Keep aside the other test papers as they are not required in the item analysis.
4. Tabulate the number of pupils in the upper and lower group who selected each alternative for each test item. This can be

done on the back of the test paper or a separate test item card may be used.

Item Difficulty

Measure whether an item was too easy or too hard.

Item difficulty is the percentage of students that correctly answered the items. The range is from 0% to 100%, the higher the value, easier is the item. p value above 0.90 are very easy items and below 0.20 indicate difficult items and should be reviewed for possible confusing language. Optimum difficulty level is 0.50 for maximum discrimination between high and low achievers. Generally, items of moderate difficulty are to be preferred to those which are much easier or much harder. Difficulty value of item provides information whether test was too easy or too difficult.

Calculate item difficulty index by using the following formula:

$$\rho = \frac{R}{T} \times 100$$

Where R= Total number of students got the item correct.
T = Total number of students tried the item.

In the following table out of 32 students from both the groups 20 students have answered the item correctly and 30 students have tried the item.

group	No. of Pupils	Alternatives					Total number of pupils responded	Item Difficulty
		A	B	C*	D	E		
Upper	16	0	0	15	1	0	16	66.67
Lower	16	5	4	5	0	0	14	
* correct answer								

The item difficulty is as following:

$$\rho = \frac{R}{T} \times 100 = \frac{20}{30} \times 100 = 66.67$$

The formula for calculating difficulty value was:

$$DV = \frac{R_U + R_L}{N_U + N_L}$$

Where,

DV = Difficulty value

R_U = the number of students in the upper group who respond correctly

R_L = the number of students in the lower group who respond correctly

N_U = the total number of students in the upper group

N_L = the total number of students in the lower group

Henning (1987) criteria and guidelines for categorizing difficulty indices is a widely quoted set of guidelines were used in this analysis:

Henning's Guidelines (Difficulty Value)		
High Difficult	Medium	Low (easy)
≤ 0.33	0.34-0.66	≥ 0.67

Also, we can use the following interpretation of difficulty index:

parameter / range (Difficulty Index)	Interpretation
≤ 30	Difficult
31 - 40	Good
41 - 60	Excellent
≥ 60	Easy

Correction for guessing:

When the students do not have sufficient time to answer the test or the students are not ready to take the test at that time they guess the correct answer, in recognition type items. In that case to eliminate the effect of guessing the following formula is used:

$$Score = R - \frac{W}{n-1}$$

Where R = No. of right responses
W = No. of wrong responses
n = No. of alternatives

But there is a lack of agreement among psychometricians about the value of the correction formula so far as validity and reliability are concerned. In the words of Ebel "neither the instruction nor penalties will remedy the problem of guessing."

Guilford is of the opinion that “when the middle is excluded in item analysis the question of whether to correct or not correct the total scores becomes rather academic.” Little said “correction may either under or over correct the pupils’ score.” Keeping in view the above opinions, the test-maker should decide not to use the correction for guessing. To avoid this situation he should give enough time for answering the test item.

Item Discrimination

Measure whether an item discriminate between students who know the material well and students who did not.

Is that ability of an item on the basis of which the discrimination is made between superiors and inferiors.

Discrimination Power of an item may be defined as how adequately an item separates or discriminates between high scores and low scores on an entire test. For assessing such ability of an item related to its discrimination power is known as discrimination index. A good item should discriminate between those who score high (top 27% cases) on the test and who score low on the test (bottom 27% cases).

Discrimination index of each item was found by applying following formula:

Item discriminating power =

$$\frac{R_U - R_L}{.5T}$$

Where:

R_U= Students from upper group who got the answer correct (Number of students of higher group answering the item correctly).

R_L= Students from lower group who got the answer correct (Number of students of lower group answering the item correctly).

0.5 T = half of the total number of pupils included in the item analysis.

In our example 15 students from upper group responded the item correctly and 5 from lower group responded the item correctly.

$$D = \frac{R_U - R_L}{.5T} = \frac{15-5}{0.5\times 32} = \frac{10}{16} = .625 \approx 0.63$$

A high positive ratio indicates the high discriminating power. Here .63 indicates an average discriminating power. If all the 16 students from lower group and 16 students from upper group answers the item correctly then the discriminating power will be 0.00. It indicates that the item has no discriminating power. If all the 16 students from upper group answer the item correctly and all the students from lower group answer the item in correctly then the item discriminating power will be 1.00 it indicates an item with maximum positive discriminating power.

Ebel's (1979) criteria and guidelines for categorizing discrimination index is a widely quoted set of guidelines were used in this analysis:

Discrimination Index / range	Item evaluation / verbal description
0.40 and above	Very good items (The item is quit satisfactorily)
Between 0.30 – 0.39	Reasonably good but possibly subject to improvement (Less or no revision is required)
Between 0.20 – 0.29	Marginal (fair) items, usually needing and being subject to improvement (The item is marginal and need revision)
0.19 and below	Very Poor items, to be rejected or improved by revision (The item should be eliminated or completely revised)

Example: Calculate the discrimination index using the following data:

Item No.	Number of correct answers in group	
	Upper 1/4	Lower 1/4
1	90	20
2	80	70
3	100	0
4	100	100
5	50	50
6	20	60
Number of students per group = 100		

The answer:

Item No.	Number of correct answers in group		D
	Upper 1/4	Lower 1/4	
1	90	20	0.70
2	80	70	0.10

3	100	0	1.00
4	100	100	0
5	50	50	0
6	20	60	-0.40
Number of students per group = 100			

Distractor Analysis

One important element in the quality of a multiple choice item is the quality of the item's distractors. However, neither the item difficulty nor the item discrimination index considers the performance of the incorrect response options, or distractors. A distractor analysis addresses the performance of these incorrect response options. Just as the key, or correct response option, must be definitively correct, the distractors must be clearly incorrect (or clearly not the "best" option). In addition to being clearly incorrect, the distractors must also be plausible. That is, the distractors should seem likely or reasonable to an examinee who is not sufficiently knowledgeable in the content area. If a distractor appears so unlikely that almost no examinee will select it, it is not contributing to the performance of the item. In fact, the presence of one or more implausible distractors in a multiple choice item can make the item artificially far easier than it ought to be. In a simple approach to distractor analysis, the proportion of examinees who selected each of the response options is examined. For the key, this proportion is equivalent to the item p-value, or difficulty. If the proportions are summed across all of an item's response options they will add up to 1.0, or 100% of the examinees' selections. The proportion of examinees who select each of the distractors can be very informative. For example, it can reveal an item mis-key. Whenever the proportion of examinees who selected a distractor is greater than the proportion

of examinees who selected the key, the item should be examined to determine if it has been mis-keyed or double-keyed. A distractor analysis can also reveal an implausible distractor. In CRTs, where the item p-values are typically high, the proportions of examinees selecting all the distractors are, as a result, low. Nevertheless, if examinees consistently fail to select a given distractor, this may be evidence that the distractor is implausible or simply too easy.

Item Analysis for essay Tests

Step 1: identify the upper and lower 25% of the students.

Step 2: compute for the following:

$$\text{Disc.} = \frac{\text{(Sum of scores for highs} - \text{sum of score for lows)}}{N \times \text{(max. possible score on item)}}$$

$$\text{Diff.} = \frac{\text{(Sum of scores for highs} + \text{sum of score for lows)}}{2N \times \text{(max. possible score on item)}}$$

N = **25%** of the number tested

Practical example:

Item score	High group		Low group	
	No. of students	No. of students × score	No. of students	No. of students × score
10	9	90	1	10
8	6	48	0	0
6	2	12	4	24
4	3	12	7	28
2	0	0	8	16
total	20	162	20	78

So, the discrimination index =

$$\text{Disc.} = \frac{\text{(Sum of scores for highs} - \text{sum of score for lows)}}{N \times \text{(max. possible score on item)}}$$

$$= (162 - 78) / [(0.25 \times 80) \times 10] = 0.42$$

Satisfactory discrimination

and the difficulty index =

$$\text{Diff.} = \frac{\text{(Sum of scores for highs + sum of score for lows)}}{\text{2N x (max. possible score on item)}}$$

$$= (162 + 78) / [(2 \times 0.25 \times 80) \times 10] = 0.60$$

Satisfactory difficulty

Sample Problem

The mean score of a class on an essay is 16.5 out of a total maximum score of 20. What is the difficulty index of the essay?

The difficulty index in this case =

$$p = \frac{\text{mean score of the class}}{\text{maximum possible score}}$$

$$\rho = \frac{16.5}{20} = .825$$

Practice Exercises

(1) Assume that you are in the process of revising a series of items and secured the following sets of information. Review each solution, and decide first whether you should revise the item and, if so how you should revise the item.

a. An item on a norm-references aptitude test has a ρ value of 0.26 and a discrimination index of 0.10. Is it likely that the item should be revised?

b. Although judged incongruent with its test-item specifications by three independent reviewers, item 47 on a norm-references test of history has a ρ value

of 0.52 and a discrimination index of 0.49. Is it likely that item 47 should be altered?

c. Here is the distractor analysis for a multiple-choice item on a norm-references test of secondary school pupils' reading abilities. Review the data, then suggest what sorts of changes, if any, you would urge for the item?

Item 14	A	B	C*	D	E	Omit
(ρ =0.69), (D=0.15)						
Upper 27%	1	0	15	0	4	0
Lower 27%	3	0	12	3	1	1

* correct answer

……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……

……
……
……
……
……
……
……
……
……
……
……
……
……
……

Item Analysis Worksheet

Ten students have taken an objective assessment. The quiz contained 10 questions. In the table below, the students' scores have been listed from high to low (Joe, Dave, Sujie, Darrell, and Eliza are in the upper half). There are five students in the upper half and five students in the lower half. The number"1" indicates a correct answer on the question; a "0" indicates an incorrect answer.

Student Name	Total Score (%)	Questions									
		1	2	3	4	5	6	7	8	9	10
Ahmed	100	1	1	1	1	1	1	1	1	1	1
Nagy	90	1	1	1	1	1	1	1	1	0	1
Michael	80	1	1	0	1	1	1	1	1	0	0
Ebtsam	70	0	1	1	1	1	1	0	1	0	1
Eliza	70	1	1	1	0	1	1	1	0	0	1

Sara	60	1	1	1	0	1	1	0	1	0	0
Samah	60	0	1	1	0	1	1	0	1	0	1
Hannah	50	0	1	1	1	0	0	1	0	1	0
Ramy	40	1	1	1	0	1	0	0	0	0	1
Yara	30	0	1	0	0	0	1	0	0	1	0

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

……

....
....
....
....
....
....
....
....
....
....
....
....
....
....
....
....
....
....
....
....
....
....
....
....
....

Calculate the Difficulty Index (p) and the Discrimination Index (D) for each question.

Item	Correct (Upper group)	Correct (Lower group)	Difficulty (p)	Discrimination (D)
Question 1				
Question 2				
Question 3				

Question 4				
Question 5				
Question 6				
Question 7				
Question 8				
Question 9				
Question 10				

Answer the following questions:

1. Which question was the easiest?
2. Which question was the most difficult?
3. Which item has the poorest discrimination?
4. Which questions would you eliminate first (if any) – why?

……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……
……

……
……
……
……
……
……
……

Practice Exercise

Answers given by 370 subjects at 3 alternatives (A, B, C) of an item where B is the correct option. In rows are the frequency of subjects who selected each alternative and have received scores above and below 27% of their sample in the total test, and the group formed by the central 46%.

	A	B*	C
Upper 27%	19	53	28
Intermediate 46%	52	70	48
Lower 27%	65	19	16

Calculate the corrected index of difficulty and the discrimination index.

……
……
……
……
……
……
……
……
……

……
……
……
……
……
……
……
……
……

Sample Problem

In a Science test administered by Mr. Phlopateer, seven students answered word problem # 1 correctly. A total of twenty-five students took the test. What is the difficulty index for word problem # 1?

……
……
……
……
……
……
……

For Practice

(1) Compute p, the difficulty index, for the following items. Interpret your results. (The asterisk indicates the correct option)

options			
A	B*	C	D
10	5	8	0

options			
A	B	C*	D
4	2	16	3

……
……
……
……
……
……
……
……
……

(2) Compute D, the discrimination index, for the following items. Interpret your results:

(Class size = 40)				
	options			
	A	B	C	D*
Upper	3	0	7	10
Lower	5	4	9	2

……

(Class size = 30)				
	options			
	A*	B	C	D
Upper	3	5	7	0
Lower	8	5	1	1

……
……
……
……
……
……
……
……
……

(3) Identify which of the following items are likely to be miskeyed, which are likely to be susceptible to guessing, and which are probably ambiguous.

a.	Upper half	A*	B	C	D
		8	0	7	2
b.	Upper half	A	B	C*	D
		10	2	4	3
c.	Upper half	A	B*	C	D
		3	11	3	10
d.	Upper half	A*	B	C	D
		9	8	11	9
e.	Upper half	A*	B	C	D
		0	1	6	2

……
……
……
……
……
……
……
……
……
……
……
……
……
……

EXERCISE

Question 1

The administration of a test before the beginning of a learning period (formative pre-testing) has the following advantages *except one*:

A. To modify educational objectives of that period.
B. To provide ways for less welt prepared students to catch up.
C. To modify the required pass level (mark).
D. To provide a base from which to measure real progress.
E. To exclude weak students from the learning period.

Question 2

All the following stages, *except one*, are recommended for scoring tests of the "essay" type:

A. Write the elements of the answer for each of the questions asked.
B Correct the answers question by question rather than student by student.
C. Determine the pass score on the basis of a sample of answers.
D. Correct the answers while preserving the anonymity of the students.
E. Identify three levels only: honor, pass, fail.

Question 3

The *content validity* a written test is *usually* obtained by means of:

A. Collective and careful review of the questions.
B. Pearson's correlation coefficient.
C. Factor analysis.
D. An "inter-rater" reliability coefficient
E. A mean discrimination index.

Questions 4 to 6

A test with 50 questions is administered to a group of 45 students. There is a choice of five answers to every question. Only one of these choices is the correct answer. One point per correct answer is allocated in calculating the total score.

Question 4

Assuming that none of the students have any knowledge of the test subject (i.e. they choose their answers by guessing), which of the following will be closest to the mean score of the group?

A.0; B.5; C.10; D.15; E.25.

Question 5
On dividing this group of 45 students into 3 groups of 15 each, on the baste of the total score of each student it is found that, for the first question, nine students out of 15 in the high group and three out of 15 in the low group have given the right answer. For this question the *difficulty index* is:
A. 12%; B. 27%; C. 30%; D. 40%; E. 60%.

Question 6
Under the same conditions, the *discrimination index* is:
A. 0.12; B. 0.27; C. 0.30; D. 0.40; E. 0.60.

Question 7
On the basis of these indexes, which of the following decisions would you take concerning this question?
A. It should be discarded from the question bank.
B. It should be referred to a drafting committee for revision.
C. It should be retained in the bank as it is.
D. A decision other than A, B or C.

Questions 8 and 9
The following data concern a multiple-choice question set to 300 students, the correct answer being D.

	Choice of answers					
	A	**B**	**C**	**D**	**E**	**No answer**
High group (100)	22	1	10	67	0	0
Low group (100)	46	5	16	33	0	0

Question 8

These data show that:
A. half the students answered the question correctly;
B. all the distractors were of good quality;
C. the question was of high validity;
D. the question was not very relevant.

Question 9
In view of these data, the examination board may decide:
A. that this question should be reviewed since it is insufficiently discriminatory;
B. that this question should be discarded from the question bank;
C. that this question is of low validity;
D. none of the above.

Question 10
What could *generally* expected on doubling the length of a test whose mean discrimination index is 0.52 (by adding questions more or less equivalent to the previous ones)?
A. A certain increase in the reliability and the validity of the test.
B. Only a certain increase in the reliability of the test.
C. Only a certain increase in the validity of the test
D. A certain decrease in the reliability and validity of the test.
E. No effect on either the reliability or the validity of the test.

Questions 11 to 16
Use the following key in answering this series of six matching type questions:
A = traditional oral test
B = written test of the essay type
C = so-called written "objective" test (MCQ)
D = standardized practical test, or written and oral simulation tests (programmed examination)
Indicate the type of test *most suitable* evaluating each of the following performances:

Question 11
Recall of concepts.

Question 12
Ability to solve problems.

Question 13
Ability to communicate satisfactorily with the patient.

Question 14
Verbal expression.

Question 15
Skill in examining the patient

Question 16
Ability to make a synthesis.

Question 17
The system of "relative" criteria of competence implies the following consequences, *except one*. Which?
A. Leads to an embarrassing disagreement among those responsible for applying the resultant decisions.
B. Leads to the failure of certain students in a particularly competent group.
C. Enables one group to become the arbiter of the standards according to which it is judged.
D. Enables "low group" students, who are however superior to the mean of the whole group to which they belong, to pass.
E. Creates an arbitrary fluctuation in the desirable level of competence at a given moment.

Questions 18 and 19
The author of the following multiple-choice question was asked to establish its acceptability index.

The diameter of a normal erythrocyte (according to Wintrobe) expressed in μm (microns) is equal to:

A. 4.5 C. 7.5 D. 8.5
B. 6.5 E. 10.5

He felt that a student who “knew just enough to pass” should be able to reject right away choices A and E.

Question 18

Indicate which among the following values of the acceptability index corresponds to the author's choice:

A. 0.10 C. 0.25 D. 0.33

B. 0.20 E. 0.50

Question 19

If item C was not included, what then would be the acceptability index?

Question 20

According to the theories about absolute or relative criteria tests, all the following statements are correct *except one*. Indicate which is false:

A. The calculation of the discrimination index provides a statistical datum applicable to absolute criteria tests.

B. The calculation of the acceptable level of performance (ALP) of a test is applicable to criterion-referenced tests.

C. The acceptable level of performance (ALP) of a test is equal to the sum of the acceptability indexes of each question.

D. The value of the difficulty index influences the value of the discrimination index.

Reference

Allen, M. J. & Yen, W. M. (1979). Introduction to Measurement Theory. Lon Grove, Illinois: Waveland Press, INC.

Bean, K.L. (1953). Construction of Education & Personal Tests, McGraw-Hill Book Co., New York.

Bhagat, Pooja & Balia, J.N. (2016). Construction and Validation of Achievement test in Science. International Journal of Science and Research, 5(6) 2277-2280.

Biswas SS, Jain V, Agrawal V, Bindra M, Small group learning: effect on item analysis and accuracy of self-assessment of medical students, Educ Health (Abingdon). 2015;28(1):16-21.

Brown, F. G. (1983). Principles of Educational and Psychological Testing. New York: Holt, Rinehart and Winston.

Clemans WV. Multiple-choice Examinations in Medicine: A Guide for Examiner and Examinee. London: Lea & Fabiger, WHO Library. 1965;18:61.

Crisp GT, Palmer EJ. Engaging Academics with a Simplified Analysis of their Multiple - Choice Questions (MCQ) Assessment Results. J University Teach Learn Practice. 2007;4(2):88-106.

De Champlain AF, Melnick D, Scoles P, Subhiyah R, HoltzmanK, Swanson D, et al. Assessing medical students' clinical sciences knowledge in France: a collaboration between the NBME and a consortium of French medical schools. Acad Med. 2003;78:509-17.

Devi, S. & Sharma, H.L. (2013).Construction of an Achievement Test for the students of VIII class in the Subject of Mathematics. International Journal of Scientific Research, 2(7).

Eaves S, Erford B. The Gale group: The purpose of item analysis, item difficulty, discrimination index, characteristic curve. 2007 edition, Online Source: Available at: www.education.com/reference/ article/itemanalysis/. Accessed on 21 April 2017.

Ebel RL. (1979). Essential of Educational Measurement (3rd Edition). Englewood Cliffs, NJ: Prentice Hall.

Ebel, R. L. (1972). Essentials of Educational Measurement. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Ebel, R. L. (1965). Measuring educational achievement. Englewood Cliffs, New Jersey: Prentice-Hall.

Fowell SL, Southgate LJ, Bligh JG. Evaluating assessment: the missing link? Medic Education. 1999;33:276-81.

Gajjar S, Sharma R, Kumar P, Rana M. Item and test analysis to identify quality multiple choice questions (MCQS) from an assessment of medical students of Ahmedabad, Gujarat. Indian J Community Med. 2014;39 (1):17-20.

Gronlund, N. E. (1968). Constructing Achievement Tests. Englewood Cliffs, NJ: Prentice-Hall, Inc.

Gronlund, N. E. (1976). Measurement and evaluation in teaching. New York: Macmillan Publishing Co.

Guilbert JJ. Educational Hand-Book for health professionals, WHO offset Publication 35, 1 ed. Geneva: World Health Organization; 1981: 461-465.

Hawkes, H.F. (1936). The Construction & Use of Achievement Examinations, Houghton Mifflin, Boston.

Henning G. (1987). A guide to Language Testing-Development, Evaluation, Research. London: Newbury House Publisher.

Hingorjo MR, Jaleel F. Analysis of one-best MCQs: The difficulty index, discrimination index and distracter efficiency. J Pak Med Assoc. 2012;62:142-7.

Hubbard JP, Clemans WV. Multiple-choice Examinations in Medicine: A Guide for Examiner and Examinee; Lea & Fabiger, Libraries Australia; 1961: 186.

Jayanthi, J. (2014). Development and Validation of an Achievement Test in Mathematics. International Journal of Mathematics and Statistics Invention,2(4),40-46.

Kehoe J. Basic item analysis for multiple-choice tests. Practical Assessment, Research & Evaluation. 1995;4(10):20-4.

Kelley TL. The selection of upper and lower groups for the validation of test items. J Educ Psychol. 1939;30:17-24.

Kelley, T. L. (1939). The selection of upper and lower groups for the validation of test items. Journal of Educational Psychology, 30, 17-24.

Knight, P. The local practices of assessment. Assessment & Evaluation in Higher Education. 2006;31(4):435-52.

Kubiszyn, T. & Borich, G. (1990). Educational testing and measurement: classroom application and practice (3rd ed.). Glenview, Illinois: Scott, Foresman and company.

Kumar N. (2016). Construction and Standardization of an Achievement Test in English Grammar. International Journal of Current Research and Modern Education,1(2).

Matlock-Hetzel S. Basic concept in item and test analysis, Presented at annual meeting of the Southwest Educational Research Association, Austin; 1997: 1-27.

Maunder P. In support of multiple choice questions: some evidence from Curriculum 2000, Paper presented at the Annual Conference of the British Educational Research Association, University of Exeter, England, Education Line; 2002: 12-14.

Mehrens, W. A. & Lehmann, I. J. (1973). Measurement and evaluation in education and psychology. New York: Holt, Rinehart & Winston, Inc.

Micheels, W.J. (1950). Measuring Educational Achievement, McGraw-Hill Book Co., New York.

Mishra, A. (2017). Construction and Standardization of Commerce Achievement Test for Higher Secondary Level. IRA International Journal of Education & Multidisciplinary Studies,7(3),250-260.

Mukherjee P, Lahiri SK, Analysis of Multiple Choice Questions (MCQs): Item and Test Statistics from an assessment in a medical college of Kolkata, West Bengal, IOSR. J Dental Med Sci. 2015;14(12):47-52.

Nelson, C. H. (1970). Measurement and evaluation in the classroom. New York: Macmillan Publishing Co., Chapters 5-8. Measurement and Evaluation Division, 247 Armory Building. Especially useful for science instruction.

Norman G. Evaluation methods: A resource handbook. In: Shannon S, Norman, G, editors. Chapter 4.1. Multiple choice questions, The Program for Educational Development, McMaster University. Hamilton, Canada: McMaster University; 1995: 47-54.

Nunnally, J. C. (1964). Educational Measurement and Evaluation. New York: McGraw-Hill Book Company.

Oosterhof, A. (1990). Classroom Applications of Educational Measurements. Merrill, Columbus, OH.

Patil VC, Patil HV, Item Analysis of Medicine Multiple Choice Questions (MCQs) for Under Graduate (3rd year MBBS) Students. RJPBCS. 2015;6(3):1242

Payne, D. A. (1974). The assessment of learning. Lexington, Mass.: D.C. Heath and Co.

Peitzman SJ, Nieman LZ, Gracely EJ. Comparison of "fact-recall" with "higher-order" questions in multiple-choice examinations as predictors of clinical performance of medical students. Acad Med. 1990;65:59-60.

Pellegrino J, Chudowsky N, Glaser R, (Book) Knowing What Students Know: The Science and Design of Educational Assessment. Washington, DC: National Academic Press; 2001: 103-108.

Popham, W. J. (1981). Modern educational measurement: practitioner's perspective. Englewood Cliffs, N.J: Prentice-Hall.

Popham, W. J. (2014). Classroom assessment: What teachers need to know (7th ed.). Boston: Pearson.

Rani, R. & Anisha (2017). Construction and Standardization of Mathematics Achievement Test for IXth Grade Students. An International Journal of Education and Applied Social Science,8(2),651-655.

Ross MM, McDonald B, McGuinness J. The palliative care quiz for nursing (PCQN): the development of an instrument to measure nurses' knowledge of palliative care. J Adv Nurs. 1996;23:126-37.

Scannell, D. P. & Tracy, D. B. (1975). Testing and measurement in the classroom. New York: Houghton-Mifflin Co.

Sharma, H.L. & Poonam (2017).Construction and Standardization of an achievement test in English Grammar. International Journal of Advanced Educational Research,2(5),230-235

Sim S, Rasiah RI, Relationship Between Item Difficulty and Discrimination Indices in True/False Type Multiple Choice Questions of a Para-clinical Multidisciplinary Paper. Ann Acad Med Singapore. 2006;35:67-71.

Thorndike, R. L. (Ed.). (1971). Educational measurement (2nd ed.). Washington, D.C.: American Council on Education, Chapter 9 (Performance testing) and Chapter 10 (Essay exams).

Walker, H.M. (1943). Elementary Statistical Methods, Henry Holt & Co., New York.

Whitney, F.I. (1950). Elements of Research, Prentice-Hall, New York.

https://www.edglossary.org/assessment/

http://www.vkmaheshwari.com/WP/?p=181