EDUCATIONAL MEASUREMENT & EVALUATION

Code: 6507/8307/8259

Unit 1-9

Department of Secondary Teacher Education
ALLAMA IQBAL OPEN UNIVERSITY, ISLAMABAD

EDUCATIONAL MEASUREMENT & EVALUATION

Code: 6507/8307/8259

Unit 1-9

Department of Secondary Teacher Education
ALLAMA IQBAL OPEN UNIVERSITY, ISLAMABAD

Course Team

Chairperson : Pro. Dr. Rehana Masrur
Chairperson
Department of Secondary Teacher Education

Course Development Coordinator: S.M. Shahid

Writers:

S.M. Shahid
Muhammad Arshad Bandesha
Prof. Salamat Ali Dogar
Jamil Hussain Shah
Muhammad Arshad Javed

Reviewer:

Dr. Maqsood Alam Burkhari

Editor:

Umar Siddique Khattak

Course Coordinator: Dr Muhammad Tanveer Afzal

Contents

Foreword

Preface

Acknowledgement

Unit-1: Measurement and Evaluation: An Introduction

Unit-2: Qualities of Good Test

Unit-3: Integrating Objectives with Evaluation & Measurement:

Unit-4: Types of Tests

Unit-5: Methods of Measurement

Unit-6: Planning and Administering for Classroom testing

Unit-7: Marking and Reporting

Unit-8: School Testing Trends and Issues

Unit-9: Statistics Used in Evaluation and Measurement

FOREWORD

In any field measurement always involves three common steps: (1) Identifying and defining the quality attribute that is to be measured; (2) Determining a set of operations by which the attribute may be made manifest and perceivable, and (3) Establishing a set of procedures or definitions for translating observations into quantitative statements of degree or amount. An understanding of each of these steps and the difficulties that it presents provides a sound foundation for presenting the procedures and problems of measurement and evaluation.

Tests and measurements are of value to the extent that they permit, and are used for better decision making. If this end is kept always in view, a sound appraisal of testing as one means to that end seems possible.

Evaluation of the pupil is often regarded essential for the benefit of teachers and administrators. This attitude overlooks the direct contribution which evaluation can make in respect of the pupil. Properly used evaluation procedure can contribute directly to improve pupil learning by (1) Clarifying the nature of the intended learning outcomes (2) Providing short term goals to work towards measuring the students' outcomes. (3) Providing feedback concerning learning progress, (4) providing information for overcoming learning difficulties and for selecting future learning experiences. Although these purposes are probably best served by the periodic evaluation during instruction, the final evaluation of intended outcomes should also contribute to these ends.

Test and other procedures for measuring pupil learning are not intended as replacement for teachers' informal observations and judegments. Rather, they are intended to complement and supplement the teachers' informal methods of obtaining information about the pupil. The teacher is still the observer and observer and decision maker. Measurement and evaluation procedures merely provide more comprehensive, systematic, and objective evidence on which to base instructional decision...

Evaluation of pupil learning requires the use of a number techniques for measuring pupil achievements. However, evaluation is not merely a collection of techniques; it is rather a systematic process that plays a significant and effective role.

The Philosophy which guides the development of this text is that measurement and evaluation are not separate disciplines. Measurement is an integral part of the evaluation process. The purpose of this text is to present measurement concepts within the framework of contemporary evaluation.

In the end, I extend my heartfelt gratitude to the course team, Chairman, Dr. M. Zafar Iqbal and his associates for developing this course book despite time constraint. Any suggestions for the improvement of this course will be warmly welcomed.

May 2002

(Prof. Dr. S. Altaf Hussain)
Vice-Chancellor

PREFACE

Broadly conceived, the main purpose of classroom instruction is to help pupils achieve a set of intended learning objectives. These outcomes should typically include all desired changes in the intellectual, emotional, and physical spheres. When classroom instruction is viewed in this light, evaluation becomes an integral part of the teaching-learning process. The intended learning objectives are established by the instructional objectives, the desired changes in pupils are brought about by the planned learning activities, and the pupils learning progress is periodically evaluated by tests and other evaluation devices.

The instructional philosophy behind this course is that a comprehensive course in measurement and evaluation should be skill oriented rather than knowledge oriented. And application oriented rather than theory oriented. Thus, the purpose of this book is not simply to have students become familiar with evaluation procedures or acquire a body of knowledge. The purpose of this book is not to mystify students with theoretical and statistical jargon. The purpose of this book is not simply to have students acquire in knowledge needed to be a competent consumer and producer of measurement and evaluation tools and techniques. The emphasis is not on what the students know but rather on what the student can do with what he or she knows. Expertise involves more than the acquisition of skills and knowledge, through experience one acquires insight, intuition, and strategies related to measurement and evaluation. Experience has meaning however, only if it is related to a foundation of basic skills and knowledge.

Information from carefully developed tests and other evaluation techniques may also be used to improve instruction. Such information can aid in judging (1) the appropriateness and attainability of the instructional objectives, (2) the usefulness of the instructional materials, and (3) the effectiveness of the instructional methods. Thus, evaluation procedures can contribute to improvements in the teaching learning process itself, as well as contributing directly to improved pupil learning.

The current proliferation of "different" evaluation models has reinforced this assumption. All systems and models for evaluation involve the same essential components, namely: specification of goals and objectives: selection and/or development of measure-ement tools; delineation and execution of strategies for obtainment; and analysis and interpretation of results. Second, evaluation should not be an afterthought; it should be planned for at the beginning of an endeavor, not at the end. We never measure a thing or a person. Measurement is always of a quality or attribute of the thing or person.

(Dr. M. Zafar Iqbal)
Chairman

ACKNOWLEDGEMENT

I express my sincere appreciation to all the unit writers and reviewers specially the course coordinator Mr. S.M. Shahid, for constant revision and refinement of the content and organization of this text. I also wish to thank Editing cell, Academic Planning and Course Production (AP & CP) unit and print production unit (PPU) of AIOU for their expertise in organizing and printing of the book in this form.

April, 2002

(Dr. M. Zafar Iqbal)
Chairman

Unit-1

MEASUREMENT AND EVALUATION: AN INTRODUCTION

BY:
S.M. Shahid

INTRODUCTION

When people talk about the "falling standard of education" they can be assumed to be referring to forms of critical skills and knowledge which the young members of a society are expected to possess but , are unable to demonstrate after a period of formal instruction. Parents complain, that their children are unable to write their names or to write letters after primary education. Many secondary school leavers find it difficult to write or to express them selves well in Urdu. These problems point to the need to overhaul our educational programmes; but before doing so, we must be well acquainted with measurement and evaluation procedures through which reliable data about the status of our educational system can be objectively gathered. Through measurement and evaluation we know where we stand and what changes are to be made.

After a programme has been implemented, the teacher assesses his pupils' performance through locally made tests or public examinations. If desired results are obtained, efforts are made to maintain the programme under the best working conditions. If undesirable outcomes are observed, corrective or remedial measures are selected and applied. If all such efforts have been carried out, and the pupils still show poor performance, the programme is either modified or abandoned completely. Thus, testing or measurement and evaluation are inextricably linked to each other in the process of curriculum planning.

Testing is usually associated with student achievement relative to specified classroom objectives, Evaluation, on the other hand, is the measurement of the over-all success or worth of a programme. We shall attempt to discuss the concept, type and role

of Educational Evaluation, and the difference between and evaluation in this unit.

OBJECTIVES

After an intensive study of the unit, you are expected to be able to:

1. Describe the concept of Educational Evaluation and Measurement.
2. List the points of difference between Evaluation and Measurement.
3. Describe the major type of Evaluation
4. Explain the role of Evaluation and measurement in teaching process.

CONTENTS

Introduction

Objectives

1 The Nature Of Evaluation And Measurement

1.1 Concept of Evaluation

1.2 Concept of Measurement

1.3 Concept of Test

1.4 Relationship between Evaluation and Measurement

1.5 Summary

2 **Types Of Evaluation**

2.1 Programme Evaluation

2.2 Student Evaluation

2.3 Difference Between the Suimmative and Formative Evaluation

2.4 Summary

3 Need And Role Of Evaluation

3.1 Need for Evaluation in Education

3.2 Role of Evaluation and Measurement in Education Process

3.3 Summary

4 Self Assessment Questions

5 Bibliography

1. THE NATURE OF EVALUATION AND EASUREMENT

The purpose of evaluation is to make a judgement about the quality or worth of something-an educational programme, worker performance or proficiency or student attainments. That is what we attempt to do when we evaluate students' achievements, employees productivity, or prospective practitioners competencies, In each case the goal is not simply to describe what the students, employees or other personnel can do. Instead we seek answers to such questions as: How good is the level of achievement? How good is the performance? Have they learned enough? Is their work good enough? These are questions of value that require the exercise of judgement. To say simply that evaluation is the process of making value judgements understates the complesity and difficulty of the effort required. Once it has been determined that evaluation is needed, the evaluator must decide what kind of information is needed, how the information should be gathered, and how the information should be synthesized to support the out come-the value judgement. Thus, evaluation is as concerned with information gathering as it is with making decisions. In addition, the term is used to refer to the product or outcome of the process. That is, we might, for example, submit our evaluation (the product) of public's school performance to his parents, following our evaluation (the process) of his accomplishments. In this respect evaluation has a dual connotation.

1.1 Concept of Evaluation

Educational Evaluation is broader in scope and more objective than measurement. It is the process of carefully appraising the individual from a variety of information giving device. Besides testing and other tools of measurement, evaluation seeks additional evidences from various sources of information

supplementing each other; like interviews, questionnaires, anecdotal records, cumulative records, case conferences, mechanical or electronic recorders, case studies, or projective techniques, etc; and the selection, through careful analysis of data, most pertinent to a wise just and comprehensive interpretation to make value judgement of the individual, or group under study.

Evaluation is based two philosophies. One, traditional philosophy is that ability to learn is randomly distributed in the general population. It means that if some learning task is assigned to a class and then a test is administered to study their performance. The result of the test shows that some students' score is very high and some students' score is low and majority of the students, score falls between these two extremes. It was the opinion of old educators that all are not endowed with same intellectual abilities to benefit from schooling. Generally, teachers weeded out students who tended to learn less well than their peers. This was the old philosophy based on the superiority of heredity.

This gave birth to norm-referenced measurement of intellectual abilities. It has been used in schools to differentiate among individuals of some defined group or whatever is being measured. In norm-referenced measurement, an individual's score is interpreted by comparing the score to those of a defined group, often called the normative group. The comparison is relative rather than absolute. The Philosophy of measurement has recently emerged. The new philosophy of measurement is based on democratic values and gives importance to the environment. It is based on the univer-salisation of education. It assumes that if education is thought universal, the responsibility of the teacher is to help as many students as possible to learn. It has discarded the selection philo-sophy of norm-referenced measurement. All

individuals can attain mastery of a learning task, provided they are given opportunities and time. It assumes that with property developed instructional sequence every child could reach 100 per cent mastery of any objective. It suggests that an absolute standard be used as reference for evaluation. These standards are the objectives specified for instruction. Each student's status is determined by how he achieves and satisfies its objectives for example, before a unit begins, the teacher may have decided that three objectives were essential for every student. A student has to satisfy each in order to receive a passing grade.

Thus we see that the two philosophies of evaluation are based on different concepts of human potentialities and their development. One believes that human abilities are not evenly distributed in the population. Achievement of individual learner differs greatly whereas the other believes that all learners can attain the mastery of learning task irrespective of individual differences among them.

William Wiersma and Stephen G Jurs (1990) remarks that Evaluation is a process that includes measurement and possibly testing but it also contains the notion of value judgement. If a teacher administers a test to a class and computes the percentage of correct responses, it is said that measurement and testing has taken place. The scores must be interpreted which may mean converting them to values like As Bs Cs and so on or judging them to be excellent, good, fair or poor. This process is called evaluation.

So we can say, evaluation is concerned with making judgements about things. When we act as evaluators, we attribute 'value' or 'worth' to behavior, objects and processes. In the wider community, for example, one may make evaluative comments about a play, clothes, a restaurant, a book or someone's behaviour.

We may enjoy a play; admire someone's clothes, rave about a restaurant and so on and so forth. Invariably these are rather simple, straightforward comments of value or worth.

According to William Wiersma and Stephen. G. Jurs (1990). To be more effective, however, evaluation requires that judgements be based on appropriate and relevant data. Ineffective evaluation is made upon whim or fancy, even in the broader community context. To say, for example, that a film was 'good' or 'bad' says little unless the basis of these judgements is made. An enjoyable or good film may have a well-written script tight direction, mood-enhancing music and so forth. These are characteristics of the evaluation upon which judgement can be made subsequently.

Norman E. Gronlund (1990) defines this term as follows:

> Evaluation is a systematic process of collecting analyzing and interpreting information to determined extent the pupils are achieving instructional objectives. (Answers) the question "How good"?)

In the light of above discussion, Evaluation in our schools is essentially concerned with two major approaches to making judgements.

1. Product evaluation is an evaluation of student performance in a specific learning context. Such an evaluation essentially seeks to determine how well the students have achieved the stated objectives of the learning situation. In this sense the student's performance is seen as a product of the educational experience. A school report is an example of product evaluation.

2. Process evaluation examines the experiences and activities coevolved in the learning situation i.e. making judgements about the process by which students acquired learning or examining the learning experience before it has been concluded. In most cases, process evaluation is used when making judgements about school effectiveness, classroom interactions, and the curriculum and the effective-ness of specific programmes. For example, process evaluation may be conducted upon the nature of student-teacher interaction, instructional methods, school curricula, and a programme for gifted students, and so forth.

Robert L. Ebel and David A. Frisible (1986) Observe:

The difference between 'product' and 'process' evaluations is some thing of a fine line. Students usually pass through a school, experience a curriculum and then depart. In that sense we can refer to product, just as we can refer to a student's progress as 'the proof is in the product' curriculum evaluations in the activity involved rarely comes to a conclusion in schools, i.e. the curriculum is ongoing. How-ever, if a curriculum or a particular programme had been terminated, then a form of product evaluation would be con-ducted.

The subcategories of process evaluation are frequently referred to in the literature: curriculum evaluation, teacher evaluation and programme evaluation.

In this unit we shall concentrate upon the evaluation of student performance, as this is a fundamental task of any teacher:-

(a) According to L.R. Hay, (1985, p-6):

(i) Evaluation is the systematic process of collecting and analyzing data in order to determine whether, and to what degree, objectives have been or are being achieved.

(ii) Evaluation is the systematic process of collecting and analyzing data in order to make decisions.

A systematic process or data collection, that is measurement and the analysis of collected data, is common to both definitions, although some definitions seem to equate measurement with evaluation, most recognize that measurement is one of the essential components of evaluation.

The basic difference between the two definitions is the issue of decisions, or judgements, whether they are an integral component of evaluation or not. Proponents of definition (i) agree that the results of evaluation may be used for decision-making: proponents of definitions (ii) consider decision making to be a part of evaluation. For two major reasons, the second definition would seem to be preferable. First of all definition I is more inclusive. Second the notion that evaluation can be conducted for strictly descriptive purposes of evaluation implies, is naive at best perhaps ideally the sole purpose of evaluation, should be to provide feedback in order to improve the object of the evaluation, as the first definition (if any) between where we are and where we would like to be.

b. Evaluation has been broadly defined by Stifle Beam 1971 As:- The process of delineating, obtaining and providing useful information for judging decision alternatives (p-15)

c. **According to Thorndike and Hagen (1977, pp-1-5),**

Measurement provides only information such as a test score and not the judgement of insight that is required for reaching a sound conclusion or plan of action. The judgement of insight is to considered as the set of evaluative procedure used to interpret information into an appraisal.

d. **Ebel (1979 clarified the difference between measurement and evaluation as described follow:**

An evaluation is a judgement of merit, sometimes based solely on measurements such as those provided by test scores but more frequently involving the synthesis of various measurement critical incidents, subjective impressions, and other kinds of evidence. (p-376)

e. In practice, evaluation is specific in terms of function and each type of evaluation uses this general definition in a special way. Common to all evaluation is the use of adequate information to make judgement about someone or something.

f. Each of these interpretations point out that judgement and introspection are necessary when evaluating. One should clearly understand that evaluation goes beyond measurement measuring and measured. To formalize the definition.

Evaluation is the continuous inspection of all available information concerning the student, teacher, educational programme and the teaching-learning process to ascertain the

degree of change in students and form valid judgement about the students and the effectiveness of the programme.

Value judgement on an observation, performance test or any data whether directly measured or inferred is called evaluation.

1.2 Concept of Measurement

The term "Educational Measurement" refers to any device for the general study and practice of testing, scaling, and appraising the outcomes of educational process. It includes administration and scoring or tests, scale construction, validation and standardization, and application of statistical techniques in the interpretation of obtained measures or test results.

Measurement is the process of assigning numbers to individuals or their characteristics according to specified rules. Measurement requires the use of numbers but does not require that value judgements be made about the numbers obtained from the process. We measure achievement with a test by counting the number of test items a student answers correctly, and we use exactly the same rule to assign a number to the achievement of each student in the class. Measurements are useful for describing the amount of certain abilities that individuals have. For that reason, they represent useful information for the evaluation process. But can we measure all the important outcomes of our instructional efforts.

Education is an extensive, diverse, and complex enterprise, not only in terms of the achievements it seeks to develop, but also in terms of the means by which it seeks to develop them. Our understanding of the nature and process of education is far from perfect. Hence it is easy to agree that we do not know how to measure all-important educational outcomes. But, in principle, all-important outcomes of education are measurable. They may not

even be measurable in principle, using only paper and pencil tests. But if they are known to be important, they must be measurable.

To be important, an outcome of education must make an observable difference. That is, at some time, under some circumstances, a person who has more of it must behave differently from a person who has less of it. If different degrees or amounts of an educational achievement never make any observable difference, what evidence can be found to show that it is in fact important? But if such differences can be observed, then the achievement is measurable, for all that measurement requires is verifiable observation of a more-less relationship. Can integrity be measured? It can if verifiable differences in integrity can be observed among individuals. Can mother love be measured? If observers can agree that a hen shows more mother love than a female trout, or that Mrs. "A" shows more love for her children than Mrs. "B" then mother love can be measured.

The argument, then, is this: To be important an educational outcome must make a difference. If it makes a difference, the basis for measurement exists. To say that Asma shows more "Spunk" than Omer, may not seem like much of a measurement. Where are the numbers? Yet out of a series of such more-less comparisons, a scale for measuring people's spunk can be constructed. The Ayres, scale for measuring the quality of handwriting is familiar example of this (Ayres, 1912). If a sequence of numbers is assigned to the sequence of steps or intervals that make up the scale, than the scale can yield quantitative measurements. If used carefully by a skilled judge, it yields measurements that are reasonably objective (that is, free from errors associated with the use of a particular set of test items or tasks).

Are some outcomes of education essentially qualitative rather than quantitative? If so, is it reasonable to expect that these qualitative outcomes can be measured? This person is a man; that one is a woman. This speaks only Punjabi; that one speaks only Urdu. But we can express these qualitative differences in quantitative terms, too. This person has more of the characteristics of a man; that one has less. This person has more eye-blueness; that one has less. This Person has more ability to speak Punjabi; that one has less.

We may think of the weight of a man, his age, or the size of his bank account as quantities, while regarding his health, his friendliness, or his honesty as qualities. And if they serve to differentiate him from other men because he exhibits more or less of them than other men, they become quantitative qualities. It is difficult to think of any quality that interests us that cannot also be quantified, "Whatever exists at all exists in some amount," said E.L. Thorndike (1918, p-16). And William A. Me Call (1939) has added, "Anything that exists in amount can be measured" (p-18).

William Wiersma and Stephen G. Jurs (1990) define easure-ement.

Measurement: For all practical purposes assessment and measurement can be considered synonymous. When assessment is taking place, information or data are being collected and measurement is being conducted.

Measurement could also involve of data about teacher performance or about the performance of a curriculum. However, regardless of what is being measured, the data obtained has little value in itself and require interpretation by someone skilled in evaluation procedures. Indeed, measurement of data in the hands of unskilled persons may be grossly misinterpreted. For example,

what does a student's score of 12/20 on a test indicate? By itself, it means very little and it requires interpretation before it is considered meaningful. It could mean that on that test, the student has performed quite poorly as the mean score on the test was 15/20 or perhaps it means that the student has performed quite well as the median score was 8/20. Thus the score by itself has little meaning and it requires interpretation through the use of assessment procedures.

(a) L.R. Gay says (1985, p-8):-

Measurement is the process of quantifying the degree to which someone or something possesses a given trait, i.e. quality. Characteristics or feature. Measurement permits more objective description concerning traits and facilitates compare-sons. Thus instead of saying that Aslam is underweight for his age and height, we can say that Aslam is 18years old, 5' 8" tall, and weight only 85 pounds. Further, instead of saying that Aslam is more intelligent than Ali, we can say that Aslam has a measured IQ of 125 and Ali has a measured IQ of 88. In each case, the numerical statement is more precise, more objective, and less open to interpretation than the corresponding verbal statement.

(b) Norman E. Gronlund (1990, p-5) states in his book. "Measurement and Evaluation in Teaching"

Measurement is the systematic ascertaining of a characteristic property or attribute through a numerical device. The device may be an inventory, a checklist, questionnaire, scale or test. Measurement is limited to quantitative descriptions of behaviour and does not include qualitative descriptions or judgement of the desirability of

the behaviour being measured. In this respect measurement differs from evaluation.

(c) **Robert I. Thorndike and Elizabeth P. Hagen (1977)** pointed out that three steps are involved in developing a measurement device. First we must identify and define the quality or attribute that is to be measured. We never measure a person, only a quality or attribute of the person like intelligence or emotional maturity. Similarly. We do not measure a table but the temperature of the fire: not the automobile tire but the durability of the tire. Having identified the quality or attribute that interests us, we need to define it. For instance, if we are concerned with the durability of a tire, do we mean its resistance to puncture, its endurance against road wear, or its ability to hold up against deterioration.

(d) **According to Robert L. Thorndike and Elizabeth P. Hagen (1977, p-137).**

The first step in developing a measurement device is to devise a set of operations to isolate the attribute and make it apparent to us. Take the durability of an automobile tire. Once we have identified and defined the attribute that interests us, we need to develop some standard to allow us to gauge or index it. If our concern is with the tire's resistance to roadway abrasion, we need to develop a procedure for ascertaining the rate at which the rubber wears away.

Similarly, various educators and psychologists have developed the Stanford-Binet and other tests that include operations for eliciting behaviour that we take to be indicative of intelligence. But as Thorndike and Gagen noticed the fact that

there is no single universally accepted test, and that different test vary somewhat in the tasks they include and in the order in which they rank people. This is evidence that we do not have complete consensus as to what intelligence is on the on hand, or what are the appropriate procedures for eliciting it on the other.

The second step in measurement is to express the result of the operations established in the second step in numerical or quantitative terms. This involves an answer to the question. How many or how much? For example, we may employ millimeters as the units for indicating the thickness of the tread on the face of the tire and hence express the amount of wear on the tire in terms of millimeters.

Similarly, educators and psychologists require numerical units for gauging anxiety, emotional maturity, intelligence, and other attributes. In the case of intelligence, they may have individuals, perform a number of tasks and count the total number of successes which they then convert into IQ units.

Clearly each step in measurement rests on human-fashioned definitions. In the first step, we define the attribute that interests us in the second step; we define the set of operations that will allow us to identify the attribute. And in the third step we define the units in which we will state the results of our operations. Thus what is measured is always function of our definitions and they have their own inherent limitations.

1.3 Concept Of Test

In education a test consists of a question or series of questions or exercises or other devices for measuring the mental ability, capacity, skill, knowledge, achievement, progress, aptitude, attitude, interest, social and emotional adjustment or personality, etc, or an individual or group.

Tests represent one particular measurement teaching. A test is a set of question each or which has a correct answer that examinees usually answer orally or in writing. Test questions differ form those used in measures of attitudes, interest or preference, or certain other aspects of personality. Ideally, the questions in tests of achievement or many tests of intelligence have answer that content experts can agree are correct; correctness is not determined by the particular values, preferences, or dislikes of a group of judges.

All tests are a subset of the quantitative tools or techniques that are classified as measurements. And all measurement techniques are a subset of the quantitative and qualitative techniques used in evaluation. A major concern in this text, but certainly not the only one, will be with the development of tests that can contribute to summative evaluation of student learning. Other measurement and evaluation techniques are useful for other evaluation purposes, but test that measure relevant school learning with precision are the most useful tools available to teachers for most classroom summative evaluation needs.

1.4 Relationship between Evaluation and Measurement

(A) Stevens (1951) explains the difference between evaluation and measurement in these words. In its broadest sense, measurement is the assignment of numerals to objects, or events, according to rules. We measure height and weight following certain rules and then assign some numerical value to the measurements. We do not assign numbers in all cases of measurement, especially when using criterion-referenced measuring instruments. Here the symbols assigned may be equivalent to (+) or (-) since the measuring instrument

set a single standard and the individual either meets or fails to meet the absolute standard set by the objective. When evaluating data we go beyond the concept of measurement and make a judgement about the measurements taken".

(B) Dubois, Alverson and Staley (1979) explain the distinction between evaluation and measurement in these words, "As with any assessment process, the evaluation of entering behaviour involves the collection and evaluation of data. Psychologists working in the field of tests and measurements use the term measurements to refer to the collection portion of the process.

(C) A Dictionary of Education (1981) explains the concept measurement as "Fundamentally we can say that measurement entails certain rules and procedures for assigning numbers to attributes in such a way that the numbers represent the quantity of the attribute. It is necessary to be clear that it is not the object, organism or event itself which is being measured. For example, we don't measure a 'piece of wood' but we measure one of its attributes such as its length or weight. In educational measurement we are faced with attributes that do not lend themselves to such intuitive procedures (as used in physical sciences).

(D) In the words of Lester O. Cron and Others, "Evaluation is a broader term than measurement. Evaluation not only is concerned with the determination of learning results but it also involves value judgement of the desirability of these results. It is a continuous process

in which various techniques of testing or measurement can be utilized. Evaluation is a cooperative activity in which the principal, the teacher, the pupils and the parents participate."

(E) H.H Remmen and N. L Gage point out, "It is the felt need that has caused the shift form the term 'measurement' implying mathematically precise mensuration of know-ledge –to the term 'evaluation' which widens the areas to be studied to include subjective opinions and qualitative changes as well as objective and quantitative changes to include changes in attitudes, appreciation and understandings as well as acquisitions of knowledge and skills."

(F) Prof. Adediran A Taiwo (1995) distinguishes measurement and evaluation in these words, "While measurement is concerned with only the amount, quantity or frequency of a variable, evaluation matches such an amount, quantity or frequency with relevant criteria for the purpose of making some value judgement about the measured or the observed amount.....In essence, the term evaluation involves both quantitative and qualitative description of events, behaviours, things, parameters, variables as well as value judgement of things or events being described. It therefore, follows that any time one talks of evaluation, be it in the realm of achievements of students, the effectiveness of a teaching method or the appropriateness of a curriculum, one is concerned with both numerical and verbal description as well as value judgement of what is being described."

(G) Evaluation is a comprehensive and continuous process, which covers every aspect of an individual's achievement in the educative programme. It is an integral part of education in which students and teachers are partners. It signifies a wider process of judging students,progress, in various aspects. Measurement, on the other hand, implies only a precise quantitative assessment of instructtional outcomes.

Evaluation is integrated with the entire task of education and not only with examinations, tests and measurement.

Evaluation encompasses tests and measurement but also goes beyond them.

Evaluation depends upon measurement but is not synonymous with it.

Measurement is a quantitative determination of how much an individual's performance has been, while evaluation is a qualitative judgement of how good or how satisfactory an individual's perfomance has been.

Measurement describes a situation; evaluation judges its worth or value.

Measurement is only a tool to be used in evaluation. By itself, it is meaningless, but without it evaluation is likely to be of little significance.

Sound evaluation is based upon the results of accurate and relevant measurement. It is also to be remembered that not all uses of a test or measurement in education can be considered evaluation, for evaluation is always in the light of some particular goal, purpose or value.

Evaluation is not only quantitative but also qualitative and includes value judgement. Mathematically it may be said that:

Evaluation=Measurement (quantitative description of students' achievements) +Qualitative description of students' abilities +value judgement about students' achievements and abilities.

(H) The difference between evaluation and measurement may be explained with the help of following examples:

(1) A teacher measures Aslam's height to be 180 cm. He evaluates his height when he says that he is 'long'.

(2) A teacher measures Ali's achievement in Economics to be 50%. He evaluates his achievement when he says that Ali's achievement in Economics is 'satisfactory'.

(3) A teacher measures the size of a classroom and finds that it is 4mx3m. He evaluates the classroom dimensions when he reports that the classroom is 'too small' for 40 students.

(4) Aslam and Ali study in the same class. In the first test they obtain 50 and 70 marks respectively in English. In the second test, both of them obtain 80 marks. Now in the second measurement (test scores), achievement in English is the same, yet the evaluation will differ. When the teacher states that the rate of progress of Aslam is comparatively better than that of Ali.

Measurement helps in evaluation. This may be clarified by taking one example. Aslam and Ali study in the same class. They take two tests. In the first test, they obtain 45 and 65marks respectively in Civics. In the second test, both of them obtain 80 marks. Now, in the second test, the measurement (test scores) of their achievement

in Civics is the same, yet the evaluation will differ, when the teacher says that the rate of progress of Aslam is comparatively better than that of Ali.

1.5 Summary

Measurement is principally concerned with quantitative descriptions of student achievement. Unlike evaluation, it does not imply judgements about the worth of an educational programme. Measurement involves the assigning of numbers that represent the amount of a property possessed (that is, value) by an object or system. Scales associated with measurement include nominal, ordinal, interval and absolute.

Testing is measuring device concerned with specific achievement of a student in terms of given objectives. Evaluation, on the other hand, deals with finding out as far as possible the worth of a process, system or programme. When, on the basis of test results, a teacher decides on what should be done to improve the outcomes of instruction, he is assuming the role of an evaluator. Thus there is a continuous interplay between testing or measurement and evaluation.

Measurement and evaluation play an important role in the instructional prgramme of the school. Basically, they provide information that can be used in a variety of educational decisions. The main emphasis in classroom evaluation, however, is on decisions concerning pupil learning and development.

From an instruction standpoint, evaluation may be defined as a systematic process of determining the extent to which instructional objectives (i.e., intended learning outcomes) are achieved by pupils. The evaluation process includes both measurement procedures (e.g., test) and nonmeasurement procedures (e.g., informal observation) for describing changes in

pupil performance as well as value judgement concerning the desirability of the changes.

The process of evaluation is likely to be most effective when guided by a set of general principles. These principles emphasize the importance of (1) clearly specifying what is to be evaluated, (2) selecting evaluation techniques in terms of their relevance, (3) using a variety of evaluation techniques, (4) being aware of their limitations, and (5) regarding evaluation as mean to an end, and not an end in itself.

2- TYPES OF EVALUATION

Basically, evaluation can be carried out at two main levels; programme and student. At pointed out in section I, evaluation is concerned with the overall effectiveness of a programme. Thus the different types of tests discussed in unit 4 form a part of student evaluation. Evaluation can further be subdivided into formative and summative evaluation, and we shall discuss these later.

2.1 Programme Evaluation

In order to determine whether any programme has or has not been successfully implemented, certain salient questions must be posted and answered. These questions include the following, among others.

(i) Is the content of the programme is of desirable quality?

(ii) Is there a positive relationship between actual learning outcomes and intended learning? Are intended learning outcomes achieved?

(iii) Are unintended learning outcomes identified and corrections made for them?

(iv) Is the content relevant to the needs of the students? Is the content relatively simple and comprehensible, and is it able to be extended and generalized to situations within and outside the school.

(v) Are casual and functional relationships identified and analysed?

(vi) Does the implemented programme continue to be effective?

(vii) Are the materials available locally or obtainable from other areas with relative ease and convenience and at a reasonable cost?

(viii) Are necessary supportive or maintenance services provided?

Not until these and other relevant questions are answered can the success or failure of the prgramme be objectively determined. It is at the implementation and evaluation stages (see Figure1) that these questions are asked. It is an economic reality that a low-quality product cannot remain for long on the competitive marker no matter how efficient the maintenance service provided. Thus the quality of the products of any programme cannot supersede its design.

When our concern is in judging the compatibility between the aims and the learning outcomes of a programme, the emphasis is on the efficacy of that programme. On the other hand, a 'good' prog-ramme may be badly implemented. The task of quality control is to maintain and maximize the efficiency of a programme.

The quality content of a programme is determined, among other factors, by i) its conceptual quality; ii) logical relevance to the needs of the students (consumers); iii) simplicity and compre-

hensibility in terms of readability and literacy level of the content; iv) relative stability and survival value in the literature; and v) appli-cability to familiar and novel situations. No Matter how good a programme may be; the maintenance system must be well facilitated. The school administrators, head of subject unit, supervisors, the teacher and the pupils must be actively involved if successful implementation of the programme is to be realized. The teacher, being the main executor of the programme must be well trained not just to be able to teach facts but to select facts that relate to other facts and principles. The teacher education programmes in the advanced teachers colleges and the universities must prepare teachers to be able to teach their subjects effectively.

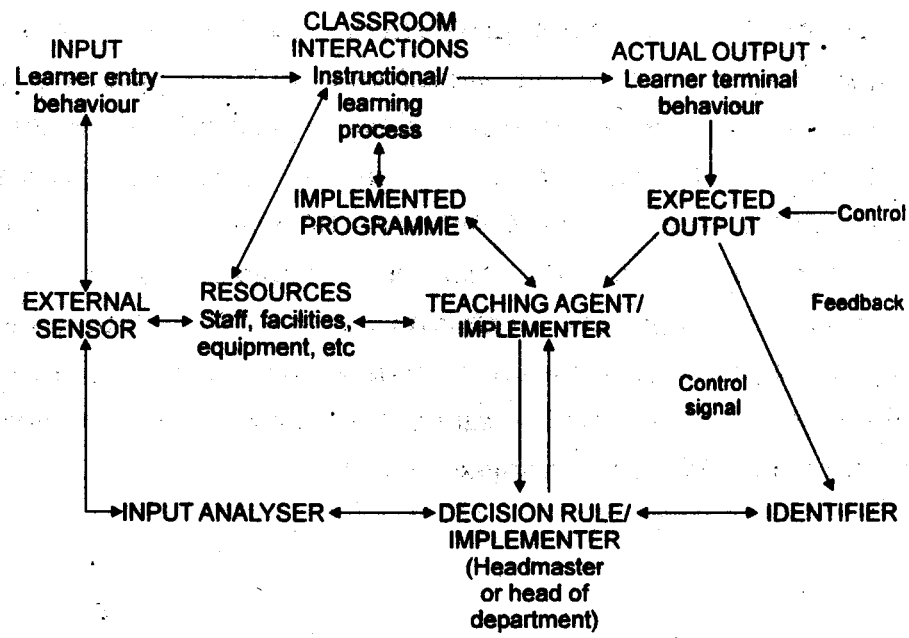


Figure i: A scheme of the implementation-maintenance process of curricular programme.

In order to be implemented a programme should be designed in such a way that under favourable conditions certain intended learning outcomes will emerge. The school teacher, the headmaster

and supervisors must gather information from time to time in order to determine the success or weakness of the programme. If desirable outcomes are observed, the focus of all concerned with instruction should be to improve the programme through an effective mainte-nance system (see Figure II). If the products (students) produced are of poor quality, corrective measures (for example, tutorials, use of audio-visual materials, limited or extensive modify-cations of the programme) are selected and applied in order to achieve the desired results. If after all these efforts, the products are still found to be poor, the programme is usually abandoned.

In order to be an efficient 'maintenance engineer' the teacher must not only have relevant professional qualifications but must also be enthusiastic and well acquainted with the programme for the design to the evaluation stage. If he is not actively involved at all levels of the planning process, a gap of knowledge is created, thus paving the way to a wide disparity between the intentions of the program and what the teacher actually delivers in the classrooms. Granting that conditions are normal, it must not be assumed that the implementation- maintenance process within the school setting is as simple as figure II a show. As can be seen in figure the process in very complex indeed.

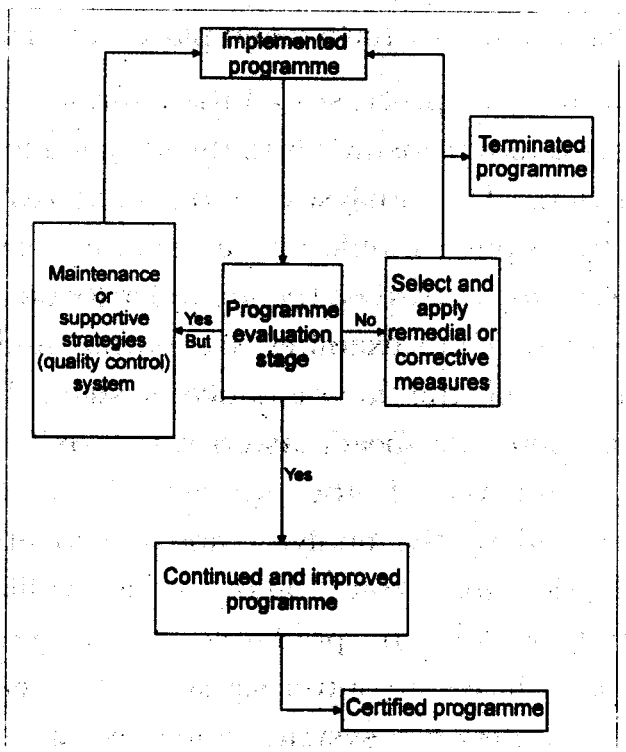


Figure ii: A scheme of the basic maintenance system within the school setting.

Several processes are involved in the input – output process. The teacher (or the teaching agent) is the most important component of the maintenance process of the programme. He interacts with the students, with other staff, experts and administrators and forms a bridge between them and learning materials. Often he acts as the input analyzer and an identifier (although these are primarily the job of curriculum experts) as well as the teaching agent of the programme. The external sensor (usually the supervisor or curriculum expert) examines the learning environment to identify changes-perhaps economic, political, and psychological or social-within the environment that

can destabilize the system. In addition he scans the input system to identify changes that may have taken place in learner.

The input analyzer processes all the information supplied by the external sensor and transmits it to the school administrator for appropriate action. He analyses and organizes information obtained from the input variable into a comprehensible structure to be used in planning activities; the identifier (usually the teacher or his head of department) examines the output and the internal working conditions of the maintenance system. It is he who provides the decision rule (headmaster) with a reliable picture of the internal conditions of the system. The input -output information provided by the analyzer and the identifier becomes the input of decision rule, and it is utilized by the headmaster/implementer to produce a decision, policy or instruction (sent in form of control signal) to the teacher. One of the aims of the maintenance system relates as far as possible the actual pupil performance and the expected output. The closer the relationship, the better the maintenance system. Thus the quality control system is to be construed as a dynamic, on-going and self-regulating process, sensitive to changes and with the goal of producing the best learning outcomes of instruction.

Factors militating against successful implementation of any programme (and particularly of many of the new programmes) are factors such as I) poor teacher preparation; ii) lack of teaching experience by the young teachers usually assigned to teach the programme; iii) too many activities to be performed and with the materials so diverse as to make a through completion of the program impossible; iv) overemphasis on process to an almost total exclusion of products of knowledge and vice versa; v) the dependence of a program on in-service teacher training course; and vi) the tendency to prepare materials that appear to favour

average and above average students at the expense of slow learners.

Too often data gathered about outcomes of programme deal mainly with incremental knowledge which relates only to learner's cognitive developments whereas the programme was designed to produce broader behavioral outcomes. Besides, although it is one thing to acquire information, it is another to behave in line with the information acquired (Romberg, 1975). Whatever evaluation techni-ques we employ, cognizance must be taken of unintended outcomes. Such outcomes must be identified and corrected. This is where quality control becomes important as a maintenance process. Similarly, Parlett and Hamilton's 'illuminative evaluation' has an advantage over the usual paper- pencil summative evaluation techni-ques. The aim of illuminative evaluation is to identify how a curri-culum operates, how it is influenced by divers school setting in which it is applied, what those directly involved regard as its merits and demrits and how pupil' intellectual tasks and academic experience are mostly affected. Thus, parlett and hamiltion (1972) contend that;

If and evaluation hinges on the supposed perpetuation of the instrumental system in more or less its original form, it makes an arbitrary and artificial distinction; it treats the innovation as a self-contained and independent system, which in practice is manifestly not.

Any given programme introduced into the school setting is not left in its naked form but assumes a different from for that setting. Its contents are emphasized or minimized, explicated or attenuated as teachers, administrators and students see fit for their particular setting.

If it is one thing to design a programme, it is quite a different matter to make it work. To a large extent, cluster such as the society, the school environment teacher and leaner characteristics, structure of the discipline, economic reality, political leadership, instructional techniques availability of materials and so on will influence the success of failure of the programme. There is nothing wrong in substituting one programme for another, but an uncontrolled passion to accept anything regarded as 'new' before its relevance is determined is unjustifiable. The unparalleled enthusiasm with which we embrace any new curricular package is symptomatic of what care (1976) call 'innovationists' - the problem of coping with one curricular package after another before-the last has been will understood. Carre was not against the implementation of new science programmes, but held the view that successful implementation of such program required careful planning and execution.

Programme evaluation can be carried out through the use of surveys, interviews, experimental studies and so on, these topics will be discussed in unit No,5.

2.2 Student evaluation

As pointed out earlier, testing forms an integral part of student evaluation. The purpose of this type of evaluation is to determine how well a student is performing in a programme. Through a series of oral questions, paper-pencil tests, manipulative skill tests, Tutorials, discussions, Tutorials, individualized instruction, assignments, projects and so on, the student is gradually guided towards a desired goal. Basically there are two types of students evaluation: formative and summative. The former is guidance-oriented, while the latter is judgemental in nature. We shall concentrate on the former in this section.

2.2.1 Formative Evaluation

Formative evaluation aims at ensuring a healthy acquisition and development of knowledge and skills by students. Formative evaluation is also used to identify student needs in order to guide them towards desired goals. As student needs and difficulties are identified, appropriate remedial measures are taken to solve such problem. The purpose is to find out whether after learning experience students are able to do what they were previously unable to do. A short-term objective of formative evaluation may be to help students pass the end-of-year promotional examination or, long term, the school certificate examination. Whatever activities are set into motion under these types of evaluation, the ultimate goal is to help students perform well at the end of the programme. As Yoloye (1976) puts it, the primary purpose of formative evaluation is to help as much as possible to ensure that summative evaluation comes out positive'. It is a process of channeling input variables through a process that will yield expected outputs. The classroom teacher is variables through a process that will yield expected outputs. The classroom teacher is the best formative evaluator. Because of his involvement with his students he is able to:

1. Draw a more reliable inference about his students than an external assessor although he may not be as objective as the latter;
2. Identify the levels of cognitive process of his students;
3. Choose the most suitable teaching techniques and materials;
4. Determine the feasibility of programme within the class-room setting;

5. Determine areas needing modifications or improvement in the teaching- learning process; and

6. Determine to a great extent the outcome of summative evaluation.

Questions asked under student evaluation differ from those asked by a programme evaluator The teacher is specifically concerned with the behavioural outputs of his students after they have been exposed to a form of instruction. The questions below are representative.

1. What is the objective of the lesson?

2. What materials will be needed to teach this lesson?

3. In what sequence will the different aspects of the topic be treated?

   How much time should be given to different aspects of the topic?

4. What teaching techniques will be most suitable to transmit this Knowledge or skill?

5. What evaluation techniques would be used to assess student achievement? Will they be effective or not?

6. What assignment or projects should be given as part of or apart from class work?

7. Has the objective been achieved?

8. What progress the students are making? What difficulties are they encountering relative to the topic?

9. What additional facilities or resources would enhance the knowledge or skills gained by the student?

10. Are students' needs and interests being met? Are the students able to transfer their knowledge or skills to other areas?

Formative evaluation attempts I) to identify the content (i.e. knowledge or skills) which have not been mastered by the student; ii) to appraise the level of cognitive abilities such as memorization, classification, comparison, analysis, explanation, quantification, application and so on; and iii) to specify the relationships between content and levels of cognitive abilities. In other words, formative evaluation provides the evaluator with useful information about the strength or weakness of the student within an instructional context.

Definition and Meaning of formative evaluation

I. In the words of A.J. Nitko, (1983), "formative evaluation is concerned with judgements made during the design and or development of a programme which are directed towards modifying, forming or otherwise improving the programme before it is completed."

II. According to N.E. Gronlund (1985) "Formative evaluation is used to monitor learning progress during instructtion and to provide continuous feedback to both pupil and teacher concerning learning successes and failures. Feedback to pupils reinforces successful learning and identifies the learning errors that need correction. Feedback to the teacher provides information for modifying instruction and prescribing group and individual remedial work".

III. In the views of R.L.Ebel and D.A. Frisbie (1986), "Form-ative evaluation is conducted to monitor the

instructional process, to determine whether learning is taking place as planned."

IV. Gilbert Sax (1989) states: "Formative evaluation takes place during instruction by letting the teacher or evaluator know if students are meeting instructional objectives, if the programme is on time and if there are ways that the programme might be improved. Formative evaluation helps current students to learn more effectively".

V. W. Wiersma and S.G. Jurs write, "Formative evaluation occurs over a period of time and monitors student progress.

Following are the implications of the above definitions for the classroom teacher:

1. Formative evaluation is done during an instructional programme.
2. The instructional programme should aim at the attainment of certain objectives during the implement-tation of the programme.
3. Formative evaluation is done to monitor learning and modifying the programme if needed before its completion.
4. Formative evaluation is for current students.

Characteristics of Formative Evaluation:

1. It relatively focuses on molecular analysis.
2. It is because seeking.
3. It is interested in the broader experiences of the programme users.

4. Its design is exploratory and flexible.

5. It tends to ignore the local effects of a particular programme.

6. It seeks to identify influential variables.

7. It requires analysis of instructional material for mapping the hierarchical structure of the learning tasks and actual teaching of the course for a certain period

1.1.2 Summative Evaluation

Summative evaluation is primarily concerned with purposes, progress and outcomes of the teaching-learning process. It attempts as far as possible to determine to what extent the broad objectives of a programme have been achieved. It is based on the following assumptions.

1. That the programmer's objectives are achievable;

2. That the teaching-learning process has been conducted efficiently;

3. That the teacher-student-material interactions have been conducive to learning.

4. That the teaching techniques, learning materials and audio-visual aids are adequate and have been judiciously dispensed; and

5. That there is uniformity in classroom conditions for all learners.

This last assumption is often far from the truth. The fact is that conditions under which teaching and learning take place are often unsatisfactory.

Unlike formative evaluation, which is guidanece-oriented, summative evaluation is judgemental in nature. Promotion examination, the first school leaving certificate examination, the public examinations belong to this form of evaluation. Students, performance in such examinations determine to a large extent their job career or prospects of further education. Summative evaluation carries threat with it in that the student may have no knowledge of the evaluator. In class tests the students often can predict to reasonable extent what they would be asked. Summative evaluation is more objective in nature than formative evaluation; we shall discuss summative evaluation further in the form of standardized tests in unit 5.

I. ***Summative Evaluation***: According to A.J. Nikto (1983). Summative evaluation describes judgements about the merits of an already completed programme, procedure or product.

II. ***N.E. Gronund (1985)***, observes, "Summative evaluation typically comes at the end of a course (or unit) of ins-tructtion. It is designed to determine the extent to which the instructional objectives have been achieved and is used primarily for assigning course grades or certifying pupil mastery of the intended learning out-comes."

III. ***In the views of Ebel, R.L. and Frisbie (1986)***, "Summative evaluation is conducted at the end of an instructional segment to determine if learning is sufficiently complete to warrant moving the learner to the next segment of instruction."

IV. ***In the words of Gilbert Sax (1989),*** A summative evaluation can provide evidence that the programme is

satisfactory and should be continued for next year's students or that student learning and learning attitudes are so negative that a new programme is needed."

V. ***W.Wiersma and S. G. Gurs (1990)*** **state**, "Summative evaluation is done at the conclusion of instruction and measures the extent to which students have attained the desired outcomes."

A perusal of the above definitions shows that the summative evaluation has following chief elements.

1. There should be some instructional programme before summative evaluation.
2. The instructional programme should be for the attainment of some objects.
3. Summative evaluation is done at the end or completion of a particular instructional programme whose duration may vary from a semester to whole year.
4. Summative evaluation should check whether there has been earning or not. If the answer is yes, then what is the quantity and quality of the learning in relation to pre-determined objectives?
5. Summative learning provides feedback to the classroom teacher for the success or failure of the programme and of the student.

Chief Characteristics of summative Evaluation

1. It lends to the use of well-defined evaluation designs.
2. It focuses on analysis.
3. It provides descriptive analysis.
4. It tends to stress local effects.
5. It is unobtrusive and non-reactive as far as possible.
6. It is concerned with broad range of issues.
7. Its instruments are reliable and valid.

2.3 Difference Between The Summative And Formative Evaluation

In the beginning these terms applied for the evaluation of curricular work only. M. Seriven explains the difference between these terms as follows in his book Evaluation Thesaurus (1980), "Formative evaluation is conducted during the development or improvement of a programme or product (or person). It is an evaluation conducted for in-house but it may be done by an internal or external evaluator (preferably) a combination. Summative evaluation, on the other hand, is conducted after ompletion of a prog-ramme (or a course of study) and for the benefit of some external audience or dicision maker (e.g. funding agency or future possible users) though it may be done by an internal or an external evaluator or by a combination."

Gloria, Hitchok and others (1986) state the difference between the summative and formative evaluation in these words, "It is fairly straight forward to produce an 'ideal' type of either a summative or a formative profile. It is far more difficult to combine the two into one unified system. The underlying philosophies of the two appear diffi-cult to reconcile."

Following are the main differences between these two types of evaluation.

1 They differ in purpose, nature and timing.

2 Summative evaluation is the terminal assessment of performance at the end of instruction but formative evaluation in the assessment made during the instructional phase to inform the teacher about progress in learning and what more is to be done.

3 The summative evaluation limits the use of profiles and record of achievement but they are regularly used in formative evaluation.

4 The main consideration in summative evaluation is to determine to what extent the examinee has mastered the knowledge and skills associated with a course. On the other hand, the main consideration in formative evalu-ation is to reveal the processes by which the examinee achieved these outcomes.

5 In summative evaluation, the assessment is done to test learning outcomes against a set of objective criteria with-out revealing the details of the route to the teacher, which the student followed in reaching that point. Formative evaluation takes the form of a dialogue between the student and teacher in which both determine the task.

Broad Differences Formative and Summative Evaluation

Characteristics	Formative	Summative
Purpose	To monitor progress of students by getting feedback	To check final status of students
Content focus	Detailed Narrow scope	General Broad scope
Methods	Daily assignments observations	Tests Projects
Frequency	Daily	Weekly, quarterly etc

Alkin (1974) pointed out that a formative evaluation study uses a great variety of instruments, which are either locally developed or standardized; it relies on observation and informal data collection devices, mostly locally chosen. In contrast, summative evaluation studies tend to use well defined evaluation designs, as unobtrusive and non-reactive as possible, they are comparative and concerned with a broad range of issues, for example, implications, politics, costs, competing options. The instruments used in summative evalu-ation are publicly accepted, reliable and valid instruments, reflecting concerns of the sponsor and of the decision maker.

Formative tests are administered at the completion of each unit of learning and help students to pace their learning and put forth necessary effort at the appropriate time. They provide immediate and continuous feedback to the student via instruments that are essentially brief, so that they do not take up inordinate amounts of instructional time. Thus, it forces and

reinforces learning mastery by providing data that can direct remedial teaching. Summative evaluations are in real sense 'final' tests of students' achievement typically covering relatively large blocks of instructional material. In formative evaluation scoring is based on criterion reference approach but in summative evaluation, scoring is generally norm referenced though it can be criterion referenced also. In formative evaluation, the method of reporting scores is individual pattern of pass-fail scores on each task in hierarchy, whereas in summative evaluation, attainment is reported in terms of total scores.

Seriven adheres to the view that there are no basic logical and methodological differences between formative and summative evalu-ation. Both are intended to examine the worth of a particular entity. Only timing, the audience requesting it, and the way its results are used can indicate whether a study is formative or summative. More-over, the same study may be viewed by one client as formative and by another as summative.

2.4 Summary

Evaluation deals with the appraisal of value or worth of a thing process or programme. Basically, evaluation can be carried out either at the student or at programme level or both. Generally, classroom tests come under student evaluation.' While surveys, interviews, supervision and so on come under 'programme evaluation.' The questions raised at both levels of evaluation are, therefore, different. Student evaluation can further be classified into formative and summative evaluation. While the former is concerned with student progress or lack of it and how to achieve success, the latter is judgemental and terminal in nature. A summative evaluation declares which students have or have not succeeded in a programme and no more. Supervision at both the

student and programme level is an indispensable maintenance process in curriculum planning and development.

3- NEED AND ROLE OF EVALUATION

Evaluation is to provide relevant information that decision makers need about input, output operations of programmes and placement of students in programs. Levels of understanding can be assessed and future educational objectives set, based on student needs. Appropriate activities can be planned buy the teacher based on the knowledge of the attributes of the student and their achievement of objectives that have been set for them. Formation of objectives, selection of content, and planning for learning experiences appropriate to student achievement can be facilitated.

3.1 Need For Evaluation In Education

As long as there is need for the educator to make some instructional decisions, **curricular** decision, **and selection** decision. **Placement** or **classification** decisions based on the present or anticipated educational status of the child0 so long will there be need for education in educational enterprise.

To the modem educator, the ultimate goal of evaluation is to facilitate learning. This could be done in a number of ways, in each way a separate type of decision is required. The evaluation decision also determines which of tests is to be used for evaluation. Thus there is a close relationship between the purpose of evaluation, evaluative decisions and types of tests to be used for them. The purposes of evaluation are as follows:

(i) Selection decision

Whenever there will be choice, selection decision is to be made. In our daily life we see that institutions and organization need persons for their work, they get responses from several people

but they cannot take all of them They have to make selection out of them. Evaluation of these persons is to be made on the bases of tests given to them. Tests will provide information, which will help in selection decision. Some persons will be acceptable while others will not be acceptable. Similarly the universities have to make section decisions for admitting the students to various courses. Courses in which hundreds of candidates are applicants, Selection decision is to be make on stronger footing. Naturally some tests are given to the candidates to help in selection decision. Aptitude tests, Intelligence tests, Achievement tests or Prognostic tests are generally given for the purpose of selection decision. There has been ruling from the judiciary that the scores on these tests should have a good relationship with the success in the job or the course for which the tests has been given. If any selection tests does not fulfill this requirement it needs to be improved or replaced by a better one I Although perfection of such tests cannot be guaranteed but any institution or organization which is interested in the best students or workers will continue to make efforts in improving the tests being used for the purpose of selection

(ii) Placement decision

Since school education should be provide to all in a welfare stat, the schools must make provision for all, they cannot reject the candidates for admission as the universities or colleges can do. How these candidates placed in different programmes of school education is to be determined on the basis of their evaluation. Such school determinations are called placement decision. These decisions are required not only in the case of those who are with some disadvantage but also with those who are gifted and talented. The schools have to find one or the other programme for all school age children depending upon their weakness or strength. Placement tests have to be different and more useful from selection

tests because they improve the decision to differentially assign students to teaching programmes. Achievement test and interview are generally used for placement decision.

(iii) Classification decisions

Evaluation is also required to help in making decisions in regard to assigning a person to one of several different categories, jobs or programmes. These decision are called classification decisions because in one particular job or programme, there may be several levels or categories. To which level or category a particular person or child be assigned, depends upon the results of the test. Aptitude tests, achievement tests, interest inventories value questionnaires attitude scale and personality measures are used for classification decision. There is a minor difference in classification, placement and selection. Classification refers to the cases whare categories are essentially unordered, placement refers to the case where the categories represent level of teaching or treatment and selection refers to the case where the persons can be selected or rejected.

(iv) Diagnosis and remedial decisions

Evaluation is required to locate the students who need special remedial help. For example what instructional strategies the teacher should use to help a particular student or a group of student so that the opportunities are maximized to achieve the objective. Aptitude tests, intelligence tests, diagnostic achievement tests, diagnostic personality measures etc. may be used to achieve the purpose.

(v) Feed back

It is not sufficient to evaluate student through a test and doing nothing after that. A good teacher will use tests for the

purpose of providing feed back to students. Feedback may be effective or ineffective depending upon the circumstances. Feed back will facilitate learning if it confirms the learner's correct responses or identifies errors and corrects them. Test results made available to parents may be used for making feedback as evaluation device. It is also to be remembered that feedback are both for the student and teacher because it provide information to both and help in knowing how will students have learnt and how well the teacher has taught.

(vi) Motivation and guidance of learning

Evaluation is also used to motivate the students for more study and providing for learning. However motivation device can be used positively as well as negatively. Unfortunately most of the school-teacher use motivation as a negative device rather than positive one. Threat to failure in the examination or refusing to grant annyual promotion to next class can motivate the student but if they are motivated with using such evaluation techniques which provide more confidence to the students in the subject, they will be more effective and lasting. Aptitude tests, achievement tests, attitude scales, perso-nality measures, interest inventories, surprise quizzes encourage student for more study and understanding.

(vii) Programme evaluation

If any education programme is introduced it is necessary to evaluate its worth from time to time so that it should be promoted if found useful and effective and should be stopped if found useless and ineffective. Since all programme cannot be suitable in all situations, so it is necessary that they should be evaluated in the same situation in which they have been introduced. Giving

Aptitude tests, Intelli-gence tests Attitude scale and interest inventories to he learners.

(viii) development

The education theories with which we are conversant and which will be developed in future are developed through researches. If the educationists and psychologists do not make researches, knowledge will not grow and the next generation will learn only those theories which were developed in the past. The researchers make use only of all evaluation devise and techniques for the purpose of finding out results and developing theories. Thus one of the main purpose of evaluation is to develop educational theories.

(ix) Assigning makers to students

The instructional programme remains incomplete if it is not followed by evaluation. Although no teacher chooses teaching profession because he is interested in evaluating the students but no teacher confines his job to teaching only. He regularly evaluates his students and assigns them makers. Actually most of the teachers are giving most of their time to this purpose. If teachers do not evaluate their students, do not assign those marks or grades, how can they check their effectiveness of teaching and learning outcome of the students?

3.2 Role Of Evaluation And Measurement In Education Process

The evaluation of learning takes place in an instructional context and,

Consequently, that learning environment shapes the reasons why we evaluate, influences the purpose for evaluating as well how we evaluate and determines how we should use the

outcomes of our evaluation. Evaluation is an integral part of instruction; it is not a separate entity that somehow is loosely attached to the teaching process. The instruction process and the role of evaluation in it both must be understood as background to the study of educational measurement. To that end, the role of evaluation in instruction will be described using a model that explains how the teaching process works.

(A) There are many models that describe the variety of approaches to teaching found in schools, but the Basic Teaching Model (BTM), introduced by Glaser (1962), accounts for the fundamental components of most other specific teaching models, such as the Socratic approach, the individualized instruction approach, or the computer dominated instructional approach (joyce and weil, 1980). Few teachers probably follow the BTM steps explicitly to guide their instructional activities. And though we do not specifically endorse the use of the BTM or any other particular model, we do advocate instructional approaches, by whatever name, that account for the fundamental functions represented in the BTM as described next.

The main purposes of the BTM are to identify the major activities of the teacher and to describe the relationship between activities figure III is a diagram of the mode. Our primary interest is the Performer Assessment component, but we cannot understand completely the role of evaluation without understanding how Performance Assessment affects, and is affected by, other teaching activities. Instructional Objectives, the first component of the BTM, represents the teacher's starting point in providing instruction. What should students learn? What skills and knowledge should be the focus of instruction? What is curriculum and how is it defined? The second component, Entering Behavior, indicates that the teacher must try to assess the

students' level of achievement and readiness to learn prior to beginning instruction. What do the students know already and what are their cognitive skills like? How receptive to learning are they? Which ones seem self- motivated? This component indicates a need for evaluation information before instruction actually begins.

Once the teacher has decided what will be taught and to whom the teachings is to be directed, the "How?" must be determined. The Instructional Procedures component deals with the material and methods of instruction the teacher selects or develops to facilitate student learning. Does the text need to be supplemented with illustration? Should small group projects be developed? Is there computer software available to serve as a refresher for prerequisites? At this point instruction could begin, and often it does, but unless the teacher makes plans to evaluate students' performance, the students and teacher will never be sure when learning is complete. The performance Assessment component helps to answer the question, "Did we accomplish what we set out to do? Tests, quizzes, teacher observations, projects, and demonstration are evaluation tools that help to answer this question. Thus evaluation should be a significant aspect of the teaching process; teaching does not occur, according to the model, unless evaluation of learner performance occurs.

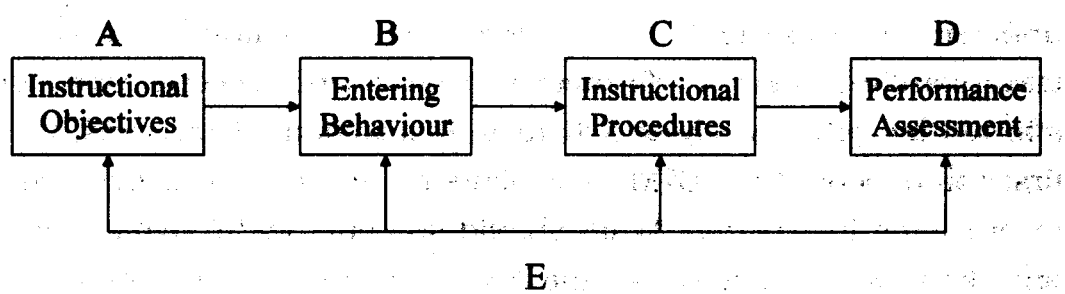


Figure: III The Basic Teaching Model (DeCecco and Crawford, 1974)

The model shows a fifth component, the Feedback Loop that can be used by the teacher as both a management and a diagnostic procedure. If the results of evaluation indicate that sufficient learning has occurred, the loop takes the teacher back to the Instructional Objectives component, and each successive component, so that plans for beginning the next instructional unit can be developed. (New objectives are needed, entering behavior is different, and methods will need to be reconsidered,) But when evaluation results are not so positive, the Feedback Loop is a mechanism for identifying possible explanations. (Note the arrows that return to each component.) Were the objectives too vaguely specified? Did students lack essential prerequisite skills or knowledge? Was the film or text relatively ineffective? Was there insufficient practice opportunity? Such questions need to be asked and frequently are. However, questions need to be asked about the effectiveness of the performance assessment procedures also, perhaps more frequently than they are. Were the test questions appropriate? Were enough observations made/ Were directions clear to students? The Feedback Loop returns to the Performance Assessment component to indicate that we must review and assess the quality of out evaluation procedures, after the fact, to determine the appropriateness of the procedures and the accuracy of the information. Unless the tools of evaluation are developed with care, inadequate learning may go undetected or complete learning may be misinterpreted as deficient.

In sum, good teaching requires planning for and using good evaluation tools. Furthermore, evaluation does not take place in vacuum. The BTM shows that other components of the teaching process provide cues about what to evaluate, when to evaluate, and how to evaluate. Our purpose is to identify such cues and to

take advantage of them in building tests and other assessment devices that measure achievement as precisely as possible.

(B) Evaluation helps the decision maker who is concerned about all aspects of the educational endeavor. The key point to consider and keep in mind is that evaluation involves appraisal of particular goals or purposes. Useful information may be obtained for evaluation procedures by both formal and informal mean and should include information collected during instruction as well as in the end of the course date. According to Ahmann and Glock (1985).

> School Administrators, guidance personnel, classroom teacher, and individual students require information that will allow them to make informed and appropriate decision regarding their respective educational activities. Ideally, they should be aware of all the alternatives open to them, the possible outcomes of each alternative, and the advantages and disadvantages of the respective outcomes, Educational and psychological measurement can help individuals with these matters.

(C) Tyler, 1966: Airasian and Madaus, 1972: Gronlund 1976: Thorndike and Hagen, 1977: rightly observe. That the data secured through testing procedures may have uses as give below:

> First, measurement data may be employed in the placement of students on one or another instructional programme. Usually pupils take a pretest to measure whether they have mastered the skills that are prerequisite to admisttance to a particular course or instructional sequence. For instance, foreign language and mathematics programmes are usually arranged in some hierarchical order so that achievement at each level of learning depends on mastery of the preceding

level. The student is lead from the entering position in the hierarchy to the terminating phase via intermediate steps, based upon the information provided by a pretest, a student can be placed

(1) At the most appropriate point in the instructional sequence

(2) In a programme with a particular instructional strategy on

(3) With an appropriate teacher.

Second, measurement data can be used in formative evaluation. Tests are administered to students to monitor their success and to provide them with relevant feedback. The information is employed less to grace a student than to make instructions responsive to the student's strengths and weaknesses as identified by the measurement device. Mastery learning procedures emphasize the use of formative tests to provide detailed information about each student's grasp of a unit's objectives.

Third, measurement data has a place in diagnostic evaluation. Diagnostic testing takes over where formative testing leaves off. When a student fails to respond to the feedback corrective activities associated with formative testing a more detailed search for the source of the learning difficulty is indicated. Remediation is only possible when a teacher understands the basis of a student's problem and then designs instruction to address the need.

Forth, measurement data may be used for summation purposes. Such testing is employed to certify or grade students at the completion of a course or unit of instruction. Often the result is 'final' and follows the student throughout

his or her academic career (as in the case or college and university transcripts). It is this aspect of evaluation that some educators final particularly objectionable.

Fifth, measurement data are used by employers educational institutions in making the selection by decisions. Many jobs and slots in educational programme are limited in number, and there are more applicants than positions. In order to identify the most promising candidates standardized tests may be administered to the applicants. The information provided by the tests presumably increases the accuracy and objectivity of administrator's decisions. College Board examinations are used by many universities in admitting students to graduate and professional schools likewise employ data from standardized testing programme make their entrance decisions.

Sixth, measurement data are used by school officials in making curricular decisions in order to evaluate existing programme and to decide among instructional alternative. School administrators need to assess their students' current levels of performance, the strengths and weaknesses of the evidence.

Seventh, measurement data finds a place in personal decision making. Individuals confront a variety of choices at any number of points in their lives. Should they attend college or pursue some other type of post-high school training? What kind of job seems most suited to their needs? What sort of training programme should they enter? Measures of interest, temperament, and ability can give individuals insights that can prove helpful in the decision-making process.

3.3 Summary

The interrelated nature of teaching, learning, and evaluation can be seen in the following sequential steps in the instructional process: (1) preparing instructional objectives, (2) preassessing learners' needs, (3) providing relevant instruction (monitoring learning progress and diagnosing difficulties), (4) evaluating the intended learning outcomes, and (5) using the evaluation results to improve learning and instruction. In addition to the direct contribution testing and evaluation make to classroom instruction, they also play an important role in marking and reporting, curriculum development, educational and vocational guidance, and evaluating the effectiveness of the school programme.

The vast array of evaluation procedures used in the school can be classified and described in many different ways. The following are especially useful designation for describing the various procedures:

Nature of the Measurement

1. Maximum performance (what a person can do).
2. Typical performance (what a person will do)

Use in Classroom Instruction

1. Placement evaluation (measures entry behavior).
2. Formative evaluation (monitors learning progress).
3. Diagnostic evaluation (identifies causes of learning problems)
4. Summative evaluation (measures end-of-course achievement).

Method of Interpreting the Results

1. Norm referenced (describes pupil performance in terms of the relative position held in some known group).
2. Criterion referenced (describes pupil performance in term of a clearly defined and delimited domain of learning tasks).

4 SELF ASSESSMENT QUESTIONS

1. Describe the meaning of the following terms: test, measurement and evaluation.
2. What is testing? How will you distinguish between measurement and evaluation? How are measurement and evaluation related?
3. What is evaluation? Why do educators evaluate curricular programme?
4. Distinguish between programme evaluation and student evalu-ation?
5. What is formative evaluation? How does formative evaluation differ from summative evaluation?
6. What main factors affect success of evaluation?
7. What factors determine to success of an implemented programme?
8. What is the difference between testing and evaluation?
9. What roles should a supervisor play in the teaching-learning process?
10. Identify the major points to be considered when evaluation a lesson.
11. Discuss the role of evaluation and measurement in education.

5 BIBLIOGRAPHY

1- Gay L.R. (1985) Educational Evaluation and Measurement Columus Charles E. Merril Publishing Co.

2- E. Gronlund Norman and Linn Robert. (1990) Measurement and Evaluation in Teaching. New york, MacMillan Publishing Company.

3- Ebel and Robert L. David A. Frisbie, (1991) essentials of Educational Measurement New Delhi, Prentice Hall of India.

4- Thorndike, R.L. and Hagen E.P. (1997) Measurement and Evaluation in Psychology and Education, New york: John Wiley & sons, Inc.

5- Vahist S.R. (1993). Theory of educational Evaluation, New Delhi; Anmol Publications PVT. Ltd.

6- Wiersma, William and Jurs, G. (1990) Educational Measurement and Testing London, allyn and Bacon.

Unit-2

QUALITIES OF GOOD TEST

Written By:
Muhammad Arshad Bandesha

INTRODUCTION

Whenever we would like to use a test or other measurement procedure to provide information to help in some decision, we face the problem of which test or procedure to use or whether there is any instrument that will really help in the decision. There are usually several tests that have been designed to help or that seem at least to have the possibility of helping with the decision. We would like to know whether any test will indeed provide useful information .And if any will, which is the best one to use?

Tests and other evaluation instruments serve a variety of uses in the school. For example, tests of achievement might be used for selection, placement, diagnosis, or certification of mastery, aptitude tests might be used for predicting success in future learning activities or occupations; and appraisals of personal-social development might be used to understand better pupil's learning problems or to evaluate the effects of a particular school programme.

When a person is faced with the responsibility of choosing among two or three tests, all of which are available from reputable sources, how does he proceed to select one of them for use? For the sake of this discussion we shall assume that the tests under consideration appear to be equally suited to local conditions and that the strengths and weaknesses of the tests are fairly well balanced as far as the obvious and non-technical features are concerned. What, then, are the basic criteria of a more technical nature that may be used as guides in the selection of a test or other measuring device?

All good measuring instruments have certain primary qualities which differentiate good tests from inferior ones whether they

be for use of the educator, the psychologist, the medical technician, the physicist, or people in other fields.

A test which lacks known and substantial degree of these primary qualities is not a measuring instrument in any true sense, and little or no dependence can be placed upon results obtained by its use. The two universals generally upon are ***reliability and validity***.

Besides these two universal requirements for a good test, whatever the field, there are certain secondary characteristics which are desirable in all good educational and psychological tests: ***adequacy, objectivity and usability.*** These are less crucial than reliability and validity, since a test may function efficiently without the presence of the secondary characteristics as long as is valid and reliable. However, the secondary qualities to some extent affect validity and reliability and in any event make the use of a test much simpler.

Reliability refers to the accuracy and precision of a measurement procedure. Indices of reliability give an indicator of the extent to which the scores produced by a particular measurement procedure are consistent and reproducible. ***Validity*** has to do with the degree to which the test scores provide information that is relevant to the inferences that are to be made from them. Thus, a judgment of validity is always in relation to a specific decision or use, and evidence that test scores are appropriate for one purpose does not necessarily mean that the scores are appropriate for another.

Reliability and validity are both required of any test that we would choose, regardless of how practical it is to use. Validity of test scores for their intended use is the absolutely essential quality for a test to have, but in a sense, reliability is a necessary precon-

dition for validity. Test scores must be at least moderately reliable before they can have any validity, but a reliable test may devoid of validity for the application we have in mind. Although the true bottom line in selecting a test is the test's validity for our proposed use. We discuss reliability first as a necessary condition for validity to exist.

OBJECTIVES

After reading this unit you are expected to be able to:

1. Define reliability and validity of a test.
2. Explain the nature of reliability.
3. Differentiate between reliability and validity.
4. Identify various techniques of estimating reliability.
5. Specify the significance of validity in good test.
6. Compute and apply standard error of measurement to different test scores.
7. Enlist basic considerations effecting reliability and validity.
8. Describe and explain the need of adequacy, objectivity and usability in testing.

CONTENTS

Introduction

Objectives

1. Reliability

1.1 Methods of Estimating Reliability
- 1.1.1 Test-Retest Method
- 1.1.2 Equivalent-Form Methods
- 1.1.3 Split-Half Method
- 1.1.4 Kuder-Richardson Method

1.2 Standard Error of Measurement

1.3 Factors Affecting Reliability

2. Vlidity

2.1 Content Validity

2.2 Concurrent Validity

2.3 Predictive Validity

2.4 Construct Validity

2.5 Nature of Validity

3. Adequacy

3.1 Factors Influencing Validity

3.2 Factors in the Test Itself

3.3 Factors in Test Administration and Scoring

3.4 Factors in Pupils' Responses

3.5 Nature of the Group and the Criterion

4. Objectivity and Usability

4.1 Usability or Practicability

4.2 Ease of Administration

4.3 Time Required for Administration

4.4 Ease of Scoring
4.5 Ease of Interpretation and Application
4.6 Availability of Equivalent or Comparable Forms
4.7 Cost of Testing

5. Self Assessment Questions

1. RELIABILITY

When we ask about a test's reliability, we are asking about what the test measures instead, how accurately it measures whatever it does measures. What is the precision of the resulting score? How accurately will the score be reproduced if we measure the individual again?

Since reliability is essential to validity and the opposite is not so, there is something to be said for placing reliability at the head of the list. A test may be reliable without being valid. Whereas the validity of a test depends in part on its reliability; therefore, a test is only as valid as it is reliable. Reliability refers to the consistency with which a test measures. We shall clarify the meaning of consistency in as test with an illustration. It is observed that when an individual measured the diameter of a very accurately turned steel ball several times with an exceedingly accurate pair of calipers, he did not get exactly the same result every time. Even with the most accurate instruments available and the best possible control of conditions, the successive measurement of the diameter of the steel ball always somewhat varies. The extent of such variations is a measure of the consistency, or the lack of it, in this measuring situation. Some degree of inconsistency is present in all measurement procedures.

In educational and psychological measurement we have another way of expressing and gauging consistency. In measuring the qualities of human beings it is seldom possible or even appropriate to determine the consistency of measurement by many repeated measurements of the same thing, as was done in the illustration of the ball. Since we are dealing with living organisms, we cannot expect repeated measurements to show such close agreement. Therefore, when dealing with people, we determine

consistency by measuring a number of individuals only twice as a rule_ and comparing the relative standings of the individuals on the two set of measurements or scores. It should be noted that the two successive measurements are usually not more than few days apart.

To illustrate with a simple example, let us suppose we had given a group of seven students a test in social studies and ranked them according to there scores. A day or two later we repeated the test on the same group of students and ranked them again. The results might be as follows:

Table -2.1

Comparison Of Scores Made By Seven Students On The Same Test Administered Twice

Pupil	First Testing		Second Testing	
	Score	Rank	Score	Rank
A	52	4	55	4
B	60	2	65	2
C	45	5	48	5
D	68	1	69	1
E	57	3	60	3
F	29	7	40	6
G	31	6	35	7

The degree of consistency of measurement can be judged here by the extent to which the pupils tend to hold the same relative position in their group. We can see that this tendency is high in this case since all pupils except F and G hold the same rank in both applications of the test, and those two pupils shift slightly.

It should be pointed out that in this example all students show a gain in score between the first and the second testing, but their relative standings or ranks change in only two cases. If all individuals made the same score both times, or made lower scores the second time, the test would still show a high degree of consistency provided that the ranks of the individuals did not change. This, then, is what we mean by consistency.

1.1 Methods of Estimating Reliability

In determining reliability it would be desirable to obtain two sets of measures under identical conditions and then to compare the results. This procedure is impossible, of course, because the conditions under which evaluation data are obtained can never be identical. As a substitute for this ideal procedure, several methods of estimating reliability have introduced. The methods are similar in that all of them involve correlating two sets of data, obtained either from the same evaluation instrument or from equivalent forms of the same procedure.

The chief methods of estimating reliability are shown in Table 2.2. Note that different types of consistency are determined by different methods_ consistency over a period of time, over different forms of the instrument, and within the instrument itself. The reliability coefficient resulting from each method must be interpreted according to the type of consistency being investigated. Each of these methods of estimating reliability will be considered in further detail as we proceed. Although these methods will be mainly with reference to testing procedures, they are also applicable to other types of evaluation techniques.

Table 2.2

Methods of Estimating Reliability

Procedure	Types of Reliability Measure	
Test-retest method	Measure of stability	Give the same test twice to the same group with any time interval between test, from several minutes to several years
Equivalent-form method	Measure of equivalent	Give two forms of the test to the same group in close succession
Split-half method	Measure of internal	Give test once. Score two equivalent halves of test(e.g. odd items and even items): correct reliability coefficient fit whole test, by Spearman –Brown formula
Kuder-Richardson method	Measure of internal consistency	Give test once. Score total test and apply Kuder-Richardson formula.

1.1.1 Test Retest Method

To estimate reliability by means of the test-retest method, the same test is administered twice to the same group of pupils with a given time interval between the two administration. The resulting scores are correlated, and this correlation coefficient provides measure of stability; that is, it indicates how stable the test results are over the given period of time. If the test will tend to be high on other administrations, and the remaining pupils will to stay in their relatives positions on both administrations. Such stability is indicated by a large correlation coefficient. A perfect correlation is indicated by 1.00 and a zero relationship by .00. Measures of stability in the .80s and .90s are commonly reported for standardized tests.

This method has certain disadvantages. Repeating the test at too short an interval introduces the memory factor and tends to make the self-correlations of the test too high, unless of course, the memory factor is what one wants to measure. On the other hand, repeating the test after a longer time interval permits such factors as growth, intervening learning, and unlearning to come into play so as to make the self-correlation lower than it should be.

In general, the longer the time interval is between test and retest, the more the results will be influenced by changes in the pupil characteristic being measured, and the smaller the reliability coefficient will be.

The best time interval between tests will depend largely on the use to be made of the results. For some decisions we are interested in reliability coefficients based on a long interval between test and retest, and for others, reliability coefficients based on a short interval may be sufficient. The important thing is to seek evidence of stability that fits the particular interpretation to be made.

1.1.2 Equivalent-Form Method

The equivalent-forms (also called parallel or alternate forms) method of estimating reliability makes it possible to avoid the disadvantages of too short or too long a time interval between successive administrations of the evaluating device. Two equivalent forms of the test must be constructed so they are as similar as possible (but not identical) in the kind of content, mental processes required, number of items, difficulty, and all other respects. The pupils take one form of the test and then, as soon as possible, the other form. The agreement between the two is again determined by means of a correlation coefficient, which for this method is sometimes called a coefficient of equivalence. This coefficient of

correlation indicates the degree to which both forms of the test are measuring the same aspects of behavior.

The equivalent forms method is sometimes used with a time interval between the administrations of the two forms of the test. Under these test-retest conditions, the resulting reliability coefficient provides a measure of ***stability and equivalence***. This is the most rigorous test of reliability because it includes all possible sources of variation in the test score. The stability of the testing procedures, the constancy of the pupil characteristic being measured, and the representative ness of the sample of tasks included in the test all are taken into account. Consequently, this is generally recommended as the soundest procedure for estimating the reliability of test scores. As with the ordinary test-retest method, the reliability coefficient must be interpreted in light of the time interval between the two forms of the test. For longer periods, we should ordinarily expect smaller reliability coefficient.

1.1.3 Split-Half Method

It is often impossible to employ either of the methods described above to determine reliability of a test. It may not be feasible to test twice; there may not be equivalent forms available. In such cases we generally use what is known as the split-half technique. In this procedure, the test whose reliability we wish to measure is given in the ordinary manner, the papers are scored as usual, and then two scores for each individual are obtained by scoring alternate halves of the test separately. Such scoring can be done in several ways. Probably the most commonly used method of obtaining two scores for each person is to base one score on only the odd-numbered items of the test and the other on the even-numbered items. Thus in a test of 100 items one score for any individual would be based on the 50 items numbered 1,3,5,.....99,

and on the 50 items numbered 2,4,6,.....100. Therefore if a pupil missed 10 of the 50 odd-numbered items and 12 of the 50 even-numbered items, his two scores would be 40 and 38. His total score would, of course, be 78, the sum of 40 and 38, or 100-(10+12).

Having obtained two scores for each person tested, we can then calculate the coefficient of correlation between the two sets of scores. This is, in effect the correlation between the two equivalent halves of the test administered at one setting. If the two halves are truly equivalent and if the test is a reliable one, the correlation thus achieved is likely to be very high. One step more is necessary to enable us to determine the reliability of the entire test, in this example a test consisting of 100 items. The correlation coefficient is based on scores which represent only halves of the test, scores on only 50 items. Since we wish to know the reliability of the 100 item test, we now apply the Spearman-Brown Prophecy Formula.

$$\text{Reliability on full test} = \frac{2(\text{reliability of half test})}{1 + (\text{reliability of half test})}$$

The simplicity of the formula can be seen in the following example, in which the correlation coefficient between the test's two halves is 60:

$$\text{Reliability on full test} = \frac{2 \times 0.60}{1 + 0.60} = \frac{1.20}{1.60} = 0.75$$

This correlation coefficient of .75, then, estimates the reliability of a full test when the half-tests correlated .60.

The applicability of the Spearman-Brown Prophecy Formula depends on well the test meets certain assumptions. The two

halves of the test must be as equivalent as possible in average score, variability of scores, and type of items.

1.1.4 Kuder-Richardson Method

Another method of estimating the reliability of test scores from a single administration of a single from of a test is by means of formulas such as those developed by Kuder and Richardson. As with the spilt-half method, these formulas provide a measure of ***internal consistency,*** or homogeneity but do not require splitting the test into halves and rescoring and calculating a correlation coefficient. The only data required for this simpler method are the number of items in the test, the standard deviation of the test, and the arithmetic mean of the total scores on the test.

One of the formulas, called the Kuder- Richard Formula 20, is based on the proportion of persons passing each item and the standard deviation of the total scores. The computation is rather cumbersome, unless information is already available concerning the proportion passing each item, but the result is equal to the average of all possible spit-half coefficients for the group tested.

A less accurate but simper formula to compute is the Kuder-Richardson Formula 21, which can be applied to the results of any test that has been scored on the basis of the number of correct answers. A modified version of the formula is

$$\text{Reliability estimate (KR21)} = \frac{K}{K-1}\left(1 - \frac{M\,(K - M)}{Ks}\right)$$

Where

K = the number of items in the test

M = the mean (arithmetic average) of the test sores

S = the standard deviation of the test scores

This formula will yield approximately the same results as will Kuder-Richardson Formula 20, but in most cases the reliability estimate will be smaller. Its chief advantage is the ease with it can be applied.

1.2 Standard Error of Measurement

The variability of the obtained score in relation to the individual's "true score" is one of the most important determiners of how high a reliability coefficient should be for various purposes. If we gave the same test to a pupil many times and could assume that he had not learned anything or forgotten anything between testing, we would expect him to obtain slightly different scores on test. We would also expect that his "true score" would lie somewhere within the range of the scores he actually made. But this kind of testing is impossible; hence we give him only one test for our purpose. We expect that his obtained score will fall somewhere near his true score. We can estimate the standard error of his score, and this figure will tell the range within which scores on the same test would be expected to fall approximately two-thirds of the time if a very large number of the tests, equivalent in all respects, were given to the pupil. The ***standard error of a score*** is computed as follows:

$$S.E._{(Mears.)} = s_t \sqrt{1 - r_{tt}}$$

Where s_t = standard deviation of the scores on the test

r_{tt} = reliability coefficient of the test

1.3 Factors Affecting Reliability

The reliability coefficient may be affected by the length of the test, the range of talent among pupils, and the conditions under which the test is administered.

Length of Test: Other things being equal, the reliability of a test is a function of its length. Longer tests tend to be more reliable than shorter tests. This is because a longer test will provide a more adequate sample of the behavior being measured, and scores are apt to be less distorted by chance factors such as guessing. Suppose that to measure spelling ability, we asked pupils to spell one word. The results would be patently unreliable. Pupils who were able to spell the word perfect spellers and pupils who could not would be complete failures. If we happened to select a difficult word, most pupils would fail: if word was an easy one, most pupils would appear to be perfect spellers. The fact that one word provides an unreliable estimate of a pupil's spelling ability is obvious. It should be equally apparent that we add more words to the list. We come closer and closer to a good estimate of each pupil's ability. Scores based on a large number of spelling words thus are more apt to reflect real differences in spelling ability and therefore to be more stable. By increasing the size of the sample of spelling behavior, therefore, we increase the consistency of our measurement.

This can be explained by using the formula

$$\text{Reliability of lengthened test} = \frac{nr}{1+(n-1)\,r}$$

For a test of 50 items with a reliability coefficient of .80, we could double the length of the test to 100 items and get a whole-set reliability of .888, triple for a coefficient of .923, or quadruple it for a coefficient of .941.

From 50 to 100 items the reliability improved considerably, but from 150 to 200 items there was very little increase in the reliability coefficient.

Range of Talent: The range of talent, achievement, or ability of the pupils on whom the reliability is based has direct effect on the reliability coefficient. The greater the variability in the group of pupils, the higher the reliability coefficient

Spread of scores: Reliability coefficient are directly influenced by the spread of scores in the group tested. Other things being equal, the larger the spread of scores is, the higher the estimate of reliability will be.

Testing Conditions: The conditions of administering and scoring the test may raise or lower the reliability of given test.

Difficulty of Test: Tests that are too easy or too difficult to the members taking it will tend to produce scores of low reliability. This is because both easy and difficult tests result in a restricted spread of scores. For easy test, the scores are close together at the top end of the scale. For the difficult test, the scores are grouped together at the bottom end of the scale.

2. VALIDITY

The Validity of an evaluation device is the degree to which it measures what it is intended to measure. Tests are used for several types of judgment, and each type of judgment requires somewhat different type of validating evidence.

The definition of validity in a testing situation may be elucidated by such questions as these. *What does this test actually measure?* To *what extent does it measure this particular ability, quality, or trait? In what situation or under what conditions does it have this degree of validity?*

Validity refers to the appropriateness of the interpretations made from test scores and other evaluation results, with regard to a particular use. For example, if a test is to be used to describe pupil achievement we should like to be able to interpret the scores as a relevant and representative sample of the achievement domain to be measured. If the results are to be used to predict pupil's success in some future activities, we should like our interpretations to be based on as accurate an estimate of future success as possible. If the results are to be used as a measure of pupil's reading comprehension, we should like our interpretations to be based on evidence that the scores actually reflect reading comprehension and are not distorted by irrelevant factors. Basically, then, validity is always concerned with the specific use of the results and the soundness of our proposed interpretations.

A test can reliable without being valid but that the converse is not true. In other works, it is conceivable that a test can measure some quality with a high degree of consistency without measuring at all the quality it was actually intended to measure. For example, a test might be devised that would require sorting of cards, as in dealing a pack of regular playing cards for a four-

handed game. The test scores might take into accounts both speed and accuracy and have a high degree of reliability. Yet the test might have little or no validity for any specific purpose, such as ability to play cards or ability to work on delicate machinery.

On the other hand, an unreliable test will never show a high degree of validity, for the validity of a test cannot exceed its reliability. An unreliable test cannot be expected to rank the same individuals twice in the same or nearly the same order. Obviously, if no dependence can be placed on the consistency for results obtained by use of a particular test, one can never arrive at any sound judgment of what the test actually measures, and the test therefore has little or no validity.

The discussion of the relationship of validity and reliability leads to another important fact regarding validity, namely, that validity is specific to the purpose and situation for which a test is used. A test might be a highly valid measure of intelligence for third-grade children and decreasingly valid for this purpose with fifth-graders, ninth-graders, high school graduates, and college seniors. Again, a test of manual dexterity might be a highly valid measure of probable success in assembling parts of small electric motors, but decreasingly valid for predicting success in farming, selling automobiles, managing a printing establishment, or teaching higher mathematics. A test which measures the thinking ability of seventh-grade pupils might well be almost a pure memory test for older persons who have been out of school for a long time. The validity of a test, assuming it is reliable, is a measure of the extent to which it serves its intended purpose. A test may be highly valid for one purpose and almost wholly lacking in validity for another. In the same way that a thermometer is used to measure temperature only, and a barometer to measure atmospheric pressure

only, each testing instrument provides valid measurement for specific purposes.

There are several widely accepted methods of assuring or determining the validity of measuring instruments used in educational or psychological work. These methods are classified for the sake of convenience into four categories which we may call content validity, concurrent validity, predictive validity, and construct validity.

2.1 Content Validity

Content validity is evaluated by showing how well the content of the test samples the class of situations or subject matter about which conclusions are to be drawn. It is especially important in the case of achievement and proficiency measures. It is also known as "face validity" and is described by the relevance of a test to different types of criteria, such as analyses of courses of study and jobs, statements of instructional objectives, analysis of textbooks, analysis of teachers, final-examination questions, pooled judgments of competent persons, concepts of social utility, and logical or psychological analyses of mental processes, motor performances, or other behaviors.

2.2 Concurrent Validity

Concurrent validity is evaluated by showing how well test scores correspond to already accept measures of performance or status made at the same time. For example, we may give a social studies class a test on knowledge of basic concepts in social studies and at the same time obtain from its teacher a report on these abilities as far as pupils in the class are concerned. If the relationship between the test scores and the teacher's report of abilities is high, the test will have high concurrent validity.

2.3 Predictive Validity

Predictive validity is evaluated by showing how well predictions made from the test are confirmed by evidence gathered at some subsequent time. It is quite similar to concurrent validity, except that the evidence on the criterion measures used is collected later. This type is especially important, for example, when the tester wants to estimate how well a student may be able to do in college courses on the basis of how well he has done on tests he took in secondary school. It is also applicable in long-range prediction of such factors as vocational success or reaction to therapy. Because of the time between giving the test and obtaining the criterion measures, as in predicting success in college from tests in high school, predictive validity is not as useful with achievement test as are content and concurrent validity.

2.4 Construct Validity

Construct validity is evaluated by investigating what psychological qualities a test measures, or in other words, by demonstrating that certain explanatory constructs account for performance on the test. It is ordinarily used when the tester has no definitive criterion measure of what he is concerned with and hence must use indirect measures. By its very nature, construct validity is inferential rather than conclusive; it is used primarily when the other three types of validity are insufficient to indicate the degree to which the test measures what it is intended to measure. However, the other three types may be considered specialized aspects of construct validity when other criteria are available. Construct validity may be illustrated by the relationship of the items in a test to a description of the universe of items from which the selection was made. This type of validity is usually

involved in such tests as those of study habits, appreciations, understandings, and interpretation of data.

For instructional uses of tests-diagnosis of achievement, planning of remedial work, within-subject determination of instructional materials-content validity is most important. For various administrative uses-pupil classifications, prediction of success, comparison of curricula-concurrent, predictive, and construct validity are important.

2.5 Nature of Validity

When using the term validity in relation to testing and evaluation, there are a number of cautions to be kept in mind.

1. Validity refers to the appropriateness of the interpretation of the results of a test or evaluation instrument for a given group of individuals, and not to the instrument itself. We sometimes speak of the "validity of a test" for the sake of convenience, but it is more correct to speak of the validity of the interpretation to be made from the results.

2. Validity is a matter of degree; it does not exist on an all-or-none basis. Consequently, we should avoid thinking of evaluation results as valid or invalid. Validity is best considered in terms of categories that specify degree, such as high validity, moderate validity, and low validity.

3. Validity is always specific to some particular use. No test is valid for all purposes. For example, the results of an arithmetic test may have a high degree of validity for indicating computational skill, a low degree of validity for indicating arithmetical reasoning, a moderate degree of validity for predicting success in future mathematics courses, and no validity for predicting success in art or

music. Thus, when appraising or describing validity, it is necessary to consider the use to be made of the results. Evaluation results are never just valid; they have a different degree of validity for each particular interpretation to be made.

4. Validity is viewed as a unitary concept based on various kinds of evidence. The three basic ways of accumulating evidence to support the validity of an interpretation are (content, criterion related, and construct). Numerous factors tend to make test results invalid for their intended use.

3. ADEQUACY

Adequacy is the degree to which a test samples sufficiently widely into the subject that the resulting scores are representative of relative total performance in the areas measured.

The careful test maker never assumes that the instrument he has constructed is capable of measuring all of the factual knowledge or skills that a pupil has acquired in a school course. There are too many by-products and incidental learnings to make this possible. Good teaching should never stress a certain restricted body of facts to the exclusion of all other knowledge. When not only factual knowledge and skills but also concepts, understandings, applications, and tastes and preferences are considered, all significant types of instructional outcomes, the task of measuring all of the outcomes from any course, any instructional unit, or even any single class period becomes hopeless. At best, a test is a sample of certain portions of the total behaviour, which the examiner considers vital to pupil mastery in the field. Just as a grain buyer samples a carload of wheat by taking samples from different places in the car and grading the samples in order to obtain a measure of quality for the whole

carload, a test constructor measures the educational attainments of pupils by constructing test items that represent widely the types of pupil outcomes expected and accepts the scores resulting from their use as representative of the pupils' relative achievements for the entire area sampled by the test items.

3.1. Factors Influencing Validity

Numerous factors tend to make test results invalid for their intended use. Some are rather obvious and can be easily avoided. No teacher would think of measuring knowledge of social studies with an English test. Nor would a teacher consider measuring problem-solving skills in third-grade arithmetic with a test designed for sixth graders. In both instances, the test results would obviously be invalid. The factors influencing validity are of this same general nature but much more subtle in character. For example, a teacher may overload a social studies test with items concerning historical facts, and thus the scores are less valid as a measure of achievement in social studies. Or a third-grade teacher may select appropriate arithmetic problems for a test but use vocabulary in the problems and directions that only the better readers are able to understand. The arithmetic test then becomes, in part, reading test, which invalidates the results for their intended use. These examples show some of the more subtle factors influencing validity, for which the teacher should be alert, whether constructing classroom tests or selecting published tests.

3.2. Factors in the Test Itself

A careful examination of test items will indicate whether the test appears to measure the subject matter content and the mental functions that the teacher is interested in testing. However, any of the following factors can prevent the test items from functioning as

intended and thereby lower the validity of the interpretations from the test scores.

1. **Unclear directions:** Directions that do not clearly indicate to the pupil how to respond to the items, whether it is permissible to guess, and how to record the answers will tend to reduce validity.

2. **Reading vocabulary and sentence structure too difficult:** Vocabulary and sentences structure that is too complicated for the pupils taking the test will result in the test's measuring reading comprehension and aspects of intelligence, which will distort the meaning of the test results.

3. **Inappropriate level of difficulty of the test items:** In norm-referenced tests, items that are too easy or too difficult will not provide reliable discriminations among pupils and will therefore lower validity. In criterion-referenced tests, the failure to match the difficulty specified by the learning outcome will lower validity.

4. **Poorly constructed test items**: Test items that unintentionally provide clues to the answer will tend to measure the pupil's alertness in detecting clues as well as those aspects of pupil performance that the test is intended to measure.

5. **Ambiguity:** Ambiguous statements in test items contribute to misinterpretations and confusion. Ambiguity sometimes confuses the better pupils more than it does the poor pupils, causing the items to discriminate in a negative direction.

6. **Test items inappropriate for the outcomes being measured:** Attempting to measure understanding, thinking skills, and other complex types of achievement with test

forms that are appropriate only for measuring factual knowledge will invalidate the results.

7. **Test too short:** A test is only a Sample of the many questions that might be asked. If a test is too short to provide a representative sample of the performance we are interested in, its validity will suffer accordingly.

8. **Improper arrangement of items:** Test items are typically arranged in order of difficulty, with the easiest items first. Placing difficult items early in the test may cause pupils to spend too much time on these and prevent them from reaching items they could easily answer. Improper arrangement may also influence validity by having a detrimental effect on pupil motivation. This influence is likely to be strongest with young pupils.

9. **Identifiable pattern of answers:** Placing correct answers in some systematic pattern (e.g. T,T,F,F, or A,b,C,D,A,B,C,D) will enable pupils to guess the answers to some items more easily, and this will lower validity.

In short, any defect in the test's construction that prevents the test items from functioning as intended will help invalidate the interpretations to be drawn from the results. Much of what is written in the following chapters is directed toward helping teachers improve the validity of their interpretations of test scores and other evaluation results.

3.3. Factors in Test Administration and Scoring

The administration and scoring of a test may also introduce factors that have a detrimental effect on the validity of the interpretation from the results. In the case of teacher-made tests, such factors as insufficient time to complete the test, unfair aid to

individual pupils who ask for help, cheating during the examination, and the unreliable scoring of essay answers tend to lower validity.

In the case of published tests, failure to follow the standard directions and time limits, giving pupils unauthorized assistance, and errors in scoring similarly contribute to lower validity.

For all types of tests, adverse physical and psychological conditions at the time of testing may also have a negative effect.

3.4. Factors in Pupils' Responses

In some instances, invalid test interpretations are due to personal factors influencing the pupil's response to the test situation rather than to any shortcomings in the test instrument or its administration. Some pupils may be bothered by emotional disturbances that interfere with their test performance. Others may be frightened by the test situation and thereby are unable to respond normally, and still others may not be motivated to put forth their best effort. These and other factors that restrict and modify pupils' responses in the test situation will obviously distort the test results.

3.5. Nature of the Group and the Criterion

Validity is always specific to a particular group. An arithmetic test based on story problems, for example, may measure reasoning ability in a slow group and a combination of simple recall of information and computation skill in a more advanced group. Similarly, s cores on a science test may be accounted for largely by reading comprehension in one group and by knowledge of facts in another. What a test measures is influenced by such factors as age, sex, ability level, education background, and

cultural background. Thus, in appraising reports of test validity, it is important to determine the nature of the validation group.

4. OBJECTIVITY AND USABILITY

The objectivity of a test refers to the degree to which equally competent scorers obtain the same results. Most standardized tests of aptitude and achievement are high in objectivity. The test items are of the objectivity type (e.g., multiple choices), and the resulting scores are not influenced by the scorer's judgment or opinion. In fact, such test are usually constructed so that they can be accurately scored by trained clerks and scoring machines. When such highly objective procedures are used the reliability of the test results is not affected by the scoring procedures.

For classroom test constructed by teachers, objectivity may play an important role in obtaining reliable measures of achievement. In essay testing and various observational procedures the results depend to a large extent on the person doing the scoring. Different persons get different results, and even the same person may get different results at different times. Such inconsistency in scoring has an adverse effect on the reliability of the measures obtained, for the test scores now reflect the opinions and biases of the scorer as well as the differences among pupils in the characteristic being measured.

The solution is not to use only objective test and to abandon all subjective methods of evaluations, as this would have an adverse effect on validity, and as we noted earlier, validity is the most important quality of evaluation results. A better solution is to select the evaluation procedure most appropriate for the behavior being evaluated and then to make the evaluation procedure as objective as possible. In the use of essay tests, for example objectivity can be increased by careful phrasing of the questions

and by a standard set of rules for scoring. Such increased objectivity will contribute to greater reliability without sacrificing validity.

4.1. Usability or Practicability

In selecting tests and other evaluation instruments, practical considerations cannot be neglected. Tests are usually administered and interpreted by teachers with only a minimum amount of training in measurement. The time available for testing is almost always limited and is in constant competition with other important activities for its allotted time in the school schedule. Likewise, the cost of testing, although a minor consideration, is as carefully scrutinized by budget conscious administrators, as are other expenditures of school funds. These and other factors pertinent to the usability of tests and evaluation procedures must be taken into account when selecting evaluation instruments. Such practical considerations are especially important when selecting published tests.

4.2. Ease of Administration

For this purpose, the directions should be simple and clear, the subtests should be relatively few, and the timing of the test should not be too difficult. Administering a test with complicated directions and a number of subtests lasting but a few minutes each is a taxing chore for even experienced examiner. For a person with little training and experience, such a situation is fraught with possibilities for errors in giving directions, timing, and other aspects of the administrations that are likely to affect the results. Such errors of administration can have of course, an adverse effect on the validity and reliability of the test scores.

4.3 Time Required for administration

With time for testing at a premium, we always favor the shorter test, other things being equal. But in this case, other things are seldom equal, because reliability is directly related to the test length. If we attempt to cut down too much on the time allotted to testing, we may reduce drastically the reliability of our scores, For example, tests designed to fit a normal class period usually produce total test scores of satisfactory reliability, but their part scores, obtained from the subtests, tend to be unreliable. If we want reliable measures in the areas covered by the subtest, we need to increase our testing time in each area. On the other hand, if we want a general measure in some area, such as verbal aptitude, we can obtain reliable results in 30 or 40 minutes, and there is little advantage in extending the testing time. A safe procedure is to allot as much time as is mecessary to obtain valid and reliable results and no more. Somewhere between 20 and 60 minutes of testing time for each individual score yielded by a published test is probably a fairly good guide.

4.4 Ease of scoring

Traditionally, one of the most tedious and troublesome aspects of a school testing programme has been scoring the tests. In the past, many an overworked teacher has spent hours upon hours at this task. To make the procedure even more burdensome than it needed to be, scoring directions were frequently complicated, the tests contained numerous subtests, and the scoring keys were cumbersome. Although scoring tests is still a problem to be reckoned with recent developments in testing have eased the burden considerably. These developments include (1) improved directions for scoring and simpler scoring keys, (2) separate answer sheets and (3) machine scoring.

In selecting tests those that require a minimum amount of time, skill, and expense for the scoring should be given preference. Separate answer sheets, for example, not only will ease scoring but also will reduce the cost of testing, because the same test booklets can be used over. In addition, if machine scoring is available at a reasonable cost, separate answer sheets can relieve teachers of an irksome clerical task. Such factors should be taken into account when the test is being evaluated, and no test should be selected until the provisions for scoring have been given careful thought. Other things being equal, we favour the test that offers ease and economy of scoring without sacrificing scoring accuracy.

4.5 Ease of Interpretation and Application

In the final analysis, the success or failure of a testing programme is determined by the use made of the test results. If they are interpreted correctly and applied on the other hand, if the test results are misinterpreted or misapplied or not applied at all, they will be of little value and may actually be harmful to some individual or group.

Information concerning the interpretation and use of test results is usually obtained directly from the test manual or related guides, Attention should be directed toward the ease with which the raw scores can be converted into meaningful derived scores, the clarity with which the tables of norms are presented, and the comprehensiveness of the suggestions for applying the results to educational problems. When the test results are to be presented to pupils or parents, ease of interpretation and application are especially important.

4.6 Availability of Equivalent or comparable forms

For many educational purposes, *equivalent* forms of the same test are often desirable. Equivalent forms of a test measure

the same aspect of behavior by using test items that are alike in content. Level of difficulty, and other characteristics. Thus, one form of the test can substitute for the other, making it possible to test pupils twice in rather close succession, without their answers on the first testing influencing their performance on the second testing. The advantage of equivalent forms is readily seen in mastery testing in which we want to eliminate the factor of memory while retesting pupils on the same domain of achievement. Equivalent forms of a test may also be used to verify a questionable test score. For example, a teacher may feel that a scholastic aptitude or achievement score is spuriously low for a given pupil and may easily check this by administering an equivalent form of the test.

Many tests also provide comparable forms. Published achievement test, for example, are commonly arranged in a series that cover different grade levels. Although the content and level of difficulty very, the test at the different levels are made comparable by means of a common score scale. Thus it is possible to compare measurements in grade 4 with measurements in grade 6 on a more advanced form of the test. Comparable forms are especially useful in measuring development in the basic skills.

4.7 Cost of Testing

The factor of cost has been left to the last because it is relatively unimportant in selecting test. The reason for discussing it at all is that it is sometimes given for more weight than it deserves. Testing is relatively inexpensive, and cost should not be major consideration. In large-scale testing programme in which small savings per pupil add up, using separate answer sheets, machine scoring, and reusable booklets will reduce the cost appreciably. To select one test instead of another, however because

the test booklets are a few cents cheaper is false economy. After all, validity and reliability are the important characteristics to look for, and a test lacking these qualities is too expensive at any price. On the other hand, the contribution that valid and reliable test scores can make to educational decisions seems to indicate that such tests are always economical in the ling run.

Activities

1. Obtain an estimate of the reliability coefficient of the vocabulary test by computing Spearman Brown Prophecy formula between "odd" and "even" scores for 40 ninth-grade pupils and give your comments on results.

2. Construct a test of 30 items in the subject of Pakistan studies IX. Administer and evaluate it. Discuss the results of this test with an educationist of your area and prepare a report of the outcomes of your discussion.

5. SELF ASSESSMENT QUESTIONS

1. Define reliability, and describe its importance in testing.
2. Briefly describe the methods of estimating reliability. Can one method substitute for another? Why or why not?
3. For which purpose is each of the following most useful and why?
   a. Reliability coefficient
   b. Standard error of measurement.
4. Enlist and briefly describe as many things as you can, think of that might be done to increase the reliability of a classroom test.
5. What is meant by the validity of a test? Define and explain its different ways.
6. Is a valid test necessarily reliable? Explain.
7. What is meant by objectivity? How are measuring instruments made more objective?
8. Compute standard error of measurement by taking different test scores.
9. Review the criteria for a good test and show why a good test must be properly balanced in all respects if it is to serve its purpose efficiently?

Unit-3

INTEGRATING OBJECTIVES WITH EVALUATION & MEASUREMENT

By:

Prof. Salamat Ali Dogar

INTRODUCTION

In teaching process a teacher has to assess his pupils achievement and his own effectiveness of teaching. If he does not do so he will not be able to proceed progressively:

The key to effective teaching is effective achievement testing and key to effective achievement testing is careful planning. Planning of an achievement test can take many forms but both the testing specialist and classroom teachers have found the following series of steps to be most useful.

1. Determing the purpose of the test.
2. Identifying the objectives and educational outcomes.
3. Preparation of content outline.
4. Preparation of table of specification.
5. Practical test construction.

In this unit these steps are explained in detail.

OBJECTIVES

After studding this unit you will be able to:

1. Explain the purpose of test.
2. Define the term objective.
3. Explain taxonomy of objectives and educational outcomes.
4. Prepare the content outline.
5. Prepare a table of specification.
6. Explain practical consideration in planning a test.
7. List the principles of constructing test items.

CONTENTS

Introduction

Objectives

1. PURPOSE OF A TEST

1.1. Difination of objectives
1.2. Objectives & educational out comes
1.3. Taxonomy of objectives
1.3.1 Cognitive Domain
1.3.2 The Affective Domain
1.3.3 The Psychomotor Domain

2. PREPARTION OF CONTENT OUTLINE

2.1. Principles of preparing Content out line selection
2.2. Table of specification
2.3. Practical considerations in planning a test
2.3.1 Test specification

3. Construction of test items

3.1. Construction of essay items
3.2. Constructing short answer items
3.3. True-False or alternative response items
3.4. Matching items
3.5. Constructing multiple choice (MC) items

4. Summary

5. Self assessment questions

6. Bibliography

1. PURPOSE OF A TEST

The first and foremost step in teaching process is to identify and formulate the objectives of instruction. There are numerous day to day decisions that a teacher has to make that require some knowledge of the pupils' attitudes, achievement and personal development. This knowledge can only be acquired through various tests the teacher can decide "what is achieved and what is to be achieved by his pupils and in this way he approaches to an educational need of learners. Here is a model of determining the educational need of pupil and this is main purpose of a test.

Desired status of Learner	MINUS	Current status of Learner	EQUALS	Educational need

To explain further, purpose of a test, to answer the following 10 questions would be helpful. (techniques of acquiring of knowledge of pupil's attitudes achievement and personal development is given in parentheses).

(i) How realistic are my teaching plans for these group of pupils (mental ability test, past record of achievement).

(ii) How should the pupils be grouped for more effective learning? (Range of mantel ability scores. Past records of achievement).

(iii) Are pupil ready for the next learning experience? If so to what exten? (readiness test pretests /over needed skills, past record of achievements).

(iv) To what extent are pupils attaining the minimum essentials of the course content? (Mastery tests).

(v) To what extent are pupils progressing beyond the minimum essential (periodic tests, general achievement tests).

(vi) What types of learning difficulties the pupils are facing? (Diagnostic tests).

(vii) Which pupils are under achieve? (mental ability test, achievement test).

(viii) Which pupils need special classes (mental ability test, achievement test, diagnostic test).

(ix) What school mark should be assingned to each pupils (Review of all data).

(x) How effective was my teaching (Achievement test pupils rating supervisor's tating).

From the above discussion we note that test can be used in an instructional programme to assess.

a. Entry behaviour (placement test).

b. Monitor learning progress (formatives test).

c. Diagnose learning difficulties (diagnostic test).

d. Measure performance at the end of instruction (summative test).

1.1 Definition of Objectives:

Goals are general statement of purpose, or desired outcomes and these are not directly measurable. Each goal may be translated into one or more specific measurable objectives. Hence objectives may be defined as specific statements of what is to be accomplished and how well and are expressed in terms of quantitatively measurable outcomes. For example the following objective states that what is to be a accomplish and how well and quantitatively measurable.

Class

Minimal level	Atleast 75% of the pupils	Will be able	
To correct Incorrect	90% or More Student Minimal Level	Of ten grammatically simple sentences	

Second Example in Technical Field:

At the end of training period the learners will be able to drive a car for half an hour in rush hours without damaging the car and voilating the traffic rules [to be able to drive car efficiently is a goal].

Third example;

In math "95% of the class will be able to solve 8 out of 10 questions of multiplication of four digits to three digits in ten minutes.

[To improve multiplication in arithmatics is a goal]. Objectives may be process oriented or product oriented. Process oriented objectives; state outcomes desired during the execution of the effort i.e. They relate to development and execution. Where as product oriented objectives describe outcomes intended as a result of the effort. Thus, the objectives give direction to all subsequent activities and achievement of objectives is ultimately measured. Objectives whether instructional or programme objectives, from the foundation of all subsequent testing activities and hence they can be evaluated in terms of relevance, measurability, substance and technical accuracy. Objective are "Guiding light for teachers for the process test construction".

1.3. Objectives & educational outcomes.

The objectives provide direction in a number of ways i.e. They suggest general strategies and specific activities for their attainment. Different sets of objectives will generate different strategies. For example there are three English teachers A,B,C, teacher A has no specific objectives. He is simply teaching English. Teacher B has a set of objectives related to students being able to speak conversational English. Teacher C has a set of objectives related to students being able to read & write English. Just think how their class room activities differ. Teacher A may just read an English text boot loud every day. Teacher B on the other hand may speak English, playing records of conversation in English and having students converse in English. Teacher C will have to devote a lot of time to having students translate written passages from English into urdu either orally or in writing. So instructional objectives also give direction to educational outcomes. The teacher and students know what is expected of them. In the example give above; 'Teacher A will not expect specific educational outcomes. Where as teacher B and C expect specific educational outcomes. The existence of objectives definitely determine the educational outcomes.

1.3.1 Taxonomy of objectives:

As described earliar, specific objectives definitely determine the specific educational outcomes. Intelligent teachers identify specific instructional objectives to get required educational outcomes.

Educational objectives:

In 1956 Benjamin S Bloom a group of associates mostly college examiners brought out a taxonomy (classification system) dealing with educational objectives. A particular important

contribution of the 1956 taxonomy was its division of educational objectives into three domains. The cognitive affective and psychomotor. Educational have used this division widely.

1. **COGNITIVE DOMAIN** is that in which we focus on those behaviours of learners that we believe reflect inteclectual skills.

2. **AFFECTIVE DOMAIN** is concerned with the dehaviours of learner which reflect attitudes, values and interests.

3. **PSYCHOMOTOR DOMAIN** is concerned with the physical and motor skills we hope learner will require. Details of these three domains are as follow!

1.3.2. Cognitive domain:

The taxonomy o cognitive objective has definitly made educators aware of wide range of abilities involved in cognitive learning. Each of the six major categories given below represent a different kind of learning process. Description of each category is given with examples so that educators may be able to insure higher order of learning to be considered and planned for six categories of cognitive domain.

(The topic of "inflation" in the subject of Economics is taken as example)

Categery name	Brief Description
Knowledge	Memorization of previously learned material. The lower level of learning outcomes. For example the ability to define the term inflation from memory.
Comprehension	Ability to grasp the meaning of material. Interpreting paraphrasing, Exampling. For example, explain ways in which inflation affects economy.
Application	Ability to use learned material in new and current situation. For example the ability to apply general principles of inflation to the current Pakistan economy.
Analysis	Ability to break down material into component parts so that organizational structure may be understood. For example the ability to determine the common element in three different plans for reducing inflation in Pakistan.
Synthsis	Ability to put ideas together to form a new whole. Proposes integrate, designs. For example the ability to develop an original plans for reducing inflation in Pakistan.
Evaluation	The ability to judge the value of material for a given purpose. For example, the ability to give a critique of a proposed plan for reducing inflation.

1.3.3. The affective domain:

The effective domain was devised by Krath Wohl and his associates. It deals with outcomes that are very difficult to promote and to measure. It assumes that the pattern involved in a requiring values moves from very low level of awareness toward the higher level of internalization. The five categories are explained with example from the lowest level of awareness toward the kigher level of internalization.

Dategory name	Brief Description
Receiving (lowest level)	The students willingness to attend to particular phenomena or stimuli. This category is subdivided into awareness, willingness to receive and controlled or selected attention. Examples; Rustum is attention when the teacher reads a poem to the class.
Responding (Second level)	Active participatation. The learner not only attends but also reacts favourably. Example; Answers the questions related to the poem which has been read to the class.
Valuing (third level)	Demonstration of a preference belief or commitment with respect to certain behaviour phenomena, Example, chooses reading poetry as a free period activity.
Organization	Bringing together different values into a consistent and coherent value system. Example; checks reading poetry out of the library.
Characterization (higher level)	The total adoption of a value system so that it becomes a person's life style. Example; At this level it would be widely recognized that Saqib loves poetry or Saqib read poetry every chance he gets.

1.3.4. The psychomoter domain:

The psychomotor domain is concerned with the development and use of muscles and body's ability to coordinate its movements. The main element of taxonomy is as follows.

Perception	Process of becoming aware of objects, qualities, or relations through the sense organs
Set	It is preparatory adjustment for a particular kind of action or experience. Three aspects of set has been identified (mental, physical and emotional).
Guided response	Guided response is the over behavioural act of an individual under the guidance of another individual. It is an early step in the performance of an act.
Mechanism	At this level the learner has achieved a certain confidence and degree of skill in the performance of an act.
Complex overt Response	At this level the learner can perform a motor act that is complex. The act can be carried out efficiently and smoothly i.e with maximum consumption of energy and time.

Gagne,s revided model of learning outcomes.

Capability	Verb	Example
Intellectual Skill Discrimination	Discriminates	Discriminates, by matching, the French sounds of 'u' and 'ou'
Concrete concept	Identifies	Identifies, by naming, the root, leaf, and stem of representative plants.
Defined concept	Classihes	Classilies, by using a definition the concept' family'
Rule	Demonstrates	Demonstrates by solving verbally stated examples, the addition of positive and negiative numbers.
Higher-order rule (problem-solving)	Generates	Generates, by synthesizing applicable rules, a paragraph describing a person's action in a situation of fear.
Gognitive strategy	Organize	Originates a solution to the reduction of air pollution, by applying model of gascous diffusion.
Information	Status	Status orally the major issues in the presidential campaign of 1932.
Motor	Executes	Executes teaching a car into driveway.
Attitude	Chooses	Chooses playing golf as a leisure activity.

Gagne and Briggs' model of outcomes.

Type of learning Outcome	Essential Prerequisites	Supportive Presequisites
Intellectual skill	Simper component Intellectual skills (rules concepts, determination)	Attitudes Gognitive strategies Vertial information.
Verbal information	Meaningfully orgarfized Sets of information.	Language skills Gognitive strategies Attitude
Cognitive strategies	Specits mettectual skills	Intellectual skills Verbal information attitudes
Attitudes	Intellectual skills (sometimes)	Other attitudes Verbal information
Motor skills	Park skills (sometimes) Procedural rules (sometime)	Attitudes.

Essential and supportive prerequisiles for five kinds of learning outcomes.

Preparation of content outline.

When instructional objective and outcomes have indentified, the next step in test construction is planning the test. The sooner you begin planning and the more care you take, the more likely it is that you will develop a good test. Given a set of related objectives to major task is to develop items which measure only behaviours specified in that objective which adequately sample the domain of possible behaviors. The next step in planning attest is to prepare a detailed content outline. For example a teacher decides that the

test would cover the topic. "The United State Congress," the outline may include "the senate and the house of representatives" as main headings. The outline may be i. developed by the teacher's ii. Provided in a curriculum guide. The outline need to be detailed enough to include all knowledge and skill outcomes desired of learner. For example, the heading "the senate" might include such sub heading as, i. "Number of senator from each province, ii. Eligibility requirements,

iii. Electoral cooege, constitinol power and limitations.

2.1. Principles Of Preparing Content Outline Selection

1. Identify the objectives
2. Compare the objectives with educational outcomes.
3. Be specific in three domains of taxonomy of objectives
4. Formulate the objectives.
5. Categories the main topics into as many sub-heading as possible
6. Compare the sub-headings with the specific objectives formulated before.
7. Put boundaries what is to be covered in the test.

The more structured and well defined the content is, the easier is this to accomplish.

(L. R. Gay)

These principles lead us to the table of specification. When you have practiced in preparation of table of specification you will fully be skilled in preparation of content outline according to the three domains of objective.

2.2. Table Of Specification

To ensure the valid measure of the instructional objectives and course content we are interested in testing. We should use representative sample of pupils performance in each of the areas to be measured. The instrument that is widely being used for this purpose is the table of specification. Table of specification is a chart which relates the instructional objectives to the course content and specifies the relative emphasis to be given to each type of learning out- comes, usually expressed in percentage.

Preparation of table of specification includes

1. Obtaining the list of instructional incluses (identification and formulation methods have already been discussed)
2. Out lining the course content. (Already discussed in 3,4)
3. Preparing the two way chart.

Preparing the two way chart relates the instructional objectives to the course content and thus specifies the nature of the test sample. This chart shows what percentage of items is to be allotted to what course content according to specific objective.

Table 3.2 Table of specifications for a Summative third Grade Social Studies Test (in percentage)

The tables of specification are givben as samples by Morman F. Gronlund.

Objectives						
Content Area	Knows Common Terms	Know specific Facts	Understa-nds Principle and Generali-zations	Applies Principle and Generali-zations	Interprets Chart and Graph	Total
Food	2	6	2			10
Clothing	2	6	2			10
Transportation	4	2	2	2	5	15
Communications	4	2	2	2	5	15
Shelter			5	5		10
City life	4	2	6	8		20
Farm life	4	2	6	8		20
Total	20	20	25	25	10	100

Table 3.2. Table of specifications for a Wether Unit in Junior High School Science

Objectives							
	Knows		Understa-nds	Interprets		Skill in	
Content	Symbols and terms	Specific facts	Unfluence of each factor on weather formation	Weather maps	Use of measu-reing devices	Construc-ting weather maps	Total Number Of items
Air pressure	2	3	3	3	Observe pupils using measu-ring devices (rating scale)	Evaluate maps construct-ed by pupils (checkinst)	11
Wind	4	2	8	2			16
Temperature	2	2	2	2			8
Humidity and precipitation	2	1	2	5			10
Clouds	2	2	1				5
Total number of items	12	10	16	12			50
Percent of evaluation	12%	10%	16%	12%	25%	25%	100%

2.3. Practical Considerations Is Planning A Test

If a teacher wants to prepare a test which is valid, reliable and useful he should carefully plan the test. The practical considerations in planning a test include the following areas.

1. Determining the purpose of the test.
2. Preparing the test specifications.
3. Selecting the appropriate item type.
4. Preparing relevant test items.

As determining the purpose of the test is the first step in planning procedure, the teacher should be specific in deciding the type of test out of four basic types. In other words, he should decide whether he is going to prepare, (i.)The placement test, (ii.) formative test,

(iii.) Diagnostic test, (IV) summative test (These four basic types have already been explained in unit-1).

2.3.1. Test specification:

When the purpose of the test (type of test) has been decided, the next step is development of test specification. The procedure of developing test specification is as follows.

1. Identification of instructional objectives (explained in 1.1, 1.2)
2. Preparation of outline (content outline) explained in section 2.
3. Matching the instructional objectives with course out line (Content outline).
4. Preparation of table of specification (explained in 2.2)

5. Deciding the number of items to be included in the test. The following table may serve as guide line for deciding the number of items.

Subject **math** **topic** **addition of fractions**

Instructional Object Content Area	Add fraction	Add fraction & mixed No	Add mixed number	TOTAL ITEMS
Denominators are alike	5	4	4	13
Denominators are unlike (with common factors)	5	4	4	13
Denominators are unlike (with out common factors)	6	4	4	14
TOTAL Items	16	12	12	40

Obtain a representative sample of items. When a teacher has decided about the number of items to be included in the test he should expect that his pupils know many facts but be can test for only a limited number of them. In each area content and for each specific learning outcomes a teacher can merly select a sample of pupil's performance and accept it as evidence of achievement in that area. Hence what of type of test may be, our sampling is most likely to be representative when test preparation is guided by a

carefully prepared set of specification. Length of a test will also be an important factor in obtaining a representative sample.

3. CONSTRUCTION OF TEST ITEMS:

The tests constructed by class room teacher are classified as objective test and essay tests. These may be further subdivided as follows.

1. Objective test:

A: Supply Tupe (i) Short Answer (ii) Completion

B: Selection Type (i) True false or Alternation response item

(ii)Matching (iii) multiple choice item

2. Essay type:

A: Extended Response

B: Restricted Response

In unit 4 you will see this classification s follows;

1. SLECTION TYPE (objective type)

i) Multiple Choices
ii) True false question
iii) Matching items
iv) Completion

2. SUPPLY TYPE: (essay type)

i) Extended response
ii) Restricted response
iii) Short answers

L.R.Gay and many other authors put "short answers type" under objective tests.

Now we discuss the Essay Type Tests and Objective Type Tests in detail

ESSAY TESTS: An essay test is one which contains items which require the student to compose responses, usually lengthy. Essay tests are not really appropriate for knowledge level outcomes. These items typically require both the recall and use fo information for demonstrating a higher order outcomes. The scoring of essay type items is some what subjective because there is no one and only one correct answer or even two and only two correct answers. Essay responses are Pengthier than objective test limited questions can be asked in a given of time. So minimum course outline can not be included in the test. Thus these tests are generally less content valid and reliable.

3.1 Construction of essay items:

Essay items are easy to construct but tough to score so very carefully preparation of essay items over come the scoring difficulties. Here are some suggestions for preparation essay items.

1. Avoid giving students a choice. Giving choice in essay items is not good practice.
2. The response to the item should be clear. Do not ask students simply "discuss" but tell them what to discuss. For example: Do not ask students "discuss the rules for preparing a time table". But ask "develop a time table for class X when time duration is 6 hours and elective subject are.....".
3. One essay item may be useful to measure more than one objective.

Example: Assume that you are going to prepare a formative test for a unit of work in a course in your major teaching area.

How would you proceed? How would you procedure differ it were to be an end of course?

4. Aspects of expected performance should not be left to the student's imagination.

5. If you want students to give examples or draw diagrams include instructions in the item. Allocate marks for examples and diagrams. For example describe the parts of flower. Draw diagrame where necessary.

6. All the question should carry equal marks. One question may subdivide and marks should be specified to these subdivisions.

7. Maximum restricted items be included in essay type test.

_____Some examples of restricted items_____

1. Given a list of foods indicate which are high in protein.

2. Given a list of modern novels indentify the athor of each.

3. List four major differences between a standardized test and a locally developed test.

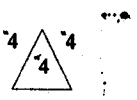
3.2. Constructing short answer item:

The short answer items are those items which can be answered by a word, phrase, number or symbol. For example: What is the name of the man who invented telephone?

The name of the man who invented telephone is.....

What is the symbol of Hydrogen?

What is the number of degrees in each angle of this triangle?



Suggestion for constructing short answer itmes:

Following suggestions will provide greater assurance that the item will function as intended.

1. Work the item so that the required answer is both frief and definite. The answer to an item should be a word, phrase, number or symbol.

2. Do not take statements directly from text books to use a basis for shout answer items.

3. Avoid as many incomplete statements as possible.

   Poor: - Allama Iqbal was born in..... .

   Better: - Name the city where Allama Iqbal was born?

4. Where the answer is required in numerical unit, indicate the type of answer wanted

   Poor: - If oranges weight 52/30 each, what will be the weight of a dozen oranges.

   Better: - If oranges weight 52/30each, What will be the weight of a dozen oranges.

5. Blanks of answer should be equal in length and in a columm to the right of the question.

6. Where completion items are used do not use too many blank (It is better to require one answer in one blank).

   Poor: - Quid-I-Azam was born in..... .

   Better: - Quid-I-Azam was born at..... .

2nd Example:-

Poor:-.....blooded animal that are born alive and.....their young are.....

Better: - Warm Blooded animals that are born alive and suckle, their young are called..... (mammals)

3.3. True false or alternative response items:

In this type of items declarative statement is given and there are only two possible answers. Pupil is asked to mark true or false, right or wrong, correct or incorrect, yes or no, etc. Most common items are true or false option. These items, measures the ability to identify the correctness of statement of fact, definitions of items, statement of principle and the like. It is very difficult to formulate the statements which are free from ambiguity and irrelevant clues. Here are some suggestions for the guidance of teachers.

Suggestion for Constructing True-False Item

1. Avoid board general statement. In true statements words such as "usually" "generally" "aften" and "sometime" are most likely to appear. Similarly words "always" "never" "all" and "only" most likely to appear in false statements.

Example:-

T-F the President of Pakistan is elected to the office. (poor)

T-F the President of Pakistan is usually elected to the office. (Better)

2. Avoid the use of negative statements and specially double negative.

Example:-

T-F Distance covered in a unit time is not velocity. (poor)

T-F Distance covered in a unit time is velocity. (better)

T-F Distance covered in a unit time is not speed. (poor)

T-F Distance covered in a unit time is speed. (better)

3. Avoid long and complex statement.

Example:-

Poor: - T-F for determining the exact PH value of a solution, it is possible to determine if a solution is acid by the red color formed on litmus paper when it is inserted into the solution.

Better: - T-F Litmus paper turns red in an acid solution.

In the above example calculation of PH value may be asked in another item.

4. Avoid including two ideas in one statement. If we include two ideas the pupil will feel difficulty to judge the statement.

5. If opinion is used then quote the source of that opinion. Similarly if definition of a term is given by many authors then quote the author whose definition is used.

6. True and false statement should be approximately equal in length and number.

3.4. Maching Items

Catching items are consisting of two vertical lists referred to as the premises and the conclusions or responses. These items can be used to measure outcomes at all levels. These are usually used when a number of related facts, such as knowledge of Titles and authors are to be measured. In two vertical lists for each premise, there is a response. These two vertical lists are named as column A and B.

Suggestion for preparing matching items

1. Use only homogeneous material in a single matching item. The premise and conclusions of matching item should be as

homogeneous as possible. In other words they should have a common theme.

Example: (Poor)

Column A	column B
1. Invented telephone	Christopher Columbus
2. Discovered Amerca	Graham Bell
3. First USA astronaut to orbit the earth	Abrahim Lincon
4. The first president of USA	John Glemn
	Ferdinand Magellan
	George Washington
	Eli Whitrey

(Better)

Column A	Column B
1. Simon Commission report	A. 1909
2. Second Round Table Conference	B. Oct 1924
3. Communal Award	C. March 1930
4. Higrat Movement	D. June 1930
5. Minto Morely Reforms	E. 1931
	F. 1932

2. Include an unequal number of responses and premises as given in above example.

3. Keep the premises on the left and responses on the right.

4. Keep the responses in the logical order as given in the above better example.

5. The basis for matching the premises and responses should be indicated in the directions. If the basis of matching is given confusion and ambigiuts will be avoided and testing time will be saved.

6. Direction should clearly be stated for matching items.

7. The premises column and the response column should be on the same page. This will prevent the disturbance of switching the page of the test back and forth.

Example:-

column A Element	Column B Symbol
Iron	**I**
Hydrogen	**F**
Iodine	**H**
Sodium	**S**
Sulphur	**Na**
	Cl
	Hg

3.5. Constructing Multiple choice (MC) ITEMS:

In multiple choice items a problem is stated in the form of a question or a statement and list of suggested solution is given. The statement or problem stated is called stem. The list of suggested solution are called alternatives. The correct alternative in each item

is called answer and remaining alternatives are called distracters. Again remember, these used in multiple choice item.

1. Stem
2. Alternative (correct is answer)
3. Distractors

The pupil is directed to read the stem and list of alternatives and to select one correct or best alternative.

The following outcomes can be measured through multiple choice items.

1. Knoledge

i. Knowledge of terminology.

ii. Knowledge of principles.

iii. Knowledge of method & procedures.

2. Understanding & Application

i. Ability to indentify application of factor and principles.

ii. Ability to interpret cause and effect of relationships.

iii. Ability to justify methods and procedures.

Following examples explain these outcomes.

i. (i)The distance covered in a unit of time is called
   a. Velocity b. Speed c. Displacement d. Acceleration

ii. Who was the first united states astronaut to orbit the earth in space.
   a. Scott Carpenter
   b. John Glenn
   c. Vergil Grissom
   d. Alan Shepard

iii. The principle of capillary action help explain how fluids.

a. Enter lower concentration solution

b. Escape through small openings

c. Pass through semipermeable membrane

d. Rise in fine tubes

iv. AC current is changed to DC by means of a

a. condenser b. generator

c. rectifier d. transformer

2. (i) Which of the following is a chemical element

a. Sodium Chloride b. Acid

c. water d. oxygen

ii. Pascal's law is used in

a. Electric fan b. Springs c. Hydraulic Braked d. levers

iii. When a fuel is burned in limited supply Oxygen quantity of Carbon Monoxide increases because

a. Corbon reacts with Carbondioxide

b. Greater Oxidation takes place

c. Carbon react with Corbonmonoxide

d. Corbonmonoxide is effective reducing agent

iv. Farmers rotate their crops

a. to conserve the soil b. to provide for strip cropping

c. to make marketing easier d. to provide more working conditions

Suggestions for constructing:

Multiple - Choice Items

The stem should be meaningful and should present a definite problem. Compare the stems in the following example!

Ammeter **(Poor)**

a. is an instrument used for measuring Amperes.

b. Is an instrument used for measuring Volts.

c. Is an instrument used for regulating Voltage.

d. is an instrument used for regulating deflection

The instrument for measuring Amperes is **(Better)**

a. Voltmeter b. Ammeter

c. Galvanometer d. Wattsmeter

2. The stems should include as much of the items as possible and should not have irrelevant material.

EXAMPLE:-

The reason for setting Spanish in south

America was **(Poor)**

a. They were adventure us.

b. They wanted lower taxes.

c. They were in seach of wealth.

d. They were seeking religious freedom.

Spanish colonists settled most in south America in search of **(Better)**

a. Lower taxes. b. Adventure

c. Wealth d. Religious freedom.

3. **Avoid negative statement in the stem:**

Example: (for **class iii)** **(Poor)**

Which one of the following city is not located north of Multan.

a. D.G.Khan b. Bahawalpurc. Muzaffar Ghar d. Vahari

Which one of the city is located south of Multan **(Better)**

a. D.G.Khan b. Bahawalpur

c. Muzaffar Ghar d. Vahari

Which one of the following is not chemical compound **(Poor)**

a. Nacl b. Hcl c. H_2SO_4 d. Fe

Which one of the following is element **(Better)**

a. Nacl b. Hcl c. H_2SO_4 d. Fe

4. **All the distractors should be homogeneous:**

Example:

An electric transformer can be used to

a. Storing electricity b. Increase the voltage of alternating current

c. Magnetise iron bar from AC d. Charge batteries

In this example distracters C & D are not homogeneous.

Example: (Better)

An electric transformer can be used to

a. storing electricity b. Increase the voltage of alternating current

c. convert Electrical energy to mechanical energy

d. Change AC to DC

5. VERBAL ASSOCIATION BETWEEN THE STEM AND THE CORRECT ANSWER SHOULD BE ASVOIDED:

Example: **(Poor)**

You can inform about the violent wind storm the following agencies in your locality.

a. Weather Station b. Local Radio Station

c. Post Office d. Police Station

You can inform about the violent wind storm the following agencies in your locality.

a. Local Weather Station b. Nearest Radio Station

c. Local Post Office d. Local Police Office

Local and locality is verbal association.

6. Length of alternatives should possibly be equal

7. Use of "None of these" or "all of the above" should be avoided in alternatives.

8. All the alternatives should be grammatically consistent with the stem of the item.

Example: 4

4. Summary

Following steps are found very useful for test construction.

1. Determining the purpose of the test.

2. Identifying the objectives and educational outcomes.

3. Preparation of content outline.

4. Preparation of Table of specification.

5. Practical test construction.

The test can be used in an instructional Programme to

1. Assess entry behavior (Placement Test)
2. Monitor learning progress (Formative Test)
3. Diagnose learning difficulties (Diagnostic Test)
4. Measure performance at the end of instruction (Summative Test)

Specific objectives definitely determine the specific educational outcomes. In 1956 Benjamin S. Bloom and his associate examiners brought out Taxonomy of educational Objectives into three domains:

The Cognitive Domain is that in which we focus on those behaviours of learners that we believe reflect intellectual skills. The subcategories of Cognitive Domain are:

i. Knowledge
ii. Comprehension
iii. Application
iv. Analysis
v. Synthesis
vi. Evaluation

The Affective Domain deals with outcomes which relate to equisition of values. The five categories of the domain are:

i. Receiving
ii. Responding
iii. Valuing
iv. Organization

v. Characterization

The Psychomotor Domain Is concerned with the development and use of muscles and body's ability to coordinate its movements. The five subcategories of this domain are

i. Perception

ii. Set

iii. Guided response

iv. Mechanism

v. Complex overt behavior

In preparing content outline consider the following principles

1. Identify the objectives.
2. Compare the objectives with educational outcomes
3. Be specific in three domains of taxonomy of objectives
4. formulate the objectives
5. Categorize the main topics into as many subheadings as possible
6. Compare the sub headings with the specific objectives formulated before.
7. Put boundaries what is to be covered in the test.

These principles lead to the table of specification.

Preparation of table of specification includes;

i. Obtaining the list of instructional objectives.

ii. Outlining the course content.

iii. Preparing the two way chart.

5. SELF ASSESSMENT QUESTIONS

(A) 1. List three domains of objectives according to Bloom's Taxonomy.

2. List sub-categories of cognitive domain.

3. Prepare a list of the main elements of Simpson's Taxonomy in psychomotor domain.

4. Make a list of seven principles for preparing content outline selection.

5. Describe three points to be considered in preparation of table of specification.

6. Select a unit from an elementary school secience and prepare a table of specification.

7. List the types of objective test item.

8. What are the classifications of essay type items.

9. Give some suggestions for improving essay type tests.

10. What outcomes can be measured through multiple choice items?

(B) From the following behaviors Tick C if it is a cognitive and A if it is an affective behavior and P if it is Psychomotor behavior

1. Recites the preamble to the ---- ---- ----
   Constellation of Islamic republic of Pakistan ---- ----
2. Balances chemical equation ---- ---- ----
3. Concentrates in a film about Mosquitoes ---- ---- ----
4. Prepare an income Tex return ---- ---- ----
5. In Terprcts a graph ---- ---- ----
6. Buys only organic foods ---- ---- ----
7. Knows the meaning of real signs ---- ---- ----
8. Defines the Term velocity ---- ---- ----
9. Types a business better ---- ---- ----
10. Knows the food Sources of Iron ---- ---- ----

For each of the following statement Tick G, if it is goal, O, if it is objective and S if it is specific objective. G O S

1. Is economically self sufficient ---- ---- ----
2. Is a good Citizen ---- ---- ----
3. Read a paragraph of twenty lives of a book in 5 minuets ----
4. Writes grammatically correct Sentences ---- ---- ----
5. Has command of basics skills ---- ---- ----

(C) Read the following items and Tick ok if you think correctly written and not ok if you think a role for item construction has been violated short answer item

1. If 5kg apples Cost Rs 90, how much one Kg of apples cost ?
Ok not Ok

2. If x+ 12 =20, then x= Ok not Ok

3. Allama Iqbal was born in ---- Ok not Ok

Multiple Choice items

4. Ammeter

A. is an instrument used for measuring Amperes

B. is an instrument used for measuring Volts

C. is an instrument used for measuring Voltage

D. is an instrument used for measuring fornicating deflection

The reason for settling Spanish in south America was

A. They were adventurous

B. They wanted lower Taxes Ok not Ok

C. They were in search of wreath

D. They were seeking religions freedom

5. Which one of the following is element

A Nacl B. HCL C-H_2So_4 D Fe Ok not Ok

7. An electric Transformer can be used to

A. Storing up electricity

B Increase the Voltage of A.C.

C. Convert electrical energy to mechanical energy

D. Change AC to DC Ok not Ok

6. BIBLIOGRAPHY

1. Gay. L. R. Educational Evaluation and Measurement. Macmillan publishing company, 1991.
2. Gronlund. E. Norman Measurement and evaluation in teaching. Macmillan Publishing. Co New York 1981
3. RaQ. V.K & Reddy R.S Instructional objectives and teacher education. Common wealth Publishers New Delhi (India) 1992
4. Romiszowski A.J. The selection and use of instructional media koganpage London, 1988.
5. Gronlund. N.E. Constructing Achievement tests Englewood cliffs. N.J. Prentice. Hall Inc 1977
6. Shahid S. M Majeed Book Depot. Urdu Bazar Lahore 2000

Unit-4

TYPES OF TESTS

BY:

S.M.SHAHID

INTRODUCTION

There are basically two broad categories of tests: essay and objective. Essay tests allow students to express themselves freely in answering particular questions. To a large extent, the emphasis is on students' over- all understanding of the subject in question. In an objective test, however, students' responses are restricted to a number of symbols, words, phrases or simple sentences, one of which is considered to be the best answer out of several plausible alternatives.

This unit deals with the type of tests and the discussion has been divided into the following topics:

- Essay type Examination
- Objective type Examination

OBJECTIVES

After reading this unit you will be able to:-

1. Describe the different types of tests
2. Explain the main features of essay type tests.
3. Describe the important items of objective tests.

CONTENTS

Introduction

Objectives

1. Essay Type Examination

1.1 Meaning of Essay tests
1.2 Classification of Essay Questions
1.3 Merits of the Essay Tests
1.4 Limitations of Essay Type tests
1.5 Suggestions for the Essay Type Tests
1.6 Short-Answer Type Tests
1.6.1 Characteristics of short - Answer test
1.6.2 Forms of Short -Answer Test
1.6.3 Advantage of short-Answer Test
1.6.4 Limitations of short Answer Test
1.6.5 Suggestions for construction of Short -Answer Test
1.7 Main point

2 Objective Type Examination

2.1 Meaning of Objective tests
2.2 Characteristics of Objective Tests
2.3 Forms of Objective Tests
2.3.1 True &False Tests
2.3.2 Matching type test
2.3.3 Multiple Choice Type Test
2.3.4 Comparison of the Essay and Objective Type Test
2.4 Main Point

3 Self Assessment Questions

4 Bibliography

1 ESSAY TYPE EXAMINATION

Examination may be classified on the basis of type of questions i.e. essay type, objective type and oral type etc. The essay question requires the student to provide a written discussion in response to the question. The traditional final examination normally contained essay type tests.

1.1 Meaning of Essay tests

Essay tests have a long history that dates back to more than four thousand years. W.E. Coffman in article entitled "Essay Examination" published in a book education Measurement (1971), edited bay R.L. Thorndike observes that essay tests were in use earlier than 2300 B.C. Until the turn of the 21st century, they were al most the only for of written examination.

Very few attempts have been made to define and clarify the concept of essay tests. They have been used so widely that it is assumed that every body understands its meaning.

(1) Robert. L Ebel and David A. Frisbie in their book Essentials of Educational Measurement (1986) writes, "An essay tests presents one or more question or other tasks that require extended written responses from the persons being tested."

(2) Glibert Sax in Principle of Educational and Psychological Measurement and Evaluation (1989) states, "Essay test is a test containing questions requiring the student to respond in writing. Essay tests emphasise recall rather than recognition of the correct alternative. Essay tests may require relatively brief responses or extended responses"

(3) William Wiersma and Stephen G. Jurs in their book Educational Measurement and Testing (1990) writes, "Essay i m is an item that requires the student to structure a rather long written response up to several paragraphs."

It is very difficult to give an exact and perfect definition of an essay test. Usually an essay test refers to any written test that requires an examination to write several paragraphs or passages. However in some cases answer may constitute even a sentence or a paragraph. Weidemann (1993) observes that an essay type question may use the following eleven words, signifying the simple to higher mental processes.

(a) What, Who, Which and Where

(b) List

(c) Outline

(d) Describe

(e) Contrast

(f) Compare

(g) Explain

(h) Discuss

(i) Develop

(j) Summarize and

(k) Evaluate

1.2 Classification of Essay Questions

According to Norman E. Godman (1985) essay question may subdivided in to two broad types -extended response and restricted response.

A. In extended response question, no restriction is placed on the student as to the points he will discuss and the type of organization he will use. An extended response type of essay question permits a student to demonstrate his ability to

   I. Recall factual knowledge;

II. Evaluate his factual knowledge;

III. Organize his ideas:

IV. Present his ideas in a coherent and logical way.

An example of an extended response question: Describe at length the state of affairs in the Golden Age of the Muslim Period in Indo-pak.

(B) Restricted Response. In restricted response, the student is restricted as to the form and scope of his answer because he specifically told the context in which his answer is to be made. For example: Describe in not more than 100 words the state of affairs in the Golden Age of the Muslim period in Indo- Pak.

Restricted response question tend to be more objective while the extend type question tend to be more subjective.

The restricted response questions are more useful in measuring more specific learning outcomes but they cannot measure learning outcomes which stress integration. The extend response question are more useful in measuring complex behaviors which cannot be done by other types of tools but how accurately they do so is controversial.

W.S. Monroe and R.E. Carter (1993) list the following types of questions.

1. Analysis
2. Application of laws, principles and rules to new situations
3. Cause or effect
4. Classification
5. Comparison of two ideas or things in general
6. Comparison of two things on a single basis
7. Criticism as to the adequacy, correctness or relevance of a statement.

8. Decisions of and against

9 Explanation of the use or exact meaning of some word, phrase of statement

10. Evaluation recall- bases given

11. Formulation of new question - problems and discussions raised

12. IIlustration of examples.

13. Inferential thinking

14. New methods of procedure

15. Outline

16. Re- organization of facts

17 Selective recall –basis given

18. Statement of an author's purpose in the selection or Organization of material

19 Statement of relationships

20. Summary of some unit of the textbook or of some article.

1.3 Merits of Essay Tests

1. Abilities like logical thinking, critical reasoning and systematic presentation, etc can be best evaluated by the essay type of tests.

2. They provide an opportunity to the student to show their originality of thought as they are permitted freedom of response.

3. They help to develop good study habits such as preparing outlines and summaries, organization arguments for and against a topic

4. They provide opportunities to students to develop abilities such as to organize ideas effectively, to criticize or justify a statement and to interpret, etc.

5. They can be successfully employed for evaluating the performance of students in all the school subjects.

6. It is relatively easier to prepare an essay type test than to prepare an objective type test.

7. It is relatively easier to administer an essay type test.

8. It takes relatively lesser time to mark an essay type test.

9 Guessing is eliminated to some extent.

10. The students cannot guess the answer because they have to supply it.

11. They give examination freedom to respond within broad limits

12. They can measure divergent thinking.

13. They require less time for typing, duplicating or printing. They can be written on the blackboard also if the number of question and student is not very large.

14. It is more economical to use essay type tests than objective tests.

15. They can measure complex learning outcomes, which cannot be measured by other means.

16. They stress integration and application of thinking and problem solving skills.

17. They can be used as an instrument for measuring and improving expression skills and language of the examinees.

18. They are more helpful in evaluation the quality of the teaching process.

19. Students focus on learning broad concept and articulating relationship, comparing and contrasting.

20. They set better standards of professional ethics for teachers because they require more time in assessing and scoring.

21. They provide less scope for unfair means.

1.4 Limitations of Essay Type Tests

1. They generally stress the length enumeration of memorized facts.

2. They have limited content validity because only a sample of questions can be asked in an essay type test.

3. They are difficult to score objectively because the examinees have wide freedom of expression.

4. There is the lack of consistency in judgments even among competent examiners.

5. They have 'halo' effect which implies that the examiners judgment in evaluating one characteristic is influenced by another characteristic.A well behaved student on account of his behavior may get more marks.

6. They have 'question to question carry effect'. A student who fives the best answer in the beginning of the answer book is likely to get more marks in the subsequent and vice versa.

7. They have 'examinee to examinee to carry effect'.Which means the a particular student may get marks not only on the basis of what he has written but also on the basis of the answer of the previous student.

8. The examiners may be influenced by the language of the examinees. The quality of handwriting of the examinees may also influence the examiners. The length of the answer rather than the depth of the content may also influence marking.

9. Some examiners are too liberal in larking and some too strict.

10. Sometimes it is said that the 'mood' of the examiners also influence marking. Immediate happy events in the family or the job may motivate the examiner to be more generous. A quarrel in the family may lead to the award of low marks.

11. Essay type of test may not provide a true picture of the comprehension level of the examinee. Some students cram answer and write the same in the examination and get good marks.

12. They are time consuming both for the examiner and the examinee.

13. The speed of writing may influence the performance of the students. Students who are slow in writing may not be able to provide answer to the entire question in the limited timetable allotted for the paper. This results in low scores although the students may be knowing the correct answers of all questions.

1.5 Suggestions for the Essay Type Tests

As Coffman (1971) has pointed out, "Essay examinations are still widely used inspite of more than a half century of criticism by specialists in educational measurement and evaluation."

There is no doubt that inspite of their limitations, the essay type questions

Have come to stay and are being extensively used. It is, therefore, very essential that all possible efforts are to improve them so that the desired results are achieved to the maximum. Suggestion for improved essay type tests are categorized under the following two heads:

1. Suggestions for constructing a good essay type test.
2. Suggestions for scoring essay type tests.

1. Suggestion for constructing a good essay type test.

(i) The test constructor or the paper setter as he is commonly called should prepare ideal answer to all the question given in the essay type test. It becomes all the more essential when different examiners are required to examine the answer books. This step is needed to ensure two things.

(a) The approximate time requires to complete the paper and

(b) To provide uniformity in marking by different examiners.

(ii) The expected length of the answer of each question should be indicated on the paper.

(iii) While preparing questions, it should be kept in mind that the maximum subject matter content is covered.

(iv) There should be no overall choice in the question paper. Choice should be given for each question or it should be section-wise.

(v) Question should be such, as they required the examines to show a reasonable command over the essential knowledge of the subject matter being evaluated.

(vi) Questions should be so worded that all the examinees interpret them in the same way as the examiner wants.

(vii) It is sometime advocated that question requiring opinion should not be asked as incase of difference in opinion of an examinee from that of the examiner an examinee may suffer.

(viii) Questions should be every explicit so that the examinee may know the intention of the examiner in asking the question and accordingly he may give the answer.

(ix) The examiner should clearly indicate the weight of each part of the question so that the examinees may determine the time to be devoted to each part of the question.

(x) The number of question may be large and the expected answers within reasonable limits.

(xi) The use of essay type question should be restricted to those learning outcomes which cannot be measured by objective type items so satisfactorily.

(xii) Reasonable amount of time should be allowed to ensure that the essay type test does not become a test of speed in writing. It should be a power test rather than a speed test.

(xiii) It should be a balanced paper in the sense that it includes different types of questions: essay type, short answer and very short answer type.

(xiv) It should include a variety of question to assess abilities like knowledge, understanding, analysis, synthesis and application.

(xv) The question should not be too general vague and comprehensive. The should define the task for the candidate and indicate clearly the scope of the answer.

(xvi) A large number of short and more specific questions requiting short or limited answer should find a prominent place in the paper and be preferred to a few long and general questions

(xvii) The time allowed for the examination should be carefully considered in relation to the amount of writing required on the part of the students. The examination should be so timed that the students. The examinations should be so timed that the students are usefully engaged for the whole duration.

(xviii) The questions may be arranged in order of difficulty, i.e., from the easier to the more difficult.

(xix) Clear instruction to the student regarding the number of question to be attempted, marks given to each part thereof, and

marks reserved of any special purpose, such as diagrams, neatness, etc., may be given in the beginning of the question paper.

(xx) As far as possible, equal marks may be allotted to each question.

(xxi) Question paper may be reviewed before it is handed over to the person concerned. Quite a few points may strike the paper setter on this second reading. A paper setter should safeguard his reputation as a setter by taking all the care that is possible at the beginning.

2. Suggestion for scoring Essay Type Questions

1. The examiner should prepare a scoring for sub-examiners, Which may contain an outline for the ideal answer.
2. Normally no weight age should be given to handwriting, spelling and better language in papers other than language ones.
3. When any weight age is given to any aspect in a language paper, it should be clearly indicted in the paper.
4. The identity of the examinees should be kept secret from the examiner so that 'halo effect' does not affect scoring.
5. The examiner should go through a sampling of the answer books of the examinees before he actually starts marking the answer books of the examinees before he actually starts marking the answer books so that he may have a general ideas of the quality of the answers.
6. It is sometime suggested that a particular question of all the examinees should be evaluated at one time then the second question of all the examinees and so on. The present practice of scoring all question of one examinee maybe abolished.

7. To avoid fatigue, only a reasonable number of answer books should be marked in one setting by an examiner.

8. There should be consistency in assessment.

3. Measures for tackling the Problem of Memorization

1. Identification of instructional objective of the subjective matter or topic in question.

2. Giving proportionate weight age to each of the instructional objective in the framing of the question paper.

3. Framing question testing different objective equivalent to the weight age given to each.

4. Measures for Tackling the Element of Subjectivity

Preparing a policy statement of the question paper by the examining agency and passing it onto the paper setter. The policy statement may cover the following aspects:

(a) Weight age to major areas of content.

(b) Weight age to objectives of each area of content.

(c) Weight age to different forms of question i.e. essay, short answer and objective type.

(d) Weigh age to different forms of objective type questions.

(e) Elimination of overall options,

(f) Providing options to individual question or within a part of the question paper. Option should be comparable in respect of objective to be tested, content area covered, from, difficulty level, expected time for answering it, etc.

(g) Wording question specifically so as to delimit the scope and length of the answer.

(h) Preparation of a scoring key and a marking scheme.

5. Steps to overcome the issue of poor coverage of the contents

(i) Giving proportionate weight age to the major area of content.

(ii) Including increasing number of short-answer and objective type tests for covering large area of content.

(iii) Eliminating overall options in the question paper.

6. Steps to overcome administrative problems

1. Appointment of those persons as paper setters who are well versed with the concept and techniques of evaluation.
2. Moderation/ reviewing of the question paper so as to ensure that no question is out of syllabus and also to ensure the balanced nature of the questions.
3. Introduction of mechanized process of results.
4. Introduction of a system of grading in place of giving marks.

1.6 Short-Answer Type Tests Meaning:

In a simple language it may be stated that a short-answer type is between an essay type test and an objective test. Here we mention a few definitions giving by experts to have a comprehensive view.

1. Anthony J.Nitku (1983) in his book "Educational tests and Measurement" writes. Short answer item require the examinee to respond to the item with a word, Short phrase, number or a symbol"
2. Norman E. Gronlund (1985) in book "Measurement and evaluation in Teaching" Writes

"The short answer item. Supply test items that can be answered by a word, phrase, number or symbol".

3. William Wiersma and Stephen G. Jurs (1990) in their book "Educational Measurement and Testing" write

"Short answer items are considered objective items in that the correct response can be secured objectively. That is, preferably there is a single correct answer so that equally important scores would agree on the correctness of a response

4. K.D. Hopins, J.C. Stanley and B.R. Hopkins (1990) in their book "Educational and psychological measurement and Evaluation". Write "The short-answer test is an objective test in which each item is the form of a direct question, a stimulus word or phrase, a specific direction, a specific problem or an incomplete statement or question. The response must be supplied by the examinee rather than merely identified from a list of a suggested answer supplied bay the teacher".

If we see all the above- mentioned definitions we find that short answer have some characteristic of essay type tests and some characteristics of objective test. However, it is not fair to call them either essay type test or objective type test.

1.6.1-Characteristics of Short—Answer Tests

The main characteristics of short answer tests are as follows:

(a) The tests can be answered by a word, a phrase, a number or symbol.

(b) The test has supply responses rather than select or identify.

(c) The test is in the form of a question or incomplete statement.

(d) In continuum, it stands between Essay type test and objective type test.

1.6.2 Forms of short-Answer Test

The short answer test can be made in the following forms;

(a) ***Question form:*** As its name shows, a question posed in the item is straightforward and there is no doubt about what for the answer is to take.

(b) ***Identification or association form:*** In this test, the examinee is given a set of words of phrases and is required to supply an association or identification for each or the word or phrase in the set. It can be in the column form in which there is one blank column for the answer and in the other column a phrase or word. The blank column is to completed by the examinee to prove association of identification

(c) ***Completion form:*** A complete or incomplete statement is provided which is be responded either by filling in the missing word or phrase or an answer is given in the forms of mathematical figure in a blank space provided outside the complete statement.

1.6.3 Advantages of Short- Answer Test

The short answer type tests have the following merits to their credit:

(a) Since they are mostly used for the recall of memorized facts and figures except those in which measurement is involved, so is very easy to construct them. Besides this they measure only simple learning outcomes.

(b) There is low probability of guessing the answer because it has be supplied by examinees rather than select or identify from the given answer.

(c) Partial knowledge can help in giving partially correct response for which some credit is given to the examinee which is not possible in objective type test.

(d) They are good to test the lowest level of the cognitive taxonomy i.e.knowledge of terminology, fact and classification etc.

1.6.4- Limitations of Short-Answer Test

Short answer tests have certain limitations. They are as follows:

(a) They are unsuitable for measuring complex learning outcomes.

(b) There are chances of several answers of varying degrees of correctness unless the questions are very carefully phrased.

(c) Since teacher cannot anticipate all possible answer and synonyms for the answer they except, the subjectivity of judgment enter into scoring which slow down the scoring process and lowers the reliability of the obtained marks.

1.6.5- Suggestions for Construction of short answer Test

Inspite of their limitations, short answer tests are widely used and will be continued to use. The following suggestion will help in constructing better short answer type tests:

(a) As for as possible the question form should be used.

(b) The question should be so worded that it has clear meaning and only one answer is possible and which can be given in a word, brief phrase or number.

(c) If there is completion form question, the blank should be at end of the sentence.

(d) The question should not be picked up exactly from the textbook.

(e) The blank space is to be completed buy an important word rather than trivial word

(f) There should not be more than one blank in completion form question

(g) All the blanks should be of equal length and their answers should be written on the right of the sentences to facilitate scoring.

(h) The question should not provide any clues.

(i) A direct question is more desirable than an incomplete statement because examinees are more familiar with it in classrooms and it better structures the situation and avoids ambiguities arising from complete statement.

(j) The unit in which the answer is to be responded should be indicated in numerical questions.

(k) The intended answer should be thought first and then an appropriate question should be formed for the answer.

(l) In numerical question, the answer should be a whole number.

(m) A scoring key should be prepared so that it contains all anticipated acceptable answers.

(n) Specific determiners such as a or an should be avoided in completion form.

1.7- Main point

1. Essay type tests govern a wide field of learning, thinking reasoning, study skill and work habits. Essay tests measure the functional aspects or the application of knowledge. We can evaluate through them the logical reasoning power of the students. The essay tests help us to test the power of expression, verbal facility, use of vocabulary and style of writing. They also help us to evaluate the capacity of the students to organize knowledge, to evaluate thoughts to create new thoughts. The objective tests fail to these powers.

2. The short- answer item requires pupils to supply the appropriate word, number, or symbol to a direct question or incomplete statement. It can be used for measuring a variety of simple knowledge outcomes, but it is especially useful for measuring

problem-solving ability in science and mathematics. However, the areas in which they can be effectively used are restricted by the relatively simple learning outcomes measured and by the fact that the scoring is contaminated by spelling errors of varying degrees of magnitude. When short-answer items are used, the question must be stated clearly and concisely, be free from irrelevant clues, and require an answer.

2 OBJECTIVE TYPE EXAMINATION

The traditional examination the essay type of examination has under heavy fire. Students denounce it because of its heavy strain; the parents criticize its injurious effect on the physical and mental health of the children: the teachers complain because of its harmful effects on school work; the practical psychologist speaks ill of it because of it unreliability and invalidity and the educational theorist attacks it because it lacks definiteness in aim and purpose.

To mitigate some of the evils of the essay type examination, objective tests seem to be very useful. Modem educationists lay much stress on this type of test to supplement the traditional type of tests.

Objective tests are of a large variety. However, only seven or eight types of objective tests are commonly used.

2.1 Meaning Of Objective Test

(1) Gilbert sax (1989) defines objective tests as: "Any test having clear and unambiguous scoring criteria. Because multiple choice and true false tests can usually be scores objectively, they are sometimes referred to as objective tests"

(2) Lou M. Carey (1988) defines objective tests as "Objective test items require students to work or select a correct or best answer. These items are called objective because they can be scored more objectively than any other type of item used to measure students,

Performance... selected response items include alternative response, matching, keyed and multiple choice items".

(3) W. Wiersma and S.G. Jurs (1990) define "objective items' as "items that can be objectively scored; items on which persons select a response from a list of options".

(4) R.L.Ebel & D.A. Frisbie (1986) define "an objective test is one that can be provided with a simple predetermined test of correct answer so that objective opinion or judgment in the scoring procedure is eliminated"

If we go through all these definitions we come to know that an Objective because it's scoring is objective and the alternative choices are already given out of which the correct one is to be selected by examinee. Besides this they are of not one form only. They have different forms.

2.2 Characteristics of objective Tests

The main characteristics of an objective type test are as follows:

(a) They can be reliably scored because there is no inconsistency in scoring.

(b) They allow for adequate content sampling because they require less time on the part of the examinee than do other item formats.

(c) They are generally written at the lowest level of cognitive taxonomy.

2.3 Forms of objective Tests

The objective types are basically of two categories only

(a) Two choice items

(b) More than two choice items

(i) True and False test

(ii) Completion type (if 2 choices are given against each blank)

In the second category we can put

(i) Matching type

ii) Multi choice

(iii) Completion type (if there is more than one blank in the passage and all the choices are written at one place and the examinee is required to select any one for each blank)

Those completion type questions which are of open nature i.e. where the examinees are to supply answer rather than select the answer; they come under the category of short answer type rather than objective type. The objective type tests also suffer from several disadvantages having difference from one of objective types test to another from of objective type test. Thus they will be discussed under a particular form of test to which it belongs.

2.3.1 True & False Test

(i) Stanley and Hopkins (1990) observe:

"The true-false test, a form that is very popular with classroom teachers has been the object of more criticism than any other form of objective test".

(ii) N.E. Gronlund (1985) observes

"The alternative response test item consists of a declarative statement that the pupil is asked to mark true of false, right or wrong, correct or incorrect, yes or no, fact or opinion agrees or degree and the like. In each case there are only tow possible answers. Because the true-false option is the most common, this item type is most frequently referred to as the true-false item".

(iii) A.J. Nitko (1983)) defines "A true-false item consists of a statement or proposition which the examine must judge and mark as either true or false". Thus it can be said true false items may or may not have true or false words with them but their nature is such that there are two answers against each

statement. One answer is Correct and the other answer is wrong. Whatever form is given to it, if one answer is true and the other answer is false, It can be called true and False test.

Advantages

Following are the advantages of using true & False test items:

(a) Since items are very short so the teacher can examine students on more material in time than he can do with any other kind of item.

(b) It takes less time to construct true- false item even it the teacher does not take statements directly from the test.

(c) The scoring is relatively mechanical, so a high degree of objectivity is possible.

(d) Certain aspects of subject matter readily lend themselves to verbal propositions that can be judged true or false.

Limitations

The True & False type items suffer from the following limitations:

(a) Generally they emphasize remote memorization because the examinations are not required to apply remote memorization because the examinees are not required to apply principle to a new situation.

(b) They believe in absolute truth or absolute false and do not admit the possibility of intermediate position. Several facts are not entirely true or entirely false. They cannot be measured through these items.

(c) They allow a high degree of guessing because there are only two choices.

(d) They are largely limited to learning outcomes in the knowledge area.

(e) They are frequently ambiguous.

(f) They expose student to error which is psychologically undesirable.

(g) The may encourage students to study and accept only oversimplified statements of truth and factual details.

Suggestions for Writing

The following suggestions are made to avoid or minimize the limitations of true & false items.

(a) Such true & false items should be constructed which measure importance objectives. Bloom's Taxonomy of Education objective can be used for this purpose.

(b) Avoid specific determiner like always all and never because such items are likely to be false and thus can be guessed by examinees.

(c) It should be seen that approximately half the statements are true and half are false.

(d) Each statement should be unequivocally true or false. It should not be partly true of partly false.

(e) Negative items create unnecessary confusion in the minds of examinees So the statements should be positive. Two negatives in the same sentences/ statement should always be avoided.

(f) Trivial statement though easy to construct, should be avoided because they compel students to memorize each and ever fact at the expense of understanding.

(g) Long & Complex statements should not be used because they measure reading comprehension also which may not be an objective of the evaluator.

(h) Only one idea should be measured in one statement. Otherwise examinee comes into trouble as to how to respond if he considers one idea is true & the other one as false in the same statement. However in the statement showing cause an effect relationship two ideas can be included.

(i) The length of the true and false statements should be equal otherwise examinees will guess that the longer one is true.

(j) Tricky statements should be avoided because they appear to be true but are made false by adding some tricky word, phrase or latter.

(k) The true of false items should neither be grouped together nor any pattern in their placement be followed.

(l) Explain which judgment is to be used true & false, yes & no, correct and incorrect and answer choice should match the necessary judgments.

(m) The directions about recording the answer should be clearly explained.

(n) Providing clues to correct answer should be avoided.

(o) The correct answer should be known to those only who are good in subject matter and should not be a matter of common knowledge.

(p) Statements should not be taken directly from the textbook and other instructional material.

2.3.2 Matching Type Test

(i) Gilbert sax (1989) defines matching type test: "A test format that requires the student to match a series of responses with corresponding terms in stimulus list"

(ii) W. Wiersma & S.G. Jurs (1990) define matching item: "An item consisting of a two column format premises and responses that requires the student to take a correspondence between the two"

(iii) A.J. Nitko (1983) defines matching exercise: "A matching exercise presents the pupil with a list of premises, a list of response and a set of direction for matching the elements of these two test"

(iv) R.L. Ebel & D.A. Frisbie (1986) define matching items: "Matching test items occur in clusters composed of a list of premises a list of responses and direction for matching the tow. In many clusters the distinction between premises and responses is simply in the names given to them. The two lists can be interchanged without difficulty"

(v) L.M. Corey (1988) defines matching items:

"Matching test items are another popular selected response format. These items require students to match information in two columns. Items in the left hand column are called premises, and those in the right hand column are called responses. Students are required to locate the correct response for each premise".

(vi) Hopkins, Stanley & Hopkins (1990) define, "A matching exercise typically consists of two columns; each item in the first columns is to be paired with an alternative in the second column."

(vii) N.E. Gronlund (1985) defines, "The matching exercise consists of two parallel columns, With each word, number of symbol in one column being matched to a word, sentence of phrase in the other column. The items in the columns for which a math is sought are

called premises and the items in the column from which the selection is made are called responses."

Advantages

There are certain advantages in using matching type test. They are as follows:

(a) They are simple to construct and score.

(b) They are will suited to measure associations

(c) They reduce the effects of guessing.

(d) They provide a space-saving, compact of and objectively measure able technique through which several important learning outcomes can be measured in one test.

(e) They can be used to evaluate examinee's understanding of concepts, principle or schemes for classifying objects, ideas or events.

Limitations

The matching type has the following limitations:

(a) They are restricted to factual information which encourages remote memorization.

(b) They generally provide clues.

(c) It is difficult to find homogenous material which may be significant for the objectives and learning outcomes.

(d) If the same number of items are written in both the columns, the matching type is converted into multiple choice at a late stage and in the end it is converted into true and False category.

(e) They result into wastage of time of the examinees because for each premise several responses are to be checked. The problem be-comes more acute if they are printed on two pages instead of

one page be-cause the examinee has to turn page for each premise.

Suggestions for construction

These limitations can be avoided to a great extent of the following suggestions are kept in view while constructing matching type test.

(a) In each matching exercise only homogenous items should be selected.

(b) No clue is provided to facilitate matching response and premise.

(c) Instructions should be very clear and definite about matching.

(d) The premises should be written in the left hand column and be numbered. Responses should be written in the right hand column and be lettered.

(e) The numbers and letter should be written in a systematic order i.e. numbers in order and letters alphabetically.

(f) All the items should be printed on the same page to avoid wastage of time in turning page every time.

(g) There should not be more than 10 items in a matching exercise to save unnecessary wastage of time.

(h) The responses should be more than premises to insure that the examinee has to think even up the last premise. It they are equal, but one premise becomes true & false and the last one does not need any thinking because there is no choice.

(i) Direction should be complete in all respects.

(j) The same types of specific determiners should not be used in matching type, which have been used in multiple choice test items.

(k) All responses should function as plausible options for each premise.

(l) Longer phrases should be written in the premise list and shorter phrases in the responses list.

(m) Incomplete sentences should not be used for premises.

2.3.3 Multiple Choice Type Test

(i) Gilbert sax (1989) defines

"Multiple choice items consist of two parts: a stem and number of options, or alternatives. The stem is question or statement that is answered or completed by one of the alternatives. All incorrect of less appropriate alternative are called distracters or foils and the student's task is to select the correct of best alternative from all the options".

(ii) N.E. Gronlund (1985) defines "A multiple- choice item consists of a problem and a list of suggested solutions. The problems may be stated as a direct question or an incomplete statement and is called the stem of the item. The test of suggested solutions may include words, numbers, Symbols, or phrases and are called alternatives (also called choices or options). The pupil is typically requested to read the stem and the list of alternatives and to select the one correct or best, alternative".

(iii) R.L.Ebel & D.A.Frisbie (1986) define "a multiple choice item has two parts: the stem consisting of a direct question or an incomplete statement and two or more options, consisting of answer to the question or completion of the statement". If we go though these definitions we come to the conclusion that a multiple choice type test consists of the following characters

(a) Once or more introductory sentences which are called stem.

(b) A list of three or more choices from which the examinee is required to choose the correct one. The suggested choices are called alternatives, responses or options.

(c) All the choice should be plausible answer those who do not possess sufficient knowledge of the matter being evaluated.

Forms

The multiple- Choice items can be written in different forms: some of them are as follows:

(a) The correct answer from

(b) The best answer form

(c) The multiple response form

(d) The incomplete statement form

(e) The negative form

(f) The substitution form

(g) The combined response form.

Lets us describe them Briefly

(a) The correct answer form: Three or more choice are there but only one of them is correct.

(b) The best answer form: One or more choice may be correct but one of them is the best answer. The examinee is required to select the best one.

(c) The multiple response form: The correct answer may consist of more than one choices and the examinee is asked to identify all those, which are correct.

(d) The incomplete statement form: the stem is incomplete and can be completed bay the correct choice. The examinee is asked to select the correct one.

(e) The negative form: the stem has negative approach. In the responses all but one may be correct response provided the stem is written in positive form. It is used by the test constructor when positive plausible choices arte not available.

(f) The substitution form: the word outlined in the stem is to substituted by the correct response. Responses are give and the examinee is asked to select the one, which can successfully substitute the underlined word.

(g) The combined response form: the choices are different phrases or sentences of a paragraph The examinee is required to select the correct order those phrases or sentences.

Advantages

Multiple-choice type tests have become very popular. They are most highly regarded because they have several advantages, which are as follows:

(a) They can measure from the remote knowledge level to the most complex level knowledge, understanding and judgment, Ability to solve problems, recommend appropriate action and make predictions.

(b) A substantial amount of course can be tested because the examinees do not require much time for writing the answer.

(c) They are objective in scoring because the key for the correct answer is prepared along with the test.

(d) They can check discriminating ability of the examinees if they are prepared in the best answer form.

(e) They reduce the effect of guessing because there are three or four choices so reliability of each item is increased.

(f) Their format is helpful in item analysis to find out the area of weakness of the examinees, evidence of item ambiguity, item difficulty and individual difference of examinees.

(g) They do not require that only homogeneous material should be tasted as in the case of matching type test.

(h) They can measure cognitive levels much better than true-false items because examinees do not score for merely knowing that whether the statement is true or false but for knowing which is the correct answer.

(i) They can be written in several forms which provide opportunities for measuring different kinds of thinking.

(j) The can provide diagnostic information if the choices are so constructed that they provide information about the kind of mistakes that examinees make. However, constructing such type of response requires a lot of time and thinking.

(k) They can be easily adapted for machine scoring and for computer administration and scoring.

(l) They can be easily compiled test items banks.

(m) They do not permit the examinees to bluff or dress up their answers because the examinees are not allowed to elaborate their answers.

Limitations

In spite of several advantages, the multiple choice type test have the following limitations.

(a) They require examinees to select the answer from a fixed list of choices and do not permit to create or express their own ideas or solutions.

(b) They can check superficial and limited knowledge only if properly constructed.

(c) Intelligent examinees may suffer if they choose incorrect answer due to ambiguities of wording while the poor examinees choose correct answer due to their limited understanding.

(d) They give the impression to the examinees that there is one correct answer to each problem that is not the case in real life of human beings.

(e) They cannot measure attitudes or motor skills, which need active performance and demonstration.

(f) Examinees who do not correct answer can succeed in guessing.

(g) It is difficult to find four choices for each item out of which three maybe plausible incorrect answers.

(h) They cannot measure the ability to organize and present ideas.

(i) They require more time to construct so they are time consuming.

(j) Examinees may be successful in selecting the correct through a process of elimination by rejecting responses, which seem unsatisfactory.

(k) They can shape the education in undesirable way because the type of examination shapes the content and nature of teaching.

(l) ***Suggestions for construction of multiple- choice type test item.***

(a) A.i.Nitko (1983) supports the following three steps of Ebel (1979):

(i) Create he stem of the item by forming a question a question or an incomplete sentence that implies a question.

(ii) Write the correct answer to the question in the stem in as few words as possible.

(iii) Write distracters that are plausible to pupils lacking the degree of knowledge you want them to assess.

(b) The stem should introduce what is expected of the examinee.

(c) Specific determiners should be avoided.

(d) Such vocabulary should be used in the test, which is suitable to the examinees.

(e) The stems and choices should be positively as far as possible.

(f) All the choices should be plausible.

(g) Test items should have a defensibly correct best choice.

(h) Opinions should not be evaluated through multiple choice test items.

(i) The correct choice should not be at the same place in all or most of the items

(j) There should not be any over lapping in choices

(k) "None of the above" choice should be used only if there is an absolutely right answer rather than the best answer.

(l) The choice like "All the above" should be avoided because all choices cannot be equally correct.

(m) The stem should agree with all the choice from the point of grammar and language.

(n) All the common words as far as possible should be taken out of choice and included in the stem.

(o) The test item through which understanding is being evaluated should contain some novelty but it should not be too for to the examinees.

(p) There should be no verbal association between the stem and the correct answer.

(q) All the choices should be of equal strength otherwise they will provide a clue to the examinee about the correct answer.

(r) Each item should pose only one problem.

(s) If negative word is used, it should be emphasized through underlining or printing bold.

(t) Each destructor should reflect a common misconception or problem that examinees have.

(u) Responses should be written in the order of alphabets, number of chronology.

(v) These points are for the guidance of the constructer. If he finds any one of them creating stumbling block in the achievement of the purpose of testing, he should ignore it.

2.3.4 Comparison Of The Essay And Objective Type Test

1. The number of question in the essay type test is very small, generally eight or ten question in the paper. There are a large number of questions in the objective tests.

2. Answer in the essay tests are large. The answers are so short in the objective test as can be written in a word or two or at the most in a sentence.

3. Easy type question are usually broad, general an indefinite. The examinees are at a loss to know how much they will have to write as answer. Objective question are specific narrow and definite.

4. In essay type personal factors like bias, temperament whims, etc, of the examiner influence marking, Scoring becomes objective in the objective tests.

5. In the answer of the essay test there is scope for partial credit. Answer may be partially wrong and right in the objective tests, there is no provision for partial credit Answer must either or be right or wrong and hence full or no credit.

6. In the essay tests examinees have to devote a long time. question can be answered in a short time in the objective tests.

7. In the essay tests, Question are selected at random and the entire course is not covered. Some chapters or questions are completely ignored. The entire course can be covered by comprehensive question in objective tests.

8. Essay tests put premium on expression. Students hardly find any time to think over the question. Objective test put premium on thinking

9. Pupils having linguistic skill are at an advantage in essay tests. Objective test are advantage to the intellectually smart students.

10. It takes a short time to frame questions in the essay tests. It takes a long time to frame question in the objective tests.

11. Student find it easier to prepare essay question. They can safely omit many portions of the subject matter in the essay tests. Students have to devote a longer time to prepare, as they have to for trough the entire course for clear conceptions.

12. Essay tests are very suitable testing composition skills and knowledge of the subject. Objective tests are not suitable for testing composition and skill as well as appreciation of the subject.

13. In the essay tests there is relatively less scope for 'guessing' questions. Students prepare themselves from 'Made easy' books. In the objective type there is more scope for guessing answers.

14. Essay tests take a longer time in examining scripts. Short answers take less time to examine scripts.

15 Printing cost is small in the case of essay tests. since there are many question, printing is costly in the objective tests.

16. Answer script cost much as answers are very lengthy in the case of essay tests. Answers being very short and usually can be written on the question paper itself, this makes the answer scripts less costly.

17. There is less scope for adopting unfair means by the students in the examination hall in the case of essay type questions. There is enough scope for adopting unfair means in the objective tests.

2.4- Main point

1. We discussed techniques for constructing short- answer items, true- false or alternative- response items, and matching exercises. These simple forms of objective test items are restricted al most entirely to measuring knowledge outcomes and are generally unsuitable for measuring understanding, thinking skills, and other complex achievements.

2. The true-false item requires the pupil to select one of two possible answers. This item type is used for measuring simple knowledge outcomes when only two alternatives are possible or the ability to identify the correctness of statements. It is also adaptable to measuring the ability to distinguish fact from opinion and the ability to recognize cause and -effect relationships.

3. The matching exercises nosiest of two parallel columns of phrases, words, numbers, or symbols that be must matched.

Examples of items included in matching exercises are persons and achievements, dates and historical events, and terms and definitions. The nature of the matching exercise limits it to measuring the ability to identify the relationship between two things. For this restricted use, it is a compact item type that can be used to measure many relationships in a short time.

4. The multiple-choice item consists item consists of a problem and a list of alternative solutions the pupil responds by selecting the alternative that provides the correct of best solution to the problem. The incorrect alternatives are called distracters because their purpose is to detract the uninformed pupil from the correct response. The problem can be stated as a direct question or an incomplete statement. In either case, it should be a clearly formulated problem that is meaningful without reference to the list of alternatives.

5. The multiple-choice form is extremely flexible and can be used to measure a variety of learning outcomes at the knowledge and understanding levels. Knowledge outcomes concerned with vocabulary, facts principle, and methods and procedures all can be measures with the multiple-choice item. Aspects of understanding such as the application and interpretation of facts, principle, and methods, can also be measured with this of item type.

3. SELF ASSESSMENT QUESTIONS

1. Define and explain the forms examination
2. Distinguish between the oral examination and practical examination.
3. Identify the main characteristics of essay tests.
4. What are the merits and demerits and demerits of an essay test?
5. Defined the statement "Short-answer items should not be classi-fied as objective items"
6. How would you handle the scoring of short-answer items when the answers are misspelled?
7. What are the merits and demerits of objective tests?
8. What factors should be considered before setting an essay or an objective rest?
9. Suggest ways of constructing any of the following forms of test: essay, multiple-choice pictorial-item, rank-order and true-false item.
10. Describe the advantages of the multiple-choice item over each of the other objective-type items. What are the comparative disadvantages?

4 BIBLIOGRAPHY

1. Clinton I. Chase (1978) ----------Measurement for education, London: Addison- Wesley publishing company

2. Ebel, Rebert L. and Frisable, David A. (1995) essential of educational measurement, New Delhi, Prentice-Hall

3. Herbert.J. Walberg and Genvad D. Haertel (1990) The International Encyclopaedia of Education, New York: Programon Press.

4. J.D.,Mcneil (1967) Concomitants of using behavioral objects in Assessment of Teachers Effectiveness. London: Journal of Experimental Education.

5. L.R. Gay (1985) Educational Evaluation and measurement. Columbus: Charles E. Merril Publishing Company.

6. Muhammad Habibullah (Nov.1967) Personnel Administration in Education in Pakistan. Project Report, Lahore: Punjab Education Extension Center.

7. Norman E. Gronlund and Robert s. (1990) Measurement and Evaluation in Teaching. New York; Mac Millan Publishing Company.

8. Wierrsms William and jurs, G. (1990) Educational Measurement and Testing. London: allyn and bacon.

9. Zais, Robert s. (1976) Curriculum: Principles and Foundations New York, Harper and Row, Publishers.

Unit-5

METHODS OF MEASUREMENT

By
S.M. Shahid

INTRUDUCTION

Selection appropriate types of question. There should be a balanced selection of it is not essential that the test items in an educational test resemble closely the behavior the test is to predict. The degree of similarly between the test sample and the behavior to be predicted may vary widely. Sometimes, the test may coincide completely with a part of the behavior to be predicted just as in a language test, the testy items relating to language proficiency are given. In vocational aptitude tests, there is a moderate resemblance between the operations performed of the job and those incorporated in the test. But in project personality tests such as (Rorschach Inkblot Tests) there is absolutely no resemblance between the test situation and the behaviour to be tested.

This unit, therefore, will discuss mostly the behavioural aspects of the individual with reference to different measurement techniques. It is hoped that this unit will help you in developing a deeper insight into the concept of behaviour measuring techniques and their applications in the field of education. It will also be of value for your professional development as an educator.

OBJECTIVE

After an intensive study of the unit you are expected to be able to:

1. Use different techniques for behaviour measuring techniques.
2. Differentiate and apply different types of observational techniques.
3. Prepare an outline showing important aspects of interview procedures.
4. Differentiate among sociograms, social distance scales, and guess whose questionnaires in term of their use.
5. Explain how self-reporting techniques measure interest and attitudes.
6. Interpret projective techniques for measuring behaviour.

Contents

Introductions

Objectives

1. Objective Methods: Standardized Tests

1.1. Achievement Tests

1.2. Aptitude Tests

1.3. Intelligence Tests

1.4. Interest Tests

1.5. Personality Inventories

1.6. Summary

2. Subjective Methods, Observational Techniques

2.1. Rating scales

2.2. Project methods

2.3. Self reporting Techniques

2.3.1 Anecdotal Record

2.3.2 Autobiography Techniques

2.3.3 Interview

2.3.4 Questionnaire

2.4. Sociometric Techniques

2.5. Summary

3- Self-Assessment Questions

4- Bibliography

1. OBJECTIVE METHODS: STANDARDIZED TESTS

A STANDARDIZED TEST is a measurement device that is commercially prepared by educational specialist for widespread use in a large number of schools. It is characterized by a fixed set of questions designed to measure a clearly defined sample of behaviour. The test is administered under uniform conditions using the same set of directions for timing constraints and scoring. This permits educators to give the identical test to students in different locations and at different times.

Most commonly test designers establish performance norms for a standardized test based upon representative groups of individuals (including various age and grade groups on a state, regional, or national level). This feature allows educators to interpret a pupil's performance in a norm-referenced manner. Accordingly, it is possible to compare the level of a student's performance with that of other groups of students. Further, a student's level of performance on one sets of questions (provided all the tests were standardized using the same sample of students).

Standardized tests can be classified in number of ways. The most popular classification is according to (Mehrens and Lehmann, 1978):

1. Achievement tests (diagnostic, single subject matter, and survey batteries).
2. Aptitude tests (general, multiple and special).
3. Personality, Attitude, interest, self-concept and adjustment inventories.

The first two categories usually seek to measure maximum performance the third, typical performance.

1.1- Achievement Tests

An achievement test has a great significance in all types of instructional progress of the individual. A classroom teacher depends upon the achievement tests for measuring the progress of his students in his subject area. Several educational and vocational decisions about students are taken on their performance in the achievement tests. It is, therefore, necessary that the teachers should be well versed with the meaning and characteristics of achievement tests.

Following definitions throw adequate light on the nature and role of achievement tests.

1. Thorndike and Hagen (1969) observe, "The type of ability test that describes what a person has learned to do is called an achievement test."
2. Gronlund (1977) defines an achievement test as "a systematic procedure for determining the amount a student has learned through instruction."
3. Popham (1981) states, "The achievement test focuses upon an examinee's attainments at a given point in time."
4. Gronlund, N.E. and Linn, R.L. in their book Measurement and evaluation in teaching (1990) have observed, "There typically have been norm- referenced tests that measure the pupil's level of achievement in various content and skill areas by comparing their test performance with the performance of other pupils in general reference group."
5. In the words of Wiersma, W. and Jurs, S.G. (1990) achievement test "is a measure of knowledge and skills in a content area."
6. Popham, W. James in his book modern Educational Measurement (1990) provides the description of achievement

tests in these words, "Tests in the cognitive or psychomotor tealm are often focused on an examinee's attainment at a given time; these tests are usually referred to as achievement tests."

Characteristics of a Good Achievement Test

Gronlund and Linn (1990) have identified the following characteristics of a good standardized achievement test:

1. A good achievement test is tried out and selected on the basis of its difficulty level and discriminating power.
2. It should have a description of measured behaviour.
3. It should contain a sufficient number of test items for each measured behaviour
4. It should be divided into different knowledge and skills according to behaviours to be measured.
5. Its instructions in regard to its administering and scoring are so clear that they become standardized for different users.
6. It is accompanied by norm, which are developed at various levels and on various age groups.
7. It provides equivalent and comparable forms of the test.
8. It carries with a test a manual for its administering and scoring

Administrator's Use of Achievement Tests

1. Tests help to evaluate the extent to which the objectives of education are being achieved.
2. Tests help to classify school objectives.
3. Tests discover the type of learning experiences that will achieve these objectives with the best possible results.

1. To evaluate, revise and improve the curriculum in the light of these results.
2. To discover backward children who need help and to plan for remedial instructions for such students.
3. To select talented pupils for special classes and courses.
4. To decide proper classification of students.
5. To gets a better understanding of the needs and abilities of pupils.
6. To select students for the award of special merits of scholarships.
7. To group pupils in a class so that students are put in such a way that individual difference are as slight as possible.
8. To help the parents in recognizing the strengths and weaknesses of their children so that they direct their energies on suitable goals only and do not put heavy demands on them.
9. To determine the efficiency of one school with the others.
10. To determine the general level achievement of a class and thus to judge of the teaching efficiency of the teacher. The level of the achievement of a class may be judged on the basis of the achievement of the class in the beginning and at the end of the school year.

The Teacher's Use Achievement Tests

1. The teacher will come to know about the general range of abilities of students in the class.
2. In the light of above, he will select appropriate materials of instruction so that all individuals benefit from instruction to the maximum

3. The teacher will determine and diagnose the weakness of the students in various subjects.
4. The teacher will spot brilliant and backward children.
5. He will determine the progress of the group in a particular subject over a period of time.
6. By studying the results of the students on achievement tests and intelligence tests, the teacher will determine whether or not the students are working at their maximum capacity.

Planning Of an Achievement Test

Shipman. M. In his books, School Evaluation (1979), suggests the following stages, which can be helpful in planning an achievement test:

1. Stating the objectives in such a form that it will enable their achievements to be verified.
2. Spelling out the actions that are necessary for the attainment of objectives.
3. Spelling out the criteria by which the attainment of objectives is to be assessed.
4. Identifying personnel whose judgments are involved in each component of the plan.
5. Stating the arrangements for any standardization procedure and the forms of reports and records that are to be maintained.

Wiersma, W. and Jurs, S.G. (1990) have identified the following stages for planning an achievement test:

1. Deciding purposes of the test.
2. Listing educational objectives of different areas of knowledge
3. Preparing table of specifications.

4. **Determining** practical consideration.

Gronlund and Linn (1990) discuss in detail the following steps in **classroom testing:**

1. **Determining** the purposes of testing.
2. **Developing** the test specifications.
3. **Selecting** appropriate type of items.
4. **Preparing** relevant test items.
5. **Assembling** the test.
6. **Administering** the test.
7. **Scoring** the test.
8. **Appraising** the test.
9. **Using the test results.**

1. **Determining the purpose of the test.** According to purpose, **Tests may be divided** into the following five categories:

   (i) **Pre- testing** which includes

      (a) **Readiness** pre-test

      (b) **Placement** pre-test

   (ii) **Testing** during placement

      (c) **Formative** test

      (d) **Diagnostic** test

      (e) **End** testing- summative test

2. **Developing test specifications.** Specifications of the test **include the following**

   (a) **A list of** instructional objectives.

(b) An outline of the course contact.

(c) A two-way chart.

Followings chart explains the specifications of the contents of the test as for weight age to different areas is concerned.

Each level of thinking and numbering of test items:

%Content	Knowledge	Comprehension	Application	Analysis	Total
Topic No.1	7	5	4	4	20
Topic No.2	6	6	4	4	20
Topic No.3	6	6	3	3	18
Topic No.4	2	5	5	2	14
Topic No.5	4	5	5	2	16
Topic No.6	5	6	1	0	12
Total	30	33	22	15	100

3. Selecting appropriate types of questions. There should be a balanced selection of essay type, short-answer type and objective type questions.

4. Preparing relevant test items. This includes the following steps:

(i) Matching the test items with the learning outcome.

(ii) Selection most representative items.

(iii) Preparing test items which are of proper difficulty level.

(iv) Avoiding all possible barriers in test items, which prevent examinees from responding.

(v) Avoiding, providing any clues to answers, which may help examinees to answer of correctly even if they lack the necessary achievement.

5. **Assembling the test.** After preparing relevant test items, the test constructor should follow the process as indicated below:

   (a) Writing each item on a separate card.

   (b) Reviewing the test items by the test constructor himself and also by some other teacher

   (c) Arranging the test items according to a well defined criteria

   (d) Providing proper instructions to the examinees.

6. **Administering the Test**. Following suggestions are made for administering the test.

   (i) Long announcements before or during the test should not be made.

   (ii) Instructions, if any, should be given in writing so that uniformity is maintained and all the examinees get the same opportunities

   (iii) The test administered should not respond to the individual problems of the examinees, otherwise any hint on their part may provide unfair chance to some examinees.

   (iv) Test should be administers in an appropriate physical and psychological environment.

7. Scoring the Test. Scoring may be done mechanically or manually; depending upon the situation.

8. Appraising the test or item-analyses. After scoring, the test should be appraised for each item.

9. Using the test results. Test results are used primarily for the following purposes.

   (a) For making decisions about the promotion of student to the next higher grade.

   (b) For bringing about improvement in teaching methods and techniques.

1.2 Aptitude Tests

According to Bingham and Freeman, aptitude tests are tests that will predict success to some degree. It may be mentioned here that aptitudes are relatively constant.

Aptitude tests help us to measure the probability of success in an activity such as playing at piano, learning a language etc.

Any definition of aptitude should be in terms of these characteristics:-

(a) Ability to acquire the skill, information etc, necessary for success.
(b) Readiness to acquire,
(c) Constancy, and
(d) Satisfaction in the job.

In many countries, tests of aptitudes for specific school subjects have been developed.

Tests of clerical aptitude, of mechanical aptitude and dexterity, of aptitude in art and music, aptitudes in professions have also been devised.

Analyzing the particular occupation or activity for which aptitude is to be measured. Aptitude tests and test items are devised for revealing the abilities. Tests of aptitude for law will include items, which measure accurate recall, reading

comprehension of legal material, skill in logic, and reasoning by analogy and by analysis.

Uses of Aptitude Tests

Aptitude tests serve the following functions.

Admissions. Aptitude tests can be used in admitting candidates for various types of professional training such as engineering, medicine and training etc.

Guidance. Aptitude tests can be used for the purpose of guidance in selecting subjects for studying in educational institutions. .

Selection for Jobs. The employer can use aptitude tests for selecting persons for jobs.

Specific Areas of Aptitude Tests

Among these may be mentioned as under:

1. Art Aptitude Test
2. Clerical Aptitude Test.
3. Manual Aptitude Test.
4. Mechanical Aptitude Test.
5. Medical Aptitude Test.
6. Musical Aptitude Test.
7. Scholarly Aptitude Test.
8. Scientific Aptitude Test.
9. Teachers Aptitude Test.

Scholastic Aptitude Tests: Scholastic aptitude tests are helpful in giving educational and vocational guidance to students. Froehick and Benson think that we can use the results of scholastic aptitude tests in Counseling students, regarding vocational

opportunities. Reading, writing and speaking abilities are involved in various jobs and a student's scholastic aptitude is a fair measure of his chances for success in such jobs.

Students with superior scholastic aptitude should be preferred in admitting to colleges.

Scientific and classical courses require a higher level of scholastic aptitude on part of the student.

How to Measure Scholastic Aptitude

1. **School Marks and Scholastic Aptitude.** This is the traditional method of measuring aptitude.

2. Occupation of Parents and Scholastic Aptitude. McNemur tested the I.Qs of children of parents following different occupations. He found that children of professional people (engineers, doctors, lawyers, etc) got higher I.Q. at all age levels than children of clinical skilled trade and retail business people. The lowest I.Q. was of the children of day labourers. This study indicated a positive relationship between the intelligence of the child and the occupational status of the father.

3. Teacher's Observation and Scholastic Aptitude. The following points may be observed:

(i) Rapidity in comprehending material of study

(ii) Rapidity and accuracy in reading

(iii) Ability in attaching new problems.

(iv) Large vocabulary

(v) Eagerness to answer questions.

(vi) Deficiency in one or more skills.

(a) negative criterion.

4. Scholastic Aptitude Tests.

The Yale Educational Aptitude Test: The battery contains tests.

Designed to measure a person's relative aptitude or ability in the areas of: (1) verbal facility; (2) linguistic ability; (3) verbal reasoning; (4) quantitative reasoning; (5) mathematical aptitude; (6) spatial visualizing; and (7) mechanical ingenuity. On the basis of these aptitude tests, students are admitted to courses like social sciences, pure sciences, mathematics and applied sciences.

Mechanical Aptitude Tests, These tests are designed to measure fundamental aptitudes of tool usage, space visualization and shop arithmetic in the area of mechanical ability .In general there are two types of mechanical aptitude tests: (1) Performance tests in which the subject is expected to do something with special equipment and (2) Paper and pencil in which the responses are given on paper .

1.3. Intelligence Test

These may be classified under three categories.

1-ndividual Tests. These tests are administered to one individual at a time. These cover age group from 2 years to 18 years. These are: (a) The Binet-Simon Tests, (b) Revised Tests by Terman, (c) Mental Scholastic Tests of Burt and (d) Wechsler Test

2.-Group Tests, Group tests are administered to a group of people. Group tests had their birth in America---when the intelligence of the recruits who joined the army in the First World War was to be calculated. These are: (a) The Army Alpha and Beta Test, (b) Terman's Group Tests, (c) Otis Self-Administrative Tests. Among the group tests there are two types: (i) Verbal and (ii) Non-verbal.

Verbal tests are those that require the use of language to answer the test items.

Non-Verbal tests do not require the use of language to respond to the items.

3- Performance Tests, These tests are administered to the illiterate persons. These tests generally involve the construction of certain patterns or solving problems in terms of concrete material. Some of the famous tests are: (a) Koh's Block Design Test, (b) The Cube Construction Tests, and (c) The Pass Along Tests.

Comparison of Individual Tests and Group Tests

Individual Test	Group Test
I. It is administered to an individual at a time	1 It is administered to a group at the same time.
2. It is costly in terms of administration and time	2. It is less costly in terms of administration and time
3. A trained tester is required to administer it.	3. No trained person is required to administer it.
4. There is face-to-face interaction between the individual and the tester	4. There is no such face-to-fa-ce interaction
5. Individual test is more reliable. Guidance can be provided to the individual on the basis of its results	5. Group test may be influenced by several factors
6. It is useful for small	. It is suitable for older

children	children and adults.
7. The tester can motivate the Individual by means of praise and encouragement as he can adapt to the needs of the individual child.	7. It is not possible to do so.
8. There is very little scope for cheating.	8. Cheating on a large-scale is possible.
9. There is no competition in individual testing.	9. Speed and reading ability may influence the test score.
10. No special formalities are observed in individual testing.	10. Several formalities are observed in administering.
11. Instructions can be made clear before testing.	11. A few members of the group may not clearly understand the instructions.

Measuring Special Ability (The I.Q. obtained as result of performance on an intelligence test indicates general status only. It does not point to the size of strength or weakness in each of the particular mental abilities that are being measured. For example, two students showing an I.Q. of 120 on a particular test may have different positions on different sub-tests; one may do very well on arithmetic and poorly on vocabulary, while the other may do well on vocabulary and poorly on arithmetic, yet both score the same total

(A) Primary Mental Abilities Test (PMA) Test). To correct this sort of error we need tests, which indicate differential success of a subject on various mental abilities. The tests of primary mental abilities prepared by Thurston meet this need. The PMA test for ages 11 to 17 is based on the group factor theory of mental ability which postulates, that intelligence is made up on certain distinct and more or less independent mental functions which Thurston called the primary mental abilities. .

The primary abilities measured by this test are:

1. Number facility
2. Verbal comprehension
3. Spatial perception
4. Work fluency
5. Reasoning.
6. Rote memory

(B) Differential Aptitude Test Battery (DATB). Another test to measure the special abilities is Differential Aptitude Test Battery (DATB). This comprises eight tests:

1. Verbal Reasoning
2. Numerical Ability
3. Abstract Reasoning
4. Space Relations
5. Mechanical Reasoning
6. Clerical Speed and Accuracy
7. Language Usage: Spelling
8. Language Usage: Grammar, punctuation and word usage.

(C) General Aptitude Test Battery (GATB) Another test of the differential aptitude type is the general Aptitude Test Battery (GATB), developed by the United State Employment Service. It consists of 15 tests, which cover 9 factors, such as intelligence, verbal aptitude, numerical aptitude, spatial aptitude, form perception, clerical perception, motor-coordination, finger dexterity and manual dexterity.

Uses of intelligence tests

1. Selection of student to a school
2. Classification of students
3. Selection of students to different courses
4. Detection of superior and inferior intelligence
5. Selection of suitable occupation
6. Award of scholarships
7. Determination of the optimum level of work
8. Discovery of unusual cards
9. Diagnosis of backwardness

It may be remembered that intelligence tests are not the only tools in measuring the capability of an individual. These should be used along with other instruments of measuring personality traits.

1.4. Interest Tests

It is not possible to measure interest as an independent entity as it is related to general intelligence and special aptitudes and is determined partly by social environment of an individual and his opportunities to find out different kinds of activities. Interest is not as consistent as aptitude.

Educational interest inventories are intended to help pupils to choose their courses, curricula, and co-curricular activities. There are also vocational interest inventories. The names of Kuder and Strong may be mentioned in this field.

Table Showing the Classification of Occupation On The Basis Of Intelligence

S. No.	Standard of Occupation	Names of Occupation	Intelligence	I.Q
1.	High professional and Executive occupations	Lawyer, principal (University and Secondary teachers), manager of big concern, Professor, Scientist, successful politician.	Genius, very superior, intelligence	Over 150
2.	Lower professional, technical and executive work	Executive of a moderately large business, doctor, engineer etc.	Superior intelligence	130 to 150
3.	Clerical, technical and highly skilled work	Shorthand typist, bank clerk, salesman photographer,	High average intelligence	115 to 130
4.	Skilled work	Typist, file clerk, carpenter, dressmaker, cashier,	Intelligence below	100 to 115
5.	Semi-skilled occupation	Number sorters, delivery man, painter, baker, barber	Average intelligence backward	85 to 100
6.	Unskilled repetition work	Manual labor, packer.	Very backward	70 to 85

1.5- Personality Inventories

A personality inventory is essentially a standard set of questions about some aspects of the individual's life history, feeling preferences or activities presented in a standard way and scored with a standard scoring. key. It resembles a questionnaire as the general device and can be used for collecting all kinds of information but a personality inventory is specially designed to collect information about the personality of an individual.

The number of available personality inventories is very large. Here we shall be concerned primarily with a few of the most widely known inventories.

Woodworth's Data Sheet, The first name in this connection may be mentioned of the Woodworth Personal Data Sheet, which was developed for use during World War-l. This inventory was essentially an attempt to standardize a psychiatric interview and to adapt the procedure for mass testing. Inventory questions were formulated with reference to common neurotic and pre-neurotic symptoms and each question was to be answered in" Yes "or "No"; a questionnaire of 116 items was assembled on the basis that no item was retained if 25% or more of a 'normal' sample answered it in the unfavorable direction. The questions dealt with such behavior deviations as abnormal fears, obsessions and compulsions, nightmares and other sleep disturbances, excessive fatigue and other psychosomatic symptoms, feelings or unreality, motor disturbance and the like. The individual score on the inventory consisted of the total number of items answered in the unfavorable direction. This was then compared with the mean scores obtained by neurotics and normal in the standardization sample. If the individual's score exceeded a certain number, it was considered to be indicative of possible neuroticism.

Cornell Index, This was developed during World War-n and the civilian revision of this questionnaire contains 101 Yes-No items pertaining to feelings of fear and inadequacy, depression and other pathological mood reactions, nervousness and anxiety, etc, and several kinds of psychosomatic symptoms. The time required to complete the inventory varies from 5 to 15 minutes.

Bell Adjustment Inventory, In this inventory, 35 items are classified into separate categories and a score is provided for each category .Two forms of this inventory are available, one for high school and college students., the other for adults. However, the .student form has been employed more widely. This form has been designed to measure adjustment in four areas: (i) Home, (2) Health, (3) Social, (4) Emotional. An additional score has been provided in the adult form to measure occupational adjustment. Encircling 'Yes' or 'No' records answers.

Allport A-S Inventory. This is also known as Ascendance Submission Inventory. This inventory seeks to assess the individual's tendency to dominate his associates or be dominated by them in face-to-face contacts of everyday life. Each item begins with a brief description of a situation, which we usually encounter at a meeting, in school on a bus or in other familiar settings. The subject is asked to indicate one of the two or four alternative ways listed for meeting the situation. Responses indicate the degree of ascendance or submission.

Separate forms of the tests are available for men and women. This inventory has greatly influenced the development of many other inventories.

Bernreuter Inventory. It consists of 125 items and is designed to measure six scores (i) Neuroticism, (2) Self-sufficiency, (3)

Introversion, (4) Dominance, (5) Confidence, (6) Sociability. The last two were added by Flanagan.

This manual provides norms on all six scores for high school, college and general adult population.

The inventory appears to be more effective with normal and near normals than with psychotics.

Heston personal adjustment inventory. This yields scores in Analytical Thinking, Sociability, Emotional Stability, Confidence,

Personal Relations and Home Satisfaction.

Guilford Inventory, The latest inventory in the Guilford series consists of 300 items, 30 for each of the following 10 scores:

G. General Activity

R. Restraint

A. Ascendance

S. Sociability

E. Emotional Stability

0. Objectivity

F. Friendliness

T. Thoughtfulness

P. Personal Relations

M. Masculinity

Thurstone Temperament Schedule, It includes 20 items for each trait and these items pertain to the behaviour of relatively normal persons. Percentile norms are reported for high school and college students.

Thurstone lists seven major factors:

Active (A)

Vigorous (V)

Impulsive (I)

Dominant (D)

Stable (E)(Emotionally Stable)

Sociability (S)

Reflective (R)

Cattell's Inventory, Cattell has devised the Sixteen Personality Factor Questionnaire. This inventory is available in two parallel forms, A and B each containing 187 items. The use of both forms is advocated for greater reliability,

The Minnesota Multiphase Personality Inventory, Published in 1943, consists of 550 items and is used for persons of 16 years of age and above. Each item is printed on a separate card. The subject has to sort the cards into 3 groups. True, False, and Cannot Say. The items are classified under 26 heads, such as health, religious attitudes, delusions, phobias, etc. Items can be grouped in separate scales to score none personality traits. These are hypochondriasis, depression hysteria, psychopathic deviate, masculinity-feminity, paravoia, psychosthenia, schizophrenia, and hypomania. This inventory is used in clinical diagnosis.

A part form its comprehensiveness, the MMPI is provided with several control keys meant for identifying untrustworthy responses. These keys give such scores as lie score (L) when the subject tries to fake good on socially approved behaviour, the 'K' score when the subject fakes bad to show himself in bad light; the question score when the subject gives a large number of 'cannot say' responses, and the 'F' score when the subject marks items

with carelessness or misunderstanding. The MMPI is one of the several most widely used inventories. A shortened version consisting of 336 items is also available for emergency use.

The Guilfore-Zimmerman Temperament Survey, It is meant to identify ten different trait dimensions of personality .Some of these are general activity, friendliness, thoughtfulness, personal relations, and masculinity .The inventory is used with adolescents and adults. These traits have been included after factor analysis and are mutually exclusive. The sample used for standardization consisted of normal persons, and not of maladjusted or neurotics.

The Edward Personal and Preference Schedule, It consists of 210 items, which assess the strength of 15 needs selected from among those listed by Murray. The items are presented to the subject in pairs and these are more or less equated for social desirability , so that the subject would respond as he really felt, and not in terms of what is the approved or desirable thing to say.

Evaluation of Personality Inventories

The construction and the use of personality inventories are beset with special difficulties in view of the following:-

1. Complex nature of personality
2. Different definitions of personality
3. Greater specificity of responses in the sphere of personality.
4. Lack of adequate criteria for the determination of empirical validity.
5. Different actions by the same individual in different situations.

Purposes of Assessment of Personality

1. To appraise the status of and changes in pupil's personality development.
2. To find out pupil's needs and possibilities of development do
3. To do student-teacher planning.
4. To familiarize the teacher with the nature of pupil learning development and progress.
5. To serve as a means of improving school community relations.
6. To facilitate the selection and improvement of assessment instruments.
7. To appraise the teacher's competence.
8. To serve as a guiding principal for the selection and application of supervisory techniques.

Difficulties in Assessing Personality

Several difficulties are encountered in the assessment of personality .Although for long psychologists have been trying to develop sound techniques for measuring personality yet even today they are far, away from this goal.

There are three basic reasons for not arriving at exact conclusions.

1. Complex nature of the individual whose personality is to be assessed.
2. Complex nature of the person who assesses personality.
3. Nature of assessment instruments.

1. **Complex nature of the personality of the individual** The personality of the individual being assessed is very complex. We

may peep into the child's personality here and there in some detail but a total peep of the personality eludes us.

Assessment of personality does not remain stable because an individual is ever growing and he may not respond to the same situation in the same manner. There are several internal and external forces which influence him constantly.

An individual who is being assessed is capable of employing hundreds of ways to evade being assessed. Even if he cooperates willingly with the best of his intentions, he may be unable to tell about the elements hidden in his unconscious mind.

2. Complex nature of the assessor of personality. Subjectivity of the person who assesses can also influence assessment. Even when the same person assesses the same individual at different intervals under the same conditions, results may differ. It is also observed that when two or more persons assess the same person, results may not be identical.

3. Nature of assessment instruments. In measuring a piece of cloth, we can use units in terms of centimeters and inches. In measuring temperature, we have units in terms of degrees and so on but in psychological measurement, we do not have any regular unit of measurement. In personality assessment there is no starting point (zero) for reference. No child is born with zero personality .This also creates difficulties in the exact assessment of personality .Above all, tools of assessment of personality are not very exact, reliable and valid in terms of their results.

1.6 Summary

Standardized achievement tests measure the common objectives of a wide variety of schools, have standard procedures for administration and scoring, and provide norms for interpreting the

scores. A test manual and other accessory materials are typically provided to aid in the administration of the test and the interpretation and use of the results. The test items are generally of high quality because specialists have prepared them, pretested, and selected on the basis of their effectiveness and their relevance to a rigid set of specification.

Despite their high technical quality , standardized achievement tests complement rather than replace teachers ' informal classroom tests. They are especially useful for measuring general educational development, detonating pupil progress from one year to the next, grouping pupils, diagnosing learning difficulties, and comparing achievement with learning ability .They are of less value for measuring learning outcomes unique to a particular course, The day-to-day progress of pupils, and knowledge of current developments in rapidly changing fields. These purposes are more effectively served by informal classroom tests.

Standardized aptitude tests are designed to predict future performance in some activity , such as school learning. Like achievement tests, aptitude tests measure learned abilities. They differ from achievement tests, however, in that the test content is broader in scope, and test performance is less dependent on any specific set of learning experiences. This makes it possible to use the tests with pupils of varying educational backgrounds and to predict performance over a wide range of learning activities.

Group tests of learning ability (or scholastic aptitude) may yield a single score, separate verbal and nonverbal scores, separate verbal and quantitative scores, or several scores based on a series of specific aptitudes. The single-score test is designed to measure general learning ability only for tests using verbal and nonverbal

scores, the nonverbal score serves as a check on the leaning ability of the poor reader .

Individual tests of learning ability de emphasize reading skill and provide more carefully controlled testing conditions. Thus, the individual test is especially valuable for testing young children and for checking on questionable scores obtained with group tests.

2 SUBJECTIVE METHODS: OBSERV A TIONAL TECHNIQUES

Observation is one of the most ancient and widely used instruments of assessing personality .

Observation has been defined as, "measurement without instruments." In education, observation is the most commonly employed of all measurement techniques. In the present as well as in the past, students have been labeled as good, fair or poor in achievement and lazy or diligent in study, etc, on the basis of observation. Similarly, teachers have listened to speeches and ranked students 1,2,3 and so on.

Rousseau wrote, "Watch nature long and observe your pupil carefully before you say a word to him".

The physicians and the psychologists depend heavily on what they observe of the patient's talk, gestures and facial expressions. Observation is one of the oldest techniques that man has made use of. Even today it is our common experience to notice that farmers feel the breeze, watch the sky, sun, moon and stars, all to determine what the weather is likely to be and what season is approaching.

Merits of Direct Observation

1. Being a record of the actual behaviour of the child, it is more reliable and objective.

2. It is a study of an individual in a natural situation and is therefore more useful than the restricted study in a test situation..
3. This method can be used with children of all ages; of course, the younger the child, the easier it is to observe him. This method has been found very useful with shy children.
4. It can be used with some training and experience and almost all teachers can use it. It does not require any special tools or equipment.
5. It can be used in every situation.
6. It is adaptable both to individuals and groups.
7. Frequent observations of a student's work and work habit can provide a continuous check on his progress.
8. The problems as they arise can be immediately detected and corrective action taken immediately.
9. Observational data provide teachers with valuable supplementary information.

Limitations and Demerits

1. There is a great scope for personal prejudices and bias of the observer .
2. Records may not be written with hundred percent accuracy as the observation is recorded after the actions of the observed. There is some time lag.
3. The observer may get only a small sample of student behavior. It is very difficult to observe everything that a student does or says.

4. It reveals the overt behavior only behavior that is expressed and not that is within.

Principles to be followed in Making Observations

1. Observe the whole situation.
2. Select one student to observe at a time.
3. Students should be observed in their regular activities, such as in classroom, on the playground or in going from class to class.
4. Observation should be made over a period of days.
5. As far as possible, observations should be colleted several teachers.

Requisites of Good Observation

As a measurement tool good observation is based on

1. Proper planning
2. Proper execution
3. Proper recording
4. Proper interpretation.

Proper Planning of Observation

1 Specific activities or units of behavior to be observed must clearly be defined.
2. An appropriate group of subjects be selected for observation .
3. Scope of observation -whether individual or group - should be decided.
4. The length of each observation period, number of periods and interval between periods should be decided .

5. The form of recording should be decided.
6. The instruments to be used should be decided.
7. Physical position of the observer should be demarcated.
8. Proper tools for recording observation should be kept handy.

Proper Execution of Observation

An expert execution demands skill and resourcefulness on the part of the investigators. This depends upon.

(i) Proper arrangement of special conditions for the subjects.

(ii) Assuming proper physical position for observing.

(iii) Focusing attention on the units of behaviour on the specific activities under observation.

(iv) Observing discreetly the length and number of periods and intervals decided upon.

(v) Proper handling of the recording instrument being used .

(vi) Utilizing suitably the training received in terms of expertness.

Recording of Observation

Generally I two methods are employed for recording observation. Which of the two methods to use depends upon the nature of the activities or behaviour of the group to be observed. The skill of the observer also plays an important role in deciding upon the method.

The first method is to record the observation simultaneously.It is useful in the sense that at time, gap may distort facts. However, at times, this may not be feasible when the action or activity performed is very swift. Moreover, this is likely to distract the subject or subjects.

Facts may be recorded soon after the observation is over. This is helpful as this does not distract the mind of the subject. But the investigator may not be able to recall facts accurately after the interval of a few minutes.

Devices Used in Observation are:

(i) Check-list

(ii) Rating Scale

(iii) Score Cards.

(iv) Blank form of tallying frequencies.

2.1 Rating Scales

Rating means the judgment of one person by another , "Rating is, in essence, directed observation," writes Ruth Strang, A.S. Barr and other define, "Rating is a term applied to expression of opinion or judgment regarding some situation, object or character. Opinions are usually expressed on a scale or values. Rating techniques are devices by which such judgments may be quantified." A rating scale is a method by which we systematize the expression of opinion concerning a trait. The ratings are done by parents, teachers, a board of interviewers and judges and by the self as well.

These rating scales give an idea of the personality of an individual.

There are two characteristics of a rating scale:

1. Description of the characteristics to be rated, and
2. Some methods by which the quality , frequency or importance of each item to be rated may be given.

Use and Advantages of Rating Scales

1. Helpful in measuring specified outcomes or objectives of education.
2. Helpful in supplementing other sources of understanding about the child.
3. Helpful in their simulating effect upon the individuals who are rated.
4. Helpful in writing reports to parents.
5. Helpful in filling out admission blanks for colleges.
6. Helpful in filling out admission blanks for colleges.
7. Helpful in making recommendations to the employers.
8. Helpful to the student to rate himself .

Limitations

1. Some characteristics are more difficult to rate.
2. Subjective element is present.
3. Lack of opportunities to rate students.
4. Raters tend to be generally generous.

The teacher or the counselor may rate each individual on each quality on a three-point, four-point or five-point scale. In a five-point scale, the description of the qualities of an individual may be 'Outstanding', 'Very good', 'Good', 'Average' and 'Poor'.

Principles Governing Rating Scales

1 . The trait to be treated should be readily observable.
2 . The specific trait or mode of behavior must be defined properly.

For example, we want to rate a child's originality in performing a task. First of all we must formulate a definition of 'originality' and then try to rate it.

3 The scale should be clearly defined, i.e. we are rating at a three, four or five-point scale.

4 Uniform standards of rating scale should be observed.

5 The rater should observe the rates in different situations involving the trait to be rated.

6 The number of characteristics to be rated should be limited.

7 In the rating scale card, some space maybe provided for the rater to write some supplementary material.

8 The directions of using the rating scales should be clear and comprehensive.

9 Several judges maybe employed to increase the reliability of any rating scale.

10 Well informed and experienced persons should be selected for rating.

Errors in Rating

(a) Generosity Error. Sometimes raters would not like to bring down their own people by giving them low ratings. The result is that high ratings are given in almost all cases. Such an error is known as "generosity error'.

(b) Stringency Error. The opposite of generosity error may be called stringency error. Some raters have a tendency to rate all individuals low.

(c) The Halo Error. I 'Halo' means a tendency to rate in terms of general impressions about the rates formed on the basis of some previous performance.

(d) The Error of Central Tendency .There is a tendency in some observers to rate all or most of the rates near the mid-point of the scale. They would like to put most of the rates as' Average, ' etc.

(e) The Logical Error. Such an error occurs when the characteristic or the trait to be rated is misunderstood .

Types of Rating Scales

1 **Descriptive.** The rater puts a check (√) in the blank before the characteristic or trait which is described in word or phrase.

- Shows marked originality.
- Willing to take initiative.
- Quite inventive
- On the whole unenterprising
- Very dependent on others.

2 **Numerical Scale.** Here numbers are assigned to each trait. If it is a seven-point scale; the number 7 represents the maximum amount of that trait in the individual; 4 represents the average amount.

3 **The Graphic Scale.** This is similar to the descriptive scale and the difference lies only in the way it is written. This is also called "behavioural statement scale". The following two examples may be noted:

(a) Example - Responsibility for completing work

Very High	High	Average	Low	Very Low

(b) Example --- social Attitude.

Anti-social	strongly self Centered Altruistic	has no positive attitude	usually considerate of others

4 **Percentage of Group Scale.** Here the rater is asked to give the percentage of the group that possesses the trait on which the individual is rated.

2.2 Projective Methods

Projective methods are those methods in which we provide the subject with relatively indefinite and unstructured material and then allow him to structure the material in anyway he likes. The subject projects his feelings, attitudes etc. In doing so, he unconsciously projects himself and reveals his personality. Examples of such materials through which the subject reveals himself are: making a story, ink blots, drawing a picture, etc. the method is designed to penetrate somewhat below the peripheral personality and to disclose latent needs, images, and sentiments which the subject would be unwilling or unable to embody in direct Communications.

"Typically, projective instruments also represent disguised testing procedures is so far as the subject is rarely aware of the type of psychological interpretation which will be made of his response. Projective techniques are likewise characterized by a global approach to the appraisal of personality .Attention is focused upon

a composite picture of the whole personality rather than upon the measurement of separate traits".

According to Anne Anastasia, "it is expected that the test material of the projective techniques" will serve as a sort of screen upon which the subject projects his characteristic ideas, attitudes, aspirations, fears worries, aggressions and the like'.

"The label projective techniques has been applied to settings or materials designed to give a person a chance to reveal his thoughts and feelings while seemingly responding to something in the external environment". Write Gates and others.

To quote Ruth Strang, "Projective techniques are a method of understanding the inner world of the individual. Rapport describes the aim of projective techniques as, "to elicit, to render observable, to record, and to communicate the psychological structure of the subject, as inherent to him at any given moment and without study of historical antecedents". An individual, it is recognized, reveals or 'projects' his personality in a free and unrestricted activity .

Murray writes, "The purpose of this procedure is to stimulate literary creativity and thereby evoke fantasies that reveal covert and unconscious complexes. The test is based on the well recognized fact that when a person interprets an ambiguous social situation, he is apt to expose his own personality as much as the phenomenon to which he is attending."

L.K. Frank was the first person to use the term 'Projective Technique' in an article that appeared in 1939 though such methods had been in use for many years prior to that date.

Projective techniques are very useful with young children, illiterates and person with language handicaps or speech defects.

Non-verbal media is readily applicable to all these groups.

Common Characteristics of Projective Techniques

1. The stimulus material is ambiguous and weakly structured and the subject is expected to supply meaning, significance, organization or in some other way leave the impression of his personality upon the under defined stimulus of situation.
2. An attempt is made to explore the psychological reality or the underlying basic personality factors of the individual which consist of his hopes, aspirations, needs motives, moods, attitudes, conflicts, complexes, fears, etc.
3. Projective techniques tap the implicit or unconscious aspects of the personality.
4. Projective techniques require sophistication in administration and interpretation.

Kinds of Projective Techniques

Some of the projective tests are standardized and are widely used. These include Ror chach's Ink Blot Test, Thematic. Test and Various other picture tests like Children's L. Apperception Test, and Blocky Pictures, the Michigan Picture Test, Rosenzweigh Picture Frustration Study, etc. Out of these Rorschach's Ink Blot and Thematic Apperception Test (TAT) are very popular and widely used.

(i) Free Association and Dream Analysis. Freud used this method to find out the repressed unconscious desires, emotions and feelings of individual men and women. It has been recognized by many psychologist that day dreams and even especially dreams are rather difficult to interpret but they do not just happen. They have their roots in the person's private world.

According to this method, the psychologist first of all develops rapport with the subject and then the subject is asked to take a comfortable position reclining on a sofa and is encouraged to talk about his troubles freely after a black cloth is tied over his eyes. The psychologist or the psychoanalyst records his responses and interprets them. After many such sittings, it becomes possible for him to have a clue of the personality of the subject.

(ii) Incomplete Sentence Technique. The method is very simple. Some incomplete sentences are presented to the child whom he completes in any way he likes. By reading the individual's responses, the counselor records observations which indicate unhealthy or conflicting, positive or healthy attitudes, etc. Stimulus in the form of the following words maybe presented. to the child.

(a) A baby...

(b) The world....

(c) My home .....

(d) My father.....

(e) Money.....

(f) My best friend...

The child may complete these sentences in this manner:

(a) A baby is very lovely.

(b) The world is full of cruel people.

(c) My home is not a good place to live in.

(d) My father is very cruel.

(e) Money is everything in life.

(f) My best friend is very poor.

The psychologist, by going through these sentences, may get some clue of the repressed wishes of the child.

(iii) Original Drawing and Paintings. Original drawings and paintings of the children give an idea of the personality of the child. The theme or subject he chooses, the color he uses, the, masses and the open spaces that the paintings contain, the length, direction, curvature of the lines -all these give an indication of the yarious traits of the personality of the child.

(iv) Play Situation. Many dolls are given to the child to play with and observation is made of his reactions. These dolls represent a father, a mother, a boy and a girl. These dolls provide the medium through which the child might act out or reveal his thoughts. impulses and feelings about himself and others. Situation should be such as to give a free play to his ideas at the unconscious level. Controlled play techniques are also used.

In a free play a boy who happens to be jealous of his sister might assign the role of the sister to baby doll and cause it punished by the father doll or mother doll, or shut the baby doll to a far corner of the room. This will indicate that the girl is preferred to him in the family.

A great variety of material may be used in play situations.

(v) Rorschach Ink Blot Tests. Rorschach was a Swiss Psychologist who experimented with the use of ink blots as a means of diagnosing mental disorders in 1921. Perceptual approach is the basis of this test. The perception of an individual is influenced by the emotional and social make up when he is asked to perceive a figure which is not well defined.

Rorschach experienced for years with thousands of ink blots and ultimately selected ten ink blots which proved to have the greatest diagnostic value. There are ten ink blots, five in black and white, two with splashes of red and three in other colors. These are printed on 7x 9 1/2 inch cards.

Each card is given to the subject in turn who is asked to tell what he sees in the ink blots, what that means to him and what that might be. The counselor notes whatever the subject says. The counselor then shows the card a second time and through well worded questions he elicits answers. The success of the counselor depends to a great extent upon his skill in asking questions that will clarify the free responses. Special techniques have been developed for recording the responses of the subject.

This technique has the following uses:

1. This helps to distinguish between 'normal' persons and persons in need of psychiatric treatment.
2. This helps to show the potential intelligence of an individual.
3. This helps to predict and plan for academic success and adjustment in college.

This may prove to be helpful in indicating personality tendencies significant for vocational success.

(vi) Story-telling and Story-Completion. An incomplete story may be told to the child and he may be asked to complete it. While completing the story, the child may reveal something about his feeling and desires. He may be asked to let his imagination go to show how imaginative he is. The situation selected should be such as touches the emotional life of the child whose personality traits we wish to assess. For instance, the child is asked to relate a story about an

innocent child who was beaten by his father. The child is likely to depict his innermost feelings, wishes and thoughts while he narrates the story .Take another case when a child is told about the beginning of a story. Father is ready to take the children to the exhibition. All of a sudden a guest comes. The child may be asked to complete the story. As the child finishes the story, the psychologist comes to know the personality of the child. This method is usually employed in the case of delinquent child.

(vii) Thematic Apperception Test (TAT). This test was designed by Morgan and Murray in 1935. The test involves a systematic use of pictures representing a number of dramatic events. The pictures are shown to the subject who identifies himself with some picture. The picture presents a stimulus to the child to talk freely. The child makes interpretations in accordance with his own past experience and present needs and attitudes. To quote Rath Strang, "To some extent, he reads into the pictures his own experiences -the responses made in the test situation also reveal the person's attitudes and ways of thinking in life situations."

The pictures selected have the following characteristics:

First, no detail in the background is provided so that this may not limit the pictures to a certain time or place. Second, the theme of these pictures is vague.

Third, the action and expression of the characters in the pictures are not very clear and are rather vague and ambiguous.

Fourth, the contents of the pictures do not give the whole story .They are incomplete.

Fifth, home and social situations relate to the ordinary situations of life.

Sixth, characters in the pictures are such as the child can very easily identify himself.

The test can be with employed suitably to persons over four years of age whose intelligence quotient is not lower than 80. The counselor shows one or more of the pictures and puts questions like these. What does the picture mean to you? What can be the factors that have led to that situation? What will be the outcome?

In interpreting pictures, it is important to note not only turn of the thoughts the child injects into the situation, but also his mood.

Advantages of Projective Techniques

1. An individual reveals himself in various situations and sometimes he is not aware of this fact. Thus we get reliable information.
2. The connection between diagnosis and the situation is very close.
3. It is not possible for the individual to give readymade, habitual or conventional responses as the tasks presented are novel and unstructured.
4. These techniques encourage spontaneous responses.
5. These enable us to have a total view of the personality of an individual, rather than in piece meal.

Limitations

First, they are very subjective.

Second, they require a lot of training in their administration. Only trained psychologists can administer them.

Third, they are time consuming.

Fourth, they are very difficult to interpret.

Lastly, there are very few standardized tests.

2.3 Self Reporting Techniques

2.3.1 Anecdotal Record

An anecdotal record is a running description of actual examples of behaviour of a student as observed by teachers and the counselor. It is followed by his comments.

According to Brown and Martin," Anecdotes are descriptive accounts of episodes or occurrences in the daily life of the student" Anecdotal record has been defined by Randall as a record of some significant item of conduct, a record of an episode in the life of the student, a word picture of the student in action. .., a word snapshot at the moment of the incident, any narration of events in

Which the student takes such a part as to reveal something which may be significant about his personality.

Raths Louis thinks that "an anecdotal record is a report of a significant episode in the life of a student". Traxler thinks, "This record, as the name implies, involves setting down an anecdote concerning some aspect of pupil behaviour which seems significant to the observer. "

Zann, D. Willard regards an anecdotal record "as a simple statement of an incident deemed by the observer to be significant with respect to a given pupil."

A Specimen of an Anecdotal Record

Name of the School

Name of pupil observed Class_____

Subject_____

Observer Dated and place

Objective Description

Characteristics and Preparation of an Anecdotal Record We cannot set any limit on the number of anecdotes to be recorded. It depends upon the time which is at the disposal of the teachers or guidance workers. The following points should be considered in connection with these records:

1 These supplement other records and should not be considered as substitutes.

2. The objective description of the behavior should not be mixed up with the subjective comments.

3. Any significant behavior, be it in the classroom, in the school or outside the school, should be recorded.

4. Behavior, whether it is favorable, unfavorable or neither favorable nor unfavorable to the child, should be recorded.

5. The facts presented in all the anecdotes must be sifted and arranged so that they may be studied in relation to one another .

6. The record should be regarded as confidential. It should not fall into irresponsible hands.

7. What is written down is what was seen or heard. Inferences, guesses, assumptions are omitted unless they are clearly labeled as inferences, guesses, or assumption.

8. The observer observes only that aspect of the behaviour which has been earlier determined and records only that.

9. Only those words and phrases are used whose meaning is clear, and so far as possible, unequivocal.

10. Words and phrases are employed that are definable in terms of things rather than other words. Concrete statements are preferred to abstract ones.

11 Words and phrases that have strong emotional connotations are avoided i.e. love, hate, insolvent courteous, loyal, dishonest, etc.

12. Words and phrases are avoided which express the observer's judgment, or his opinion, and not his perception. Among the frequently encountered,

"judgmental terms" that should be avoided are the following:

(a) Well-behaved

(b) Delinquent.

(c) Aggressive

(d) Didn't try.

(e) Industrious.

(f) Nervous

(g) Happy.

Values and uses of Anecdotal Records

1 They provide specific and exact description of personality and minimize generalizations.

2. They are very helpful in understanding the child's behaviour in diverse situations.

3. They provide a continuous record.
4. They provide data for pupils to use in self appraisal.
5. A summary of these records is valuable for forwarding with a pupil when he is transferred from one school to another.
6. The new members of the staff may use these records and acquaint themselves with the student.
7. These records aid in clinical service.
8. They stimulate teachers to use the records and to contribute to them.

Limitations

1. Anecdotal records tend to be less reliable than other tools.
2. They are time consuming to write.

2.3.2 Autobiography Technique

In this technique, a child gives an account of his life experiences from early childhood to the date of writing. This technique provides a lot of material to the teacher to measure the several aspects of the development of the personality of the child. The autobiography is also a tension releaser. It serves as a 'psychic safety valve' and allows the child to express his pent up feelings.

Organization of Autobiographies

Following outline in writing an autobiography may be used.

1 My familyz
2. My first year before school
3. My years in the elementary school
4. Places I have lived
5. Vacations I have spent

6. Trips I have taken
7. The way I usually spend the afternoon when school is over
8. The way I spend my evening after supper
9. The way I spend my Sunday
10. Subjects I like the most, the least
11 Some subjects and activities I wish our school would provide
12. Things I should like to do well
13. The work I hope to do (three-choices) and reasons for the choices.
14. Kinds of magazines and books I like to read, kinds of Radio and TV programmes I like to listen and watch, kinds of pictures I like to see.
15. If I could have all the things I wanted by asking for them, I would ask for these things.

Principles of Using this Method

1 Rapport with the student should be established so that the child expresses his feelings freely.
2. He should be ensured that the information supplied would be kept confidential.
3. Detailed instructions may be given to the child as to way in which autobiography is to be written.
4. Questions/hints may be given to the child so that he may reveal his history.

Limitations

1. Its scope is limited and it cannot be used when the child is unable to express himself correctly in speech.

2. The student may present facts in a distorted way.

Interpreting the Autobiography

1 First reading of the autobiography maybe done in a general way.

2. Second reading should be more thorough.

3. The general appearance of the document; its neatness and contents would provide some clue to the understanding of the personality of the child.

2.3.3 Interview

An interview can be define as a face to face verbal exchange in which one person i.e., the interviewer attempts to elicit information on a variety of topics from the interview Interviews are used for a variety of purposes and as such there are various types of interviews.

1. The assessment or evaluative interview for determining the fitness of a person for admission, for a job or for scholarship, etc.

2. The personality assessment interview of a student for finding out the status of development of his personality.

3. The diagnostic interview for getting some information about the home, environment and school situations of the student or the client.

4. The introductory interview for preparing for further interviews.

5. The informative interview for giving some information to the students on subjects or careers, etc.

6. The research interview for collecting data about a problem.

7. The administrative or disciplinary interview for finding out the causes of indiscipline and taking further action.
8. The counseling interview for helping the counselee or the student in gaining insight into the problem and assisting him solving the same.

It is to be remembered that in educational institutions, the primary purpose of interview of personality assessment is to find out the existing status of the development of the various facts of personality and to take further measures for bringing about improvement in the learner.

Main Points in Successful Interviewing

1. Who is to interview? The interviewer must be clear in his mind of his strengths and weaknesses, so that prejudices or bias could be minimized.
2. Whom -He should understand his client thoroughly.
3. How -He should know the technique of interviewing.
4. Why -He should be very clear in his mind for the purpose of interview.
5. Where -Suitable place should be selected for this purpose
6. When -Proper motivation should be developed in the child.
7. What to ask -He should prepare his questions thoughtfully.

Preparation for the Interview

1 .A quiet and orderly place should be provided for interview purposes.

2. The teacher or the psychologist should prepare himself for the interview by gathering all the data concerning the student.

3. The teacher or the psychologist should prepare clear objective for the interview.

4. The teacher or the counselor should keep an open mind regarding the child.

5. Each interview should be considered as a step in the process of guiding pupil.

Techniques of Interviewing

Rapport should be established. Rapport is a technical term used to denote the feelings of friendliness, security and mutual confidence between the teacher and the child.

Davis and Robinson suggest the following techniques to increase rapport:

Sympathy -The teacher counselor expresses sympathy for the child.

Assurance -The teacher or the counselor will try to ease the child's fear by encouraging him that this step is for his personality development.

Humor-The teacher or the counselor expresses his approval with something that the counselee makes with humorous remarks.

Personal Reference -The teacher or the counselor tells about some of his own experiences to illustrate his point by saying, 'I would do this'.

Non-Personal Reference -The teacher or the counselor cites an example of the experience of another person.

Question Form -The teacher or the counselor asked questions to stimulate the counselee to think further about his problem.

Threat -The teacher or the counselor sometimes threatens the counselee that unpleasant results may occur if the counselee does not follow a certain specified course of action.

Merits of an Interview

1 it is the most dynamic way of understanding the individual.

2. It is natural like conversation.

3. it can be made flexible so as to suit many situations.

4. it is relatively easy to conduct.

5. It is possible to get some most confidential information from the child, which otherwise he may hesitate to reveal through writing.

Limitations of an Interview:

1 It needs a trained and competent interviewer.

2. It is costly in terms of labour, money and time.

3. it suffers from the subjective bias of the interviewer .

4. The interviewee may not unfold himself.

5. The interviewee may have a language handicap to express his feelings freely.

2.3.4 Questionnaire

Good and Hatt point out, "In general the word questionnaire refers to a device for securing answers to questions by using a form which the respondent fills in himself."

Barr. Davis and Johnson define questionnaire as a "systematic compilation of questions that are subject to a sampling of population from which information is desired. " A questionnaire may contain two kinds of items.

(i) The Closed or Structured Form.

(ii) The Open-End or Unrestricted Form.

The Closed Form. This form requires short and 'check' responses. It may provide the marking 'Yes' or 'No', or just a 'check' from a list of suggested responses.

Main advantages

(i) It is very easy to respond to such a questionnaire.

(ii) It takes a little time to answer.

(iii) It keeps the respondent on the subject.

(iv) It is relatively objective.

(v) It is fairly easy to tabulate and analyses.

Drawback of the Closed Form. The closed form does not provide any opportunity to the respondent to express his views very clearly as there is no scope for explanatory information.

The Open-End or Unrestricted Type or Fee Responses or Unstructured Form. As the name of the form indicates, the respondent is at liberty to express his attitudes, interests, preferences, and decisions in his own words because no clues are provided. However, sometimes it becomes very difficult to tabulate, interpret and summarize such responses.

As far as possible, the questionnaire should be a balanced one and it should contain both the open and closed type items.

Questionnaire Compared with other Techniques of Evaluation

1 It is less expensive and less time consuming than interview or observation .

2. Its construction needs less technical skills as compared with those required in conducting interviews and observations.

3. Questionnaire do not permit variations in questions and as such they help in focusing the attention of the respondents on all the significant items. The interviewing situation, on the other hand, is rarely uniform.

4. A questionnaire places less pressure on the subject for immediate response. A subject has adequate time to think of the responses whereas in an interview the subject is expected to give his responses immediately.

5. In an interview the investigator records responses whereas in the questionnaire, the responses are given in the language of the subjects. This brings validity to responses in the case of a questionnaire.

2.4 Sociometric Techniques

Sociometry was designed by J. L. Moreno and Helen Jennings in 1946. Through sociometric techniques we come to know what other members of the group feel about the subject whether they like him or not The members of the group may be asked to name in order of preference one or two individuals with whom they would like to work or play.

According to Jennings, "Stated briefly, sociometry may be described as a means of presenting simply and graphically the entire structure of relations existing at a given time among members of a given group. The major lines of communication or

the pattern of attraction and rejection in its full scope, are made readily comprehensive at a glance. "

William J. Goode and others state. "These and other variants of sociometric techniques offer rather simple methods of ranking individuals on a continuum of 'acceptability' or 'outgoingness' on the part of group members. When their use is justified they may be powerful research tools since they meet the general problems of scaling very well."

Sociometric studies have been made of many types of social groups including classroom groups. Being peer rating rather than rating by superiors, sociometry adds another dimension to the understanding of social relationships.

Example -Each group consisting of eleven students is asked to write his first choice about some significant and pertinent type of social setting. He may be asked questions like this:

1. Whom would you like to be the secretary of your debating society?
2. Whom would you like to sit next to you in the class or in the bus whire going for a picnic?
3. With whom do you enjoy most?
4. With whom would you like to work in the laboratory?
5. With whom would you like to walk home?

All those questions are positive questions and hence show social acceptances.

Negative questions may also be given to show social rejections.

In the above example, the individual has to name only one person of his choice.

Data maybe tabulated as under:

(i) Let the members of the group be numbered from 1 to 10.

(ii) Write 'Choosers' in the vertical column and chosen in the horizontal column.

(iii) Total choices received by each member maybe shown at the bottom.

(iv) In the cells, check marks may be shown.

(v) Add the number of each choice.

A similar table can be prepared for social rejections. In the vertical column will be listed the 'rejectors' and in the horizontal column 'rejected'.

An individual may be asked to name more than one person

In order of preference .

- **Interpretation of a Sociogram**

1. One person should be concentrated at a time.
2. A detailed study of the choices made and received should be made.
3. The 'isolates' and the 'stars' maybe looked for. An 'isolate' is one whom nobody chooses. Of course he is not rejected. A 'star' is a member of the group who receives most of the choices.
4. Attempts should be made to discover the causes for such selections.

An individual may be isolated because:

(i) He is a new member of the group.

(ii) He is of a shy and withdrawing nature.

(iii) He does not try to make friendship with others.

(iv) He may belong to a lower or upper socio- economic level and therefore he is not acceptable in the group.

5. Look for individuals who select each other. This might be due to factors like:

   (i) Close relations.

   (ii) Neighbours.

   (iii) Common interest and the like.

6. A triangle shows three persons selecting each other.

This may be an evidence of cliques, or sharp divisions in the group.

Role of the Teacher and the Guidance Worker

In general, he can work on three points:

1. Providing opportunities for developing friendly relations.
2. Improving social skills.
3. Building up competency for accomplishing something.

Reliable results can be achieved only when all the members comprising a group are fully acquainted with each other. The worker the counselor must establish friendly relations with the members of the group so that they may give their frank opinion about an individual or individuals.

Interpretation

1. Roll No.5 is the 'star' as he has been chosen by the maximum number of students.
2. Roll Nos.7 and 9 did not get any choice. This indicates that they tended to be isolated i.e. not being social.
3. Roll Nos. 2 and 8 came next to roll no.4.

4. Roll No.5 i.e. the star preferred roll no.2.

5. Mutual choices were: 2 and 5; 3 and 4; and 2 and 10.

Notes.

1 Students can make to give their second and third choices also.

2. Choices are illustrated in Fig.

Sociogram

Practical Work

Object. To find out the social structure of a group of 1 O students.

Material. A questionnaire asking students: Whom would you choose to be the secretary of your club?

Procedure. After taking the students into confidence, one line questionnaire as indicated above was given to the students to mark against one of the 10 Roll Nos. Thereafter tabulation of the student's responses was made.

Fig.1 Sociometric Matrix showing who chooses whom

3. There are several ways of preparing a sociogram.

4. This seems to be the simplest form of a sociogram .

Using Sociometric Techniques

1. To study the relationships among members of the group and to improve them.

2. To organize classroom groups.

3. To assist those who have become isolates in the group.

4. To assign responsibility to the members °'the group.

Limitations

1 The relationships are not necessarily stable .

2. Some members of the group may not reveal their real relationships on account of some fear or other considerations.

2.5. Summary

Observational techniques are especially useful in evaluating performance skills and certain aspects of personal-social development. In addition, the results of observation supplement and complement paper-and- pencil testing by indicating how pupils typically behave in natural situations.

The least structured observational technique is the anecdotal record. This is simply a method of recording factual descriptions of pupil behaviour. To make anecdotal record keeping feasible, it is desirable to restrict observations at any given time to a few types of behaviour or too few pupils.

Anecdotal records have the advantages of (1) describing behaviour in natural settings, (2) highlighting evidence of exceptional behaviour apt to be overlooked by other techniques, and (3) being usable with the very young and the retarded. Rating methods are a

systematic procedure for obtaining and recording the observers' judgments. Of the several types of rating scales available, the descriptive graphic scale seems to be the best for school use. In rating procedures, products, and various aspects of personal-social development, certain types of errors commonly occur. These include (1) personal bias, (2) halo effect, and (3) logical errors.

Checklists perform somewhat the same functions as rating scales do. They are used in evaluating procedures, products, and aspects of personal-social development where an evaluation of the characteristics is limited to simple "present-absent "judgment.

Self-report techniques are used to obtain information that is inaccessible by other means, including reports on the pupil's attitudes, interests, and personal feelings. Such information can be obtained by means of a personal interview, but a self-report inventory is more commonly used.

The inventory is a sort of standardized written interview that produces comparable results from one person to another.

3. SELF ASSESSMENT QUESTIONS

1. Define aptitude. What are the uses of aptitude test? How will you measure aptitude?
2. Critically examine the merits and demerits of personality inventory.
3. What are the various techniques of assessing personality? Describe one of them.
4. How would you measure intelligence? Compare the Wechsler scales and culture-free tests.
5. How a standardized test is prepared? Discuss some of the good standardized tests?
6. Define a checklist and describe the steps involved in the construction of this instrument.
7. Describe the various types of rating scale and mention its uses in student's assessment.
8. Discuss the advantages and disadvantages of sociometric techniques.
9. Describe the various types of sociometric techniques.
10. Define the term interview and discuss four main types of interview.
11. Describe the advantages and limitations of an interview.
12. Define a questionnaire and discuss characteristics of a good questionnaire.
13. Discuss the uses and limitations of attitude scales.
14. Describe the various types of projective techniques.

4. BIBLOGRAPHY

1. . Ebel. L. Robert (1972), Essentials of educational measurement, New Jersey: Prentice-Hall, Inc.
2. Jum, C. Nunnally (1972), Educational measurement and evaluation, New York. Mc Graw-Hil1 Company.
3. J. Stanley. Ahmann, and Glock, Marvin D. (1981), Evaluation student progress. London. Al1yn and Bacon, Inc.
4. Larry, A. Hjelle and Ziegler Daniel (1992) personality theories, New York, Mc Graw-Hill, Inc.
5. Gronlund, Nonnan, E. and Linn, Robert L. (1990): Measurement and evaluation in teaching: New York, Macmillan Publishing company.
6. Wiersma, William and Jors. Stephen G. (1990): Educational measurement and testing London, Ally and Bacon. hlc.

Unit-6

PLANNING AND ADMINISTERING FOR CLASSROOM TESTING

BY;

Jamil Hussain Shah

INTRODUCTION

Effective teaching is not matter of whim, rather, good teachers know what changes in student behaviour they want to produce and hold themselves account able by modifying strategies until objective have been met. In practice, teaching involves setting goals that consider student need and backgrounds and selecting the most effective instructional strategies. Since many goals and teaching strategies are not static and unchanging, the process of goal setting and trying out new strategies is never ending process. In this non-static and ever changing process of teaching- learning, effective classroom testing beings with a test plan that specifically describes the instructional objectives and content to be measured and the relative emphasis to be given to each intended learning outcome. This followed by the selection of the most appropriate items formats and the preparation of test items into a test prepare directions, administer the test, score the, interpret and appraise the test results.

Our goal throughout the preparation and use of classroom tests is to obtain valid evidence of pupil learning. Valid achievement testing is the end product of a systematically controlled series of steps involved in teaching - learning process. Although validity is built in during the construction of the test item, systematic procedures of assembly, administration and scoring will provide greater assurance that the items will function as intended. Keeping the importance of planning, assembling, administering and scoring of classroom tests in mind, it is decided to point out all aspects in detail.

OBJECTIVES

After reading this unit, you should be able to:

1. Discuss the importance of construction the most appropriate and relevant test items.
2. Plan an ideal an test for an ideal purpose.
3. Assemble the classroom tests.
4. Administer and score the classrooms tests.
5. List advantages and disadvantages of different types of test items.
6. Specify the rules for writing different types of test items
7. Indicate how the effects of guessing can be reduced.

Contents

Introduction

Objective

1. General Consideration in constructions of test items

1.1. The Representativeness of the sample of items

1.2. Test Length

1.3. Proper item Difficulty

1.4. Elimination of irrelevant Barriers to the Answer

1.5. Avoidance of unintended Clues to the Answer

1.6. Focus on improving Learning and instruction

1.7. Self Assessment questions

2. Test planing considerating

2.1. The Purpose of Classroom testing

2.2. Development of Test Specifications

2.3. Establishing Priorities Among Objectives

2.4. Self Assessment questions

3. Assembling the classroom test

3.1. Recording Test Items

3.2. Reviewing Test Items.

3.3. Arranging items in the test

3.4. Preparing Direction for the test

3.5. Self Assessment questions

4. Administering and scoring the classroom tests

4.1. Administering the test

4.2. Scoring the Classroom Tests

4.2.1. Correction for Guessing

4.3. Self Assessment questions.

5. Bibliography

1. GENERAL CONSIDERATIONS IN CONSTRUCTION OF TEST ITEMS

The construction of items for a classroom test should proceed by a series of preliminary steps:

First, the purpose of the test should be determined.

Second, a set of specifications should be developed.

Third, the most appropriate item types should be selected .

Finally, the test items should be constructed in accordance with the specifications developed during the preceding steps.

A classroom test is most likely to provide a measure of the instructional objectives if the test items are designed to measure the performance defined by the specific learning outcomes, as follows:

Example Specific Learning Outcomes: Identifies the function of a given body structure.
Relevant Test item:
What is the function of kidney?
A. Eliminate waste products.
B. Improve the circulation of blood.
C. Maintain respiration.
D. Stimulate digestion.

Thus, the preparation of relevant test items means analyzing the performance described in the specific learning outcome (i.e., " Identifies the function of ") and the construction of a test item that calls forth that performance (i.e." What is the function of ...?"). Note in our example that the specific learning outcome defines the type of response the pupil is expected to make, but it does not

indicate the specific body structure (i.e., kidney) the pupil is to identify. Keeping the learning outcome free of specific course content, like this , makes it possible to key the intended response to various areas of content. For example, pupil could be asked to identify the function of heart, the lungs, the muscles, or any other body structure pertinent to the course's content. The desired pupil performance stated in the specific learning outcome can be keyed to each specific area of content by means of the table of specifications.

In some cases it may be desirable to prepare a general item pattern as intermediate step between the specific learning outcome and the test item. A general item for our illustrative test item, for example, would be as follows

What is the function of.?

An item pattern such as this could be completed by adding the name of any body structure and using it as short-answer question or, in addition, by listing appropriate alternatives and using it as a multiple-choice item. Thus, using the item pattern as a guide, we could generate a large number of relevant test items for this particular learning outcome .This procedure is specially useful when a file of test item is being prepared or when more than one form of the test is needed (e.g. pre-testing -post -retesting in mastery learning).

When item pattern are used as guide to test construction, they can be arranged by general type of learning outcome:

Knowledge Outcomes

1. What is the name of ...?
2. What is the location of.....?
3. What are the characteristics of...?

4. What is the function of.....?

Understanding outcomes

1. What is the reason for.....?
2. What is the relationship between...?
3. Which of these is an example of...?
4. Which of these best summarizes...?

Application outcomes

1. What method would be best for ....?
2 What steps should be followed to construct...?
3. Which of these indicates correct application of ...?
4. Which of these solutions is correct for...?

Item pattern like these should not, of course be developed haphazardly. Rather, they should be derived form the specific learning outcomes, they represent .Although it usually will not be possible to develop item patterns for all outcomes, listing them will help generating pools of relevant test items. The test construction time saved by using such a list can be profitable used to construct more effective items in those areas in which item patterns are infeasible.

1.1 The Representative ness of the Sample of Items

A test no matter how extensive is almost always a sample of the many possible test items that could be included. For example, we expect pupils to know thousands of facts, but we can test for only a limited number of them, we expect pupils to develop understanding applicable to innumerable situations, and we expect pupils to develop thinking skills that will enable them to solve a variety of problems, but we can test their problem-solving ability with only a limited number of problems. In each areas of

content and for each specific learning outcome, than, we merely select a sample of pupil performance and accept it as evidence of achievement in that area. We assume that the pupil, s responses to our selected set of test items are typical of what their responses would be to other test items drawn from the same area. This means, of course, that our limited samples must be selected in such a way that provide as representative a sample as possible in each of the various areas for which the test is being developed.

The problem of obtaining a representative sample, as noted earlier, is greater with norm-referenced testing than with criterion-referenced testing because of its broader coverage. For both types of tests, however, our sampling is most likely to be representative when test preparation is guided by a carefully prepared set specification. Unless a table of specification, or some similar device, is used as guide in test construction there is a tendency to overload the test the with items measuring knowledge of isolated facts and to neglect the more complex learning outcomes. In the social studies area, for example, it is not uncommon to include a disproportionately large number of items that measure knowledge of names, dates, places and the like. In Science, defining terms and structures and functions are commonly overemphasized. In mathematics, computational skill is frequently the only learning outcome measured. In language, arts and literature, the identification of parts of speech, literary characteristics, authors, and the like is frequently too prominent. These learning outcomes are, generally, not stressed because we think knowledge of isolated facts is more important than understanding, applications, and various thinking skills. Rather, they usually receive undue prominence because we find it easier to construct such items. With out a carefully developed test plan, ease of construction all too frequently becomes the dominant criterion in constructing tests

items. As a consequence, the test measures a limited and biased sample of tasks and neglects many outcomes of greater importance.

1.2 Test Length.

The length of a test is also an important factor in obtaining a representative sample. Test length is determined when the set of specifications is built and depend on such factors as the purpose of testing, the types of test items used, the age of the pupils, and the level of reliability needed for effective test use. Thus, a criterion-referenced mastery test over third-grade social studies until might contain 30 objective items, whereas a norm-referenced survey test over a tenth-grade social studies course might contain more than 100 objectives items and several essay questions. Although there are no hard and fast rules for determining test length, an important consideration from a sampling standpoint is the number of test items devoted to each specific area being measured. We want our classroom test to be long enough to provide an adequate sampling of each objective and each content area. As a rule of thumb, it is desirable when constructing a criterion –referenced mastery test to use at least ten objective test items to measure each specific learning outcome. This number, however, might be lowered to as few as five if the task is extremely limited (e.g., "Adds two single-digit numbers," "Capitalizes proper names") and the pupils are to supply the answers rather than to selection them. For a norm-refererıced test, where the sample of test items typically covers a board area and emphasis is on the total score, using several objective test items for each specific learning outcome and ten more for each general objective would probably be sufficient.

Special problem of sampling arise when complex learning outcomes are being measured, because here we must turn to more elaborate objective-type items and essay question. Both items types require considerable testing time, but a single test exercise is still inadequate for measuring intended outcome. One exercise calling for the interpretation graphs, the nature of the data or the type of graph may be the most influential factor in determining whether it is interpreted properly. When several graphs are used, the effect of such factors is minimized, and we obtain a more representative sample of the ability to inter prêt graphs. A similar situation occurs with the use of essay questions. The answer to any single question depends too heavily on the particular sample of information called for by the question and thus the only feasible is to confine each test of complex outcomes to a rather limited area (e.g., graph interpretation, problem solving)and to test more often. In any event our aim should be to obtain as representative a sample of pupil performance as possible in each area to be tested other things being equal, the greater the number of test items the greater the likelihood of an adequate sample and thus the more reliable the results.

1.3 Proper Item Difficulty

The difficulty of items to be included in a classroom in a test depends largely on whether the test is being designed to describe the specific learning tasks the students can perform(e.g. CRT) or rank the students in order of their achievement (e.g. NRT)

Item Difficulty and Criterion Referenced Testing. The difficulty of test item in a criterion referenced mastery test is determined by the nature of the specific learning tasks which to be measured. If the learning tasks are easy, the test items should be easy. If the learning tasks are modify item difficult, the test items

should be moderately difficult. No attempt should be made to moderately difficult, the test item difficulty or to eliminate easy items from the test in order to obtain a range of test scores .On a criterion referenced mastery we would expect all, or nearly all, student to obtain high scores when the instruction have been effective . Special care should be taken into account e.g. avoid irrelevant barriers (ambiguity) to the answers, unintended clues to the correct response, or any other factor that might alter the level of difficulty of the test task.

For criterion referenced test at the developmental level of learning, we need test item of varying difficulty for each instructional objectives. Ideally, the difficulty of the test tasks would be derived directly from the instructional content, rather than form some arbitrary attempt to manipulate item difficulty. It should also be kept in mind that in criterion referenced mastery testing, a wide range of scores is expected.

Item Difficulty and Norm Referenced Testing. Because norm referenced tests are designed to rank students in order of achievement, deliberate attempts are made to obtain a wide spread of scores. That is why easy items are eliminated to maximize the differences in students' performance. Maximum differentiation among students in terms of achievement is obtained when the average score is near the midpoint of the possible scores, and the scores range from near zero to near perfect. The average difficulty to try for on a 100-item test for various choice-type items would be as follows:

	Chance Score	Average Difficulty
Two Choice Item (e.g., True-false)	50	75
Three-choice multiple-choice item	33	67

Four-choice multiple-choice item	25	63
Five-choice multiple-choice item	20	60

1.4- Elimination of Irrelevant Barriers to the Answer

When constructing items for a classroom test, care must be taken to eliminate any extraneous factors that might prevent pupils from responding. If pupils have achieved a particular learning outcome (e.g., knowledge of terms), we would want them to answer correctly those test items that measure the attainment of that learning outcome. We would be very unhappy, if they answered such test items incorrectly merely because the sentence structure was too complex, the vocabulary too difficult, or the type of response called for unclear. These factors, which are extraneous to the central purpose of the measurement, limit and modify the pupil's responses and prevent them from showing their true levels of achievement. Such factors are as unfair as determining a person's running ability when an ankle is sprained. Although a measure of running ability would be obtained, the performance would be restricted by a factor we did not intend to include in our measurement.

One way to eliminate factors that are extraneous to the purpose of a measurement is to be certain that all pupils have the prerequisite skills and abilities needed to make the response. These have been called enabling behaviours because they enable the pupil to make the response but are not meant to be critical factors in the measurement. That is, they are a necessary but not a significant condition for responding correctly. Probably the most important enabling behaviour in objective testing is reading skill. In essay testing, skill in written expression is an additional factor to be considered. In measuring, understanding thinking skills, and

other complex learning outcomes, knowledge of certain facts and simple computational skills might also be necessary prerequisites.

In constructing test items, we need to strive for items that measure achievement of the specific learning outcomes and not differences in enabling behaviours. Differences in reading ability, computational skills, communication skills, and the like should not influence the pupils' responses unless such outcomes are specifically being measured. The only functional difference between those pupils who get an item correct and those who miss it should be the possession of the knowledge, understanding, or other learning outcome being measured by the item. All other differences are extraneous to the purpose of the item, and their influence should be eliminated or controlled for valid test results.

A special problem in preventing extraneous factors from distorting our test results is avoiding ambiguity Objective test items are especially subject to is misinterpretation when long, complex sentences are used, when the vocabulary is unnecessarily difficult, and when words that lack precise meaning are used. Thus, from viewpoint of both level of reading difficulty and preciseness of meaning, the antidote for ambiguity seems to be a careful choice of words and the use of brief, concise sentences. In some cases, ambiguity can be reduced by using pictures or other illustrative material in place of verbal descriptions. When this is done, the illustrative material must, of course, also be carefully checked to make sure it is clear and unambiguous.

An effort should also be made to avoid any racial, ethnic, or sexual bias in preparing the test items. The vocabulary and test situations should be acceptable to various racial and ethnic groups and to both males and females and should be free of stereotyping. For example, in presenting characters in a story problem, a

reading passage, or other test situation, minorities should not be portrayed as having subservient roles. Similarly, test situations should not always place males in such traditional roles as athlete, business executive, and professional person and females in such traditional roles as homemaker, teacher and nurse. A balanced use of different roles for minorities and males and females is necessary if we are to avoid bias as possible barrier to maximum test performance.

Some Possible Barriers in Test item

Ambiguous statements.
Excessive wordiness.
Difficult vocabulary.
Complex sentence structure.
Unclear instructions.
Unclear illustrative material.
Race, ethnic, and sex bias.

1.5- Avoidance of Unintended Clues to The Answer.

Test items should be constructed so that pupils obtain the correct answer only if they have attained the desired learning outcome. This is the counterpart of the preceding principle. In that one, we were concerned with those factors that prevent pupils from responding correctly, even though they have attained the desired learning outcome. Here we are concerned with those factors that enable pupils to respond correctly, even though they lack the necessary achievement. These are the clues, some rather obvious and some very subtle, that inadvertently creep into test items during their construction. They lead the non-achiever to the correct answer and thereby prevent the items from functioning as intended. When test items are short-circuited they provide invalid evidence of achievement.

Example: A porpoise is an

A. Plant.
B. Reptile
C. Animal.
D. Bird.

Simple verbal associations may also lead to the correct answer. Note how the word wind in the following item provides a clue to the answer:

Example: Which one of the following instruments is used to determine the direction of the wind?

A. Anemometer.
B. Barometer.
C. Hygrometer.
D. Wind vane.

Rather it leads the uninformed to the correct answer such clues should lead the non-achiever away from the correct answer. In the following item the same clue makes wind vane a plausible (but incorrect) answer for those pupils who have not learned the use of the various weather instruments.

Example: Which one of the following instruments is used to determine the speed of the wind?

A. Anemometer.
B. Barometer.
C. Hygrometer.
D. Wind vane.

Verbal clues need not be as obvious as these. In fact, the clues that appear in the final version of a test are usually rather subtle, as they are based on partial knowledge and. Verbal associations not readily apparent to the casual observer. For

example, at the first glance the following item appears to be free from clues:

Example: Which one of the following is used to prevent polio?

A. Gamma globulin.

B. Penicillin.

C. Salk vaccine.

D. Sulfa. .

An examination of this item, however, will indicate that the word vaccine provides a clue to the answer. All the pupil needs to know to answer the item correctly is that vaccine is used to prevent disease. Because most pupils have been vaccinated at one time or another, they probably possess this partial knowledge needed to make the clue apparent to them. Some pupils may also have developed a verbal association between Salk and Polio and respond correctly on that basis. In either case, partial knowledge can lead to the correct answer and prevent the item from functioning as intended.

Another type of subtle clue is one based on the words used to qualify statements. For example, true-false statements that include qualifiers such as sometimes, usually, generally, and the like are most often true, where as statements containing absolutes such as always', never, none, and only are most often false. Such words have been called specific determiners. They are difficult to remove from true-false items because true statements generally must be qualified, and false statements frequently must be stated in absolute terms to make them clearly false.

Other common clues in selection-type items include:

1. Stating correct answers in textbook language or in greater detail than incorrect answers.

2. Making correct answers longer than incorrect answers
3. Placing the correct answers in some identifiable pattern (e.g., T, F, T, and F). Some of these clues are more likely to be detected by low-achieving pupils, who are desperately searching for some basis for answering, than by high achieving pupils, who are concentrating more intently on the knowledge or skill called for by the items.

Clues, which prevent test items from functioning as intended, can usually be eliminated during test construction. In fact, many of the suggestions for constructing each type of test item are aimed directly at removing such clues. It is also helpful to analyze each completed test item in terms of the apparent mental process a pupil must use to obtain the correct answer and to compare this with the item's intended purpose. Only when these two are in harm only can we be fairly certain that irrelevant factors are not operating and that correct answers to the test items indicate attainment of the desired learning outcomes.

Some Common Clues in Test Items

Grammatical inconsistencies
Verbal associations
Specific determiners (e.g., always)
Phrasing of correct responses
Length of correct responses
Location of correct responses

1.6 Focus on Improving Learning and Instruction

The ultimate purpose of testing, as with all classroom procedures, is to improve pupil learning. Thus as we construct classroom tests, we should keep in mind the extent to which it is likely to contribute, directly or indirectly, to this end. A well

constructed classroom test should increase both the quantity and quality of pupil learning.

1. **One way of ensuring that tests have a desirable influence on pupil learning is to pay particular attention to the breadth of content and learning outcomes measured by the tests.** When we select a representative sample of content from all of the areas covered in our instruction, we are emphasizing to our pupils that they must devote attention to all areas: they cannot neglect some aspects of the course and do well on the tests. Similarly, when our tests measure a variety of types of learning outcomes, the pupils soon learn that a mass of memorized facts, develop conceptual understandings, draw conclusions, recognize assumptions, identify cause-and-effect relations, and the like. This discourages them from depending solely on memorization as a basis for learning and encourages them to use more complex mental processes.

2. **The practice of constructing tests that measure a variety of learning outcomes should also lead to improved teaching procedures and, thus, indirectly to improved pupil learning.** As we translate the various learning outcomes into test items, we develop a better notion of the mental processes involved. Thus, the function of understanding, thinking skills, and other complex learning outcomes becomes clearer to us. This clarification of how achievement is reflected in mental processes enables us to plan more effectively the pupils , learning experiences. Furthermore, we also more apt to emphasize understanding, thinking skills, and other complex learning outcomes in our teaching when we include them in our testing. This may seem a case of the cart pulling the horse but a well-constructed test frequently leads to a review of teaching procedures and to the abandonment of those that encourage rote learning.

3. **Finally, a test will contribute to improved teacher-pupil relations (with a beneficial effect on pupil learning) if pupils view the test as a fair and useful measure of their achievement.** We can make fairness apparent by including a representative sample of the learning tasks that have been emphasized during instruction, by writing concise directions, by making certain that the intent of each test item is clear and free of any bias that would prevent a knowledgeable person from answering correctly, and by providing adequate time limits for the test. The pupils' recognition of usefulness, however, depends as much on what we do with the results of the test as on the characteristics of the test itself. We can make the usefulness apparent by using the result as a basis for guiding and improving learning.

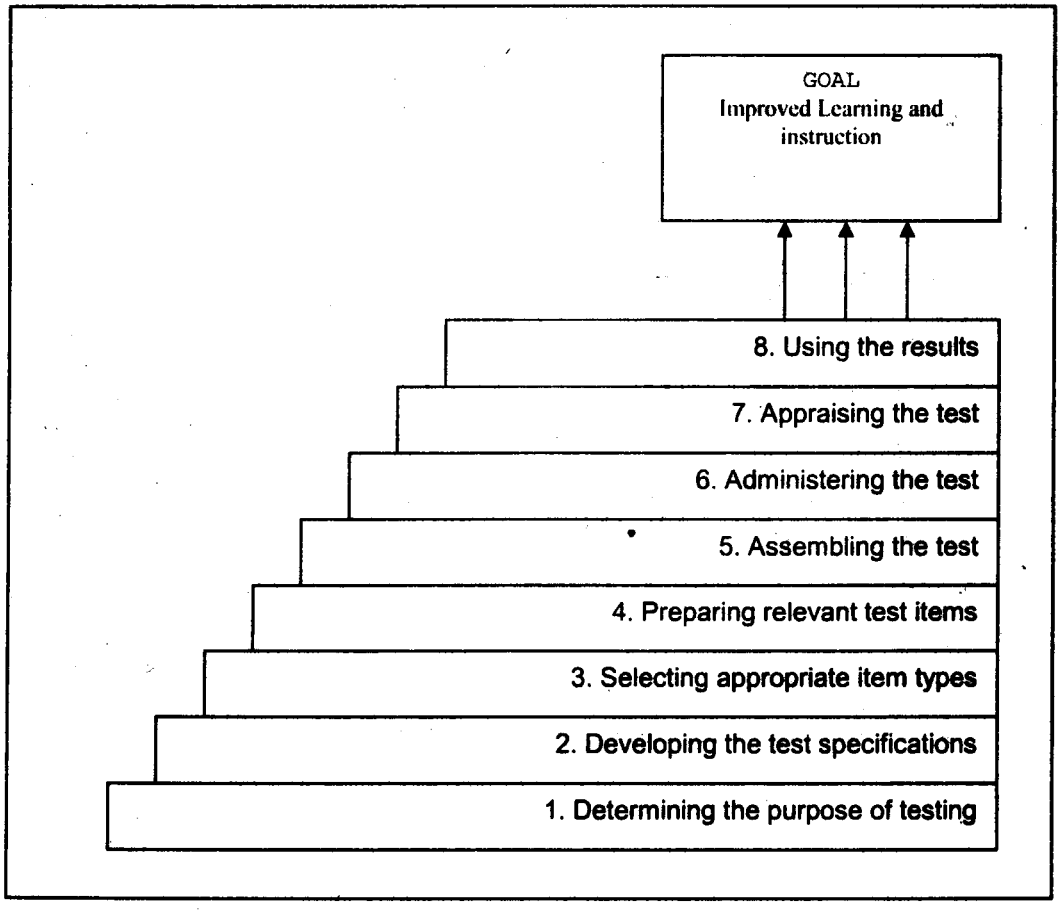
1.7 Self-Assessment Questions

1. Describe Criterion-referenced Mastery Test and Norm-referenced Survey Test.
2. List as many factors as you can think of that might prevent pupils from obtaining correct answer even though they possessed the knowledge the items were designed to measure.
3. List as many factors as you can think of that would enable pupils to answer items correctly, even though they lacked the knowledge the items were designed to measure.

2. TEST PLANNING CONSIDERATIONS

This topic will discuss those factors concerned with planning classroom tests which includes the following:

1. Determining the purpose of testing.
2. Developing the test specifications.
3. Selecting appropriate item types.
4. Preparing relevant test items.



2.1 The Purpose of Classroom Testing

Classroom tests can be used for a variety of instructional purposes. However, the various uses of tests and other evaluation instruments can be classified into four types of classroom evaluation:

1. Placement evaluation,
2. Formative evaluation,
3. Diagnostic evaluation, and
4. Summative evaluation.

Because teacher-made tests are useful in all four areas, this classification system provides a convenient basis for considering the role of test purpose in planning the classroom test.

Placement Testing

Most placement tests constructed by classroom teachers are pretests designed to measure:

1. Whether pupils possess the prerequisite skills needed to succeed in a unit or course or
2. To what extent pupils have already achieved the objectives of the planned instruction.

In the first instance we are concerned with the pupils readiness to begin the instruction. In the second we are concerned with the appropriateness of our planned instruction for the group and with proper placement of each pupil in the instructional sequence.

Pretests for determining prerequisite skills are typically rather limited in scope. For example a pretest in. Algebra might be confined to computational skill in arithmetic; a pretest in science might consist solely of science terms; and a pretest in beginning

German might he limited to knowledge of English grammar. In addition to being confined to a small area of knowledge or skill, the readiness pretest also tends to have a relatively low level of difficulty. This is because this type of pretest is used to determine whether pupils have the minimum essentials needed to proceed with the course or unit of work. Pretests of this type are typically criterion-referenced tests (i.e., tests designed to describe the learning tasks pupils can perform), because their major function is to identify the presence or absence of prerequisite skills.

Pretests for determining the extent to which pupils have already achieved the objectives of the planned instruction are no different from the tests used to measure the outcomes of instruction. Thus, a test is designed to measure final entry performance on the course objectives. Because this type of test is broad in scope and measures instructional objectives at various levels of complexity it is typically a norm-referenced test (i.e., test designed to rank pupils in order of achievement) using items with a wide range of difficulty. If the purpose of the test is to measure the achievement of a limited set of objectives for a unit of instruction, however, the pretest may be designed as a criterion-referenced test. This would be the case when mastery was the goal and the end-of-unit test was criterion referenced, as in individualized instruction and in some types of classroom instruction (e.g., Bloom's Mastery Learning Strategy).

Formative Testing

Formative tests are given periodically during instruction to monitor pupils' learning progress and to provide ongoing feedback to pupils and teacher. Formative testing reinforces successful learning and reveals learning weaknesses in need of correction. A formative test typically covers some predefined segment passes a

rather limited sample of learning tasks. The test items may be easy or difficult, depending on the learning tasks in the segment of instruction being tested. Formative tests are typically criterion-referenced mastery tests, but norm-referenced survey tests can also serve this function. Ideally, the test will be constructed in such a way that corrective prescriptions can be given for missed test items or sets of test items. Because the main purpose of the test is to improve learning, the results are seldom used for assigning grades.

Diagnostic Testing

Diagnosis of persistent learning difficulties involves much more than diagnostic testing, but such tests are useful in the total process. The diagnostic test takes up where the formative test leaves off If pupils do not respond to the feedback-corrective prescriptions of formative testing, a more detailed search for the source of learning errors will be indicated. For this type of testing, we will need to include a number of test items in each specific area, with some slight variation from item to item. In diagnosing pupils' difficulties in adding whole numbers, for example we would want to include addition problems containing various number combinations, with some not requiring carrying and some requiring carrying, to pinpoint the specific types of error each pupil is making. Because our focus is on the pupils' learning difficulties, diagnostic tests must be constructed in accordance with the most common sources of error that pupils encounter. Such tests are typically confined to a limited area of instruction, and the test items tend to have a relatively low level of difficulty.

Summative Testing

The summative test is given at the end of a course or unit of instruction, and the results are used primarily for assigning grades

or certifying pupil mastery of the instructional objectives. The results can also be used, for evaluating the effectiveness of the instruction. The end-of-course test (final examination) is typically a norm-referenced survey test that is broad in coverage and includes test items with a wide range of difficulty. The more restricted end-of-unit summative test might be norm referenced or criterion referenced, depending on whether mastery or developmental outcomes are the focus of instruction.

2.2 Development of Test Specifications

The only assurance we have that a classroom test validly measures the instructional objectives and course content, we are interested in testing, is to use some systematic procedure for obtaining a representative sample of pupil performance in each of the areas to be measured. One device that has been widely used for this purpose is the two-way chart, called a table of specifications. This chart relates the instructional objectives to the course content and specifies the relative emphasis to be given to each type of learning outcome.

Building a Table of Specifications

A table of specification in includes:

1. Obtaining a list of instructional objectives,
2. Outlining the course content, and
3. Preparing the two-way chart.

1. Obtaining a list of Instructional Objectives. Many aspects. of pupil performance can be measured by means of paper-and-pencil tests. Infect a novice in the area of measurement is frequently surprised at the variety of learning outcomes that can be measured in this manner. Thus, all of the intended outcomes of instruction should be considered when planning the classroom

test. If a comprehensive list of instructional objectives and specific learning outcomes has been prepared, it is simply a matter of selecting those outcomes that can be measured by paper-and-pencil tests. If such a list is not available, a set of instructional objectives can be prepared for the classroom test.

2. Outlining the Course Content. The list of instructional objectives describes the types of performance the pupils are expected to demonstrate (e.g., knows, understands, applies) and the course content indicates the area in which each type of performance is to be shown. Thus the second step in preparing the test specifications is to outline the course content. This may be simply a list of major topics to be covered during the course or a more detailed list of topics and subtopics. The amount of detail in the content outline depends on the purpose of the test, the segment of the course covered, and the type of test interpretation to be used. A criterion-referenced test (used to describe the learning tasks pupils can perform), for example, will require a much more detailed description of both objectives and content than will norm-referenced test used to rank pupils in order of achievement. 111

3. Preparing the Tow-Way Chart. The final step in building a table of specifications is to prepare the two-way chart that relates the instructional objectives to the course content and, thus, specifies the nature of the test sample.

The relative emphasis to be given to each instructional objective and content area should reflect the emphasis of the instruction. In assigning relative weights, both the importance the teacher attaches to the learning outcome and the amount of instructional time devoted to it can serve as guidelines. Typically, the weighting is done by first assigning percentages across the

bottom row (for each objective), then assigning percentages down the right-hand column (for each content area), and finally allotting the percentage, or number, of test items to each of the two-way cells in the table. Proper weighting will make it possible to construct a test that measures a reasonably representative sample of the intended outcomes of instruction that can be evaluated by paper-and-pencil tests.

2.3 Establishing Priorities Among Objectives

1. Effective test planning requires that objectives be determined and written.
2. Some attempts have been made to establish priorities among objectives. In some instances priorities are established by state legislatures and boards of education. Priorities can also be established by asking community members to estimate how much money the accomplishment of a goal might be worth, the amount of time that should be spent in pursuing some goal, tasks are also ways of establishing priorities among objective.
3. Objectives may be categorized by four dimensions:

a) Process and Product Objectives Process Objectives are used to monitor activities to make sure that someone is responsible for accomplishing an objective within specified time limitations. Product objectives indicate what the student must do or know as a result of instruction.

b) Behavioral and Implicil Objectives, Behavioral objectives state what the student must do to demonstrate the attainment of an objective; implicit objectives refer to covert behaviors. Many implicit objectives can be made behavioral by stating the conditions under which testing is to take place. Behavioral objectives are important in the following conditions:

i) When the subject matter is sequential and mastery of prerequisite skills is necessary before the student can be allowed to advance in the sequence.

ii) If teachers are to hold themselves accountable for student achievement.

iii) If objectives are to serve as a broad base from which behavioral objectives can be written.

iv) If objectives are to be communicated to parents.

v) If an item pool is available to be used as a criterion to demonstrate and clarify the meaning of an objective.

c) Immediate and Ultimate Objective, Immediate objectives are those over which the school exerts some influence; ultimate objectives are long-term goals that the school may consider desirable but which teachers cannot control directly.

d) Restricted and Inclusive Objectives, Restricted objectives are stated so specifically that they are equivalent to the items given to students; inclusive objectives are more broadly stated so as to include more than one specific item, Inclusive objectives specify the universe of items that could be asked on a topic.

4. Some studies have shown that learning is improved when students are allowed to see the teacher's objectives; other studies have failed to show any advantage in making these objectives available. Since no studies have shown that objectives are detrimental to students, the position taken here is that the weight of evidence slightly favours making objectives available to students.

5. Because measurements are samples of behaviour, the test constructors need to determine how many items they will formulate for each objective. In addition, they need to determine how complex each item should be. Complexity of

items can be defined by using a taxonomy or classification system such as that proposed by Bloom:

a) Knowledge, The least complex level refers to items that require simple memorization.

b) Comprehension, This level requires students to translate from one level of abstraction to another by restating a problem in their own words.

c) Application, Items in the third level of complexity requires the student to relate principles or concepts in new situations.

d) Analyze Students who analyze a problem critically examine information to determine the interrelationships among the parts, such as in finding assumptions or fallacies in an argument.

e) Synthesis, at these level students constructs something unique, such as writing a story or developing a hypothesis.

f) Evaluation, The students makes a value judgment using internal or external criteria, such as in evaluating conflicting evidence.

6. Krathwhol taxonomy of the affective domain has also described.

a) Receiving or Attending, The lowest level of this taxonomy requires the individual to be aware of something and to at least tolerate its presence.

b) Responding, at this level the individual responds to stimuli in various degrees of zeal.

c) Valuing, Valuing implies the judgment of worthiness of some belief along with some internalization and commitment.

d) Organization, at this level values become organized and interrelated to form a system of beliefs that is internally consistent.

e) Characterization by a Value or Value Complex, at the highest level the individual has developed a coherent philosophy of life and lives according to that philosophy.

7. Harrow's taxonomy of the psychomotor domain includes six levels:

a) Reflex Movements, these include motor responses that are present at birth.

b) Basic Fundamental Movement, At a higher level than reflexes are those movements that lead to specialized complex skills of ambulation, the movement of the body in place around an axis, and the ability to be prehensile and dexterous.

c) Perceptual Abilities, This level combines both cognitive and motor abilities.

d) Physical Abilities, This level includes endurance, strength, flexibility, and agility. ,

e) Skilled Movement, These are the movements that require learning and skill development in sports, dancing, and recreation.

f) Non-discursive Communication, The highest level of Harrow's taxonomy includes the ability to communicate by means of facial expressions, posture, dance choreographies, and the like.

8. Barrett, has constructed a taxonomy of objectives in reading comprehension that includes both cognitive and affective domains. This taxonomy could also be adapted to art and music.

9. A two-way grid or test blueprint can be prepared to determine if desired students behavior is being adequately sampled. Content areas might be listed along one dimension of the grid, and any or all of the six levels of Bloom's

taxonomy can be listed along the other. The total number of items measuring each content area could then be allocated among various levels of complexity. The purpose of this is to make certain that a reasonable sampling of items from all levels of complexity and from all topics studied is included in the test.

10. Students generally expect tests to be composed of items roughly proportional in number to importance and to the amount of time devoted to each topic in a text or in class discussions. If this is not the case, students should be informed in advance.

11. No hard-and-fast rules can be developed to determine how many items should come from each level of complexity. Some objectives can be measured only by asking students to memorize knowledge, such as specific facts, places, or dates, but more complex items demonstrate a greater understanding of these facts.

12. Once objectives have been written, the teacher still faces a number of decisions:

a) Open-Book and Closed-Book Examinations, Although most examinations do not allow students to use notes or the textbook, there are good reasons why their use might be encouraged. If the students are to apply information or use special reference materials, open-book exams are often a good choice. If recall of information is desirable or if the amount of time saved in relearning the material is the criterion, closed-book examinations are probably more useful.

b) The Take-Home Examination, This is an extension of the open-book test. In some cases, allowing students to obtain

help from others may be useful. The take-home test has been used successfully as a teaching device or exercise.

c) When tests are used to motivate students, frequent testing seems to be of some benefit to students of lower ability.

d) Surprise or unannounced tests are not recommended because they tend to create anxiety, reduce preparation time, and do not lead to improved learning.

e) The reading of test items aloud to students is not recommended except perhaps for true-false or completion items and only if it is necessary. Tests can be administered successfully on slides or on closed circuit television.

f) Items should be arranged from easy to hard if time limits are restricted when subtest scores are wanted, items should be arranged within each subset in ascending order of difficulty. Essay questions should be placed at the end of the examination to discourage students from spending inordinate amounts of time on one or two of these items and being penalized by not attempting a greater number of true-false and multiple-choice items.

2.4 Self-Assessment Questions

1. Enlist the purposes of classroom testing.
2. How will you develop a table of specifications for a particular subject to test in the classroom?
3. Explain the process of categorization of the educational objectives.
4. Enlist different decisions, which a teacher has to make after finalizing the educational objectives.

3. ASSEMBLING THE CLASSROOM

The preparation of test items for use in a test is generally facilitated if the items are properly recorded, if they are written at least several days before they are to be used, and if extra items are constructed.

3.1 Recording Test Items

When constructing the test items, it is desirable to write each one on a separate index card. In addition to the test items, the card should contain information concerning the instructional objectives, the specific learning outcome, and the content measured by the item. A space should also be reserved for item-analysis information, usually on the back of the card, to allow room to record the data each time the item is used.

Item cards provide flexibility. As items are reviewed and edited, they can be eliminated, added, or revised with very little difficulty. The same holds true when arranging the items for the test: they can be arranged and rearranged merely by sorting cards. The flexibility of this recording system also makes it possible to build a card file of effective items for future use.

3.2 Reviewing Test Items

No matter how carefully test items have been prepared, defects inadvertently creep in during construction. As we concentrate on the clarity and conciseness of a question, a verbal clue slips in unnoticed. As we attempt to increase an item's difficulty, we unwittingly introduce some ambiguity. As we rework an item to make the incorrect choices more plausible, the behaviour called forth by the item is unintentionally modified. In short, we focus so closely on some aspects of item construction that we overlook others. This results in an accumulation of

unwanted errors that may distort the item's function. Such technical defects can most easily be detected by:

1. Reviewing the items after they have been set-aside for a few days
2. Asking a fellow teacher to review and criticize them,

In reviewing test items, we should try to view the item from the pupil's viewpoint, as well as from that of the test maker. From these two vantage points, each item should be read carefully and its possible functioning effectiveness judged. The following questions will help you analyze the quality of each item.

I. **Is the item format appropriate for the learning outcome being measured?** If the learning outcome calls for the definition of a term, for example, a supply-type item (e.g., short-answer item) would be appropriate, and a selection-type item (e.g., multiple-choice) would be clearly inappropriate. On the other hand, if the intended outcome was simply the identification of the correct definition, then a selection-type item would be adequate. Thus, the first step is to check whether the item format is suitable for the type of pupil performance described in the test plan. The action verb in the statement of each specific learning outcome (e.g., defines, describes, identifies) indicates which item format is more appropriate.

2. **Does the knowledge, understanding, or thinking skill called forth by the item match the specific learning outcome and subject-matter content being measured?** When a table of specifications has been used as a basis for constructing the test items, this is merely a matter of checking to see whether the item is still relevant to the same cell in the table. If the item's functioning content has shifted during construction, the item should either be modified so that it serves its original purpose or

reclassified in light of the new purpose. In any case, the response called forth by an item should agree with the purpose for which the item is to be used.

3. **Is the point of the item clear?** A careful review of test items often reveals ambiguity, inappropriate vocabulary, and awkward sentence structure that were overlooked during their construction. Returning to test items after they have been set aside for a few days provides a fresh outlook that makes such defects more apparent. The difficulty of the vocabulary and the complexity of the sentence structure must, of course, be judged in terms of the pupils' maturity level. At all levels, however, ambiguity should be removed. In its final form, each item should be so clearly worded that all pupils understand the task. Whether pupils respond correctly should be determined solely by whether they possess the knowledge or understanding being measured.

4. **Is the item free from excessive verbiage?** Often, items become excessively wordy, because of awkward sentence structure -mentioned in the previous section -or the inclusion of nonfunctional material. Some teachers prefer to justify the use of an item by including a statement or two concerning the problem's importance. Others expand a simple question into an elaborate story situation to make the item more interesting. Although adding such nonfunctional material may be useful in some instances, items are generally more effective when the problem is stated as concisely as possible. When reviewing items, the content of each item should be analyzed to determine the functional elements leading to the correct response. If there are any elements that the pupils may disregard entirely and still respond correctly, they probably should be removed.

5. **Is the item of appropriate difficulty?** If the test item is intended for a criterion-referenced mastery test, its difficulty should, of course, match the difficulty indicated by the statement of the specific learning outcome. No attempt should be made to alter item difficulty to fit some predetermined level. If the test item is to be used in a norm-referenced test, however, a difficulty level of 50 percent is desirable. In reviewing items for a test, all we can do is make our best judgment about item difficulty (unless item analysis data are available), taking into account the nature of the test and the educational background of the Pupils

6. **Does the item have an answer that would be agreed upon by experts?** This is seldom a problem with factual material, which usually can be judged as correct or incorrect. It is mainly a problem with selection-type items that ask for the best reason, the best method, the best interpretation, or whatever, in which the judgment of experts determines the answer. If experts agree on the best answer, fine, but do not include items that require pupils to endorse someone's unsupported opinion.

7. **Is the item free from technical errors and irrelevant clues?** The checklists for reviewing each of the item types be prepared and list points to consider in searching out technical errors and irrelevant clues. Most of the clues can be removed merely by trying to detect them during the item review. They somehow seem more obvious after the items have been set-aside for a while.

8. **Is the item free from racial, ethnic, and sexual bias?** A final check should be made to make certain that the vocabulary and problem situation in each item would be acceptable to the members of all groups and would have a similar meaning to them. An effort should be made to remove any type of stereotyping such

as always portraying minorities in subservient roles, women in homemaking roles, and the like. A judicious and balanced use of different roles for minorities and males and females should contribute to more effective testing.

When it is possible to get fellow teachers to review the test items, they should be asked to read each item, indicate the answer, and note any technical defects. If an answer does not agree with the key, it may be because it is ambiguous. Asking another teacher to "think out loud" when deciding on the answer will usually, reveal the misinterpretation of the question and the source of the ambiguity. This is how other persons can be most useful. Reviewers will be less helpful in evaluating the types of responses called forth by the items, because this requires knowledge of what the pupils have been taught. Only the teacher who prepared the item knows for sure whether an item measures understanding or merely the retention of a previously learned answer.

When the test items have been revised and those to be included in the test have been tentatively selected, the following questions should be asked:

1) Do the test items measure a representative sample of the learning outcomes and course content included in the test plan?
2) Are there enough test items for each interpretation to be made?
3) Is the difficulty of the test items appropriate for the purpose of the test and for the pupils for whom the test is intended?
4) Are the test items free from overlapping so that the information in one does not provide a clue to the answer in another?

The first question can be answered by comparing the final selection of items with the table of specifications or other test plan. Answers to the last three are determined by reviewing the test items in each section of the test and the test as a whole. Affirmative answers to these questions mean the items are ready to be arranged in a final test form.

3.3 Arranging Items in the Test

There are various methods of grouping items in an achievement test, and the method will vary somewhat with the use to be made of the results. For most classroom purposes, the items can be arranged by a systematic consideration of:

1. The types of items used.
2. The learning outcomes measured.
3. The difficulty level of the items.
4. The subject matter measured.

First and foremost, the items should be arranged in sections by item type. That all true false items should be grouped together, then all matching items, then all multiple-choice items, and so on. This arrangement requires the fewest sets of directions; it is the easiest for the pupils because they can retain the same mental set throughout each section; and it greatly facilitates scoring. When two or more item types are included in a test, there is also some advantage in keeping the simpler item types together and placing the more complex ones in the test, as follows;

1) True-false or alternative-response items.
2) Matching items.
3) Short-answer items.
4) Multiple-choice items.
5) Interpretive exercises.
6) Essay questions.

Arranging the sections of the test in this order produces a sequence that roughly approximates the complexity of the learning outcomes measured, ranging from the simple to the complex. It is then merely a matter of grouping the items within each item type. For this purpose, items that measure similar outcomes should be placed together and then arranged in order of ascending difficulty. For example, the items in the multiple-choice section might be arranged in the following order

1. Knowledge of terms.
2. Knowledge of specific facts.
3. Knowledge of principles.
4. Application of principles.

Keeping together the items that measure similar outcomes is especially helpful in determining the types of learning outcomes causing pupils the greatest difficulty.

If, for any reason, it is not feasible to group the items by the learning outcomes measured, it is still desirable to arrange them in order of increasing difficulty. Beginning with the easiest items and proceeding gradually to the most difficult has a motivating effect on pupils. Also, encountering difficult items early in the test often causes pupils to spend a disproportionate amount of time on such items. If the test is long, they may be forced to omit later questions that they could easily have answered.

With the items classified by item type, the sections of the test and the items within each section can be arranged in order of increasing difficulty. Some shifts in the first four item types may be warranted by the difficulty of the items used, but the interpretive exercises and essay tests should certainly be last.

In constructing classroom achievement tests, there is little to be gained by grouping test items according to subject-matter

content. When it appears desirable to do so, such as in separating historical periods, these divisions should be kept to a minimum.

To summarize, the most effective method for organizing items in the typical classroom test is to:

1. Form sections by item type.
2. Group the items within each section by the learning outcomes measured.
3. Arrange both the sections and the items within sections in an ascending order of difficulty. Use subject-matter groupings only when needed for some specific purpose.

3.4 Preparing Directions for the Test.

Teachers sometimes devote considerable time and attention to the constructing and assembling of test items and then dash off directions with very little thought. In fact, many teachers include no written directions with their tests, assuming either that the items are self-explanatory or that the pupils are conditioned to answering the types of items used in the test. Some teachers also use oral directions, but they frequently leave much to be desired. Whether written, oral, or both, the directions should include at least the following points:

1. Purpose of the test.
2. Time allowed for answering.
3. Basis for answering.
4. Procedure for recording the answers.
5. What to do about guessing.

The amount of detail for each of these points depends mainly on the pupils' age level, the test's comprehensiveness, the test items' complexity, and the pupils' experience with the testing procedure used. Using new item types and separate answer sheets,

for example, requires much more detailed directions than do familiar items, requiring pupils merely to circle or underline the answer.

1. **Purpose of the Test.** The purpose of the test is usually included when the test is announced or at the beginning of the semester when the evaluation procedures are described as a part of the general orientation to the course. Should there be any doubt whether the purpose of the test is clear to all pupils, however, it could be explained again at the time of testing. This is usually done orally. The only time, a statement of the purpose of the test needs to be included in the written direction is, when the test is to be administered to several sections taught by different teachers. Then a written statement of purpose ensures greater uniformity.

2. **Time Allowed for Answering.** It is helpful to tell the pupils how much time they will have for the whole test and how to distribute their time among the parts. When essay questions are included, it is also good to indicate approximately how much time should be allocated to each question. This enable the pupils to use their time most effectively and prevents the less-able ones from spending too much time on questions that are particularly difficult for them. Classroom tests of achievement should generally have liberal time allowances. Except for special purposes, such as measuring proficiency in shorthand, typing, and simple computational skills, speed is not important. Our main concern is the level of achievement each pupil has attained. Were it not for practical considerations like the length of class periods and the pressure of other school activities, there would be no need for any time limits with most classroom achievement tests.

Judging the amount of time that pupils will need to complete a given test is not simple. It depends on the types of time used, the

age and ability of the pupils, and the complexity of the learning outcomes measured. As a rough guide, the average high school student should be able to answer two true-false items, one multiple-choice item, or one short answer item per minute of testing time. Interpretive test items take much more time; the exact amount depends on length and the complexities of the introductory materials. Also, elementary school pupils generally require more time per item than high school students do, and reading skill is an important determiner of the amount of time needed by a specific group. Experienced teachers familiar with the ability and work habits of a given group of pupils are in the best position to judge time allotments. It is better to err in the direction of allotting too much time than to deprive some of the slower pupils from demonstrating their maximum levels of achievement.

3. **Basis for Answering.** The directions for each section of the test should indicate the basis for selecting or supplying the answers. With true-false, matching, and multiple-choice items, this part of the directions can be relatively simple. For example, a statement like "select the choice that best completes the statement or answer the question" might be sufficient for multiple-choice items. When interpretive exercises are used, however, more detailed directions are necessary, because the basis for the response is much more complex. The directions must clearly indicate the type of interpretation expected. Each interpretive exercise requires its own directions.

It is sometimes good to include sample test items correctly marked so that pupils can check their understanding of the basis for answering This practice is especially helpful to elementary school pupils and to pupils at other levels when complex item types are used.

Essay questions frequently require special directions concerning the type of response expected. If the selection and organization of ideas are emphasized, for example, this should be indicated to the pupils so that they have a more adequate basis for responding.

4. Procedure for Recording Answers. Answers may be recorded on the test form itself or on separate answer sheets. If the test is short, the number of pupils taking the test is small, or the pupils are relatively young, answers are generally recorded directly on the test paper. For most other situations, separate answer sheets are preferred because they reduce the time needed for scoring, and they make it possible to use the test papers over again. The latter feature is especially useful when the test is to be given to pupils in different sections of the same course.

Directions for recording the answer on the test paper itself can be relatively simple. With selection items, it is merely a matter of instructing the pupils to circle, underline, or check the letter indicating the correct answer. For pupils in the primary grades, it is usually better to ask them to mark the answer directly by drawing a line under it. With supply items, the directions should indicate where to put the answer and the units in which it is to be expressed if the answer is numerical.

Separate answer sheets are easily constructed, and the directions for their use can be placed on the test paper or on the answer sheet itself. The directions on the sheet should be general and they must cover instructions for recording various types of answers. Pupils can be instructed to cross out rather than circle the letters indicating the correct answers, to facilitate scoring with a stencil key. Circled letters cannot be readily seen through holes in a stencil.

Special answer sheets for machine scoring can be used with classroom tests, but there is no advantage in using them unless machine scoring facilities are readily available and number of papers to be scored warrants the expense. When machine scoring is used, special directions should be obtained from the company supplying the scoring service.

5. **What to do About Guessing.** When selection-type items are used, the directions should tell pupils what to do when they are uncertain of the answer. Should they guess or omit the item? If no instructions are given on this point, the bold pupils will guess freely, whereas others will answer only those items of which they are fairly certain. The bold pupils will select some correct answers just by lucky guesses, and thus their scores will be higher than they should be. On the other hand, if the pupils are merely instructed, "Do not guess" or " Answer only those items of which you are certain," the more timid pupils will omit many items they could answer correctly. Such pupils are not very certain about anything, which prevents them from responding even when they are reasonably sure of the answers. With these directions, the bold pupils will continue to guess, although possibly not quite so widely.

Course_____ Name_____

Section_____ Date_____

Test_____ Score: Part-1_____

Part-2_____

Total_____

Directions: Read all directions on the test paper carefully and follow them exactly. For each test item, indicate your answer on this sheet by crossing out the appropriate letter (X) or filling the

appropriate blank. Be sure that the number on the answer sheet is the same as the number of the test item you are answering.

True-False			Multi-Choice		Short-Answer	
Item	Answer		Item		Item	Answer
1	T	F	21	ABCDE	41	_____
2	T	F	22	A B C D E	42	_____
3	T	F	23	A B C D E	43	_____

As Cronbach (1984) pointed out, the tendency to guess or not to guess when in doubt about an item is determined by personality factors and cannot be entirely eliminated by directions that caution against guessing or that promise penalties to those who do guess. The only way to eliminate variations in the tendency to guess is to instruct pupils to answer every item. When this is done, no pupil is given a special advantage, and it is unnecessary to correct for guessing in the scoring. Directions such as the following are usually sufficient to communicate this to the pupils: "Because your score is the number right, be sure to answer every item."

Some teachers object to such directions on the grounds that encouraging guessing is undesirable from an educational standpoint. Most responses to doubtful items are not wild guesses, however, but are guided by some information and understanding. I this respect, they are not too different from the informed guesses we make when we predict weather, judge the possible consequences of a decision, or choose one course of action over another. Problem solving always involves a certain amount of this type of informed guessing.

A more defensible objection to directions that encourage guessing is that the chance errors introduced into the test scores lower the accuracy of measurement. Although this is certainly

objectionable, it probably has less influence on the validity of the results than does the. Systematic advantage given to the bold guessers by the do not guess directions.

For liberally timed classroom test, the answer-every-item directions are favored. But for speed tests and when teachers want to discourage guessing directions such as the following good compromise: Answer all the items for which you can find some reasonable basis for answering, even though you are not completely sure of the answer. Do not guess wildly, however because there will be a correction for guessing.

There seems to be a tendency in standardized testing to use the make informed guesses but not wild guesses type of directions. Speed, however, is more significant in standardized testing than in ordinary classroom testing, and the test items are not as closely keyed to the pupils, learning experiences. When the pupils are familiar with the test's content and have ample opportunity to consider every item, there is generally no need to warn against wild guesses or to correct for them.

Reproducing the Test

In preparing the test materials for reproduction, it is important that the item be spaced and arranged so that they can be read, answered, and scored with the least amount of difficulty. Cramming too many test items onto a page is poor economy. What little paper is saved will not make up for the time and confusion that result during the administration and scoring of the test.

All test items should have generous borders. Multiple-choice items should have the alternatives listed in a vertical column beneath the stem of the item, rather than across the page. Items should not be split, with parts of the item on two different pages. With interpretive exercises, the introductory materials can

sometimes be placed on a facing page or separate sheet, with all of the items referring to it on a single page.

Unless separate answer sheet is used, the space for answering should be down one side of the page, preferably the left. The most convenient method of the response is circling the letter of the correct answer. With this arrangement scoring is simply a matter of placing a strip scoring key beside the column of answers.

Test items should be numbered consecutively throughout the test. Each test item will need to be identified during discussion of the test and for other purposes such as item analysis. When separate answer sheets are used, consecutive numbering is, of course, indispensable.

The duplication of the classroom tests is usually by mimeograph or Ditto machine. When many copies are needed, the photo-offset method is used. Regardless of the process selected, it is desirable to proofread the entire test before it is administered. Charts, graphs, and other pictorial material must be checked to ensure that the reproduction has been accurate and the details clear.

Helping Pupils Prepare for Tests

A. General preparation
   1. Suggest ways of studying.
   2. Give practice tests like those to be used.
   3. Teach test-taking skills.
   4. Teach how to write well-organized essay answers.
   5. Stress the value of tests for improving learning.

B. Preparation for each test
   1. Announce in advance when the test will be given.
   2. Describe the conditions of testing (e.g., one hour, closed book).
   3. Describe the test's length and the types of items to be used (e.g., 20 multiple-choice and 3 essay items).
   4. Describe the content and type of performance to be covered (a table of specifications is use full for this).
   5. Describe how the test will be scored and how the results will be used.
   6. Give the students sample items similar to those in the test (use a short practice test or present items orally and discuss answer).
   7. Relieve anxiety by using a positive approach in describing the test and its usefulness.

3.5 Self -Assessment Questions

1. What special precautions might be taken to avoid ambiguity , irrelevant clues, and other errors in objective test items?
2. Under what conditions a correction for guessing be used to score a test?
3. What are some of the reasons a correction for guessing should not be used with the typical classroom test?
4. What major points should be considered while preparing directions for the tests?

5. Enlist the major factors to be kept in mind, when a teacher is reviewing the test items.

4. ADMINISTERING AND SCORING THE CLASSROOM TESTS

The same care that went into the preparation of the test should be carried over into its administration. Here we are concerned with (1) providing optimum conditions for obtaining the pupils' responses and (2) selecting convenient and accurate procedures for tabulating the results.

41. Administering the Test

The guiding principle in administering any classroom test is that all pupils must be given a fair chance to demonstrate their achievement of the learning outcomes being measured. This means a physical and psychological environment conductive to their best efforts and the control of factors that might interfere with valid measurement.

Physical conditions such as adequate work space, quiet, proper light and ventilation, and comfortable temperature are sufficiently familiar to teachers to warrant little attention here. Of greater importance, but frequently neglected, are the psychological conditions influencing test results. Pupils will not perform at their best if they are tense and anxious during testing. Some of the things that create excessive test anxiety are:

1. Threatening pupils with tests if they do not behave.
2. Warning pupils to do their best "because this test is important".
3. Telling pupils they must work fast in order to complete the test on time.
4. Threatening dire consequences if they fail the test.

The antidote to test anxiety is to convey to the pupils, by both word and deed that the test results are to be used to help them improve their learning. They should also be reassured that the time limits are adequate to allow them to complete test. This, of course, assumes that the test will be used to improve learning and that the time limits are adequate.

The time of testing can also influence the results. If tests are administered just before the "big game", the results may not be representative. Furthermore, for some pupils, fatigue, the onset of illness, or worry about a particular problem may prevent maximum performance. Arranging the time of testing accordingly and permitting its postponement when appropriate can enhance the validity of the results.

The actual administration of the test is relatively simple, because a properly prepared classroom test is practically self-administering. Oral directions, id used should be present clearly. Any sample problems or illustrations put on the blackboard should be kept brief and simple.

Beyond this, suggestions for administering a classroom test consist of many things to avoid:

1. **Do not talk unnecessarily before the test.** When a teacher announces that there will be "a full forty minutes" to complete the test and then talks for the first ten minutes, pupils feel that they are being unfairly deprived of testing time. Besides, just before a test is no time to make assignments, admonish the class, or introduce next week's topic. Pupils are mentally set for the test and will ignore anything not pertaining to the test for fear it will hinder their recall of information needed to answer the questions. Thus, the well-intentioned remarks fall on "deaf ears" and merely

increase anxiety toward the test and create hostility toward the teacher.

2. **Keep interruptions to a minimum during the test.** At times a pupil will ask to have an ambiguous item clarified, and it may be beneficial to explain the item to the entire group at the same time. Such interruptions are necessary but should be kept to a minimum. All other distractions outside and inside the classroom should, of course, also be eliminated when possible. It is sometimes helpful to hang a "Do not disturb- TESTING" sign outside the door.

3. **Avoid giving hints to pupils who ask about individual items.** If the item is ambiguous, it should be clarified for the entire group, as indicated earlier. If it is not ambiguous, refrain from helping the pupil to answer it. Refraining from giving hints to pupils who ask for help is especially difficult for beginning teachers. But giving unfair aid to some pupils (the bold, the apple polishers, and so on) decreases the validity of the results and lowers class morale,

4. **Discourage cheating, if necessary.** When there is good teacher-pupil rapport and the pupils view tests as helpful rather than harmful, cheating is usually not a problem. Under other conditions, however, it might be necessary to discourage cheating by special seating arrangements and careful supervision. Receiving unauthorized help from other pupils during a test has the same deleterious effect on validity and class morale as does receiving special hints from the teacher. We are interested in pupils doing their best; but for valid results, their scores must be based on their own unaided efforts.

4.2 Scoring The Classroom Tests

Here we shall discuss scoring objective items.

If the pupils' answers are recorded on the test paper itself, marking the correct answers on a blank copy of the test can make a scoring key. Scoring then is simply a matter of comparing the columns of answers on this master copy with the columns of answers on each pupil's paper. A strip key, which consists merely of strips of paper on which the columns of answers are recorded, may also be used if more convenient. These can easily be prepared by cutting the columns of answers from the master copy of the test and mounting them on strips of cardboard cut from manila folders.

When separate answer sheets are used, a scoring stencil is most convenient. This is a blank answer sheet with holes punched where the correct answers should appear. The stencil is laid over each answer sheet, and the number of answer checks appearing through the holes are counted. When this type of scoring procedure is used, each test paper should also be scanned to make certain that only one answer was marked for each item. Any item containing more than one answer should be eliminated from the scoring.

As each test paper is scored, make each item that is answered incorrectly. With multiple-choice items, a good practice is to draw a red line through the correct answer of the missed items rather than through the pupil's wrong answers. This will indicate to the pupil those items missed and at the same time will indicate the correct answers. Time will be saved and confusion avoided during discussion of the test. Marking the correct answers of missed items is especially simple with a scoring stencil. When no answer check appears through a hole in the stencil, a red line is drawn across the hole.

In scoring objective tests, each correct answer is usually counted as one point, because an arbitrary weighing of items

makes little difference in the pupils' final scores. If some items are counted two points, some one point, and some one-half point, the scoring will be more complicated without any accompanying benefits. Scores based on such weightings will be similar to the simpler procedure of counting each item one point.

When pupils are told to answer every item on the test, a pupil's score is simply the number of items answered correctly. There is no need to consider wrong answers or to correct for guessing. When all pupils answer every item on a test, the rank order of the pupils' scores will be same whether the number right or a correction for guessing is used. Some teachers prefer to correct for guessing because they feel the resulting scores are a more accurate indication of the pupil's actual achievement. As we shall see in the following section, however, this is debatable.

4.2 Correction for Guessing.

Correcting for guessing is usually done when pupils do not have sufficient time to complete all items on the test and when they have been instructed that there will be a penalty for guessing. The most common formula used for this purpose is the following:

$$\text{Score} = \text{Right} - \text{Wrong}/n-1$$

In this formula, n is the number of alternatives for an item. Thus, the formula applies to various selection-type items as follows:

True-False items:

$$S = R - W/2-1$$

(or)

$$S=R-W$$

Multiple-choice items:

(A) Three alternatives S = R-W/2

(B) Four alternatives S = R-W/3

(C) Five alternatives S = R-W/4

Using a correction formula in the scoring makes it necessary to count both right and wrong answers. Omitted items are not counted in the scoring.

These correction-for-guessing formulas assume that when pupils do not know the answer to an item, they guess blindly among all alternatives and select the correct answer a given number of times on the basis of chance alone. Thus, if a pupil has 60 items right and 15 items wrong on a true-false test, it is assumed that there was blind guessing on 30 items on the test and that there was chance success in guessing (15 right and 15 wrong). The formula merely removes the lucky guesses from the score by subtracting the number wrong from the number right: Correct score = 60 -15 = 45

The same assumption is made in applying the formula to multiple-choice items, but the possibility of selecting the correct answer is smaller because there are more alternatives to choose from. For example, when a pupil has 60 items right and 15 item wrong on a four-alternative multiple-choice test, it is assumed that the pupil guessed blindly on 20 items and guessed successfully one fourth of the time. Thus the blind guessing resulted in 5 right answers and 15 wrong answers. To remove the lucky guesses from the score, it is simply a matter of subtracting one third of the wrong answers. This is what the correction formula does, as illustrated below:

$$S=R-W/3$$

$$S=60-15/3=55$$

The correction-for-guessing (or correction-for-chance) formula provides a suitable correction when the basic assumption can be satisfied -that is, that pupils guess blindly when they do not know the answer. Such blind guessing seldom occurs in classroom testing, however. Some correct guesses are informed guesses based on partial information, and some wrong answers are due to misinformation or extremely plausible distracters. When pupils can eliminate some of the alternatives in items and make informed guesses among those remaining, the formula under-corrects for chance success. When pupils select incorrect alternatives because of misinformation or the plausibility of distracters, the formula overcorrects for chance success. Consequently, when the correction formula is used with classroom tests, an unknown amount of error is introduced into the scoring. Although it is hoped that the two types of error will cancel each other out, there is no way of determining the amount of distortion in the test scores.

Because of the questionable assumption on which the correction-for-guessing formula is based, it is recommended that it not be used with the ordinary classroom test. The only exception is when the test is speeded to the extent that pupils complete different numbers of items. Here its use is defensible, because pupils can increase their scores appreciably by rapidly (and blindly) guessing at the remaining untried items just before the testing period ends.

4.3 SELF-ASSESSMENT QUESTIONS

1. What is test anxiety? What will you consider in your classroom to avoid test anxiety?
2. What ate different methods for scoring a classroom test?
3. How will you apply the correction for guessing in different test items?

5. BIBLIOGRAPHY

1. Cronbach, L. J. (1984); Essentials of Psychological Testing, 4th ed., Harper and Row Publishers New York.

2. Gronlund, N. E. (1985); Measurement and Evaluation in Teaching, 5th ed., Macmillan Publishing Company, New York.

3. Murphy, K. R, and Davidshofer, C. 0. (1994); Psychological Testing-Principles and Applications, Prentice-Hall Inc., Englewood Cliffs, New Jersey.

4. Sax, G. (1980); Principles of Educational and Psychological Measurement and Evaluation,2nd ed., Wads Worth Publishing Company Belmont, California.

5. Wiersma, w. and Jurs, S. G. (1985); Educational Measurement and Testing, Allyn and Bacon Inc., Boston.

Unit-7

MARKING AND REPORTING

BY:
Jamil Hussain Shah

INTRODUCTION

A philosopher Michael Scriven (1970) has written:

> "Like so many other everyday practices, grading has often seemed to humble to merit the attention of high powered test and measurement people. My feeling is that it is far more important and in more need of help than any thing else they work on".

The process of measurement I sonly one aspect of evaluation. At regular intervals the quality of students' performance should be conveyed to them and their parents. Converting scores and performance into grades is at best a rather arbitrary process, which is further complicated by public relations problems in reporting to parents. Frequently these difficulties produce double talking teachers and confused students and parents.

The conversion of performance data into meaningful ratings of quality has been hotly debated topic for many decades. Prior to the use of objective tests, marking and grading were usually synonymous and the infallibility of he teacher's judgement was rarely questioned. In China's "Cultural Revolution," initiated in 1966, grading was greatly de-emphasized to help create a "classless society". In United Stats, a survey found that both faculty and students believe grading serves several necessary and useful purposes.

OBJECTIVES

Having intensively gone through this unit the students should be able to

1. Learn the basic principles of a good marking system.
2. Identify the functions of grades and progress reports.
3. Describe different types of marking reporting systems.
4. Assign Letter Grades.
5. Conduct Parent-Teacher Conferences.
6. Report Test results to the parents.

CONTENTS

Introduction

Objectives

1. Basic Principles of Good Marking System

1.1 Educational reporting

1.2 Prerequisites for reporting

1.3 Functions of marks and progress reports

1.4 Types of marking and reporting system

1.5 Suggestions for improving marking and reporting

1.6 Multiple marking and reporting systems

1. Grades An Purposes Of Grades

2.1 Assigning letter grade

2.1.1 Nor-referenced grading

2.1.2 Criterion-referenced grading

3 Conducting Parent-Teacher Conferences

3.1 Preparing for the conference

3.2 Establishing and maintaining rapport during the con-ference

3.3 Sharing information with parents during the con-ference

3.4 Planning a course of action with parent during con-ference

4 Reporting Test Results To Parents

4.1 Describing what the test measures

4.2 Explaining the meaning of test Scores

4.3 Clarifying the accuracy of test Scores

4.4 Discussing use of the test results

5 Self Assessment Questions

6 Bibliography

1. BASIC PRINCIPLES OF GOOD MARKING SYSTEM

If instructional objectives have been clearly defined in performance terms and evaluation procedures have been effectively applied, the task of reporting pupil's progress will be greatly simplified. It is still a rather perplexing one, however, as the evaluative data usually must be summarized into a single letter grade or, at best, a very brief form. The process is highly subjective on, for which there are relatively few helpful guidelines. This has led to the use of marks and progress reports which vary widely in composition and meaning.

The greatest confusion arises when pupil progress is summarized as a single letter grade (e.g., A.B.C.D.E.F.). It can be explained as under:

1 Should the assigned mark represent level of achievement, gain in achievement, or some combination of the two?
2 Should effort be included, or should high achievers be given good marks regardless of their effort?
3 Should pupils be marked according to their own potential lear-ning ability or in relation to their classmates' achievements?

There are no simple answers to such questions. Practice varies from school to school, and frequently from teacher to teacher with the same school system. Therefore, it is better to point out basic principles of good marking system.

According to Chand (1990) basic principles of a good marking system are as under:

1 A marking system should be clear and definite, which can easily comprehended by the pupils, teachers and parents.

2 A marking system should be realistic, reasonable, and as true to human life patterns as possible.

3 A marking system should provide sufficient range of grades, so that, various degrees of attainment can be indicated reliably.

4 A marking should be based on objective measures or standards that can be checked objectively or rated consistently with high degree of reliability.

5 A marking should utilize statistical procedure in converting scores into grades.

6 A marking system must be used a means to an end and not as an end in itself.

When a teacher is aware of the basic principles of a good marking system, then the next step should be the awareness of the teacher about Educational Reporting.

1.1 **Educational Reporting**

Educational reporting has been generally defined as the communication of educational outcomes. When, how, why, to whom, and under what conditions should outcomes be reported? Answers to these questions are based on the assumptions that educators have threefold responsibility.

a) To report outcomes accurately.

b) To make sure that these outcomes are properly interpreted.

c) To respect every person's right of privacy.

To share the above mentioned threefold responsibility, it is necessary to let the teacher know about the Prerequisites for Reporting.

1.2 Prerequisites for Reporting.

Logic dictates the following prerequisites for proper reporting:

1. The first requirement of educational reporting is having infor-mation to repot.
2. In addition to getting information to repot, teacher must also sort or categorize information. This sorting will make the reporting task much easier and more useful.
3. Academic achievements in each course or subject should be kept separate.
4. Attendance, social behaviour and study characteristics should not be mixed.
5. The next prerequisite for effective reporting is knowing about the characteristics of one's observations or recorded scores. For example, parents should not be expected to understand the technical aspects of reliability and validity. But a teacher's knowledge of these concepts will influence his or her reporting procedures. Therefore, it is better for a teacher to know more about the quality of his or her observations, which will enable him or her for reliable and valid reports.
6. The nature or characteristics of the course might also influence the valisity of reporting. For example, in a skill subject such as basketball, terminal performance is far more important than a student's performance in the early quizzes or unit tests. Therefore, low scores on the early tests may be ignored.
7. Knowledge about the "marketing" of reports is another prerequisite for effective educational reporting. It means, the form of a report is likely to have a marked effect on how it is received and interpreted. Therefore, the teacher must select proper form of the repot after a survey and through consul-

tation with the administrators, policy matters of the school system, parents, age and learning abilities of the students.

8. Before reporting, a survey may be conducted to find out what parents and students care about.
9. In order to provide effective educational reporting you must understand the statistical characteristics of the information, and find the best way to market it.

Precautions for Reporting

To increase the probability that students and parents will correctly interpret school reports and respond to them in ways that make it easier to reach educational goal. You must make sure that reports are clear and consistent, provide desired information, and are based on reliable and valid measurement. Following three general precautions might provide special guidance in reporting negative aspects of performance and behaviour.

a) Negative reports, such as such as declines in grades or problem with conduct, might be read more carefully and responded to in more helpful fashion if positive reports can also be given. Do not let a student's positive points go unnoticed if you want parents to attend seriously to weak points.

b) Negative reports are likely to be responded to constructively if possible causes and solutions can be identified.

c) Negative reports are likely to be responded to constructively if they are seen as fair.

Keeping the above-mentioned points in mind, following points regarding the privacy of the student should also be given proper consideration.

1. Try if possible to convince the person about whom you wish to reveal information that communication is likely to be of

more benefit.than harm to him or he, then get the person's consent.

2. never divulge personal information about a student to a group o fteachers when only a few of them are directly involved with the student.
3. When seeking advice about how to deal with students and their problems, present and discuss the situation without divulging the person's identify if possible.
4. Be sure, personal comments entered in permanent records are supported by facts and are not accusations based on sus-picions.
5. When giving out personal information about a student to someone from outside the school such as a researcher, social worker, police officer, or community activities director, be sure to obtain written permission from the parents and possibly the student. Be sure to explain the nature and purpose of the disclosure before consent is obtained.
6. When discussing a particular student in a reference letter or conversation with a counselor, make certain that all comments are reasonably objective and can be supported with examples of behaviour.
7. Encourage the professional personnel of your school to discuss and formulate explicit guidelines for protecting the confidences and records of students if such guidelines are not yet in effect.

After knowing about the basic principles of good marking system, educational reporting, its prerequisites, precautions and student privacy the functions of marks and grades should also be made clear to the teacher. So that, an ideal progress report could be presented to report to the parents and the pupils. Gronlund (1985)

has explained the functions of the marks, grades and the progress reports as under:

1.3 Functions of Marks and Progress Reports

School marks and other reports of pupil's progress serve a variety of functions in the schools, which can best be described in relation to the reports' users, including.

1. Reports to the Pupils and Parents
2. Use of repots by Teachers and Counselors
3. Use of Reports by Administrators

1. Reports to the Pupils and the Parents

The main reason for reporting to pupils and parents is to facilitate the pupils' learning and development. Therefore, the reports should.

a) Clarity the school program's objectives.
b) Indicate the pupil's strengths and the weaknesses in learning.
c) Promote greater understanding of the pupil's personal-social development and,
d) Contribute to the pupil's motivation.

From the standpoint of pupil learning most of the functions are probably best served by the day-to-day evaluation and feedback during instruction. However, there also a need for a periodic summary of progress. Pupils find it difficult to integrate test scores, ratings and other evaluation results into an overall appraisal of their success in attaining school objectives, and so the periodic progress report supplies this summary appraisal. In addition it giving pupils a general picture of how they are doing. Such reports also give them a basis for checking the adequacy of their own self-estimates of learning progress.

The desirability of using school marks and progress reports for motivational purposes is often questioned. As with other evaluation procedures, it seems to depend largely on how they are used. If a bad report is held out as a threat to stimulate pupils to work harder, the consequences are apt to be undesirable. However, when the reports are viewed as opportunities to check on learning progress, they are likely to have the same motivational values as properly applied test do, in that they provide shout-term goals and knowledge of results. Although the feedback concerning progress is not as immediate as that obtained from testing, properly prepared reports have advantage of providing a more comprehensive and systematic picture of the pupil's strengths and weakness in learning.

Reports to parents should inform them of the school's objectives and the progress of their children are making toward those objectives. This is important from several viewpoints.

(i) What the school is attempting to do, to enable the parents to cooperate with school personnel in promoting their children's development.

(ii) Information about the success, failures and problems their children are experiencing in school. It enables parents to five them needed emotional support and encouragement.

(iii) Summary repots of learning progress give parents a basis for helping their children to make sound educational plans.

To serve these purposes adequately, the reporting system will need to contain as much information and detail as we can reasonably expect parents to comprehend and use.

2. Uses of Reports by Teachers and Counselors.

Marks and progress reports contribute to the school's instructional and guidance programmes by providing more information about

pupils. Such reports supplements and complement test scores and other evaluation data in the cumulative records. If pupil's past achievements are known, we can better understand their present strengths and weaknesses and can better predict the areas in which they are likely to be successful. The increased information supplied by progress report is especially useful to teachers when they are planning instructions, diagnosing the learning difficulties, and coping with special problems of personal-social development. Counselors use the reports, along with other information, to help pupil's develop better self-understanding and make more realistic educational and vocational plans. Many progress reports also contain information useful in counseling pupils with emotional problems.

The school's instructional and guidance functions seem to be best served by a reporting system that is both comprehensive and diagnostic. To guide learning effectively, aid in personal-social development, and help with future planning teachers and counselors, need detailed information concerning the pupil's abilities and disabilities in each area of development.

3 Use of Reports by Administrators.

Marks and progress reports serve a number of administrative functions. They are used for determining promotion and graduation, awarding honors, determining athletic eligibility, and reporting to other schools and prospective employers. For most administrative purpose, a single letter grade is preferable, largely because such marks are compact and can easily recorded and averaged. With the increased use of machines for routine clerical work, this advantage will probably become even more important in the future.

There is little doubt that the convenience of the single mark in administrative work has been a major factor in retarding the development of more comprehensive and useful progress reports. This need not be the case, however. When a new reporting system is developed , it is possible to use letter grades for administrative purposes and supplement them with the type of information needed by pupils, parents, teachers and counselors. At the high school level, the retention of letter grades is almost mandatory, because most college admission officers insist on them.

1.4 Suggestions for Improving Marking and Reporting,

1. The marking and reporting system should be carefully planned and guided by sated objectives, such as school related motivation, student, parent and teacher understanding; and home-school cooperation.
2. Students, parents, teachers, and administrators, usually with the aid of a technical expert should develop the reporting system and forms.
3. Informal teacher-student reporting and direct communication should be an ongoing process. Student-teacher conferences should not be a 'Last Resort' and should be encouraged as a normal part of the reporting system.
4. Parent-teacher conferences can be very effective. Released time for teacher is often necessary for this activity to be practicable. It is very difficult to make parent conferences practical for every student at the secondary school level, but the conferences are especially desirable for students whose academic perfor-mance begins to decline. Report forms for parent conferences, with copies for each party, are desirable.
5. The reporting system should include feedback on school behaviour, attitudes, work habits, and attendance as well as describe performance in school subjects.

6. Marks and parent conferences are necessary to fully describe the performance at least at the upper elementary and secondary levels. Reporting systems in the primary grades can be less standardized, with greater reliance on parent conferences. Some parents need time to adjust to the reality of their child's ability.
7. Failing marks are rarely justified or needed in the primary or even upper elementary grades. A failing mark should rarely be given in elementary or middle school grades.

1.5 Types of Marking and Reporting System

An effective system of marking and reporting will

I Provide the type of information needed by the report 's users and

II. Present it in an understandable form. These seems relatively simple criteria to satisfy, but most reporting systems fall far short. Much of difficulty is caused by the variety of purposes such- reports are expected to serve. As already discussed, for some uses we prefer comprehensive and detailed reports and for others, a single mark. An additional problem arises from variations in the users educational backgrounds; Information that is understandable to teachers and counselors may be confusing to many parents. Most marking and reporting sys-tems represent some type of compromise between the need for detailed information and the need for conciseness' and simplicity.

a. Traditional Marking, System

The traditional method of reporting pupil progress, which is still Wide used today, is to assign a single letter grade (e.g., A, B, C, D, F) or a single number (e.g., 5, 4, 3, 2, I) to represent a pupil's

achievement in each subject. This system is concise and convenient, but it has several shortcomings:

I. The meaning of such marks is often unclear because they are a conglomerate of such diverse factors as achievement, effort, and good behavior.

2. Even when it is possible to limit the mark to achievement only, interpretation is difficult. A mark of C may mean either average work in all areas or high performance in some areas and low performance in others. An overall summary appraisal in the form of a single mark tells us nothing about the pupils relative success in achieving the various course objectives.

3. As typically used, letter grades have resulted in an undesirable emphasis on marks as ends in themselves. Many pupils and parents view them as goals to be achieved, rather than as means for understanding and improving pupil learning. Although this is not entirely the fault of marking system, the lack of information provided by a single letter grade probably contributes to this misuse.

Numerous attempts have been made to improve the traditional marking system by changing the number and meaning of the symbols used. One common procedure is to reduce the number of symbols to two or three. Typical reports of this type use letters such as H (honor), S (satisfactory), or simply S and U. these variations have been generally unsatisfactory because they provide even less infor-mation concerning the pupil's learning progress.

b. Pass-Fail System

Although a two-category marking system has been used in some elementary schools for many years (e.g., satisfactory-unsatisfactory, pass-fail), it has just recently found widespread use

in many high schools and colleges. As typically used at these levels, it is offered as an option to the traditional letter grade (A, B, C, D, F) in a limited number of courses. Students are permitted.

1.6 Multiple Marking and Reporting Systems

Schools have used traditional letter grades (A, B, C, D, E, F) to report pupil progress for over fifty years, despite efforts to replace them with a more meaningful report. Their continued use indicates that they are serving some useful functions in the school (e.g., adminis-trative). In addition, they are a simple and convenient means of maintaining permanent school records. Thus, rather than replace letter grades (or number grades), it seems more sensible to try to improve the letter grade system and supplement it with more detailed and meaningful reports of pupil learning progress. Some schools already use multiple marking and reporting systems.

The typical multiple reporting system, retains the use of traditional marking (letter grades or numbers) and supplements the marks with checklists of objectives. In some cases, two marks are assigned to each ıbject: one for achievement and the other for effort, improve-ment, or growth.

Guidelines for Development a Multiple Marking and Reporting System

No marking and reporting system is likely to be equally satisfactory in all schools. Each school system must develop methods that fit its particular needs and circumstances. The following principles for devising a multiple marking and reporting system are guidelines for this purpose:

1. The development of the marking and reporting system should. Function is guided by the jimction.5 to be served.

The type of information most needed by the report's users should be included. This typically requires a study of the functions for which the reports are to be used by pupils, parents.

teachers, counselors, and administrators. Although it is seldom possible to meet all of their needs, a satisfactory compromise is more likely if they are known. It is helpful to supplement letter grades in each subject with separate reports on course objectives, effort, personal and social characteristics, and work habits the letter grade should be retained as a pure measure of achievement, and any marks for improvement, effort, or growth should be made separately.

2. The marking and reporting .system should be developed cooperatively by parents, pupils, and school personnel.

School reports are apt to be most useful when all users have some voice in their development. This is usually done by organizing a committee consisting of representatives of parent groups, pupil's organizations, elementary and secondary school teachers, counselors, and administrator and the members carry back to their own respective groups, for modifications and final approval, the committee's tentative plans. This cooperative, participation will not only result in a more adequate reporting system, but It also Increases the likelihood that the reports will be understood by those for whom they are intended.

3. The marking and reporting system should be based on a clear statement of educational objectives.

The same objectives that have guided instruction and evaluation should serve as a basis for marking and reporting. Some of these will be general school objectives and others will be unique to particular courses or areas of study. Nevertheless, when devising a reporting system, the first question should be, "How can we best

report pupil progress toward these particular objectives?" The final report form will be limited and modified by a number of practical considerations, but the central focus should be on the objectives of the schools and course and the types of performance that represent the achievement of these objectives.

4. The marking and reporting system should be based on adequate evaluation.

Teachers should not be expected to report on aspects of pupil performance when evidence is lacking or is very unreliable. By the same token, including items in a report form assumes that the performance will be evaluated as objectively as possible. Rating on such items as critical thinking, for example, should be the end product of testing and controlled observation, rather than depend on snap judgements based on hazy recollections of incidental happenings. Therefore in planning the marking and reporting system, it is necessary to take into account the types of evaluation data needed. The items included in the final report form should be those on which teachers can obtain reasonably reliable and valid information.

5. **The marketing and reporting system should be detailed enough to be diagnostic and yet compact enough to be practical.**

For guiding pupil's learning and development, we should like as comprehensive a picture of their strengths and weaknesses as possible. This needs detail; however, it must be balanced by such practical demands as:

a) A reasonable amount of time is required to prepare and use the reports

b) Reports that are understandable to pupils, parents, employers, and school personnel

c) Reports that are easily summarized for school records. As noted earlier, a compromise between comprehensiveness and practicality is probably best obtained by supplementing the letter grade system with more detailed reports on other aspects of pupil development.

6. The marking and reporting system should provide for parent-teacher conferences, as needed.

At the elementary school level, regularly scheduled conferences with parents might constitute part of reporting system at the high school level, such conferences typically arranged, as needed, to deal with specific problems. At both levels, however, such conferences should supplement a more formal reporting form, rather than replace it. A uniform method of reporting pupil progress is needed for school records, and this is difficult to obtain from conference notes.

In summary, multiple marking and report system takes into account the varied needs of pupils, parents, teacher, and other school personnel. The letter grade system (A, B, C, D, F) provides a simplified method of keeping a record of pupil achievement, the checklist of objectives provides a detailed report of pupil strengths and weaknesses in learning and development, and the parent teacher conference helps maintain cooperation between home and school When letter grade are supplemented by other methods of reporting, these grades themselves become more meaningful. Rather than being a conglomerate of achievement, effort, improvement, and personal behavior, letter grades can be confined to measuring achievement only. Multiple marking makes this

Possible by reporting separately on the other aspects of pupil development.

Activities:

1. What do you mean by multiple marking?

2. Suggest guidelines for a new teacher to develop a multiple marking and reporting system.

3. Enlist the possible drawbacks of multiple marking and reporting system.

2. GRADES AND PURPOSES OF GRADES

The individual teacher must think about grades in the context of learning and reality, must formulate a philosophy and approach to grades and report cards to act accordingly. The grades are with us and it is our job to make them as accurate and effective as possible. After accepting the need for grades, let us review some of their historical purposes as given by Karmel and Karmel (1978):

I. They provide data for parents on their children's progress.
2. They certify promotional status and graduation.
3. They serve as an incentive to do school lessons
4. They help in educational and vocational guidance by presenting a realistic basis for future choices.
5. They serve as a reference point for personal development.
6. They provide a basis for awarding honors.
7. They enable the school to ascertain the amount of extra-curricular activities, if any, in which the student should participate.
8. They may be used as a source for communication to prospective employers.

9. They provide information for curriculum research.
10. They provide data to a school that the student may later attend through transfer or graduation.

2.1 Assigning Letter Grade

Because most schools use A, B, C, D, F. system, most teachers will be faced with the

problem of assigning letter grades. This involves the question such as the following:

a) What should be included in a letter grade?
b) How should achievement data be combined in assigning letter grades?
c) What frame of reference should be used in grading?
d) How should the distribution of letter grades be determined? Each of these issues will be discussed in turn.

a) Determining What to Include in a Grade

As noted earlier, letter grades are likely to be most meaningful and useful when they represent achievement only. If they are contaminated by such extraneous factors as effort, amount of work completed (rather than the quality of work), personal conduct, and so on, their interpretation will become hopelessly confused. When letter grades combined various aspects of pupil development, not only do they lose their meaningfulness as a measure of achievement, but they also suppress information concerning other important aspects of development. A letter grade of B, for example, may represent average achievement with outstanding effort and excellent conduct or high achievement with little effort and some disciplinary infractions. Only by making the letter grade as pure as measure of achievement is possible and reporting on these other

aspects separately, can we hope to improve our descriptions of pupil learning and development.

If letter grades are to serve as valid indicators of achievement, they must be based on valid measures of achievement. This involves the process described earlier- defining course objectives as intended learning outcomes and developing or selecting tests and other evaluation devices that measure these outcomes most directly. How much emphasis should be given to tests, ratings, written reports and other measures of achievement in the letter grades is determined by the nature of course and the objectives being stressed. Thus a grade in English might be determined largely by tests and writing projects. A grade in science by tests and evaluations of laboratory per-formance, and a grade in music by tests and ratings on per-formance skill. The types of evaluation data to include in a course grade and the relative emphasis to be given to each type of evidence are determined primarily by examining the instructional objectives. Other things being equal, the more important the objective is, the greater the weight it should receive in the course grade. In the final analysis, letter grade should reflect the extent to which pupils have achieved the learning outcomes specified in the course objectives, and these should be weighed according to their relative importance.

b) Combining Data in Assigning Grades

When the aspects of achievement (e.g., tests, written reports, performance ratings) to be included in a letter grade and the emphasis to be given to each aspect have been decided, our next step is to combine the various element so that each element receives its intended weight. If we decide, that the final examination should count 40%, the mid-term 30%, laboratory performance 20%, and written reports 10%, we will want our

course grades to reflect this emphasis. A typical procedure is to combine the elements into a composite score by assigning appropriate weights to each element and then uses these composite scores as basis for grading.

Combining data into a composite score in order to produce the desired weightage is not as sample as it may appear at first glance. This can be illustrated by a simple example. Let us assume that we want to combine scores on a final examination and a term report and that we want them to be given equal weight. Our ranges of scores on the two measures are as follows:

	Range of Scores
Final Examination	80 to 1OO
Term Report	10 to 15

Because the two sets of scores are to be given equal weight, we may be inclined to simply add together the final examination score and the term report score for each pupil. We can check on the effectiveness of this procedure by comparing the composite score of a pupil who is highest on the final examination and lowest on the term report (1OO + 10 = 110) with a pupil who is lowest on the final examination and highest on the term report (80 + 50 =130). It is obvious from this comparison that simply adding together the two scores will not give them equal representation. Likewise there may be other erroneous methods to equate scores without considering a larger difference between the two composite scores. This is due to the fact that the influence each component has on the composite score depends on the variability, or spread, of scores and not on the total number of points. Thus, to weigh properly the components in a composite score, the variability of the scores must be kept into account.

The range of scores in the example provides a measure of score variability, or spread, and this can be used to equate the two sets of scores. We can give the final examination and the term report equal weight in the composite score, by using a multiplier that makes the two ranges equal. Because the final examination scores have range of 20 (100 -80) and the term report scores a range of 40 (50 -10), we would need to multiply each final examination score by 2 to obtain the desired equal weight. If we wanted our final examination to count twice as much as the term report, it would be necessary to multiply each final examination score by 4 rather than by 2.

A more refined weight age system can be obtained by using the standard deviation as the measure of variability, but the range is satisfactory for most classroom purposes.

The components in a composite score can also be weighted properly by converting all sets of scores to STANINES (Standard Scores, 1 through 9). When all scores have been converted to the Stanine system, the scores in each set have the same variability. They then are weighed by simply multiplying each Stanine score by the desired weight. Thus, a pupil's composite score would be determined as follows:

	Desired Weight	Pupil's Stanines	Weighed Score
Examination	2	9	18
Laboratory work	1	7	7
Written reports	1	8	8
			Composite Score = 33

These composite scores can be used to rank pupils according to an overall weighed measure of achievement in order to assign letter gra-des.

c) Selecting the proper Frame of Reference for Grading

Letter grades are typically assigned on the basis of one of the follow-ing frames of reference:

I. Performance in relation to other group members (norm referenced).
II. Performance in relation to prespecified standards (criterion re-ferenced).
III. Performance in relation to learning potential or amount of im-provement.

Assigning grades on a norm-referenced basis involves comparing a pupil's performance with that of a reference grouped, typically one's classmates. With this system, the grade is determined by the pupil's relative ranking in the total group, rather than by some absolute standard of achievement. Because the grading is based on relative performance, the grade is influenced by both the pupil's performance and the performance of the group. Thus, one will fare much better, grade wise, in a low-achieving group than in a high-achieving group.

Although norm-referenced has the disadvantage of a shifting frame of reference (i.e., grades depends on group ability), it is widely used in the school, because much of classroom testing is norm referenced. That is, the tests are designed to rank the pupils in order of achievement, rather than to describe achievement in absolute terms. Although relative position in group is the key element in the norm-referenced system of grading, the actual grades assigned are also likely to be influenced to some extent by the achievement expec-tations that the teacher has acquired from

teaching other groups. Thus, a high-achieving group of pupils is likely to receive a larger proportion of good grades than a low-achieving group.

Assigning grades on a criterion-referenced basis involves comparing a pupil's performance to prespecified standards set by the teacher. These standards are usually concerned with the degree of mastery to be achieved by pupils and may be specified as

1. Tasks to be performed (e.g., type 40 words per minute without error) or
2. The percentage of correct answers to be obtained on a test designed to measure a clearly defined set of learning tasks.

Thus, with this system, letter grades are assigned on the basis of an absolute standard of performance rather than a relative standard. If all pupils demonstrate a high level of mastery, all will receive high grades.

The criterion-referenced system of grading is much more complex than it first appears. To use absolute level of achievement as a basis for grading requires that

1. The domain of learning tasks be clearly defined,
2. The standards of performance be clearly specified and justified, and
3. The measures of pupil achievement be criterion referenced. These conditions are difficult to meet except in a mastery learning situation. When complete mastery is the goal, the learning tasks tend to be more limited and easily defined. In addition, percentage-correct scores, which are widely used in setting absoluter standards, are most meaningful in mastery learning because they include how far a pupil is from complete mastery. All too frequently, schools use absolute

grading based on percentage correct scores (e.g., A = 95 - 100, B = 85 - 94, C = 75 -84, D = 65 -74, F = below 65) but the domain of learning tasks has not been clearly defined and the standards have been set in a completely arbitrary manner. To fit the grading system, teachers attempt to build test (norm referenced) with scores in the 60 to 100 range. If the test turns out to be too difficult or too easy, they somehow adjust the scores to fit the absolute grading scale. But such grades are difficult to interpret because they represent adjusted level of performance on some ill-defined conglomerate of learning tasks. 4. Although reporting pupil performance in relation to learning potential or amount of improvement shown has been fairly widely used at the elementary school level, this type of grading is fraught with difficulties. Making reliable estimates of learning potential, with or without tests, is a formidable task, because judgements or measurements of potential are likely to be contaminated by achievement to some unknown degree. Similarly, improvement (i.e., growth in achievement) over short spans of time is extremely difficult to estimate reliably with classroom measures of achievement. Thus, the lack of reliability in judging achievement in relation to potential, and in judging degree of improvement will result in grades of low dependability. If used at all (e.g., to motivate low-ability pupil's), such grades should be used as supplementary. In dual marking, for example, one letter grade might be used to represent achievement in relation to potential, or the degree of improvement shown since the marking period.

d) Determining the Distribution of Grades

As noted in the previous section, there are two ways of assigning letter grades to measure the level of pupil achievement the norm-

referenced system based on relative low level of achievement and criterion-referenced based on absolute level of achievement.

2.1.1 Norm-Referenced Grading.

The assignment of norm-referenced grades is essentially a matter of ranking the pupils in order of overall achievement and assigning letter grades on the basis of each pupil's rank in the group. This ranking might be limited to a single classroom group or might be based on combined distributions of several classroom groups taking the same course. In any event, before letter grades can be assigned, the proportion of As, Bs, Cs, Ds and Fs to be used must be deter-mined.

One method of grading that has been widely used, in the past is to grade on the basis of the norm curve. The procedure results in a distribution of grades like that shown on the next page Grading on the normal curve results in an equal percentage of As and Fs, Bs and Ds. Thus, regardless of the group's level of ability, the proportion of high grades is balanced by an equal proportion of low grades. Such grading is seldom defensible for classroom groups because (1) the groups are usually too small to yield a normal distribution: (2) classroom evaluation instruments are usually not designed to yield normally distributed scores, and (3) the pupil population becomes more select as it improve through the grades and the less-able pupils fail or drop out of school. It is not only when a course or combined courses have a relatively large and unselected group of pupils that grading on the normal curve might be defined. Even then, however, one might ask whether the decision concerning the distribution of grades should be left to a statistical model (i e., normal curve) or should be made on a more relational basis.

The most sensible approach in determining the distributions of letter grades to be used in a school is to have the school staff set general guidelines for the approximate distributions of marks This might involve separate distributions for introductory and advanced courses, for gifted and slow learning classes, and the like. In any event, the distributions should be flexible enough to allow for variation in the caliber of pupils from one course to another and from one time to another in the same course. Indicating ranges rather than fixed percentages of pupils who should receive each letter grade offers this flexibility. Thus, a suggested distribution for an introductory course might be as follows'

A = 10 to 20 percent of the pupils
B = 20 to 30 percent of the pupils
C = 30 to 50 percent of the pupils
D = 10 to 20 percent of the pupils
F = 0 to 10 percent of the pupils

These percentage ranges are presented for illustrative purposes only; there is no simple or scientific means of determining what these ranges should be for a given situation. The local school staff, taking into account the school's philosophy, the pupil population, and the purposes of the grades, must make the decision. A11 staff members must understand the basis for assigning grades and this basis must be clearly communicated to the users of the grades.

In setting an approximate distribution of grades for teachers to follow, the distribution should provide for the possibility of no failing grades. Whether pupils pass or fail a course should be based on their absolute level of learning rather than their relative position in some group. If all low-ranking pupils have mastered enough of the material to succeed at the next highest level of instruction, they all probably should pass. On the other hand, if

some have not mastered the minimum essentials needed at the next highest level, these pupils probably should tail Whether minimum performance has been attained can be determined by reviewing the low-ranking pupils , performance on tests and other evaluation instruments or by administering a special mastery test on the course's minimum essentials. Thus, even when grading is done on a relative basis, the pass-fail decision must be based on an absolute standard of achievement if it is to be educationally sound.

2.1.2 Criterion-Referenced Grading

Criterion referenced grading is most useful when a mastery learning approach is used, because mastery learning provides the necessary conditions for grading on an absolute basis. This includes delimiting the domain of learning tasks to be achieved, defining the instruct-tional objectives in performance terms, specifying the standards of performance to be attained, and measuring the intended outcomes with criterion referenced instruments.

If the course's objectives have been clearly specified and the standards for mastery appropriately set, the letter grades in a criterion referenced system may be defined as the degree to which the objectives have been attained, as follows:

A = Outstanding. Pupil has mastered all of the course's major and minor instructional objectives.

B = Very Good. Pupil has mastered all of the course's major instructional objectives and most of the minor objectives.

C = Satisfactory. Pupil has mastered all of the course's major objectives but just a few of the minor objectives

D = Very weak. Pupil has mastered just a few of the course's major and minor instructional objectives and barely has the essentials needed for the next highest level of instruction. Remedial work would be desirable.

F = Unsatisfactory. Pupil has not mastered any of the course's major instructional objectives and lacks the essentials needed for the next highest level of instruction. Remedial work is needed.

If the tests and other evaluation instruments have been designed to yield scores in terms of the percentage of correct answers, criterion referenced grading then might be defined as follows.

A = 95 to 100 percent correct
B = 85 to 94 percent correct
C = 75 to 84 percent correct
D = 65 to 74 percent correct
F = below 65 percent correct

As noted earlier, defining letter grades in this manner is defensible only if the necessary conditions of a criterion-referenced system have been met. Using percentage-correct scores when the measuring instruments are based on some undefined hodgepodge of learning tasks produces uninterruptible grades. With criterion-referenced grading systems such as these, the distribution of grades is not predetermined. If all pupils demonstrate a high level of mastery, all will receive high grades. If some pupils demonstrate a low level of performance, they will receive low grades. Thus, the distribution of grades is determined by each pupil's absolute level of performance, and not by the pupil's relative position in the group.

There are basically two procedures for determining what letter grade a pupil should receive in a criterion-referenced system. The

first, called the one-shot system, provides a single opportunity to achieve the prespecified standards. The pupil is assigned whatever grade is earned on the first attempt. The second procedure, which is widely used in mastery learning, permits the pupil to make repeated attempts to achieve the prespecified standards. In this approach, the pupil is given corrective help and enough additional learning time to achieve a satisfactory level of mastery. Thus, the first approach may result in some failing grades, where as the second approach eliminates failure. Typically, only the letter grades A, B, C, D, and F are used, and pupils are permitted to repeat examination until a satisfactory level of performance is achieved.

The criterion-referenced system for reporting on pupil progress seldom uses letter grades alone. A comprehensive report generally includes a check list of objectives to inform both pupil and parents which objective have been mastered and which have not been mastered by the end of each marking period. In some mastery learning programme letter grades are assigned to each objective to indicate the level of mastery achieved.

Individual differences among students are inevitable, regardless of the type of instruction or evaluation measures employed. Even with individualized, mastery learning approaches, students will differ in rate and degree of mastery. As Eble (1979) has pointed out:

"The notion that criterion-referenced testing will avoid problems of marking seems to be based on quite unrealistic expectations of uniform achievements in learning by all students.... Programmes of mastery learning can not abolish individual differences in ability, interest, and determination."

Students will continue to have strengths and weaknesses irrespective of the instructional strategies employed, and these should be diagnosed and communicated to the students and their parents.

Objections To marks. R.L. Thorndike (1969) carefully observed the grading and different allegations which were offered as support for the abolition of marking. The following summary is hereby produced:

1. Marks are inaccurate and not comparable across instructors, departments, or schools.
2. Marks focus on false and inappropriate objectives and have little relationship to important educational objectives.
3. Marks have limited value as a medium of communication between teachers, students and parents.
4. Marks are responsible for a variety of detrimental side effects such as anxiety, dishonesty, hostility, and poor mental health; they produce negative attitudes resulting from chronic failure, encourage undesirable value patterns, are incompatible with democracy and so on.....

3. CONDUCTING PARENT-TEACHER CONFERENCES

Regardless of the type of marking and reporting system used in the school, the parent-teacher conferences are an important supplement to the written report of pupil progress. The face-to-face conference makes it possible.

(1) To share information with parents
(2) To overcome any misunderstand between home and to school
(3) To plane cooperatively a programme of maximum benefit to the pupil.

At the elementary school level, where parent-teacher cooperation is most important, conferences with parents are regularly scheduled at the secondary level; the parent-teacher conference is typically used only when some special problem situation arises.

Conferences with parents are most likely to be productive when they are preceded by careful planning and the teacher has skill in conducting such conferences. Many schools offer in-service training for teacher to help them to develop effective conference techniques. Typically such training includes knowledge of how to conduct a parent-teacher conference and role playing to practice the use of conference skills. The following guide lines list the types of things that contribute to the effective use of parent teacher conference for reporting pupil progress.

3.1 Preparing for the Conference

1. Have a clear grasp of the purpose of the conference
2. Review the pupil's school records for general background information.
3. Assemble a folder of specific information concerning the pupil's present learning progress along with other curricular and co-curricular activities and work habits.
4. Organize the information to be presented to parents in a systematic manner.
5. Make a tentative list of questions to ask the parents.
6. Anticipate parents' questions.
7. Provide a comfortable, informal setting, free from interruption.
8. Prepare a written plan of activities of the conference.
9. Encourage two-way communications.
10. Accept some of the responsibility for problems.
11. Conclude the conference with an overall summary.

12. Keep a written record of the conference, listing problems and suggestions, with a copy for the parents.

3.2 Establishing and Maintaining Rapport during the Conference

1. Create a friendly informal atmosphere.
2. be professional and maintain a positive attitude.
3. Use language that is understandable to parents.
4. be willing to listen to parents.
5. be honest and sincere with parents and do not betray confidences

3.3 Sharing Information with Parents during the Conference

1. Begin by describing the pupil's strong points
2. Describe the area needing improvement in a positive and tactful manner
3. Encourage parents to participate in the conference.
4. be cautious about giving advice-

3.4 Planning a Course of Action with Parents during Conference

1. Begin the concluding phase of the conference with a brief overall summary.
2. Have parents participate in planning a course of action.
3. Review your conference notes with parents.
4. End the conference on a positive note.

In addition to these positive things to do, V. Bailard and R. Strange suggest a list of important don'ts:

1. Don't blame the parents.
2. Don't put the parent on the defensive about anything.
3. Don't talk about other children or compare this child with other children. It is unprofessional.

4. Don't talk about other teachers to the parents unless the remarks are of a complimentary nature.
5. Don't belittle the administration or make derogatory remarks about the school direct.
6. Don't argue with the parent.
7. Don't try to outtalk a parent.
8. Don't interrupt the parent to make your own point.
9. Don't go too far with a parent, who is not ready and able to understand your purpose,
10. Don't ask parents questions which might be embarrassing to them. Only information pertinent to the child's welfare is important. Questions asked out of mere curiosity are unforgivable.
11. After the conference, don't repeat any confidential information which the parent may volunteer. It is most unprofessional and can be very damaging to the parent or the child.

Although one cannot expect to conduct an effective conference by simply reading a list of do's and don'ts, such guidelines provide a good frame work for in-service training and also are a useful reminder of what to do and to avoid when developing conference skills.

Activities

1. What is the Importance of Parent-Teacher-Conferences in our educational setup?
2. as an in charge for conducting Parent- Teacher-Conference.

a) How will you prepare yourself for a conference?

b) How will you share information with parents during the conference?

3. Which precautions should be kept in mind for conducting Parent -Teacher-Conference?

4. REPORTING TEST RESULTS TO PARENTS

Test results are commonly reported to parents during the parent-teacher conference. Although parents have a legal right to all information the school has, concerning their children, it should be presented to them in an understandable and useable form. This means avoiding technical jargon and presenting test results to parents in language that is meaningful to them. There also will be less chance of misunderstanding and more chance of being viewed in proper perspective if the test results are presented as part of the total pattern of information about pupil.

In preparing for the parent-teacher conference, review the test results, and decide when and how they will be introduced into conference. The meaningful communication of test results to parent includes:

1. Describing what the test measures.
2. Explaining the meaning of the test scores.
3. Clarifying the accuracy of the test scores.
4. Discussing the use of the test results.

4.1 Describing what the test Measures.

In reporting on the results of a scholastic aptitude test, saying something like "this test measures skills and abilities useful in school learning" may be sufficient. If the test contains several scores (e.g., verbal, quantitative, and nonverbal), each section of the test may be described in similarly general terms. Test manuals usually contain general descriptions of the tests and subtests that

can be used to explain the test to parents. The following list of things to avoid should help to prevent misinterpretation.

1. Do not refer to scholastic aptitude tests as intelligence tests. The term intelligence is emotionally charged and often misunderstood.
2. Do not describe scholastic aptitude test as measures of fixed abilities. They are not! They measure learned abilities.
3. Do not say, "These test scores predict how well your child will do in school." They would not! Predictions for individuals are hazardous at best, and many factors determine school success. It is better to say something like "Pupils with scores like these usually do well in school," or, for low scores "usually finding difficult."

Most of the test publishers have removed the term intelligence from the titles of their scholastic aptitude tests and now call them tests of school ability, cognitive ability and so on. This makes it easier to describe the tests according to what they actually measure and to avoid the confusion and misinterpretation associated with the term intelligence

Achievement tests are easily described in terms of the test content, and the names of the subtests indicate what the test measures. To say that a reading test measures "vocabulary and reading comprehension" or that a Math test measures "computation and problem solving" is frequently sufficient. In some cases it may be desirable to describe the test results by objective or item clusters, and these are typically identified on the pupil's individual report form. When na-rrative report forms are used, the test content is included as part of the narrative report.

Interpretations of vocational interest inventories, personality invent-tories, and other guidance oriented assessment devices are

best interpreted by the school counselor or other guidance personnel. Parents should be referred to the appropriate staff member if they have questions about scores on these instruments.

4.2 Explaining the Meaning of Test Scores

In making norm-referenced interpretations of test scores, both the meaning of the score and the nature of the norm group should be explained to the parents, both simply and understandably. Both percentile ranks and Stanines are widely used in reporting to parents because they are easy to explain and misinterpretations are less likely to occur.

EXAMPLE

Interpreting percentile ranks: "On the reading vocabulary test, Shaz scored higher than 85 % of a national group of eighth- grade pupils." (It might also be necessary to point out that the 85% does not refer to the percentage of the items answered correctly, but only to the percentage of pupils scoring lower.)

Interpreting Stanines. "On a scale 1 to 9, on which the average score is 5, Shaz received a score of 7 on a reading vocabulary test, when compared with a national group of eighth grade pupils."

Above mentioned figure shows the relationship among Stanines, percentile ranks, and verbal descriptions that can be used in reporting to parents. It is advisable that Grade Equivalents typically should not be used in reporting test results to parents. If it is necessary to use them, point out their shortcomings that a grade equivalent score does not necessarily indicate the grade-level work that the pupil can do.

It is wise to report both achievement results and aptitude results in terms of the same type of score. This makes it easier to explain the scores to parents and makes comparisons of aptitude and achieve-ment more understandable.

Although the parents are interested to know how their child's test performance compares with that of other children, they also want to know what the child has learned and has yet to learn. This type of criterion-referenced interpretation is more readily understood by parents and is typically reported in terms of relative degree of mastery.

4.3 Clarifying the Accuracy of Test Scores.

It is important to communicate to parents that all test scores contain some error. This can be done most easily if confidence bonds (e.g., error bonds) are used in interpreting test scores. Profiles using percentile rank frequently include confidence bonds. If these are not available, percentile rank should be interpreted as estimates that may vary up or down by several points on retesting.

Stanines contain broad units that allow for measurement error because, each Stanine is at least one-half standard deviation wide, a difference of two Stanines usually represents a significant difference in test performance, e.g.

A student scores as under:

Mathematics	8
Reading	6
Science	5

The performance of the student is higher in Mathematics than reading and science, but there is no difference in performance between reading and science. The parents should be told that a

difference of 1 stanine is so small that it can be accounted for by errors of measurement alone.

When interpreting test results by objective or by item-cluster, attention should be paid to the number of items on which each interpretation is based. If the number of items is small (say less than 10), make only tentative interpretations, and explain to the parents that these are simply clues to be verified by further study. When combined with the results of teacher-tests and other classroom work, more dependable interpretations may be possible. It is always good practice to interpret test scores to parents in light of the other available data concerning the pupil.

4.4 Discussing use of the Test Results

The interpretation of the test results should be accompanied by an explanation of how the test results are to be used in the instructional programme and a discussion of what action should be taken by both teacher and parent to improve the pupil's learning and development. This discussion should not be limited to the test results but should be based on all of the evidence concerning the pupil's learning and development. The value of test scores becomes clearer to parents when they are coordinated with all of the other information about the pupil and when they are seen as contributing to plans for constructive action.

Activities:

1. As a good teacher, which major points you will consider while communicating the test results to the parents?

2. Define the test scores.

---

3. With the help of an example clarify the accuracy of test scores to the parents.

5. SELF ASSESSMENT QUESTIONS

1. What are the advantages and limitations of a multiple marking and reporting system?
2. If you were helping set up a multiple marking and reporting system for the grade level you are teaching or plan to teach, what types of marks and reports would you want included? Why?
3. What are the advantages and limitations of assigning letter grades (A, B, C, D, F) on a relative basis (norm-referenced grading)?
4. What are the advantages and limitations of assigning letter grades on the basis of an absolute level of achievement (criterion-referenced grading)? What are the prerequisites for an effective criterion-referenced system?
5. What problems might arise when interpreting letter grades based on performance in relation to learning potential? What problems are involved in computing such grades?
6. Describe the procedure to be followed in (a) determining what types of evaluation data to use in assigning grades, (b) determining what relative weight should be given to each type of data, and (c) computing composite scores.

7. What are the limitations of assigning letter grades strictly on the basis of the normal curve? What are the advantages, if any?

8. Some educators have suggested that letter grades be abolished at the elementary and secondary levels. What are the advan-tages and disadvantages of abolishing grades?

9. What factors should be considered when deciding whether to pass or fail a pupil? Do you think the decision should be based on a relative or an absolute standard? Why?

10. What types of information should you have at hand during a parent-teacher conference? How would you explain to parents that their child was performing poorly in school? Describe the general approach that you would use in explaining low test scores to parents.

6. BIBLIOGRAPHY

1. Gronlund, N.E. (1985); Measurement and Evaluation in Teaching, 5th ed., Macmillan Publishing Company, New York

2. Chand, T. (1990); Educational Technology, Anmol Publications, New Dehli.

3. Scriven, M. (1970); Discussion. Proceeding of the 1969 Invitational Conference on Testing Problems; Educational Testing Service, Princetion, New Jersey.

4. Ebel, R.L., (1979); Measuring Educational Achievements, 3rd ed., Prentice-Hall, Englewood Cliffs, New Jersey.

5. Thorndike, R.L., (1969) Perspectives in Educational and Psychological Measurement, Prentice-Hall, Englewood Cliffs, N.J.

6. Bailard,V., and Strang, R., (1964); Parent Teacher conferences, McGrow Hill, New York.

7. Karmel, L.J. and Karmel, M.O. (1978); Measurement and Evaluation in the Schools, 2nd Edition, Macmillan Publishing Co., Inc., New York.

8. SAX, G. (1980); Principles of Education and Psychological Measurement and Evaluation, 2nd. ed; Wadsworth Publishing Company Belmont, California.

Unit-8

SCHOOL TESTING, TRENDS AND ISSUES

BY:
JAMIL HUSSAIN SHAH

INTRODUCTION

Testing has expanded rapidly in the schools. In classroom instructions, the focus oh tests for improving learning has led to their increased use. This is especially apparent in programmes that emphasize mastery learning and individualized instruction. With the advent of minimum-competency testing and mainstreaming, additional testing has been required to meet the needs for certification, screening, and placement. Similarly, the increasing number of students going to college, and the need to monitor educational progress has resulted In expanded college admissions, national and international testing Programmes. Some of these Programmes are likely to have direct influence on our work as a teacher. There, the teacher should know about the testing Programmes, so that we may serve as an informed professional in talking to pupils, parents, and the public.

The increased use of tests in the schools and the varied functions they are expected to serve have contributed to a changing emphasis in educational measurement. They have also created public concern about the role of testing in the schools. Trend and issues discussed in the unit are from the global perspective, particularly USA.

OBJECTIVES

After going through unit the students should be able to:

1. Understand the role of testing in classroom instructions.
2. Differentiate between Mastery Learning and Individualized Instructions.
3. Define Minimum Competency Testing and Mainstreaming.
4. Write about different Admission Testing Programmes.
5. Have the awareness about the National Assessment of Educational process.
6. Differentiate between Trends. and Issues.
7. Enumerate current Measurement Trends and Issues.

CONTENTS

Introduction

Objectives

1. Testing in Classroom Instruction

1.1. Mastery Learning

1.2. Individualized Instruction

2. Minimum Competency Testing

3. Mainstreaming

3.1 Strategy

4. Admission Testing Programmes

4.1 College Entrance Examination Board (CEEB)

4.2 American College Testing Programme (ACT)

5. National Assessment of Educational Progress

5.1 Current Measurement Trends and Issues

6. Concerns and Issues

7. Self Assessment Questions

8. Bibliography

1. TESTING IN CLASSROOM INSTRUCTION

Testing can-help to improve learning by assisting in:

1. Pre-assessing learners' needs,
2. Monitoring learning progress,
3. Diagnosing learning difficulties, and
4. Determining the degree to which the intended learning outcomes have been attained.

Using tests for these various functions has increased testing in the classroom, especially when mastery learning and individualized instruction are emphasized.

1.1 Mastery learning

Mastery learning can be viewed as one aspect of conventional classroom instruction that a course's intended learning outcomes are divided into *mastery* outcomes and *developmental* outcomes. The former are concerned with the minimum essentials of the course (e.g., basic skills) that must be mastered by all pupils and the latter with the more complex transfer-type objectives (e.g., application, problem solving) toward which pupils can show varying degrees of progress. Another way of viewing mastery learning in classroom instruction is a basic strategy to be applied to all intended learning outcomes. This is the approach taken by B.S. Bloom, J.H. Block, and others. Here, we shall describe Bloom's mastery learning strategy to illustrate how tests are used in monitoring and guiding pupil learning in such a programme.

Essentially, Bloom's mastery learning approach is an instructional strategy designed to bring all, or nearly all, pupils to a specified level of mastery on all course objectives. It combines regular classroom instructions with feedback corrective techniques

for overcoming individual learning errors. Additional learning time is provided for those pupils who need it. Thus, the Bloom approach uses regular group-based instruction that is supplemented by carefully prescribed corrective study for those pupils who fail to achieve mastery during the grouped instruction.

The following steps outline the essential features of Bloom's mastery learning strategy:

1. The course is subdivided into a series of learning units that includes a week or two of learning activity. These units might be chapters in a textbook or some other meaningful segment of course content.

2. The instructional objectives are identified and specified for each learning unit. A wide range of learning outcome is stressed (e.g. knowledge, comprehension, application), and the objectives are defined in specific terms.

3. Mastery standards are set for the objectives in each learning unit, often as the percentage of test items a pupil is expected to answer correctly. Although the setting of mastery standards is somewhat arbitrary, the performance of pupils who have previously taken the course is used as a guide. Mastery is frequently set at 80 to 85 percent correct for each unit, but this must be adjusted to fit various learning and testing conditions.

4. The learning tasks in each unit are taught using the regular materials and methods of group-based instruction.

5. Diagnostic-progress tests (formative tests) are given at the end of each learning unit. The results of these formative tests are used to reinforce the learning of the pupils who have mastered the material and to diagnose the learning

errors of those who have failed to achieve mastery. They are generally not used for assigning grades.

6. Procedures for correcting learning errors and additional learning time are prescribed for those pupils who do not demonstrate unit mastery. Those prescriptive-corrective techniques include reading particular pages in an alternate textbook, using programmed materials, using audiovisual aids, individual tutoring, and small group study sessions. If one method does not prove successful with a particular learning problem, the pupil is encouraged to use another method. The pupil is usually retested after corrective study.

7. After completing all the course units, an end-of-course test (summative test) is administered. The results of this test are used primarily to assign course grades. All grades are assigned on the basis of absolute standards that were set in the beginning of the course. Thus, if all pupils achieve the level of mastery prescribed for an A grade, all will receive that grade.

8. The results of the formative tests and the summative tests are used as a basis for evaluating and improving the instructions. Typically, the methods, materials, and sequencing of instructions are examined closely whenever a majority of pupils experience difficulty in mastering the learning tasks.

In summary, Bloom's mastery learning strategy is a group-based method that uses specials techniques for adapting instruction to the needs of individual pupils. It differs from conventional classroom instruction in that:

1. It emphasizes the mastery of all objectives in each of a series of learning units.

2. It uses diagnostic-progress tests (formative tests) to identify each pupil's learning errors.

3. It uses systematic feedback-corrective procedures and alternate learning resources (e.g. programmed material) for helping pupils to overcome learning difficulties.

4. It provides additional learning time for those pupils who need it. Thus, instead of holding learning time constant and accepting a wide range of achievement, Bloom's approach allows for variation in learning time and emphasizes a high level of achievement for all pupils. Learning effectiveness and course grades are determined by the level of mastery pupils achieve, rather than by how their performance compares with that of their classmates.

As can be seen from the description of Bloom's mastery learning strategy, formative tests playa key role by:

1. Reinforcing the learning of high achievers.

2. Pinpointing the errors of low achievers. ,

3. Indicating the type of corrective prescriptions needed to improve learning.

Therefore, the summative tests are used principally for assigning grades and evaluating the effectiveness of the instructions.

1.2- Individualized Instruction

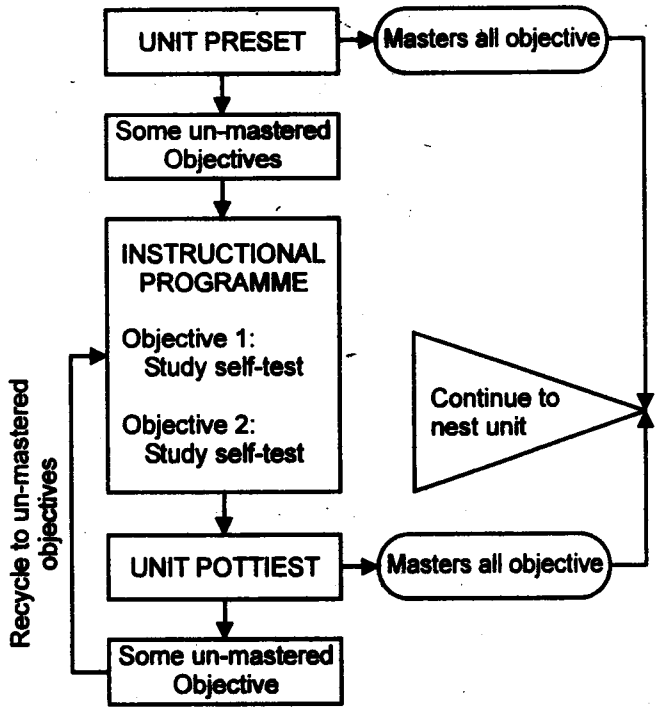
Individualized instruction permits pupils to work on a series of individual learning units at their own pace and level of achievement. The pupils' work on each unit of study is commonly directed by a learning guide called an Individual Study Unit, a Teaching-learning unit, a Learning Package, a Self Instructional Unit, a Student Learning Contract, or some similar name. Despite

the variation in titles, learning guides are similar in content and design typically include the following key elements:

1. One or more instructional objectives.
2. Pretest
3. List of learning activities and materials.
4. Self-tests to aid the pupils in monitoring learning progress.
5. Post test.

Thus, each learning guide consists of a self-contained instructional unit that permits a pupil to work through a programme of study, unit by unit, with only a minimum of teacher guidance.

As noted in the learning guide outline, testing is a prominent part of each individual learning unit. The pretest aids in determining pupils' readiness for studying the unit and serves as a placement guide. If pupils do well on the pretest, they may be permitted to skip some of the objectives in the unit or may be directed to move on to the next unit. The self-tests are designed to measure mastery of each of the unit's instructional objectives. These tests help pupils decide when they should move on to the next objective and when they are ready to take the unit post test. The post test determines whether pupils have mastered the unit's objectives. If pupils perform satisfactorily on the post test, they will move on to the next unit of study. If their performance is unsatisfactory, they will continue work on the unit and retake a second form of the posttest at a later date through units of individualized instruction. To read the chart, start with the UNIT **PRETEST** and simply follow the arrows.



(Flow Chart For Individualized Instructional Unit)

Learning guides for individualized instructional units can be developed locally, or a school can participate in one of the comprehensive individualized instructional programmes, such as Individually Prescribed Instruction (IPI). Regardless of the approach, there will be heavy emphasis on testing and evaluation. All such programmes typically involve:

1. Pre-testing to determine entry behaviour and to place the pupil properly in the programme.
2. Testing during the programme to monitor pupil progress.
3. Post-testing to determine pupil's final mastery of the instructional objectives.

Thus, the same types of testing used in conventional classroom instruction are also used in programmes for individualizing instruction. Here, however, the testing function is

even more crucial because periodic testing is used to guide and direct pupil progress on self-contained, individual learning units.

Activities

1. Define Mastery Learning

6. Enlist essential features of Bloom's Mastery Learning Strategies
7. Define Individualized Instruction.

4. Explain About the key elements Individualized Instruction

2. MINIMUM COMPETENCY TESTING

Minimum competency testing is another area where the use of tests has potentially great benefits, but not without some costs to individuals. The minimum competency testing movement usually refers to the requirement that individuals who have completed high school must pass a test of competency in the basic educational skills before being awarded a diploma. The movement supporting minimum competency testing gained impetus with increasing evidence concerning the failure of a high school education to guarantee even the most basic skills or reading and using arithmetic.

Based on research data provided by the National Assessment of Educational Progress, it has been estimated that about 13 percent of American high school graduates could be classified as illiterate, and another 44 percent as semiliterate, leaving a shocking figure of only 43 percent qualifying as functionally literate. Among black students, the illiteracy rate was 40 percent, and only 20 percent qualified as functionally literate.

Unfortunately, the failure rate on minimum competency tests was much higher among black students than among white students, meaning that proportionately more blacks were denied high school diplomas on the basis of test scores.

Although minimum competency testing is under continued scrutiny, several arguments in its favor, as well as against it, should be noted. First of all, minimum competency testing was intended to have, in the long run, a beneficial effect on students' educational and occupational attainments. An external requirement, such as a test of basic educational skills, is designed to have the effect of increasing time and effort on these basic skills on the part of both teachers and students. It has been hoped, therefore,

that in the long run minimum competency testing will reduce rates of illiteracy and increase student's master of educational skills. In turn, students should later find greater occupational and economic success as well as the personal rewards from some thing so basic in this society as knowing how to read.

But even as minimum competency testing is designed to have positive effects and should be considered a matter of educational policy rather than civil rights, it does adversely affect blacks' ability to receive a high school diploma. Assuming evidence for content validity with reference to the instruction offered, until the quality of the education received by blacks is more consistently comparable to the quality of that received by whites and unless students and teachers are given time to prepare themselves for the fact of competency examinations, it may not be fair to deny diplomas on the basis of testing. Yet the alternative, which is to allow the serious problem of illiteracy to continue without the multifaceted attention that would include testing, does not sound appealing either. Again, there are not totally costless approaches to the problems.

One of the most rapidly developing educational trends in the late 1970s and early 1980s is the movement toward minimum-competency testing. Typically this involves testing pupils on the basic skills of communication and computation and setting a satisfactory level of performance as a graduation requirement. Some of the more elaborate programmes also include testing for competency in such basic life skills as citizenship, consumer education, career education, and health. In some programme, competency tests are also used as a basis for promotion at the lower grade levels.

By the beginning of the 1980s, the majority of states had either passed legislation or adopted state board-of-education rulings mandating some type of minimum-competency testing. In some states the programmes are developed by local school districts but in others they are centrally controlled, and statewide competency tests are used. Pupils who failed may be given remedial work, and those who fail repeatedly may be given a certificate of attendance or a special diploma rather than the regular high school diploma.

Although minimum-competency testing programmes differ widely in origin and nature, they have some common problems that are not easily resolved. The following questions pertain to some of the more serious concerns:

1. What competencies should be assessed, and when?
2. What types of tests and other evaluation instruments should be used?
3. How can we determine whether the evaluation instruments are providing dependable data?
4. How should we set the standards for passing or failing?
5. What type of remedial instruction is needed for those who fail?
6. How can we ensure fairness to minority group members and handicapped pupils?
7. How can we ensure that the programme will contribute to pupil learning rather than detract from it?

Although the orientation is different, these are the same types of issues encountered in classroom testing and evaluation. Thus, whether you find yourself helping develop a local competency-

testing programme, participating in a statewide programe merely serving as an informal critic, your knowledge of testing can be useful in making this type of testing program more effective.

A number of states have followed up minimum-competency testing of pupils with minimum-competency tests for teachers. These tests typically cover the basic skills in addition to various areas of professional knowledge and are required for teacher certification. The Educational Testing Service recently added a section on communication skills (included reading, writing, and listening) to its National Teacher Examinations. The communication skills section can be administered as part of the battery of tests for certification or be used separately as a screening device at the beginning of teacher education programmes. It seems likely that in many states both teachers and pupils will be expected to pass basic skills-competency in the future.

Activities

1. What do you mean by Minimum Competency Testing?

_____

_____

_____

_____

_____

2. Critically discuss about Minimum Competency Testing.
3. Enlist different aspects, which are dealt by Minimum Competency Testing.

3. MAINSTREAMING

The Education for All Handicapped Children Act (Public Law 94-142) passed by the United States Congress in 1975 and put into effect in the fall of 1977 is having far-reaching effects on school practice. This law has numerous provisions, but it is essentially designed to give the handicapped, from age 3 to 21, access to equal educational opportunity by placing them. in the *least restrictive environment* and providing adequate resources and support. Handicapped pupils must be placed with their peers in the regular classroom unless other placement can be justified by suitable criteria. The term mainstreaming has been widely used to represent the educational provisions for helping handicapped pupils attain their rights under this law. Although mainstreaming requires shifting many pupils from special classes to regular classes, the term does not simply mean the placement of handicapped pupils in regular classrooms. Instead, it encompasses all activities for establishing an environment that is most conducive to the learning and development of the handicapped.

As might be expected, testing and evaluation playa significant role in the mainstreaming of handicapped pupils. They aid in assessing their special needs, planning individualized educational programs (IEPS) fro them, determining their proper placing in the school programme and evaluating their learning progress and adjustment. Thus, as classroom teachers and other school personnel work more closely with the handicapped, they are likely to be exposed to a broader range of test data than in the past. The results of various psychological tests that were formerly used only by the school psychologists and special education teacher may now also need to be interpreted and used by regular

classroom teachers. In addition, classroom teachers will likely be expected to assess and report periodically on the progress and adjustment of handicapped pupils placed in their classes. In short, mainstreaming requires more testing and demands that classroom teachers learn more about testing and evaluation. Keeping in view, the importance of the testing and evaluation, Ministry of Education, Government of Pakistan emphasized upon introducing **National Testing Service** in National Educational Policy 1992. For this purpose following background was provided:

Evaluation makes an important link in the teaching-learning process. Apart from a weak delivery system, the examination system is excessively flawed on account of a variety of malpractices. Neither the semester system nor the annual system have been able to stand up to the widespread corruption in exami- nations. Under a variety of circumstances, the examiners, the paper-setters, the invigilators and the examination departments appear to be equal partners in maintaining the vicious circle of corruption around public examination. Even the Boards of Intermediate and Secondary Education (BISE) have vied with one another in the awarding of unjustifiably inflated grades to their student. Also, in their anxiety to bring their awards to .professional colleges, the parents have Joined the race of nefarious practices. This state of affairs has caused two types of damages: (a) lack of confidence in the results of public examinations, and (b) distortion in admission to professional colleges.

Implementation of examination reforms suggested in various reports have met with little success. While corruption continues to erode the sanctity of our examinations, "there are other issues with also deserve serious attention. For instance, (i) the equality of-question papers needs improvement so that these are able to test various abilities of students, (ii) objectives of education other than

the cognitive ones need to be taken into account in the process of evaluation, and (iii)the time lag between holding of examinations and declaration of results ought to be reduce.

The policy has attended to the problem with the seriousness it deserves, and proposes that efforts shall be made to improve the present system of examination in all respects. It also proposes that for the purposes of admission in colleges including professional colleges and universities, undue importance attached to the present examination results shall, in part, be surrendered in favour of scientifically national tests.

The development of these tests will be extremely useful in another way. Owing to establishment of a large number of teaching and examining bodies, the difference in the standards of various boards and universities has become a cause of concern. University Grants Commission (UGC) and Inter Board Committee of Chairmen (IBCC) will use these tests to monitor the standards of various universities and boards respectively, and develop a system of accreditation.

Policy Statement

A National testing service will be established for providing national level leadership in developing tests needed in educational setting, and carrying out research for this purpose. The work of test development and research will be carried by NTS on its own, or in collaboration with agencies and individuals with or outside Pakistan. The NTS will provide test as well as testing services to schools, colleges including professional colleges, universities and other institutions. This agency will also be responsible for developing a system of academic accreditation of various institutions on the basis of the performance of students of these institutions on national tests.

3.1 Strategy

National Testing service will be established at the Federal level as a high level autonomous organization in close collaboration with UGC and IBCC.

The tests will be developed by NTS in collaboration with national and international agencies.

NTS will also organize the administration of tests at various places in Pakistan under suitable supervision.

The tests will be administered a number of times during the year so that a students may take a test at any time he is ready for it.

The merit for admission based on the results of public examination will be prepared as usual. However, each person on the merit list will be required to pass the national test before the admission is finalized. Failure in national test will make the student ineligible for admission, irrespective of his position on the merit list. The institutions will introduce regulations to accommodate this provision.

The students waiting for the results of examination shall be eligible for appearing in the national test.

The students waiting for the results of examination shall be eligible for appearing in the National Test.

4-ADMISSION TESTING PROGRAMMES

There are two major college admission testing programmes; one is conducted by the College Entrance Examination Board (CEEB) and the other by the American College Testing Programme (ACT).

4.1 College Entrance Examination Board (CEEB)

The College Board, founded in 1900, is a nonprofit organization composed of schools, colleges, and educational associations. It administers tests at selected centers throughout the world, usually located in schools and colleges. The best known College Board tests are:

1. Preliminary Scholastic Aptitude Test / National Merit Scholarship Qualifying Test (PSAT / NMSQT)
2. Scholastic Aptitude Test (SAT).
3. Achievement Tests (in various school subjects).

The PSAT/NMSQT is an abbreviated version of SAT and is given to high school juniors. It contains a verbal and mathematical section, and the scores are used for early college counseling and preliminary screening for National Merit Scholarship Awards. The SAT and the Achievement tests in school subjects are typically given to high school seniors and are used for college admissions. The SAT provides verbal, mathematics, and English Scores, and Achievement Tests cover fifteen or more specific subjects in social studies, science, mathematics, English and foreign languages. A multiple-choice test, entitled Test of Standard Written English (TSWE), is administered as an adjunct to SAT, but it is not intended for admissions decisions. Rather, the TSWE is used for placement in college freshman English courses and planning remedial programmes.

Both percentile score and standard score are reported for College Boards tests. Standard scores on PSAT/NMSQT range from 20 to 80, with a mean of 50, and scores on SAT and Achievement Tests range form 200 to 800, with a mean of 500 thus, to determine what SAT score is comparable to a given PSAT/NMSQT score, simply add zero to score (e.g., 55=550). When using PSAT/NMSQT to predict what SAT score might be earned a year later, however, another 10 to 50 points need to be added to this converted score (with lower scores requiring a bigger correction) to allow for an increase in test performance from junior to senior year.

In to the college admission testing programme College Board also provides the Advance Placement Programme (APP), the College Level Examination Programme (CLEP), and a series of other testing, scholarship, and guidance services. The APP enables students to obtain college credit while still in high school, and CLEP enables individuals to obtain college credit for equivalent life experiences. CLEP provides a nationwide testing progrmme that permits colleges to grant equivalency credit by examination.

The College Entrance Examination Board was one of the organizations that founded the Educational Testing Services (ETS), in 1947, and all of the programmes are administered by ETS (Princeton, New Jersey).

4.2 American College Testing Programme (ACT)

The ACT programme is an independent, nonprofit corporation, founded in 1959. Its programmes and services are also widely used, and its tests are administered at centers throughout the world. Its testing programme is now called as the ACT Assessment Programme. Although the ACT Assessment is

likely to be modified and expanded over the years, at present it includes the following sections:

1. Academic Tests in English, Mathematics, Social Studies, and Natural Science.
2. ACT Interest Inventory covering six interest areas (Science, Creative Arts, Social Services, Business Contact Business Detail, and Technical).
3. Student Profile Section, which is a self report inventory of student characteristics.

The ACT academic tests are achievement type tests that assess general educational development. They emphasize interpretive and application skills similar to those of the Iowa Tests of Educational Development. A separate score is reported for each of the four tests, plus a composite score based on an average of the four scores. The standard score range from 1 to 36, with a median score of 20 for college-bound high school seniors.

In addition to the Assessment Programme, The ACT has a Proficiency Examination Programme (PEP) for awarding college credit by examination and other related programmes and services similar to those provided by the CEEB.

Activities

1. Define Mainstreaming

2. Enlist Best known College-board-tests.

_____

_____

_____

_____

5. NATIONAL ASSESSMENT OF EDUCATIONAL PROGRESS

The National Assessment of Educational Progress is a nationwide testing programme designed to obtain census like data for reporting to the public on the educational attainment of children and young adults in the country. Testing began in 1969 and has been continued annually, with respect to the public through the news media, professional associations, and various publications.

Tests are given in reading, mathematics, science, social studies, literature, writing, citizenship, art, music, and occupational development. In each area, test exercises are designed to measure specific objectives so that criterion-referenced interpretation of the data is possible. The exercises measure knowledge, understanding, skills, and attitudes and include a variety of item types. In addition to short-answer and essay questions, the exercises include performance tests, observations, questionnaires, and sample products.

Each year, tests in some of the areas are given to a selected sample of individuals at four age levels 9, 13, 17 and young adults from age 26 to 35. To determine change in performance over time, the test in each area is repeated after a period of several years. Although some exercises are released to the public during reporting and others may be revised, many of the exercises are retained from one testing to another to provide a common basis for measuring the growth or decline in achievement.

The results of the testing are reported separately by test item. These reports indicate what percentage of the individuals in each group gives correct and incorrect answers to each exercise. In addition to reports by age group, comparisons are also made by geographic region, type of community, sex, race, and parents' educational status. No results are given for individual students or for individual schools. The intent of the programme is not to compare individuals or schools, but to report the public and the professionals on:

1. What level of educational achievement is being attained by various age-groups.

2. What changes in achievement are taking place over time.

5.1 Current Measurement Trends And Issues

In recent years, numerous changes have taken place in educational measurement. In some instances, the various educational testing and assessment programmes have helped to create or hasten these changes, whereas in others they have simply made the changes more visible. Mounting criticism of testing and possible hazards to individuals and groups has also paralleled the rapidly expanding role of testing in education.

Some Apparent Trends

Describing trends in any areas is harzardous, and is especially so in education. There seem to be so many new developments in education that turn out to be fads-which quickly rise and even more quickly fall. Be that as it may, the following changes in educational measurement seem to have enough breadth and staying power to represent trends

1. **There is greater emphasis on the use of tests to improve learning and instruction.** The shift away from using tests primarily to evaluate students and schools can be seen at all levels of testing. In classroom testing there is increased emphasis on formative testing, the use of mastery and diagnostic tests, and the integration of testing and learning in individualized instructional programmes. With the increase use of computer scoring, tests results from local school wide testing programs, the new emphasis on accountability and competency testing has resulted in greater concern with improved learning and instruction. In fact some states has mandated that remedial instruction must be given to pupils scoring low on a competency test. The National Assessment of Educational Progress has helped shift the focus from test scores to description of pupil learning by keying its assessment exercises to clearly stated objectives and by presenting its report in terms of pupil performance on specific achievement tasks.

2. **There is increased use of criterion-referenced measurement.** This is a close corollary to the first one. The use of tests to improve learning and instruction has created a greater need for criterion-referenced testing, which describes achievement in performance terms. Likewise, the growth of criterion-referenced testing has contributed to the increased use of tests for teaching-learning purposes. Although much of the impetus for criterion-referenced testing came from programmes for individualized instruction and mastery learning, the trend is also apparent in large-scale testing programmes. The National Assessment of Educational Progress has used criterion-referenced assessment exercises from its first testing in 1969. In recent years, there has also been an increasing use of criterion-referenced exercises in the state assessment programmes. Accountability and minimum-competency testing in the state programmes are typically designed

to determine how well pupils are achieving specific learning tasks: this requires criterion-referenced measures. The trend towards criterion-referenced testing can also be seen in the larger number of such tests being published and in the objective-item pools that have been made available for preparing custom-designed criterion-referenced tests. This trend is also visible in the numerous standardized tests that now provide for both criterion-referenced and norm-referenced interpretation.

3. **There is a shift in the nature of aptitude and achievement tests.** There is noticeable tendency for scholastic aptitude tests to emphasize school-learned abilities and to avoid identification as "intelligence" tests. This can be seen in the changing of test titles (e.g., to tests of cognitive ability or school ability) and in the warnings in test manuals that these tests measure developed abilities and are designed primarily to predict school success.

As aptitude tests have shifted toward measuring school-learned abilities, achievement tests have moved toward measuring more complex learning outcomes (e.g., application and interpretation), which have made two type of tests more alike in the skills and abilities being measured. One danger of this trend lies in the detection of underachievers. The same pupils who score low on the achievement-oriented aptitude tests are likely to score low on the abilities-oriented tests, as both measure similar functions.

4. There are more types of assessment exercises being used. There seems to be an increasing awareness of the limitations of standardized testing and the need to supplement them with other types of assessment. As noted earlier, the National Assessment of Educational Progress uses short-answer and easy questions,

performance tasks, observations, interviews, questionnaires, and sample products. Similarly, a number of state assessment programme probably influenced by National Assessment, have expanded their testing programme to include various types of assessment exercises. This broadened scope has been accompanied by an increasing concern with the measurement of affective outcomes and an increased use of application exercises in the cognitive area. Many competency testing prograrmme for example, focus on skills needed in every day life (e.g., reading ads, filling out an employment application).

Classroom teachers have always supplemented paper-and-pencil testing with other methods of evaluation, but the recent emphasis on criterion-referenced testing and its use in improving learning has apparently created new interest in various measures of pupil performance. To describe the types of tasks pupils can and cannot perform requires a much broader range of evaluation techniques than needed for ranking pupils in order to assign grades. In addition, the placement of handicapped children in regular classrooms through mainstreaming has crated a special need to evaluate peer relationships, self-concepts, and other effective outcomes.

5. There is more uses of computers in testing. The use of computer has had a profound impact on the development of testing and can be expected to have even greater influence in the future. From the test user's standpoint, the most readily apparent contribution has been in the scoring and analysis of tests. Test publishers can now provide test results in almost any form desired. A summary of results by class, building, and district can be supplemented by item-analysis data for each item and by a printout for each pupil showing correct and incorrect responses. A comparison with both national and local norms can be made, and

in many cases, interpretation by objective or content cluster is available. The type of feedback made possible by the computer has been a major factor in the trend toward using test results to improve learning and instruction.

Using computers to assist with various aspects of test preparation and use within the school is also on the increase. Any schools with small computers of their own or access to large ones provide rapid-scoring services for teachers' classroom tests, including item analysis data and other analysis useful in diagnosing pupils' strengths and weaknesses. More advanced systems include item banking (e.g., storage of items) which makes it possible to select items and prepare a test with known characteristics (e.g., objectives and content measure, difficulty). It has been suggested that a complete computer assisted test construction system should include item banking, item generation, item attribute banking, item selection, and test printing. When such systems are more widely used in the schools, the teachers will be able to prepare classroom tests of high quality and avoid many of the routine chores of test construction. It appears that the use of microcomputers is rapidly making this possible.

Computers are also being used more in the administration of tests to individual pupils. Computer-assisted individualized instructional programme commonly administer conventional tests by computer. More importantly, the computer has also been used for tailored (adaptive, sequential) testing. The computer's flexibility makes it possible to administer an individual test to each pupil, with each particular set of items determined by the pupil's ability level. Typically, the individual's response to key item will determine which set of items will be presented. Such tailored testing prevents wasting time on items that are too easy or too difficult.

Computers have also been used to compose interpretive statements that are useful in teaching and counseling. Many standardized achievement tests offer a narrative report that summarizes the score profile of individual pupils. The Career Planning Report used with the Different Aptitude Tests is a verbal report describing which occupations are most closely related to an individual's measured aptitudes, interests, and educational aspirations. Some personality inventories also offer a narrative summary of the results.

6. There is increasing public concern about testing. Decisions concerning the selection, administration, and use of educational tests are no longer left to the educator alone; the public has become an active and vocal partner. At the state level, mandated assessment programmes have been Imposed on the schools as a result of the public demand for evidence of the school programmes effectiveness. In some states, the public-at-large has participated, through selected groups in determining the objectives and standard of the statewide assessment programmes. In other states in which competency testing has been made the responsibility of the local school district, parent groups have often help to shape the programmes. It is interesting to note that the concern of state legislators and the general public with the quality of school programmes has created a demand for more testing in the schools-not less.

During the expansion of testing programmes there has also been some concern that there is too much testing in the schools, especially for high school students. In addition to taking the tests in the local school progrmmes, these students may also have to take one or more state competency tests and several college admissions tests. It is feared that the heavy demand on their time and energy might detract them from their schoolwork and that

external testing programmes may cause undesirable shifts in the school's curriculum. When teachers and schools are judged by how well students perform on state competency tests and by how many students are accepted by leading colleges, direct preparation for the tests is likely to enter into classroom activities and thereby distort the curriculum.

Probably the greatest public concern has been with the social consequences of testing, that testing may threaten the rights and opportunities of individuals and groups. This concern has shown up in the form of attacks on standardized tests and the testing industry, new legislation affecting testing, calls for a moratorium on standardized testing, and charges that tests are biased and discriminatory. Although there is certainly some justification for the public's concern with the social consequences of testing, much of the furor has been caused by the misinterpretation and misuse of test scores. Some of the current issues surrounding the testing controversy will be discussed in the following section.

7. Test User Competence. It is essential not only that tests be of high quality, but also that test users be qualified to use them. Tests are sophisticated psychological tools that can be used in harmful and / or effective as well as helpful ways. They should be used only by qualified and informed individuals who will make sure that they are used and interpreted effectively and correctly. Accordingly, a major principle of the APA ethical code is that psychological knowledge and techniques are used only by those qualified to use them and, conversely, that test users operate only within the bounds of their own knowledge and competence. According to the APA code it is important that psychologists who use test results, have an understanding of psychological measurement, problems of test validation, and test research.

Therefore, it is important to ensure that tests and test scores are used only by those people who are qualified to use them.

Qualifications for test use vary according to the types of the tests in question, but in general are stricter with tests having greater potential for harm and misinterpretation. One of the earliest systems by which user qualifications were specified, provided by the first APA test standards, has classified tests according to three levels of complexity:

Level A tests are those that can be administered, scored, and interpreted by responsible non-psychologists who have carefully read the test manual and are familiar with the overall purposes of testing. Educational achievement tests fall into this category.

Level B tests require technical knowledge of test construction and use and appropriate advanced coursework in psychology and related courses (e.g., statistics, individual differences, and counseling). Vocational interest inventories, group intelligence and special aptitude tests, and some personality inventories are generally considered Level B tests. For example, Consulting Psychologists' Press limits purchase of tests such as the Strong-Campbell Interest Inventory, the State-Trait Anxiety Inventory, the Myers-Briggs Type Indicator, and the Bern Sex Role Inventory to people who have completed university Courses in tests and measurements or equivalent training. Similar requirements for access to tests such as the Jackson Vocational Interest Inventory, the Personality Research Form and the Jackson Personality Inventory are stated by their publisher, Research psychologists press.

Level C tests require an advanced degree in psychology or licensure as a psychologist and advanced training/supervised experience in the particular test. Level C tests generally include

individually administered intelligence tests and personality tests, (e.g., the Stanford-Binet Intelligence Scale, the Wechsler Adult Intelligence Scale, and the Minnesota Multiphasic Personality Inventory). Graduate students may be qualified to purchase and use Level B or Level C tests if they are being supervised in that work by some one who does possess the appropriate user qualifications.

Specific criteria for test user qualifications are contained in the catalogues distributed by the major test publishers. Responsible test publishers not only list user qualifications, but also require potential test purchasers to provide their credentials so that only qualified people are sold tests. But ensuring that test users are competent is also the responsibility of the administrators of any organization or agency using tests (e.g., schools and businesses), of test developers who should make available complete information about the technical adequacy and use of the test, and of test users themselves, who, even after earning the appropriate degrees, must engage in continued study of testing research and techniques in order to be competent in their use. A "qualified" test user not only possess the necessary education, training, and experience, but is familiar with the technical, psychometric characteristics of the test to be used, is able to defend the use of the particular test selected rather than alternative tests, is knowledgeable about both administration and interpretation of the test, and is aware of the potential misuses of the test and circumstances and / or types of individuals with whom particular care must be taken.

8. **Privacy and Confidentiality.** Principle 5 of the APA ethical code describes psychologists' obligations in maintaining the confidentiality of information obtained from individuals through testing or any other formal or informal means of gathering

information. The maintenance of confidentiality means protecting the individual's right to privacy and involves the principles of informed consent, constraints on the provision of individual information to other parties, and care in the storage and disposal of information.

The first responsibility of the test user is to minimize, as far as possible, the extent to which testing may threaten a person's right to privacy. The extent to which the use of tests constitutes an invasion of individual privacy is a complex issue involving a number of considerations. It should be noted that test data are no more invasive of privacy than are other kinds of information obtained about people. With any kind of personal information, including test data, the use of the data should be characterized as far as possible by the principles of relevance and informed content.

The concept of relevance means that test scores are collected and used only if they are relevant to some valid set of purposes; collecting a set of test scores for no other purpose than "curiosity" would constitute an unnecessary invasion of an individual's privacy. Similarly, asking a person's height and weight would be appropriate only if related to the purpose at hand. The principle of informed consent means that, as far as possible, the individual is informed about both the nature of the information collected and the purposes for which the information will be used. This principle does not imply that test takers (or their parents) necessarily have a right to information that would invalidate the use of the test or that is beyond the sophistication of the consumer (e.g., test items or methods of scoring), but does require that the test taker be informed as fully as possible about the nature and uses of the test scores.

Although the principle of informed consent is intended to provide to the test taker at least some opportunity for choice in the matter of how much information to reveal and in the uses to which that information will be put, the extent to which the individual actually does have a choice varies across different situations. For a client in counseling, revealing personal infor- mation is important to the process of getting help; unwillingness to be open with the counselor or therapist may make treatment very difficult and client improvement less likely. As an extreme example, a client wishing help with vocational decision making who refused to talk about his/her vocational interests or take a vocational interest inventory would be very difficult for a counselor to help.

When tests are used for institutional (rather than individual) purposes and decisions the individual may choose not to reveal the requested information, but that choice may be costly in terms of an educational or job opportunity. For example, an individual who chooses not to take the Scholastic Aptitude Test (SAT) might forfeit his/her opportunity to go to certain colleges and universities.

Where tests are being used in research, the principle of informed consent is to give the individual and opportunity to choose not to participate in the research. A particularly important principle in the research use of test scores is anonymity of results. This principle means that where the identity of research participants is unnecessary to the purposes of the research the participants' anonymity should be preserved so that the invasion of privacy is minimized.

Finally, cases in which test takers are minors involve special informed-consent concerns. Although minors should be informed as far as possible and at a complexity level that will facilitate their comprehension, parental or guardian consent may be necessary in

some circumstances. Guidelines for the collection and use of test data with minors should guide the use of tests in these situations.

In essence, the issue of invasion of privacy means finding a compromise between the need for meaningful, relevant information to guide problem solving in applied situations and the individual's right to personal privacy and freedom of choice. This compromise is best achieved when respect for promise is best achieved when respect for the individual is combined with sound professional judgement concerning the kinds of information needed for the purpose at hand.

Confidentiality issues pertain to decisions about access to test results on the part of those other than the original test user. For example, a school system may have collected achievement test data that are later requested by a potential employer of the student. Or a psychologist may have given his/her client a Minnesota Multiphasic Personality Inventory (MMPI), scores from which are later requested by another psychologist or the school system. Ethical and test standards mandate that, in general, the confidentiality of test results is to be protected by the original test user unless the test taker gives his/her consent for test results to be provided to someone else. The original informed consent should have covered all intended test uses, and additional consent is necessary for any use beyond those originally agreed to. In the examples given, the student would need to release achievement test scores to the potential employer, and the client would need to release his/her MMPI scores to the second psychologist or to the school system.

A final responsibility of test users involves the secure storage of test results and when test results are too old to be useful, the permanent disposal of test data. Keeping test results in place

where they are accessible to anyone but the original test user is irresponsible, as is disposing of test scores in such a way that they could be easily retrieved by unauthorized individuals. And the test user is responsible for judgements concerning the continued utility of test results collected previously. According to the AERA / APA / NCME standards, test scores should be kept in a person's file only as long as they serve a useful purpose.

To summarize, issues of privacy and confidentiality require respect for the individual client or test taker and his / her rights to privacy and informed consent, as well as care in the release, storage and disposal of test data and other types of personal information.

9. Assessment Techniques: Use and interpretation. One entire principle of the APA ethical code is devoted to the use of assessment techniques. As stated in Principle 8 "Assessment Techniques":

> In the development, publication, and utilization of psychological assessment techniques, psychologists make every effort to promote the welfare and best interests of the client. They guard against the misuse of assessment results. They respect the client's right to know the results, the interpretations made and the bases for their conclusions and recommendations. Psychologists make every effort to maintain the security of tests and other assessment techniques within limits of legal mandates. They strive to assure the appropriate use of assessment techniques by others.

Note that this principle reiterates several issues mentioned previously that is issues of promotion of client welfare, protecting

the security of test results, and ensuring the client's right to be informed about test procedures and results.

First of all, communication and interpretation of test results should always be done in ways that will best serve the original purposes for which test data were collected. For example, the interpretation of personality test data collected in counseling and therapy should be used with the goal of better understanding and helping the client. Test data collected for use in revising school curricula should be interpreted with that end in mind.

Second, the individual interpreting a test should be considerate of the needs of the test taker. As far as possible, test interpretation should occur within the context of an interpersonal situation or relationship. The test user should be prepared to be supportive of emotional reactions from the test taker, particularly when test information may be sensitive or in some way negative. Test interpretation should utilize language that is understandable to the test taker and should avoid too much technical terminology. Simple respect for the individual test taker and for his/her rights and welfare, in combination with knowledge and skill, will help to ensure responsible test interpretation.

Third, test results should, as far as possible, be communicated descriptively rather than as numbers and labels, and ideally should provide both a current understanding of the person and constructive suggestions for facilitating the person's development. For example, interpreting intelligence and/or scholastic aptitude test scores in terms of expected levels of educational performance and including suggestions for potentially facilitative educational interventions would generally be preferable to the use of a numerical IQ score or a label such as "retarded" or "genius".

The use of categories or labels based on test scores has been the subject of much justifiable criticism. Too often labels tend to stick with, and consequently , stigmatize, a person, even when they're no longer appropriate. For example, ridding oneself of the labels "retarded", "ex-convict", or "alcoholic" may prove extremely difficult. Lay persons often assume an erroneous "fixity", an unchanging-ness in the human personality. Although there is much about us that doesn't change, human beings are capable or much growth, change, and flexibility. Current behavioral description rather than labeling is more consistent with this great potential for growth and change.

Another problem with labeling is that it can influence our views of ourselves and others' views of us and thus how we actually he have or are seen to behave. The phenomenon of the "self-fulfilling prophecy" refers to the fact that our expectations of events or people can influence what actually happens. The label "mentally retarded" may cause both teachers and the child so labeled to have very low expectations of performance and potential. The child may receive less encouragement and attention from teachers do than "normal children". Worse, a child who learns to think of himself/herself as retarded may "give up" before ever testing his/her potential. Thus, whether or not a label like "retarded" or "delinquent" is initially justified, it may adversely influence the development of the individual.

Although labels have dangers, they can also be useful for some purposes. Labels often serve to efficiently summarize and communicated a great deal of information and should always be accompanied by some suggestions for treatment of positive intervention. The AERA/APA/NCME test standards require that if labels or categories are used, the categories chosen should be based on careful criterion, content, and/or construct validation, as

appropriate. Furthermore, the least stigmatizing label should always be assigned. Ultimately, the knowledge and good judgement of the test user are needed to control the use that will be made of psychological labels and diagnostic categories and to decide when the potential risks of assigning labels outweigh the benefits in terms of the facilitation of understanding.

A final point in test interpretation, and probably the most important one, is that the communication, interpretation, and use of test/assessment results should always occur within a context that includes and considers all relevant information about a person. Historical and demographic data and information from interviews and naturalistic observation serve to provide a framework from which to make any kind of test/assessment data richer, more meaningful, and more useful. It is inappropriate to interpret a scholastic aptitude or achievement test score without reference, among other things, to a person's cultural and socioeconomic background and educational opportunities and experiences. Similarly, a score one a vocational interest inventory must be interpreted in a context that includes the person's sex, racial and ethnic heritage, and other background characteristics.

10 Social Issues in Testing

a. Background

Although the testing profession has long had both ethical standards and standards regarding test quality, recent years have witnessed increasing public concern and controversy over particular kinds of test use. These areas of controversy have involved important social issues and have had far-reaching legal, legislative, and public policy implications. In essence, testing has "gone public", becoming, among other things, a consumer issue

and an object of much attack for its discrimination against minority groups.

Before describing recent controversies in testing it may be worthwhile to note that criticism of testing is not new. In a recent article of a special American Psychologist issue on testing, Haney (1981) reviews the history of social concern over standardized tests. Controversy over the use of intelligence tests dates back to 1922, when a series of debates over the use of tests such as the Army Alpha Test for Literates occurred. Discussion over the next 40 years was mainly centered around IQ and aptitude testing in the schools, although a controversy in the mid-1960s concerned the use of. Personality tests in the federal government's personnel selection procedures.

Jensen published an article entitled "How Much Can We Boost IQ and Scholastic Achievement"? In the Harvard Educational Review. In this article, Jensen asserted that genetic factors were the cause of observed differences between blacks and whites in performance on standardized intelligence tests. Research data had fairly consistently indicated that on standardized tests of intelligence the performance of whites as a group generally exceeded the performance of blacks as a group. To illustrate, the mean la of whites on a given intelligence test might be 105, in contrast to a mean IQ of 95 for a group of blacks. Jensen took this observed difference in intelligence test performance and attributed it primarily to heredity rather than to environment. In other words, it would be possible to interpret this scores difference as the result of some inherent differences between blacks and whites and/or it would be possible to attribute the difference to the fact that in American culture blacks are disadvantaged relative to whites in terms of their socioeconomic status, their educational and occupational opportunities, and so on. So although the existence of

the test score difference was not necessarily being debated, the reasons for this difference were an important focus of controversy.

Jensen's article led to widespread public outrage. Among other things, his assertions implied that blacks were genetically inferior to whites. In a society valuing equality of people and equal opportunity for all, these suggestions undermined the whole American value system.

Making matters even worse was an article entitled "I.Q." written by Richard Herrnstein and published in the Atlantic Monthly in 1971. Herrnstein argued that IQ differences across the social classes are at least in part inherited and that upward mobility is therefore less and less likely for future generations. In other words, Herrnstein was contesting the American ideal of "rags to riches" -the idea that bright people born into the lower classes can rise to the top through education and other opportunities -by suggesting that "brightness" would increasingly be a characteristic of the middle and upper classes, while the lower classes would inherit their parents' alleged low intellects.

The arguments of Jensen and Herrnstein were not only contradictory to many fundamental American ideals, but also had serious implications for the educational system. If differences in intelligence test performance across racial and socioeconomic groups are due to heredity, then there is little that can be done to improve the performance of blacks or lower socioeconomic status (SES) individuals. However, if differences are the result of disadvantaged environments, then programs such as compensatory education may be viewed as potentially helpful in allowing blacks or other disadvantaged students to develop the educational competencies (particularly verbal skills) necessary to perform well on standardized tests and in the American educational system.

Thus viewing differences as at least in part due to disadvantaged environments allows for the possibility of positive interventions and therefore, of hope for change.

In addition to many questions concerning the basic validity of Jensen's and Herrnstein's arguments, a number of responses were given. First of ail there is no question that blacks and whites in this country grow up in drastically different environments; black people have been of generally lower socioeconomic status than heave white people and have in general been culturally disadvantaged. Any group that suffers cultural disadvantage may be expected to perform less well on tests that are constructed within and reflect the majority culture or the middle class culture. Thus any discussion of race differences in la test performance must account for the fact that blacks and whites develop in drastically different socioeconomic, cultural, and educational environments. Second, it is always essential to emphasize the point that even if we observe group differences in intelligence, we cannot make any statements about the performance of a given individual. For example, let's assume that we know that John has an IQ of 150. We don't know if John is black or white, and in facto John could be black or white (or any other race for that matter). In contrast, let's assume we know that Bill has an lQ of 75. Again, Bill could be black or white. Genius and retardation are found in every population, and the group differences should not be used to make inferences about performance of any given individual.

Probably the most frequently offered argument, and that which has been the subject of most current controversy, is that intelligence and aptitude tests are by their very nature and content biased against blacks and other individuals not representative of the "majority culture". This argument suggests that because blacks and others represent a cultural, as well as numerical and

economic, minority, tests constructed by and for the cultural majority.

b) Cultural Bias in Testing

As has been mentioned, one major possible explanation for the generally poorer performance of blacks and other minority groups on standardized aptitude and intelligence tests is that tests discriminate against minority groups through cultural bias. This argument is based on the charge of cultural bias in both the nature and the content of intelligence and aptitude tests. It is assumed that tests developed by and for white, middle class stress white, and middle-class values and areas of knowledge, rather than the values and areas of knowledge within the black or other minority cultures. Cultural bias has been postulated to be manifested in tests themselves in one of three ways: content bias, bias in internal structure, and selection bias.

Content bias is probably the best known and most easily understood type of bias. Very simply, content bias refers to a test that contains content more familiar to white, middle-class examinees than to members of other racial or socioeconomic groups. Specifically, content bias may involve test questions that would be more familiar to one cultural group than to another. For, getting your bearing children, for example, questions about rural life and farm animals would be less familiar than would questions about urban life, such as traffic.

Another type of content bias is found in the item format and item presentation. For example, pictorial material that shows only white males and never females or black people is biased, as are test materials and pronounces always using "he". The use of sexist language and sexist job title (e.g., "policeman", "salesman") in

vocational interest inventories is a common example of sex bias in test content.

Test constructors have addressed the problem of content bias by obtaining expert judgements of the degree to which item content is culturally loaded and by soliciting item contributions from test professionals representing other cultural groups. The constructing of most scholastic aptitude and achievement tests is now done by panels of experts that include the representation of women, blacks, Hispanics, and so on and members of lower socioeconomic status groups These panel members both contribute items and evaluate the item pool with the objective of minimizing gender, race, cultural and class bias. Tests not originally constructed in this manner should be revised to remove content bias as far as possible, test users should include in their evaluations of tests the extent to which issues of gender and cultural diversity were incorporated into test development.

A second type of test bias postulated is that in the internal structure of the test. If the internal or factor structure of a test and/or the behavior of items in relationship to each other (i.e., internal consistency) are found to differ across cultural groups, then the test is measuring different things across groups. In other words, factor structure and item interrelationships contribute to the explication of the construct being measured by the test, and tests should not be used for the same purposes in different groups unless they are measuring the same construct as well.

Accordingly, test developers and publishers have done research investigating and comparing factor structure and item characteristics across groups. In general, available studies of factor structure have reported similarity across groups in factor structure and patterns. Studies of such characteristics of individual test

items as item difficulty, item total score correlations, and item characteristic curves have generally shown at least some group differences. However, consistent patterns of difference that would contribute to test revisions designed to eliminate bias have not yet been found. More research comparing tests' internal structural characteristics across groups is needed.

The final type of test bias postulated is a large category of bias which can be summarized as selection bias. Selection bias would be a bias caused when a test has different predictive validity across groups. For example, just because a scholastic aptitude test predicts the performance of white, middle-class students doesn't mean that it predicts the performance of blacks or lower-class individuals. It is erroneous to assume that what we know about predictive validity in one group also characterizes other groups, and if a test does not predict well for blacks or members of other groups then it shouldn't be used for that purpose in those groups.

Selection bias is examined through the comparison of regression equations and regression lines obtained within different groups. If the regression lines differ significantly across groups, then some type of selection bias is occurring. Differences in the regression lines can be either in the slope of the regression line ("slope bias") or in the interception of the line with the Y-axis ("intercept bias").

When both the predictor and the criterion scores are expressed in standard score units, the slope of the regression line ("line of best fit") is equal to the correlation (r) between predictor and criterion. If the value of the correlation in one group differs significantly from the value in another group (thereby resulting in different slopes), then the test is differentially valid in predicting the criterion. Critics of standardized testing have postulated that

tests that are significant predictors in white samples may not predict as well or at all in black samples.

c) Test in Employment Selection

The controversy surrounding the use of tests in employment selection may be summarized as the problem of adverse impact. Adverse impact means that even when standardized tests are not used with intent to discriminate against minorities and women, their use has led to the hiring of proportionately fewer minorities. Thus, whether or not discrimination was intended, the impact of using standardized tests in employment selection has been negative or adverse in terms of the representation of minorities and women.

The major concept currently underlying the justification for using tests or other selection methods that have adverse impact on certain groups is job-relatedness -that is, the biased tests or selection procedures are claimed to be related to job performance. Often the supposed job-relatedness of certain tests and procedures is highly questionable. If a test or selection requirement has little or nothing to do with job performance, its use is obviously unfair and its use is illegal. For example, imagine how unfair it would be if only tall people were admitted to colleges or accepted for jobs as computer programmers or management trainees. We would surely protest such a criterion based on its lack of relevance to doing well in college or in computer programming or management. A test's content should have some relationship to the requirements of the job.

A significant set of research findings regarding the validity of tests for employment selection is that from research on "validity generalization". A variety of research evidence -summarized, for example, by Schmidt and Hunter (1977, 1981) and Schmidt,

Hunter, Pearlman, and Shane (1979) suggests that cognitive ability test (including tests of verbal, quantitative, spatial and mechanical ability, perceptual speed, reasoning abilities) are valid predictors of performance both on the job and in training across a variety of jobs and job settings and are equally valid for majority group and minority group applicants.

The implications of research on "validity generalization" and other research on the use of tests in selection are that the use of most tests of cognitive abilities, no matter how valid, "job-related", and so forth, will continue to have adverse impact on blacks and other minorities in employment selection.

Originally, employee selection procedures were designed to allow the selection of people predicted to be most productive and successful on the job, with the intent of maximizing organizational productivity, efficiency, and profit.

The use of tests and other objective methods of employee selection was also originally thought to represent an improvement over selection based on family background and connections, race, religion, and other biased factors. In other words, the concept of a "meritocracy", which rewards the most intellectually talented and enterprising people, was seen as the best avenue to equal opportunity by giving people the chance to develop their individual talents, abilities and interests.

d) Educational Classification and Placement

Another area in which charges of race discrimination have been filed is that of the use of intelligence tests, such as the Wechsler Intelligence Scale for Children and the Standford-Binet intelligence Scale, to place children in tracks within a school. system. Two main issues have characterized the controversy: the overrepresentation of blacks in special education programs and the

cultural bias in IQ tests that is suggested to underlie this overrepresentation. In essence, these programs have tended to classify for more black than white children as mildly retarded (using the term "educable mental retardation", or EMR) and/or in need of special education. Critics view IQ tests as the cause of such differential classification.

In the 1980 California case of Larrry P. v. Riles the judge ruled that the tests were inherently culturally biased and, therefore prohibited the use of standardized tests for the EMR identification, classification, or placement of black children without court approval. The decision also made it almost impossible to restore use of the tests by removing the cultural biases thought to be present, because these biases were judged as inherent to the form and content.

Test administrators should also take special care in the administration of tests to minority children. It is essential to be aware of the possibility that, in contrast to the white, middle-class examinee, the minority child (or any culturally different person) may have had less previous experience in taking tests, may be less motivated to perform well in the situation, or may be alienated by test content emphasizing white, middle-class values. It is the responsibility of the test administrator to build rapport with the culturally different examinee and, as far as possible, to ensure that the level of familiarity with testing materials and of motivation' are as similar as possible to conditions that prevail when white, middle-class examinees are tested.

Finally, it is especially crucial in testing for educational classification and placement to ensure that the parents are informed concerning the meaning of test scores. Test users should view test scores as only one part of the data used to make

classification decisions and should therefore interpret scores in the context of another information about the child. And, finally, reassessment leading to the reconsideration of earlier decisions should be systematically incorporated into the program of classification and placement.

e) Scholastic Aptitude Testing

On of the bitterest educational controversies has involved criticisms of the use of scholastic aptitude tests. In a widely read and commented-upon 1980 report, Ralph Nader's research group charged that scholastic aptitude tests like the Scholastic Aptitude Tests (SAT) and the Law School Admission Test (LSAT) are not only biased against minority-group- and lower-income students but are not helpful, because they do not predict success in college, not to mention success in pursuits following completion of formal education. Scores, the group contended, reflect family income rather than scholastic potential. Furthermore, scores are not valid because they can be boosted significantly by coaching. The report was particularly critical of the Educational Testing Service (ETS), the publisher of many widely used scholastic aptitude test, and implied that ETS's only concern was the profits that could be gained through scholastic aptitude testing, not with either the educational utility of testing or the scientific basis or social implications of testing. As may be expected, the Nairn/Nader report elicited considerable reaction and many responses to the charges from the testing and assessment community.

In response to charges questioning the quality of the SAT and tests like it, psycho-morticians have stressed that the SAT is, psychometrically, one of the most impressive tests available. It is highly reliable and possesses a respectable degree of validity for predicting grades in the first year of college. Although validity

coefficients vary across academic majors, they range from the 0.10s to the 0.60s, and for liberal arts students average in the mid 0.40s for women and the 0.30s for men.

With respect to the alleged lack of predictive validity of these tests, the Nairn/Nader reports conclusions were based on frequent misinterpretation of validity data, for example, regression coefficients. In addition, the Nairn/Nader report failed to acknowledge the effects of restriction in range of aptitude test scores (because samples were constituted by individuals whose aptitude test scores were high enough to be selected) in reducing the maximum possible size of regression coefficients and of the less-than-perfect reliability and validity of the criterion itself, that is, grades during one's freshman year in college.

In its challenge of racial bias, the Nairn/Nader report defined a biased test as any test that results in a disproportionate selection of minorities -in other words, causes what has been referred to as "adverse impact". As has been mentioned, adverse impact is difficult to eradicate without throwing out test altogether, and it can be addressed by Simply including race as a variable in selection decisions. As long as there are group differences on standardized test of cognitive abilities, any selection procedure not explicitly taking race into account will lead to the disproportionate selection of minority applicants.

Charges that SAT scores reflect primarily family wealth were based on the inaccurate data. The correlation of SAT scores with family income has ranged between 0.23 and 0.29 (rather than 0.965, as reported by Nairn), and there is huge variability in family income at a given SAT score level. In actuality, family income is more closely related to first year grades than it is to test scores. Furthermore, high school grades, which the Nairn/Nader report

said should be used instead of test scores to make admission decisions, are almost as discriminatory against blacks as are test scores.

Other recommendations of the Nader/Nairn report restated guidelines already stated in the SAT manual. For example the Nader/Nairn report insisted that SAT scores should not be the only factor in selection, but the SAT manual has always recommended against such' practices. Any selection policy based solely on one test score would be inappropriate and irresponsible. Even as scholastic aptitude tests are excellent predictors of both grades and persistence in higher education, it is essential to remember that there are a number of other important variables that scholastic aptitude tests simply do not measure. Variables such as motivation, social skills common sense, perseverance, sense of humor, sensitivity and caring, and talent in a variety of areas are not measured by these tests. Certainly scholastic aptitude is helpful in terms of achieving goals of higher education, but there are many other personal and motivational qualities that are important to people's ability to be successful in and contribute to our society.

The Nader/Nairn criticism concerning the "coach ability" of scholastic aptitude test is based on the argument that if such scores as those on the SAT can be significantly increased through a few hours or weeks of "cramming" then they can't be very meaningful either conceptually or in the prediction of college performance. The argument is based on the following three assumptions: that commercial coaching programs significantly improve test performance; that the effects of "coaching", if any, cast doubt on the meaningfulness of the concept of scholastic aptitude; and that coaching reduces the predictive validity of test scores.

The effects of coaching cast doubt on the meaning of test scores only if coaching does not address the developed abilities of verbal and mathematical reasoning and comprehension. The aptitudes measured by the SAT and similar tests are not innate, but rather represent the impact of years of schooling on a person with some range of potential intellectual development. To the extent that coaching programs are long enough in duration and include emphasis on the basic cognitive skills in addition to or instead of "cramming" on specific test items, they should enrich the individual's developed abilities and enhance his/her capability of performing well in school. In other words, if coaching programs emphasize the important cognitive skills and abilities, they would be expected to improve both test scores and performance in college. In this way they resemble any effective educational intervention. However, compensating fully for years of inadequate schooling takes years of compensatory education and cannot be expected from the standard coaching programs available, no matter how high in quality.

Finally, evidence regarding the effects of coaching on test validity is spares and, again, the effects would depend on the nature of the coaching -that is on the position of the particular program of the "cramming" versus "education" continuum. Research that is available suggests that cramming programs lead to reductions in the predictive validity of test scores.

Thus in response to the Nadir/Nairn report it may be said that score increases due to coaching are of questionable practical significance and are certainly of less significance than implied in the report. However, when score increases do occur they can be viewed positively and as consistent with the concept of scholastic aptitude if the coaching program emphasized the acquisition of important cognitive skills. But the small effects of coaching in

general seem insufficient to warrant the attacks of the Nairn/Nader report.

f) Sex Bias in Testing

Findings of sex differences on basic dimensions of vocational interest are most evident and durable for Holland's Social and Realistic interests. Social interests are far more predominant among females, while Realistic interests are found far more frequently among males. Females' high scores on the Social and Conventional themes suggests traditionally female educational, social welfare, and office and clerical occupations. In contrast, females' lower scores on the Realistic, Investigative, and Enterprising themes result in infrequent suggestion of traditionally male professions, such as medicine, engineering, science, and of occupations in management and the skilled trades. Thus socialized patterns of interest led to interest inventory results that perpetuate females' overrepresentation in traditionally female occupations and their under representation in occupations traditionally dominated by males.

To some extent such aptitude and interest test score differences are due to different patterns of male and female socialization in our society; for example males have been encouraged to take math, science, and shop courses, and females have been encouraged to take English, home economics, and typing. But tests themselves have in many ways worsened the problem through bias in item content and wording. In aptitude tests, for example, use of predominantly male characters in word problems and reading passages, sexist language (e.g., the pronoun "he" is used to refer to the generic), and sex biased content have frequently been used. Examples of sex-biased content may be drawn from, for example the General Aptitude Test Battery (GATB).

The content of interest in inventory items is also frequently sex-biased.

In response to the criticisms of sex bias, many test developers have addressed these issues by combining the men's and women's forms, for example as was done with the Strong-Campbell Interest Inventory, (SCII). Other approaches have included eliminating sexist language, reducing the bias in item content, and discussing issues of sex-role socialization in interpretive materials. Other test developers have focused on reducing the sex-restrictive ness of the resulting scores, primarily though the provision of same-sex normative scores.

Same-sex normative scores compare the person's score to a norm group consisting only of persons of his/her own sex. Thus the comparison group consists of people sharing the same sex-role socialization experiences as the examinee. A girl with the potential for mechanical or scientific interest is much more likely to be alerted to those interests if her scores are shown in comparison to those of other girls, rather than to those of "people in general", which includes males. Similarly, a boy's social interests are more likely to be highlighted if he is compared only to other boys.

The use of same-sex normative scores and sex-balanced interest inventories is intended to increase the probability that females who could potentially be interested in Realistic, Investigative, or Enterprising occupations or males potentially interested in Social, Artistic, or Conventional occupation will obtain interest inventory profiles suggesting those areas. Thus such methods of constructing and scoring interests inventories are designed to facilitate both males and females' exploration of the full range of occupational alternatives and to minimize the extent to

which people continue to be directed toward stereotypic occupational areas.

While attempts to remove sex restrictive ness from interest inventories are important and useful, the more direct solution to the problem of sex-stereotypic score patterns involves increasing the range of experiences available to members of both sexes, rather than to just one sex or the other. Until girls and women have the opportunity to engage in activities relevant to, for example, Realistic and Investigative, as well as Social and Artistic, abilities and interests, competencies and interests in nontraditional areas will not develop in the majority of women. Likewise boys and men should be encouraged to explore their social, artistic, and domains. Assessing the abilities greater educational and occupational opportunity when based on a rich background of experience rather than a restricted, narrow range of such experiences.

Activities

1. Describe Trend.

_____

_____

_____

_____

_____

_____

2. Explain about National Assessment of Educational Progress.
3. Enlist some apparent trends.
4. Describe Cultural Bias in Testing
5. What do you understand by Educational Classification and placement?

6. CONCERNS AND ISSUES

Although educational testing has always had its friendly critics, in recent years there has been increasing concern about the

role of testing in the schools, especially the use of standardized test. Critical issues concern the nature and quality of the tests to minorities, and the potential hazards of testing to the individual's right to privacy.

Nature and Quality of Tests. Some of the strongest criticism of standardized testing occurred in the early 1960s and was directed primarily at the use of objective-test items. Critics such as Hoffman contended that the multiple choice items, which is the main item type used in the standardized tests, penalized the more intelligent original thinkers. He supported his claims by reviewing items from standardized tests and showing how the more brilliant and creative students were likely to see implications in the items that would question the correctness of the keyed answers. Although Hoffman obviously was able to discover some defective items that appeared in standardized tests, his criticism seem to go well beyond the evidence presented: On the positive side, he probably encouraged test publishers to supplement statistical Item analysis with a more careful logical analysis of test items.

A closely related criticism, that test measure only limited aspects of an individual, has also received considerable attention. Tests do measure specific and limited samples of behaviour. Aptitude test typically measure samples of verbal and quantitative skills useful in predicting school success, and achievement test measure samples of pupil performance on particular learning tasks useful in assessing educational progress. Both fulfill their limited functions quite well, but the difficulty arises when we expect more of them than was intended. For example, both the advocates and critics of college admission testing sometimes assume that the tests measure all that is needed for success in college and beyond. This tendency to read into test scores more than they really tell has been called the whole person fallacy by W. w. Trumbull, the

former president of the Educational Testing Service. A quotation from one of its publications makes clear the limited nature of these tests.

> Ability and academic achievement occupy on Olympian perch on the prestige ladder. Yet it is widely agreed that motivation, creativity, Personal honesty, intuition, even the degree of social consciousness play significant roles in the struggle for the most cherished of American ideals-"success in life". Admission tests thus measure a relatively narrow segment of the human potential.

Much of the misinterpretation and misuse of test scores would be avoided if the limited information that tests provide was more widely recognized. In college admission decisions, as well as in all other educational decisions, test scores provided just one type of information and should always be supplemented by past records of achievement and other types of assessment data. No major educational decision should ever be based on test scores alone.

The reliability and validity of tests have also been subject to criticism. This has usually been by uninformed critics who mistakenly believe that tests should be completely reliable and provide perfect predictions. These, of course are unrealistic expectations. All measurement is subject to error, and predictions in all areas are fallible. Rather than compare the tests' reliability and validity with nonexistent ideal standards, they should be compared with the alternatives. Would our judgements of aptitude and achievement be more reliable without test results? Would our predictions of future school or occupational success be more valid without the additional information supplied by test? I doubt it! Qualified users of tests take into account the possible error in test

score during test interpretation and use, and they combine test scores with other relevant information when making educational decisions. To argue that better decisions are made when less information is available. Test scores are certainly fallible, but probably less so than most of the other types of information that enter into educational decisions.

Effects of Testing on Pupils. Critics of testing have charged that testing is likely to have a number of undesirable effects on pupils. Some of the most commonly mentioned charges directed towards the use of aptitude and achievement tests are listed here with brief comments.

Criticism 1: Test create anxiety. There is no doubt that anxiety increases during testing. For most pupils, it motivates them to perform better. For a relatively few, test anxiety may be so great that it interferes with test performance. These typically are pupils who are generally anxious, and the test simply adds to their already high level of anxiety. A number of steps can be taken to reduce test anxiety, such as thoroughly preparing for the test, taking practice exercises and use liberal time limits. Fortunately, many test publishers in recent years have provided practice tests and shifted from speed tests to power tests. This should help, but it is still necessary to observe pupils carefully during testing and to discount the scores of overly anxious pupils. There is little likelihood, however, that test anxiety has any lasting influence on a pupil's mental hearth.

Criticism 2: Tests categories and label pupils. Categorizing and labeling individuals is a serious problem in education, just as it is in our general society. It is all too easy to place individuals in pigeonholes and apply labels that then determine, at least in part, how they are viewed and treated.

Classifying pupils in terms of levels of mental ability has probably caused the greatest concern in education. Critics contend that when pupils are classified as "mentally retarded," for example, it influences how teachers and peers view them, how they view themselves, and the kind of the school programme they are provided. When pupils are mislabeled as mentally retarded, as has been the case with some racial and cultural minorities, the problem is compounded. At least some of the support for mainstreaming handicapped pupils has come from the desire to avoid the categorizing and labeling that accompanies special education classes.

Classifying pupils into various types of learning groups can more efficiently use the teacher's time and the school's resources. However, any grouping system needs to take into account that tests measures only a limited sample of a pupil's abilities and that pupils are continuously changing and developing. Keeping the groupings tentative and flexible and regrouping for different subjects can avoid most of the undesirable features of grouping. It is when the categories are viewed as rigid and permanent that labeling becomes a serious problem. In such cases, it is not the test that should be blamed, but the use of the test.

Criticism 3: Tests damage pupils' self-concepts. This is one of the most serious charges against testing and requires the attention of the teachers, counselors, and other users of the tests. The improper use of tests may indeed contribute to distorted self-concepts. The stereotyping of pupils, mentioned in the previous section, is one misuse of tests that is likely to have an undesirable influence on a pupil's self-concept. Another is the inadequate interpretation of test scores that may cause pupils to over generalize from the results. It is certainly discouraging to receive low scores on tests, and it is easy to see how pupil might develop a

general sense of failure unless the results are properly interpreted. Low scoring pupils need to be made aware that aptitude and achievement tests are limited measures and that the results can change. In addition, the possibility of over generalizing from low test scores will be lessened if the pupil's positive accomplishments and characteristics are mentioned during the interpretation. When properly interpreted and used, tests can help pupils to develop a realistic understanding of their strengths and weaknesses and, thereby, contribute to improved learning and a positive self-image.

Criticism 4: Tests create self-fulfilling prophecies. This criticism has been directed primarily toward intelligence or scholastic aptitude tests. The argument is that test scores create teacher expectations concerning the achievement of individual pupils; the teacher then teaches in accordance with those expectations, and the pupils respond by achieving to their expected level -a self-fulfilling prophecy, Thus, those who are expected to achieve more do achieve more, and those who are expected to achieve less do achieve less. This so-called Pygmalion effect received strong support from a widely heralded study by Rosenthal and Jacobsen, even though the study was later challenged by other researchers. The belief that teacher expectations enhance or hinder a pupil's achievement is widely held, and the role of the testing in creating these expectations is certainly worthy of further research.

In Summary, there is some merit in the various criticisms concerning the possible undesirable effects of tests on pupils. But these criticisms should be directed at the users of the test rather than the tests themselves. The same persons who misuse the test results are likely to misuse alternative types of information that are even less accurate and objective. Thus, the solution is not to stop using tests, but to start using tests and other data sources more effectively. When tests are used in a positive manner, that is, to

help pupils to improve their learning and development, the consequences are likely to be desirable rather than undesirable.

Fairness of Tests to Minorities. The issue of test fairness to racial and cultural minorities has received increasing attention over the years. Concern with the fairness of the tests has paralleled the general public concern with providing equal rights and opportunities to all United States citizens. Critics have charged that tests are biased and discriminatory and impede educational and occupational opportunities for minorities.

The charge of test bias or unfairness can be examined from two view points:

(1) the possible presence of bias in test content and

(2) the possibly unfair use of test results.

These factors are undoubtedly related, but we shall discuss them separately. Much of the concern with bias in test content focuses on the fact that some minorities frequently earn lower test scores than do their more advantaged peers. As Gardner pointed out, however, low test scores do not necessarily indicate test bias:

> Lower scores alone on an achievement test do not signify bias. If they did, then every spelling test would be biased against poor spellers, every vocabulary test against person who had poor vocabularies, and every short hand test against persons who had never learned shorthand.

Thus, in evaluating the possible presence of bias in test content, it is important to distinguish between the performance the test is intended to measure and factors that may distort the scores unfairly. In testing arithmetic skills with story problems, for example, it is important to keep the reading level low so that the test scores are not contaminated by reading ability, if the reading

is too difficult, poor readers will obtain lower scores than warranted, and thus, the test will be biased against them. Because a particular minority group may have a disproportionately large number of poor readers, the test may be more biased for that minority group than for other pupils. But if the test of arithmetic skill is not contaminated by reading or other factors, low scores will simply indicate lack of arithmetic skills. Such a test is fair for everyone even if the test scores indicate cultural differences in the mastery of arithmetic.

In the past, standardized tests typically emphasized test content and values that are more familiar to white middle-class pupils than to racial or cultural minorities and pupils of lower socioeconomic status. Thus, the content of some ties had less opportunity to learn in their culture. Similarly, some reading tests contained stories and situations that were unrelated to their life experience. In addition, racial and cultural minorities were seldom represented in pictures, stories and other test content. And they were, it was sometimes in an offensive manner. How much these types of bias might have lowered the scores individual pupils is impossible to say, but most persons familiar with testing would acknowledged some adverse effect. Fortunately, efforts are now being made to correct the situation. Test publishers now employ staff members representing various racial and cultural minorities, and new tests being developed are routinely reviewed for content that might be biased or offensive to minority groups, statistical analysis is also being used to detect and remove biased test items.

The most controversial problems concerning the fair use of tests with minority groups' are encountered when aptitude tests are used as a basis for educational and vocational selection. Much of the difficulty here is with the definition of fair test use. One view is that a test is fair or unbiased if it predicts as accurately for

minority groups as it does for the majority group. This traditional view, which favors a common cutoff score for selection, has been challenged as being unfair to minority groups because they often earn lower test scores, and thus, a smaller proportion of qualified individuals tends to be selected. Alternative definitions of test fairness favor some type of adjustment such as separate cutoff scores or bonus points for some minorities.

Although the fair use of tests in selecting students and employees is widely debated in the professional literature, whether minority group membership is to be ignored or given special consideration in selection, the decisions will not be determined by educators or psychologists. The fair use of tests in selection is part of a larger issue that must be settled by society through court rulings. Stated in simplified form, the issue is how equal educational and occupational opportunities best be provided for members of minority groups without infringing on the rights of other individuals.

7. SELF ASSESSMENT QUESTIONS

1. List reasons for and against the use of a statewide minimum-competency testing programme.

2. What is Mainstreaming? Why does it require knowledge of testing and evaluation?

3. List as many factors as you can think of that might account for the decline in college admissions test scores.

4. What types of coaching are most likely to improve performance on college admissions tests? Should schools provide this coaching for college applicants?

5. Describe the purpose and function of the National Assessment of Educational Progress. In what ways might this programme influence school instruction?

6. What types of testing do you think should be increased or decreased in the schools? Why?

7. List possible advantages and disadvantages for each of the apparent trends in educational measurement.

8. Which criticisms of testing do you consider to be most serious? What steps should be taken to correct them.

9. What are the advantages of the legal requirement that parents and pupils be given access to school records? What problems might arise from this, and what solutions would you suggest?

8. BIBLOGRAPHY

1. Gronlund, N.E. (1985): Measurement and Evaluation in Teaching, 5th ed. Macmillan Publishing Company. New York.

2. Ebel. R.L, (1979), Measuring Educational Achievements, 3rd . ed, Prentice-Hall, Englewood Cliffs, New Jersey.

3. Thorndike, R.L, (1969); Perspectives in Educational and Psychological Measurement, Prentice-Hall, Englewood Cliffs, N.J.

4. Karmel, L.J. and Karmel, M.O. (1978); Measurement and Evaluation in the Schools, 2nd Edition, Macmillan Publishing Co, Inc. New York

5. Walsh, W.B. and Betz, N.E. (1985); Tests and Assessment, Prentice-Hall. Inc, Englewood Cliffs, New Jersey.

Unit-9

STATISTICS USED IN EVALUATION AND MEASUREMENT

BY:
Muhammad Arshad Javed

INTRODUCTION

We are living in the information age. The information about which we are concerned is in the form of numbers called data. This unit is about data-how it is collected, analyzed and interpreted. Statistical techniques are essential part of professional training. Statistical logic, thinking and operations are necessary for either profession. Educationists will depend upon statistical background in their administration and in the interpretation of results. Statistics is everywhere basic to research activities. Research work is not possible without statistical tools.

OBJECTIVES

After going through this unit the students are expected to be able to:

1. Learn basic and important concepts of Statistics.
2. Apply these concepts in practical life.
3. Interpret test scores
4. Compute measures of central tendency and apply them in educational measurements.
5. Compute and apply different measures of variation in educational field.
6. Compute and interpret percentiles and percentile ranks.
7. Understand basic concepts of probability.
8. Apply Normal distribution. ~

Contents

Introduction

Objectives

1. Interpret Measures

1.1 Interpretation by Criterion Referencing

1.2 Interpretation by Percentages

1.3 Interpretation by Norm Referencing

1.4 Ordering and Ranking

1.5 Self Assessment Questions

1.6 Exercises,

2. Frequency Distribution

2.1 General rules for the Construction of a Frequency Distribution

2.2 Frequency

2.3 Mid Point or Class Mark

2.4 Self Assessment Questions-2

2.5 Exercise

3. Measurs of Central Tendency

3.1 The Arithmetic Mean or Mean

3.2 The Median

3.3 The Mode

3.4 Empirical Relationship between Mean, Median and Mode

3.5 Comparison of Measures of Central Tendency

3.5.1 Quartiles

3.5.2 Percentiles

3.5.3 Percentile Ranks (PR)

3.6 Self Assessment Questions-3

3.7 Exercise

4. Relationsips with a distribution

4.1 Computing the Coefficient of Correlation

4.2 The Product-Moment Method

4.3 Self Assessment Questions-4

4.4 Exercise

5. Measures of variability

5.1 The Range

5.2 The Mean Deviation or Average Deviation

5.3 The Standard Deviation

5.4 Calculation of the Standard Deviation for the Grouped data

5.5 The Coefficient of Variation

5.6 Chebyhev's Rule

5.7 Other Characteristics of Distributions.

5.8 Standard Scores

5.9 Self Assessment Questions-5

5.10 Exercises

6- The Normal Curve

6.1 Relative frequency Interpretation of Probability

6.2 Some Basic Properties of the Normal Curve

6.3 Self Assessment Questions-6

6.4 Exercises

7- Bibliography

1. INTERPRETATION MEASURES

1.1 Interpretation by criterion referencing

The raw score is number of points received on a test when the test has been scored according to the instructions. Raw score is not very meaningful without further information. Criterion-referenced test interpretation permits us to describe an individual's test performance without referring to the performance of other individuals. Thus we might describe a student's performance in terms of the speed, precision with which a certain task is performed. Criterion- referenced interpretation of test scores is most meaningful when the test is designed to measure a set of clearly stated learning tasks. Enough items are used for each interpretation to make dependable judgments.

1.2 Interpretation by Percentages.

In mathematics, a relationship with 100 is called percentage (denoted by %). Often it is useful to express the scores in terms of percentages for comparison. Consider the following example.

Grade	Class A		Class B	
	No of Student's	%	No. of students	%
A	10	12.50	8	40
B	25	31.25	6	30
C	30	37.50	4	20
D	15	18.75	2	10
Total:	80	100	20	100

Ten students from class A and eight students from class B got grade A. It looks apparently that class A is better in getting A grade but 12.5% of the students from class A and 40% students from class B got grade A. It is clear from the percentages that class B is far better in getting grade A than class A.

1.3 Interpreta[illegible] Norm referencing.

Interpretati[illegible]scores by norm referencing involves ranking of scores and ex[illegible]ing a given score in relation, to the other scores. Norm-re[illegible]ed test interpretation tells us how an individual is co[illegible] with other persons who have taken the same test. The s[illegible]t type of comparison is to rank the scores from highest to l[illegible]nd to note where an individual's score falls. The rest of the [illegible]erve as the norm group. The given score is compared with t[illegible]r scores by norm referencing. If a student's score is second f[illegible]e top in a group of 20 students, it is a high score meaning th[illegible] scores of 90% of the students are less than him.

1.4 Ordering a[illegible]nking.

A first ste[illegible]ganizing scores is the listing of scores in order of magnitu[illegible] largest to the smallest score. The data so arranged are ca[illegible]red array. By scanning an ordered array, we can determin[illegible]y the largest score, the smallest score and other facts abou[illegible]ta.

Ranked d[illegible]sist of scores in a form that shows their relative position[illegible]me characteristic but does not yield a numerical value[illegible] characteristic. The order of finish of cars in a race is an[illegible]e of ranking. If we list the car as first, second, third e[illegible]o the last car, we can say that they were ranked on the c[illegible]ristic of overall speed. We know each car's position relative[illegible] other car's position but we have no precise knowledge of th[illegible] of any car. A high school teacher ranked Akram 30th in a[illegible]f 100 means that Akram did better than 70 of his classmat[illegible]oorer than 29. But nothing has been said about Akram's g[illegible]evel of achievement.

Measurement Scalesiu

Measurement scales are of great s ce in analyzing and interpreting results. The important f measurement scales are:

The Nominal Scale

The lowest measurement scale is the al scale. In this scale, each individual is put into one of the ct categories or classes. Each class has a name. The names ust labels. There is no order in these classes. We cannot say e class is larger than the other class. You cannot do etic operations (addition, subtraction, multiplication, divisi his scale.

Examples of the nominal scale are

Categorization of blood groups of th nts of a college into A, B, AB and O groups. We cannot say roup A is better than group B. Classification of books in a co ibrary according to subjects.

Distribution of the population of Pak ccording to sex, religion, occupations, marital status, literac are examples of the nominal scale.

The Ordinal Scale

When measurements are not only di rom category to category but can also be ranked according e criterion, they are to be measured on an ordinal scale. Th mbers of anyone category are considered equal but member one category are considered lower than those in an other cat The ordinal scale is one-step higher than the nominal scale se we distribute the individuals not only in classes but we als r these classes.

Examples of the ordinal scale are Categorization of schools according to their educational level into primary, middle, secondary or higher secondary is an ordinal scale. There is an order in these classes. The primary level is lower than the middle level and the middle level is lower than the secondary level. You cannot do arithmetic operations on this scale.

Individuals may be classified according to socioeconomic status as low, medium, high. Intelligence of students may be average, above average or below average.

Classification of examination results into different grades A_1, A, B, C, D, E etc.

In this measurement scale, we can say that one individual is larger than the other but we cannot say how large it is.

The Interval Scale

In this scale, it is not only possible to order measurement but also the distance between two measurements is known. We can say that the difference between two measurements 30 and 40 is equal to the difference between measurements 40 and 50. The level of the interval scale is higher than the nominal and the ordinal scales. This is truly a quantitative scale. A unit of measurement and a zero point are required for this scale. The selected zero point is not necessarily a true zero. It does not have to indicate a total absence of the quantity being measured. We measure height in meters or feet, weight in kilograms or pounds, temperature in centigrade or Fahrenheit, income in rupees and the time in seconds. Arithmetic operations can be done on this scale. You can add the income of a wife to that of his husband.

The Ratio Scale

The highest level of measurement is the ratio scale. Equality of ratios as well as equality of intervals is determined in this scale. Fundamental to the ratio scale is the true zero point. The measurement of height, weight and length makes use of the ratio scale.

1.5 Self Assessment Questions 1

Read the following statements carefully and mark whether each statement is true or false.

(i). Division of population of Pakistan with respect to religion is a nominal scale.

(ii) Division of library books according to subjects is an ordinal scale

(iii) The highest level of measurement is the ratio scale.

(iv) The level of the interval scale is higher than the nominal and the ordinal scales.

(v) The ordinal scale is one-stepping lower than the nominal scale.

Fill in the blanks

(i) Categorization of blood groups of the students of a college into A, B, AB and O groups is an example of _____scale.

(ii) Classification of examination results into different grades A_1, A, B, C, D, and E is an example of _____ scale.

(iii) The lowest measurement scale is the _____scale.

(iv) There is no order in the classes of_____ scale.

(v) _____is truly a quantitative scale.

Answers

True/False

(i) T (ii) F it is a nominal scale. (iii)T (IV) T (v) F. It is one step higher.

Fill in the blanks.

(i) Nominal (ii) Ordinal (iii) Nominal (iv) Nominal (v) Interval scale.

1.6 Exercises

1. Give at least two examples of nominal, ordinal and interval scales not mentioned in the book.
2. Specify whether the following measurement scale is nominal, ordinal or interval.

(a) Distance traveled between Faisalabad and Islamabad.
(b) Sex of a newly born baby.
(c) Weights of students.
(d) Number of shares offered for sale in a Stock Exchange on February 20,2002.
(e) Marital status of a school teacher.
(f) Level of an educational institution.
(g) Occupation of an individual.
(h) Number of pages in a book.
(i) Winter crops of Pakistan.
(j) Rating of a teacher by students.

Answers

2. (a) interval (b) nominal (c) interval (d) interval (e) nominal (f) ordinal (g) nominal (h) interval (i) nominal j) ordinal

2 FREQUENCY DISTRIBUTION

Data that have been originally collected is called raw data or primary data. It has not yet undergone any statistical technique. To understand the raw data easily, we arrange it into groups or classes. The data so arranged is called grouped data or frequency distribution.

2.1 General rules for the construction of a frequency distribution

1. Determine the Range. Range is the difference between highest and lowest scores.
2. Decide the appropriate number of class intervals: There is no hard and fast formula for deciding the number of class intervals. The number of class intervals is usually taken between 5 and 20 depending on the length of the data.
3. Determine the approximate length of the class interval by dividing the range with number of class intervals.
4. Determine the limits of the class intervals taking the smallest scores at the bottom of the column to the largest scores at the top.
5. Determine the number of scores falling in each class interval. This is done by using a tally or score sheet.

Example 1.

The marks obtained by 120 students of first year class in the subject of Education are given below. Construct a frequency distribution.

57	86	69	62	75	73	80	78	87	83
77	35	70	68	84	73	81	78	61	72
59	98	95	63	76	73	88	60	52	83
86	45	70	53	85	74	62	78	89	84
60	79	91	64	84	85	81	79	90	78
83	50	71	65	76	58	71	79	51	61
61	89	81	74	76	74	82	91	71	76
80	52	71	66	77	65	44	79	95	74
79	63	83	87	77	75	83	48	70	85
61	70	72	67	61	83	75	79	97	75
66	54	81	68	78	75	83	61	33	76
62	55	72	76	78	75	99	80	83	86

The following steps are followed to make a frequency distribution.

Step 1: Range= maximum score -minimum score=99-33=66

Step 2: Number of approximate class intervals to be taken is 7.

Step 3: Length of the class intervals, usually denoted by i, is

$$i = \frac{\text{Range}}{\text{Number of class intervals}} = \frac{66}{7} = 9.4$$

The length is usually rounded upward to whole number. Therefore 9.4 is taken as 10.

Step4: Determine the limits of the class intervals

90- 99
80- 89
70- 79
60- 69
50- 59
40- 49
30- 39

The lowest class interval is taken in which the minimum score can be included. The minimum score is 33. the lowest class interval can be started from 30, but it is convenient to start the lowest class interval from the score to which addition of the length of the class intervals is easy. So we start from 30. This is called lower limit of the class interval. Add 9 (i-1=10-1=9) to the lower limit to get the upper limit of the first class interval. Now add consecutively i=10 to the lower limits and upper limits to get the remaining class intervals.

Step 5: Distribute the scores in the class intervals by putting a tally mark in the relevant class interval and count the number of scores in each class interval.

Grade	Tallies	No.of students
90- 99	𝍸 |	8
8 0 -8 9	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	3 0
70- 79	𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸 𝍸	4 5
60- 69	𝍸 𝍸 𝍸 𝍸 | |	22
50- 59	𝍸 𝍸	10
40- 49	|||	3
30- 39	||	2

2.2 Frequency

The number of scores lying in a class interval is called the frequency of that class interval. For Example two scores lie in the class interval 30-39. Therefore 2 is the frequency of the class interval 30-39.

2.3 Mid point or class mark

The middle of a class interval is called mid point or class mark and is usually denoted by X. It is calculated as

$$\text{midpoint} = \frac{\text{lower limit} + \text{upper limit}}{2}$$

For Example, the mid point of the class interval 30-39 is

$$X = \frac{30+39}{2} = \frac{69}{2} = 34.5$$

2.4 Self Assessment Questions 2

Read the following statements carefully and mark whether each statement is true or false.

(i) To understand the raw data easily, we arrange it into groups or classes.

(ii) The class-mark is the lower limit of a class.

(iii) There is no hard and fast formula for deciding the number of class intervals.

(iv) Data that have been originally collected is called raw data or primary data.

(v) The primary data is the data on which a statistical technique has been applied.

Fill in the blanks

(i) The number of scores lying in a class is called _____.

(ii) The difference between largest and smallest scores is called_____.

(iii) The lower-Limit of the class interval 10-14 is _____.

(iv) The class-mark of the interval 20-24 is _____

(v) The data arranged in groups is called _____.

Answers

True/False

(i) T (ii) F It is middle of a class. (iii) T (iv) T (v) F It has not undergone any statistical technique.

Fill in the blanks: (i) Frequency. (ii) Range. (iii) 10 (iv) 22 (v) Frequency distribution

Activities

1. Measure the heights of the students of a class in kilograms and make a frequency distribution of the weights taking 5 as width of the classes.
2. Take a class test of the students of class ten and make a frequency distribution of the marks obtained by them taking appropriate class-width.

2.5 Exercise

The following table gives the marks received by 50 students in an examination. Make a frequency distribution taking suitable class interval.

121	166	75	65	149	49	81	114	122	159	153	136
119	147	131	168	69	57	81	109	118	143	94	158
191	79	76	150	89	92	156	179	73	93	139	85
105	138	109	L03	165	95	142	81	103	88	89	106
82	85										

3. MEASURES OF CENTRAL TENDENCY

A single score calculated to represent all the scores is Called an average. Average tend to lie in the centre of an array. That is why averages are called measures of central tendency. Since ! aver- μ ages locate the centre of a data set, these are also called measures of location.

Several types of averages can be defined. Most commonly used averages are arithmetic mean, median and mode.

3.1 The Arithmetic Mean or Mean

The arithmetic mean is the most commonly used average. It is usually called mean or average. The arithmetic mean is defined as the number obtained by dividing the sum of the scores by their number. It is denoted by putting bar on the variable symbol e.g X (read as X bar).

In symbols $\overline{X} = \frac{\sum X}{N}$

$\sum$ read as sigma is the Greek symbol means sum of

$\sum X$ mean sum of the values of variable X.

N is the number of scores or measurements.

Example 2.

Following are the marks obtained by 9 students in the subject of Education. 55,68,78,45,80, 33, 67, 57, 48. Find their arithmetic mean.

Solution

Step 1: Find the sum of the marks obtained i.e.

$\sum X = 55+68+78+45+80+33+67+57+48 = 531$

Step 2: Count the number of marks i.e. N=9.

Step 3: Divide $\sum X$ by N to obtain the arithmetic mean i.e.

$$\bar{X} = \frac{\sum X}{N} = \frac{531}{9} = 59$$

Calculating arithmetic mean from the grouped data

Example 3.

The following frequency distribution shows the grades on an examination in a college. Find the mean grade.

Grade	No. of students
90-99	8
80- 89	30
70- 79	45
60- 69	22
50- 59	10
40- 49	3
30- 39	2

Solution

1 Grade	2 X	3 f	4 fx
90-99	94.5	8	756.0
80-89	84.5	30	2535.0
70-79	74.5	45	3352.5
60-69	64.5	22	1419.0
50-59	54.5	10	545.0
40-49	44.5	3	133.5
30-39	34.5	2	69.0

Step 1: Find the midpoints = $\frac{lower\ limit + uper\ limit}{2}$ (column 2)

Step 2: Find the sum of the frequencies i.e. $\sum f=120$

Step 3: Multiply midpoints Xs (column 2) by the corresponding frequencies (column 3) to obtain fX (column 4) .Find the sum of this column i.e. $\sum fx=8810$

Step 4: Divide $\sum fx$ by $\sum f$ to obtain $\overline{X}$

$$\overline{X} = \frac{\sum fX}{\sum f} = \frac{8810}{120} = 73.42$$

Short formula for calculating arithmetic mean for common class interval width is

$$\overline{X} = A + \frac{\sum fX}{\sum f} \times C$$

A is assumed or guessed mean usually chosen from the values of**X** and c is the common class interval length.

The following steps are involved in its calculation

Grade	X	f	X'	fX'
90-99	94. 5	8	2	16
80-89	84.5	30	1	30
70-79	74.5→A	45	0	0
60-69	64.5	22	-1	-22
50-59	54.5	10	-2	-20
40-49	44.5	3	-3	-9
30-39	34.5	2	-4	-8
Total		120		-13

Step 1: Find the midpoints commonly denoted by x.

Step 2: Find the sum of the frequencies i.e. $\sum f=120$

Step 3: Select an assumed mean (A) from the column X (column 2) usually in the centre of the column against maximum frequency.

Step 4: Subtract A from each X-value and divide it by common class interval width(c) and denote it with X' (column 4).

Step 5: Multiply X' (column 4) with f(column 3) to obtain fX (column 5). Sum this column and obtain $\sum fX'$.

Step 6: Put the values in the formula and solve to obtain the arithmetic mean that

$$is\ \overline{X} = 74.5 + \frac{-13}{120} \times 10 = 74.5 - 1.08 = 73.42$$

3.2 The Median

The median of a set of scores is the middle score or the arithmetic mean of two middle most scores in an array. 50% of the scores are less than median and 50% of the scores are greater than median.

In symbols

$$\text{Median} = \left(\frac{N+1}{2}\right) \text{the score}$$

Example 4.

Find the median of the following scores

(i)5,4,8,3,7,2,10

(ii) 17.4, 18.7, 21.3, 19.5, 23.6, 20.3, 18.9, 20.0

Solution

(i) Array: 2,3,4,5,7,8,10

The number of scores is 7 which is odd, so the middle score in the array is median i.e.

median = 5.

Alternatively, by using the formula Median = $\left(\frac{N+1}{2}\right)$ the score

$$\text{Median} = \left(\frac{7+1}{2}\right) \text{ the score} = 4^{th} \text{ score} = 5$$

(ii) Array: 17.4,18.7,18.9,19.5, 20.0,20.3,21.3, 23.6.

The number of scores is 8 which is even, so the average of two middle most scores is median in the array i.e.

$$\text{Median} = \left(\frac{19.5+20.0}{2}\right) 19.75$$

Alternatively, by using the formula $\text{Median} = \left(\frac{N+1}{2}\right)$ in score

$$\text{Median} = \left(\frac{8+1}{2}\right) \text{ th score} = 4.5^{th} \text{ score}$$

$= 4^{th}$ score + 0.5 (5^{th} score-4^{th} score)

= 19.5 + 0.5 (20.0-19.5) = 19.75

Calculation of median from the grouped data

When the data is grouped into a frequency distribution,

$$\text{Median} = L + \frac{i}{f}\left(\frac{N}{2} - C\right)$$

L=lower class boundary of the median class interval.

i=length of the median class interval.

f=the frequency of the median class interval.

$N=\sum f$

C=the cumulative frequency of the class interval below the median class interval.

Example 5.

Calculate the median for the data given in Example 3.

Solution

Grade	f	C
90- 99	8	8+ 112 =120
80- 89	30	30+82 =112
70- 79	45	45+37 = 82
60- 69	22	22+ 15 = 37
50-59	10	10+5 = 15
40- 49	3	3+2 = 5
30- 39	2	= 2
Total	120	

Median = $\frac{N}{2}$ th score= $\frac{120}{2}$ = 60th score which lies in the class interval 69.5-79.5 called median class interval.

The median can be interpolated by the following formula

Median =L + $\frac{i}{f}\left(\frac{N}{2}-C\right)$ =69.5+$\frac{10}{45}$ (60-37) + 74.61

3.3 The Mode

The mode is the score which occurs greatest number of times in a data set. Mode does not always exist. If each score occurs the same number of times, there is no mode. There may be more than one mode. If two or more scores occur greatest number of times, then there are more than one **mode**.

Example 6.

Find mode for the following set of scores

(i) 10, 12, 15, 19, 20, 26, 30.

(ii) 10, 12, 15, 19, 19, 19, 19, 20, 26, 30

(iii) 10, 12, 15, 19, 19, 19, 20, 26, 26, 26, 29, 30

Solution

(i) There is no mode.

(ii) Mode=19

(iii) There are two modes 19 and 26.

For grouped data, the mode can be calculated by the formula

$$\text{Mode} = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i$$

The mode lies in the class interval having maximum frequency. This class interval is called the modal class.

L=lower class boundary of the modal class interval.

f_m=the maximum frequency.

f_1=the frequency preceding to the modal class.

f_2=the frequency succeeding to the modal class.

i=the length of the modal class interval.

Example 7.

Find the modal grade for the data given in Example 3.

Grade	f
90 - 99	8
80 - 89	30 → f_2
70 - 79	45 → f_m
60 - 69	22 → f_1
50 - 59	10
40 - 49	3
30 - 39	2

Solution

The mode lies in the class interval having maximum frequency that is 69.5- 79.5. The mode is interpolated by the formula

$$\text{Mode} = L + \frac{f_m - f_1}{2f_m - f_1 - f_2} \times i$$

$$\text{Mode} = 69.5 \frac{45 - 22}{2(45) - 22 - 30} \times 10 = 75.55$$

3.4 Empirical Relationship between Mean, Median and Mode

For moderately skewed distributions, we have the following empirical relation:

Mode = 3 Median -2 Mean.

Mode = 3(74.61) -2(73.42) = 76.99

3.5 Comparison of measures of Central Tendency

The numerical value of every score in a data set contributes to the mean. This is not true of the mode or median because only the mean is based on the sum of all the scores. In a single peaked symmetrical distribution mean = median = mode. In practice, no distribution is exactly symmetrical, so the mode, median and mean usually have different values. If a population is not symmetrical, the mean, median and mode will not be equal. The mean is affected by the presence of a few extreme scores while the median and mode are not. The mean is preferred if extreme values are not present in the data. Median is preferred if interest is centered on the typical rather than the total score and if the distribution is skewed. If some scores are missing so that the mean cannot be computed directly, the median is appropriate. Mode is preferred only if the distribution is multimodal and a multi-valued index is satisfactory.

3.5.1 Quartiles

The values that divide a set of scores into four equal parts are called quartiles and are denoted by Q_1, Q_2 and Q_3. Q_1 is called the lower quartile and Q_3 is called the upper quartile. 25% of the scores are less than Q_1 and 75% of the scores are less than Q_3. Q_2 is the median. The formulas for the quartiles are

$Q_1 = \left(\frac{N+1}{4}\right)$ th score, $Q_2 = \frac{2(N+1)}{4} = \frac{N+1}{2}$ th score and

$Q_3 \frac{3(N+1)}{4}$ th score.

Example 8.

The following data show the weekly TV watching times of 20 people in hours. Determine and interpret the quartiles for these data.

30,46,32,37,48,71,40,36,20,10,39,31,37,43,21,35,43,35,25,26.

Solution

To find the quartiles, first we arrange the data in increasing order called array.

Array: 10 20 21 25 26 30 31 32 35 35 36 37 37 39 40 43 43 46 49 71

$Q_2 = \left(\frac{N+1}{4}\right)$ th score, $= \left(\frac{20+1}{4}\right) = 5.25^{th}$ score

=5th score + 0.25(6th score -5th score) = 26+0.25(30-26) = 27

$Q_1 = \frac{2(N+1)}{4} = \frac{N+1}{2}$ th score $= \left(\frac{20+1}{4}\right) = 10.5^{th}$ score

=lOth score + 0.5(11th score -lOth score) = 35+0.5(36-35)=35.5

$$Q_3 = \frac{3(N+1)}{4} = \frac{N+1}{2} \text{ th score} = \frac{3(20+1)}{4} = 15.75^{th} \text{ score}$$

=15th score+0.75(16th score -15th score)=40+0.75(43-40)=42.25

Interpreting our results, we conclude that 25% of the watching times are less than 27 (Q_1), 25% are between 27 (Q_1) and 35.5 (Q_2) hours. 25% are between 35.5 (Q_2)and 42.25 (Q_3) hours and 25% are greater than 42.25 (Q_3) hours. 50% of the watching times are less than 35.5 (Q_2) and 50% are greater than 35.5 (Q_2)

For the grouped data

$$Q_1 = L + \frac{i}{f}\left(\frac{N}{4} - C\right) \quad Q_2 = \text{median} = L + \frac{i}{f}\left(\frac{N}{2} - C\right)$$

$$Q_3\ L + \frac{i}{f}\left(\frac{3N}{4} - C\right)$$

Example 9.

Calculate the first and the third quartiles for the data given in Example 3.

Solution

Grade	f	C
90- 99	8	8+112 = 120
80- 89	30	30+82 = 112
70- 79	45	45+37 = 82
60- 69	22	22+15 = 37
50-59	10	10+5 = 15
40- 49	3	3+2 = 5
30- 39	2	= 2
Total	120	

$Q_1 = \frac{N}{4}$ th score = $\frac{120}{4}$ = 30 th score which lies in the class interval 59.5-69.5 called the first quartile class interval.

Q_1 can be interpolated by the following formula

$$Q_1 = L + \frac{i}{f}\left(\frac{N}{4} - C\right) \quad 59.5 + \frac{10}{22}(30 - 15) = 66.32$$

$Q_3 = \frac{3N}{4}$ th score = $\frac{3(120)}{4}$ = 90th score which lies in the class interval 79.5-89.5 called the third quartile class interval.

Q_3 can be interpolated by the following fonnula

$$Q_1 = L + \frac{i}{f}\left(\frac{3N}{4} - C\right) = 79.5 + \frac{10}{30}(90\text{-}82) = 82.17$$

3.5.2 Percentiles

The values that divide a set of scores into hundred equal parts are called percentiles and are denoted by $P_1, P_2, P_3, \ldots, P_{99}$. P_{25} is the first quartile, P_{75} is the third quartile and P_{50} is the median.

The formulas for the percentiles are

$P_1 = \left(\frac{N+1}{4}\right)$ th score, $P_2 = \frac{2(N+1)}{100}$ th score, $P_3 = \frac{3(N+1)}{100}$ th score, ...,

$P_{99} = \frac{99(N+1)}{100}$ th score

For the grouped data

$$P_1 = L + \frac{i}{f}\left(\frac{N}{100} - C\right), \quad P_2 = L + \frac{i}{f}\left(\frac{2N}{100} - C\right), \quad \ldots, \quad P_{50} = \text{median} = L + \frac{i}{f}\left(\frac{N}{2} - C\right), \ldots,$$

$$P_{75} = Q_3 \; L + \frac{i}{f}\left(\frac{75N}{100} - C\right), \; L + \frac{i}{f}\left(\frac{3N}{4} - C\right), \ldots,$$

$P_{99} = L + \frac{i}{f}\left(\frac{99N}{100} - C\right)$

Example 10.

Find 6th, 20th, 78th and 95th percentile for the data given in Example 3.

Solution

Grade	f	c
90- 99	8	8+ 112=120
80- 89	30	30+82 =112
70- 79	45	45+37 = 82
60- 69	22	22+15 = 37
50-59	10	10+5 = 15
40- 49	3	3+2 = 5
30- 39	2	= 2
Total	120	

P_6 is $\frac{6N}{100} = \frac{6(120)}{100}$ th score = 7.2nd score with lies in the class interval 49.5-59.5 P_6 can be interpolated by the formula

P_6 can be interpolated by the formula

$P_6 = L + \frac{i}{f}\left(\frac{6N}{100} - C\right)$,=49.5+$\frac{10}{10}$ (7.2-5) =51.7

P_{20} is $\frac{20N}{100} = \frac{20(120)}{100}$ th score= 24 score which lies in the class interval 59.5-69.5 P_{20} can be interpolated by the formula

$P_{20} = L + \frac{i}{f}\left(\frac{20N}{100} - C\right)$ = 59.5+$\frac{10}{22}$ (24-15) =63.59

P_{78} is $\frac{78N}{100} = \frac{78(120)}{100}$ th score = 93.6th score which lies in the class interval 79.5-89.5

find the percentage of scores below a given score. Calculation of percentile rank (PR) from ranked data PR= 100 $\left(\frac{100R-50}{N}\right)$

Where R is the rank position.

N=number of individuals in the ranked data.

Suppose there are 20 student in a class. The class teacher ranked them according to their marks obtained in Statistics. That is he/she allotted # 1 to the student having maximum marks, # 2 to the students who got second position in the class, # 3 the student having third position in the class and so on. Now PR of the student having rank of 2 is

PR = 100 - $\left(\frac{100R-50}{N}\right)$ = 100 - $\left(\frac{100(2)-50}{20}\right)$ = 100-7.5=92.5.

It means that the marks of 92.5% of the students in the class are below the student who got the second position in the class.

3.6 Self Assessment Questions 3

Read the following statements carefully and mark whether each statement is true or false.

(i) 50% of the scores are less than median and 50% of the scores are greater than median.

(ii) Several types of averages can be defined. .

(iii) Averages tend to lie in the centre of an array. That is why averages are called measures of central tendency.

(iv) Mode always exists.

(v) The procedure for calculating percentile ranks is the same as for calculating percentiles.

Fill in the blanks

(i) A single score calculated to represent all the scores is called._____.

(ii) Since averages locate the centre of a data set, these are also called. _____.

(iii) The values that divide a set of scores into four equal parts are called. _____.

(iv) The values that divide a set of scores into hundred equal parts are called. _____.

(v) By calculating percentile rank, we find the _____.of scores below a given score. _____.

Answers

True/False

(i) T (ii) T (iii) T (iv) F (v) F

Fill in the blanks

(i) An average (ii) Measures of location (iii) Quartiles (iv) Percentiles (v) Percentage

Activity

Measure and record heights of the students of a class in inches. Find the arithmetic mean, median and mode of these heights.

3.7 Exercise

Given below is the distribution of weekly income of 100 households to the nearest rupee in a locality.

Income:	65-69	60-64	55-59	50-54	45-49	40-44	35- 39
No. of households:	10	15	20	30	10	8	7

Student	Marks in I st test (X)	Marks in 2nd test (Y)	Ranks of X	Y	D	D^2
1	65	77	13	10	3	9
2	75	70	10	12	-2	4
3	85	90	8	6	2	4
4	98	97	2	3	-1	1
5	97	91	3	5	-2	4
6	88	85	7	9	-2	4
7	74	75	11	11	0	0
8	68	65	12	13	-1	1
9	89	89	6	7	-1	1
10	90	98	5	2	3	9
11	100	95	1	4	-3	9
12	84	88	9	8	1	1
13	40	50	15	15	0	0
14	55	60	14	14	0	0
15	95	100	4	1	3	9
Total						56

1. Rank each set of scores (columns 4 and 5)

2. Find the difference (D) between the ranks (column 6=column 4 column 5)

3. Square each difference in ranks (column 6) and write squared differences in column 7.

4. Sum the squared differences in column 7 to obtain D_2.

5. Put the values in the formula and solve.

$$r_s=1-\frac{6\sum D^2}{N(N_2-a)}=1-\frac{6x56}{15(15^2-1)}=0.9$$

There is high positive correlation between two sets of marks.

4.2 The Product-moment method

The product-moment coefficient is usually used when the number of scores is large. Thus this method is used in most research studies. The product-moment coefficient is usually denoted by r.

Example 13

Calculate and interpret the product-moment coefficient of correlation for the following sets of scores

X:75 76 72 65 80 63 60 78 45 49 58 56 77 74 63 46 65 59 51 75

Y:80 71 75 55 76 55 65 77 40 33 55 60 80 70 45 33 48 45 60 75

Solution

The following steps will guide you to compute a product-moment correlatjon coefficient r.

X	Y	XY	X^2	y^2
75	80	6000	5625	6400
76	71	5396	5776	5041
72	75	5400	5184	5625
65	55	3575	4225	3025
80	76	6080	6400	5776
63	55	3465	3969	3025
60	65	3900	3600	4225
78	77	6006	6084	5929
45	40	1800	2025	1600
49	33	1617	2401	1089
58	55	3190	3364	3025
56	60	3360	3136	3600
77	80	6160	5929	6400
74	70	5180	5476	4900

(iii) The scores are ranked according to size or some other criterion.

(iv) r=0 means no linear relationship between sets of score.

Fill in the blanks

(i) The range of the coefficient is from -1 to _____.

(ii) The most common methods of computing the coefficient of correlation are _____. and _____.

(iii) _____is the useful method for computing correlation coefficient when the number of scores to be correlated is small or exact magnitude of the scores cannot be ascertained

Answers

True/False

(i) T (ii) F (iii) T (iv) T

Fill in the blanks

(i) + 1 (ii) Rank-difference method, Product-moment method.

(iii) Rank-difference method

Activity

Administer a test and record marks of ten outstanding students of class nine in the subjects of Physics and Mathematics. Calculate product moment coefficient of correlation between marks of the two subjects. What results would you draw?

4.4 Exercise

Calculate the rank difference coefficient of correlation for the following sets of scores and interpret the result.

X: 125 117 114 120 115 109 107 106 105 104 103 100 97 95 94 93 91 90 88

Y: 100 77 75 83 63 90 78 69 48 50 60 74 54 59 65 46 57 62 58

Answer 0.688. There is positive correlation between two variables.

5 MEASURES OF VARIABILITY

Measures of central tendency measure the centre ofa set of scores. However, two data sets can have the same mean, median or mode and yet be quite different in other respects. For example, consider the heights (in inches) of the players of two basketball teams

Team 1: 72 73 76 76 78

Team 2: 67 72 78 76 84

The two teams have the same mean height, 75 inches, but it is clear that the heights of the players of team 2 vary much more than those of team 2. If we have information about the centre of scores and the manner in which they are spread out, we know much more about set of scores. The degree to which scores tend to spread about an average value is called dispersion.

5.1 The Range

It is the simplest measure of dispersion. The range of a set of scores is the difference between maximum scores and minimum scores.

In symbols Range=X_m-X_o

Where X_m is the maximum score and X_o is the minimum score.

Example 14.

The heights, in inches, of the 8 students admitted in 11th class in a college are 60, 66, 72, 62, 65, 70,65 and 69. Find range of the heights.

Solution

1	2	3	4	5	6
Grade	X	f	fX	$\sum\|X-\overline{X}\|$	$f\|X-\overline{X}\|$
90-99	94.5	8	756.0	21.08	138.64
80-89	84.5	30	25350 5	11.08	332.40
70-79	74.5	45	3352.5	1.08	48.60
60-69	64:5	22	1419.0	8.92	196.24
50-59	54.5	10	545.0	18.92	189.20
40-49	44.5	3	133.5	28.92	86.76
30-39	34.5	2	69.0	38.92	77.84
			120		1099.68

step 1. Find the arithmetic mean $\overline{X}=\frac{\sum fX}{\sum f}=\frac{8810}{120}=73.42$

Step 2. Subtract the arithmetic mean from each class mark **X and** ignore the sign (column 5) i.e.

find $|X-\overline{X}|$

Step 3. Multiply Ix -xI (column 5) with corresponding f(column 3) and sum column 6 to

obtain $\sum f|X-\overline{X}|$

Step 4. Divide $\sum f|X-\overline{X}|$ by $\sum f$ to obtain mean deviation i.e.

M.D. = $\frac{\sum f|X-\overline{X}|}{\sum f}=\frac{1099.68}{120}=9.16$

5.3 The Standard Deviation

The standard deviation is the positive square root of the arithmetic mean of the squares of deviations of all the scores from their mean.

Example 18.

Find the standard deviation for the data given in Example 12.

Solution

X	$X-\overline{X}$	$(X-\overline{X})^2$	X^2
60	-6.125	37.5156	3600
66	-0.125	0.0156	4356
72	5.875	34.5156	5184
62	-4.125	17.0156	3844
65	-1.125	1.2656	4225
70	3.875	15.0156	4900
65	-1.125	1.2656	4225
69	2.875	8.2656	4761
529		114.875	35095

$$S = \sqrt{\frac{\Sigma\left(X-\overline{X}\right)^2}{N}} = \sqrt{\frac{114.875}{8}} = \sqrt{14.3594} = 3.79$$

Calculation of the standard deviation using the short computational formula

$$S = \sqrt{\frac{\Sigma X^2}{N} - \left(\frac{\Sigma X}{N}\right)^2} = \sqrt{\frac{35095}{8} - \left(\frac{529}{8}\right)^2} = \sqrt{\frac{114.875}{8}} = 3.79$$

5.4 Calculation of the standard deviation for the grouped data.

Example 19.

Calculate the standard deviation for the data given in Example 3.

solution

1	2	3	4	5	6	7	8
Grade	X	f	fX	$X-\overline{X}$	$\left(X-\overline{X}\right)^2$	$f\left(X-\overline{X}\right)^2$	fX^2
90-99	94.5	8	756.0	21.0833	444.5069	3556.0556	71442
80-89	84.5	30	2535.0	11.0833	122.8403	3685.2083	214208
70-79	74.5	45	3352.5	1.0833	1.1736	52.8125	249761
60-69	64.5	22	1419.0	-8.9167	79.5069	1749.1528	91526
50-59	54.5	10	545.0	-18.9167	357.8403	3578.4028	29703
40-49	44.5	3	133.5	-28.9167	836.1736	2508.5208	8941
30-39	34.5	2	69.0	-38.9167	1514.5069	3029.0139	2381
	120	8810.0			18159.1667	664960	

The arithmetic mean was calculated as $\overline{X}=\frac{\sum fX}{\sum f}=\frac{8810}{120}=73.4167$

The standard deviation is calculated as under

Step I: Find the mid-values (column 2).

Step 2: Multiply f with X to obtain the column fX (column 4) and sum the column to find $\sum fX$

Step 3: Sum the column of frequencies (column 3) to obtain $\sum f$

Step 4: Divide $\sum fX$ by $\sum f$ to calculate arithmetic mean $(\overline{X})$.

Step 5: Subtract X from each value of X (column 2) and obtain the column $X-\overline{X}$ (column 5).

Step 6: Square the column $X-\overline{X}$ (column 5) to obtain the column of $(X-\overline{X})^2$(column 6).

Step 7: Multiply each $(X-\overline{X})^2$ (column 6) with the corresponding f (column 3) to obtain the column of $f(X-\overline{X})^2$ (column 7). Sum this column and obtain $\sum f(X-\overline{X})^2$.

Step 8: Divide $\sum f(X-\overline{X})^2$ by $\sum f$ and take the square root of the result. Thus the standard deviation is obtained.

$$s = \sqrt{\frac{\sum f(X-\overline{X})}{\sum f}} = \sqrt{\frac{18159.1667}{120}} = \sqrt{151.3264} = 12.30$$

The short computational formula for calculating the standard deviation is

$$s = \sqrt{\frac{\sum f X^2}{\sum f} - \left(\frac{\sum fX}{\sum X}\right)^2}$$

After the step 4 above, do the following steps

Step 5: Multiply X (column 2) with fX (column 4) to obtain fX^2 (column 8), Sum this

column and obtain $\sum f X^2$

Step 6: Put th,e values in the short formula and solve to obtain s.

$$\text{Thus } S = \sqrt{\frac{\sum f X^2}{\sum f} - \left(\frac{\sum fX}{\sum X}\right)^2} = \sqrt{\frac{664960}{120} - \left(\frac{8810}{120}\right)^2} = \sqrt{151.3264} = 12.30$$

Yet there is another short formula for the common class interval width (c). Thus

$$S = \sqrt{\frac{\sum f X'^2}{\sum f} - \left(\frac{\sum fX'}{\sum X}\right)^2} \times C$$

Where $X' = \frac{X-A}{c}$

A is assumed or guessed mean usually chosen from the values of X and c is the common class interval length.

The following steps are involved in the calculation

Grade	X	f	X	fX	fX^2
90-99	94.5	8	2	16	32
80-89	84.5	30	1	30	30
70-79	74.5 → A	45	0	0	0
60-69	64.5	22	-1	-22	22
50-59	54.5	10	-2	-20	40
40-49	44.5	3	-3	-9	27
30-39	34.5	2	-4	-8	32
Total		120		-13	183

Step 1: Find the mid-values (column 2).

Step 2: Sum the column of frequencies (column 3) to obtain $\sum f$.

Step 3: Select an assumed mean (A) from the column X (column 2) usually in the centre of the column against maximum frequency..

Step 4: Subtract A from each X-value and divide it by common class interval width (c) and

denote it with X' (column 4).

Step 5: Multiply X (column 4) with f(column 3) to obtain fX (column 5). Sum this column and obtain. $\sum fX$

Step 6: Multiply X' (column 4) with fX' (column 5) to obtain fX^2 (column 6). Note that xQ entry in column 6 will be negative. Sum this column and obtain $\sum f X^2$

Step 7: Put the values in the formula and solve to obtain S.

$$\sqrt{\frac{183}{120}-\left(\frac{13}{120}\right)^2} \times 10 = \sqrt{105133} \times 10 = 12.30$$

5.5 The Coefficient of Variation

Karl Pearson introduced a relative measure of dispersion known as coefficient of variation (denoted by c.v), It expresses the standard deviation as a percentage of the arithmetic mean of a data set. It is number without units and is used to compare variation in two or more distributions. The smaller value of the c. v. indicates lesser variation. It is also used as a criterion for consistent performance of the students, players etc.

$$\text{C.V.} = \frac{s}{\overline{X}} \times 100$$

5.6- Chebychev' s Rule

For any data set

At least 75% of the data lie within two standard deviations to either side of the mean, that is, between $\overline{X}$ -2s and $\overline{X}$ + 2s

At least 89% of the data lie within three standard deviations to either side of the mean, that is, between $\overline{X}$ -3s and $\overline{X}$ + 3s

In general, for any number k> 1, at least $1-\frac{1}{k^2}$ of the data lie within k standard deviations to either side of the mean, that is, between $\overline{X}$ -ks and $\overline{X}$ + ks .This rule can be applied using any value of k that is greater than 1. For Example, assume k=2.5. Then

$1-\frac{1}{k^2}=\frac{1}{2.5^2}=\frac{1}{6.25}$=0.84 or 84%. Thus for any data set, at least 84% of the data lie within 2.5 standard deviations to either side of the mean.

5.7 Other characteristics of distributions.

Skewness

In case of a symmetrical distribution, both tails are equidistant from center and deviation below the mean is exactly equal to the corresponding deviation above the mean. This is known as symmetry .In a symmetrical distribution,

mean = median = mode.

Skewness is the degree of asymmetry .If the left tail of the distribution is longer than the right tail, it is called negatively skewed. In a negatively skewed distribution mode is greater than median and the median is greater than mean.

If the right tail of the distribution is longer than the left tail, it is called positively skewed. In a positively skewed distribution mean is greater than median and the median is greater than mode.

Measures of Skewness

Karl Pearson suggested the following formula for measuring

$$Skewness = \frac{mean - \mathrm{mod}\, e}{s\tan dard deviation}$$

To avoid the use of mode, we can apply the following formula

$$Skewness = \frac{3(mean - \mathrm{mod}\, e)}{s\tan dard deviation}$$

If these coefficients are 0, the distribution is symmetrical.

If these coefficients are + , the distribution is positively skewed.

If these coefficients are -, the distribution is negatively skewed.

Example 20

Calculate and interpret the coefficient of skewness for the data given in example 19.

Solution

The Arithmetic mean, mode and standard deviation were calculated as

Mean=73.4167, Mode=75.55 standard deviation=12.30.

Putting and solving these values in the formula, we get

$$\text{Skewness} + \frac{mean - \text{mod}e}{s\tan darddeviation} = \frac{73.4167 - 75.55}{12.30} = -0.17$$

As the coefficient of skewness is negative, the distribution is negatively skewed.

5.8 Standard Scores.

A frequently used quantity in statistical analysis is the standard score or Z-score. The standard score for a data value is the number of standard deviations that the data value is away from the mean of the data set.

In symbols $Z = \frac{X - \overline{X}}{s}$

5.9 Self Assessment Questions 5

Read the following statements carefully and mark whether each statement is true or false.

(i) Measures of central tendency measure the centre of a set of scores.

(ii) Two data sets can have the same mean, median or mode and yet be quite different in other respects.

(iii) The smaller value of the C. v. indicates greater variation.

(iv) Skewness is the degree of asymmetry.

(v) If the left tail of the distribution is longer than the right tail, it is called positively skewed.

Fill in the blanks

(i) The positive square root of the arithmetic mean of the squares of deviations of all the scores from their mean is called_____.

(ii) _____expresses the standard deviation as a percentage of the arithmetic mean of a data set.

(iii) At least _____% of the data lie between $\overline{X}$-3s and $\overline{X}$ + 3s

(iv) If the right tail of the distribution is longer than the left tail, it is called _____.

(v) Karl Pearson introduced a relative measure of dispersion known as _____.

Answers

True/False

(i) T (ii) T (iii) F (iv) T (v) F

Fill in the blanks

(i) Standard deviation (ii) Coefficient of variation (iii) 89 (iv) Positively skewed (v) Coefficient of variation

5.10 Exercises

Given the following data, calculate quartile deviation, mean deviation and standard deviation.

Class Interval: 48-50 45-47 42-44 39-41 36-38 33-35 30-32 27-29 24-26 21-23

Frequency 1 2 5 6 8 7 5 4 1 1

Answers

Mean 36.1 MD 4.89 SD 6.007495 Q131.9 Q3 40.5 QD 4.3

6- THE NORMAL CURVE

Before explaining the normal distribution, some basic concepts of probability is given below an event is a specified result that mayor may not occur when an experiment is performed. For example, in tossing of a coin once, appearance of head is an event, which may or may not occur. The probability of an event is a measure of the likelihood of its occurrence. A probability near indicates that the event is very unlikely to occur, whereas a probability near 1 indicates that the event is quite likely to occur.

6.1 Relative frequency interpretation of probability

Consider the, experiment of tossing a balanced coin once. There is 50-50 chances the head will appear. Consequently, we assign a probability of 0.5 to that event. The relative-frequency interpretation is that in a large number of tosses, the head will appear about half of the time.

The following table shows the ages of the students in a class. Suppose one of the students is selected at random, meaning that each student is equally likely to be selected. Find the probability that the student selected is 18 years old.

Age(yrs) X	Frequency f	Relative frequency
15	1	1/25 = 0.04
16	5	5/25 = 0.20
17	10	10/25 = 0.40
18	6	6/25 = 0.24
19	2	2/25 = 0.08
20	1	1/25 = 0.04
Total	25	1.00

It is clear from the above table that 6 out of 25 students in the class are 18 years old (second column). Thus the chances are 6 out of 25 for selecting a student of 18 years age. The probability is therefore

$$\frac{Number of 18-year-old students}{Total number of students} = \frac{6}{25}$$

Note that the probability, $\frac{6}{25}$ of randomly selecting a student of age 18 is exactly the same as the relative frequency, 0.24, of the students of age 18.

Definition

If there are n equally likely outcomes for an experiment, then the probability that a specified event occurs is equal to the number of ways that the event can occur (f) divided by the total number of possible outcomes (n).

In symbols: Probability of an event = $\frac{f}{n}$

The probability of an event is always between 0 and 1.

The probability of an event that cannot occur (impossible event) is 0.

The probability of an event that must occur (sure event) is 1.

A random variable is a quantity whose value depends on chance. For example, the height of a person chosen at random is a random variable whose values depend on chance of which person is selected. There are two main types of random variables: discrete and continuous. A discrete random can take only discrete set of integers or whole numbers. The number of books owned by a student, the number of rooms in a house are examples of discrete

random variable. A continuous random variable can take any fractional or integer value within some interval of numbers. The height of a person, the weight of a student are examples of a continuous random variable.

Random variables are usually denoted by X, Y, Z etc.

A listing of the possible values and corresponding probabilities of a discrete random variable or a formula for the probabilities is called probability distribution.

Example 21.

Two fair coins are tossed. Let X be the number of heads. Find the probability distribution of X.

Solution

When two fair coins are tossed, the following are the possible outcomes (results):

HH, HT, TH, TT

Probability distribution of X

X	Probability
No. of heads	
0	1/4=0.25
1	2/4=0.50
2	1/4=0.25
Total	4/4=1.00

The probabilities are non-negative and the sum of the probabilities is 1.

The probability that X=O is 1/4=0.25 because 1 out of 4 possible results is that there is no head (there are all tails). The probability

that X=1 is 1/2=0.50 because 2 out of 4 possible results are that there is one head. Similarly, the probability that X=2 is 1/4=0.25 because 1 out of 4 possible. Results are that there is two head.

Bernoulli Trials

Single performance of an experiment is called a trial. Repeated identical trials are called Bernoulli trials if the following conditions are satisfied:

1. Each trial has two possible outcomes usually called success and failure.

2. The trials are independent.

3. The probability of success remains same from trial to trial.

Binomial Distribution

The binomial distribution is the probability distribution for the number of successes in a sequence of Bernoulli trials. Suppose n Bernoulli trials are to be performed and the probability of success in one trial is p. Let X denote the number of successes in n trials. Then the probability of exactly x successes is

$$P(x)=\binom{n}{x} P^x q^{n-x} \quad x=0, 1, 2, \ldots, n$$

Where $\binom{n}{x}=\frac{n!}{x!(n-x)!}$ and n! (Read as n factorial)=n(n-1)(n-2)...3.2.1

For example, 5! =5.4.3.2.1=120

$$\binom{5}{2}=\frac{5!}{2!3!}=\frac{120}{12}=10$$

Example 22.

A salesman knows from the past experience that he will make a sale to 25% of his customers.

Find the probability that in any 6 attempts, he sells to

i) no customer .

ii) Exactly one customer.

iii) At most one customer.

iv) At least one customer.

Solution

Let X denotes the number of sales in 6 attempts. The following steps are performed to find the required probabilities.

Step 1. Identify success and find the probability of success in one trial (denoted by p) . In this problem, the success is a sale to a customer. Hence the probability that the salesman makes a sale to a particular customer is p = 0.25.

Step 2. Determine the number of trials denoted by n. Here the number of trials is the number of customers. Thus n=6.

Step 3. The binomial probability formula is

$$P(x)=\binom{n}{x}p^x q^{n-x} \qquad x=0, 1,2, \ldots, n$$

Puttirtg n=6and p=0.25 in the formula

$$P(x)=\binom{6}{x}(0.25)^x(0.75)^{6-X} \qquad x=0, 1,2, \ldots, 6$$

i) Here we have to fine probability that the salesman sell to no customer i.e. P(x=O)

$$P(x)=\binom{6}{0}(0.25)^0(0.75)^6 =0.178$$

ii) Here put x=l in the formula.

$$P(x)=\binom{6}{1}(0.25)\,(0.75)^5=0.356$$

iii) To find the probability that the sale is made to at most one customer is P(O)+P(l)

$$\binom{6}{0}(0.25)^0(0.75)^6+\binom{6}{1}(0.25)\,(0.75)^5=0.178+0.356=0.534$$

iv) The probability of at least one sale is P(1) + P(2) + P(3) + P(4) + P(5) + P(6)

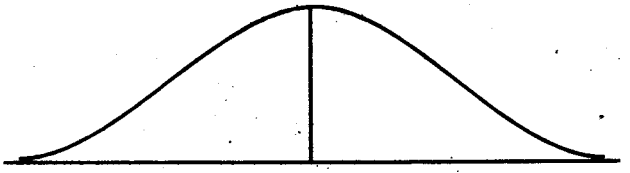
or alternatively,

$$1-P(0)=1-\binom{6}{0}(0.25)^0(0.75)^6=1-0.178=0.822$$

The probability distribution of a continuous random variable is called continuous probability distribution. The most important continuous probability distribution is the normal distribution. The binomial distribution is closely approximated by the normal distribution when n is sufficiently large and neither p nor q is close to zero. As a rule, the normal distribution provides a good approximation to the binomial distribution if both np and nq are equal to or greater than 5. The equation of the normal distribution is

$$y=\frac{1}{\sqrt{2\pi\sigma}}e^{\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2}$$

μ =mean σ =standard deviation e=2.7I83 π =3.I4I6



The normal probability curve

6.2- Some Basic Properties of the Normal Curve.

I. The total area under the normal curve is equal to 1.

2. The normal curve extends indefinitely in both directions.

3. The normal distribution is symmetric about the mean μ that is the part of the curve to the left

of μ is the mirror image of the part of the curve to the right of it.

4. The mean, the median and the mode are equal.

5. Mean deviation is 0.7979 σ

6. Quartile deviation is 0.6745 σ

7. In a normal distribution,

μ - 0.6745 σ to μ + 0.6745 σ covers 50% of the area.

μ -σ to μ+ σ covers 68.27% of the area.

μ-2σ to μ+ 2σ covers 95.45 of the area.

μ-3σ to μ+ 3σ covers 99.73 % of the area.

Example 23.

The heights of 10000 college students are assumed to be normally distributed with mean 65 inches and standard deviation 4 inches. What percentage of students fall

(i) below 58 inches. (ii) above 60 inches.

(iii) below 75 inches (iv) above 75 inches.

(v) between 60 and 75 inches. (vi) between 55 and 60 inches.

(vii) between 70 and 80 inches. (viii) How many students are below 60 inches?

Solution

The following steps are involved

Step 1 Sketch the normal curve and indicate the area (probability) to be determined on the graph.

Step 2 Compute the required z-scores and mark them on the graph below x-values.

Step 3 Find the required area (probability) using table (the area under the standard normalcurve).

Step 4: Find the percentage, frequency, if required.

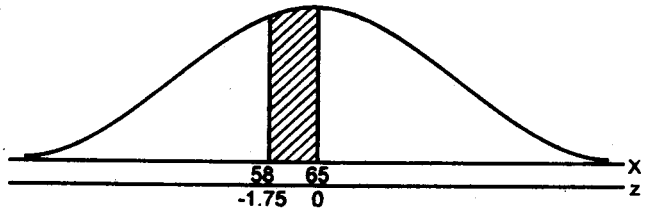
(i) Percentage of students that fall below 58 inches.

First we find the probability that a student has a height below. 50 inches. Then we convert this probability into percentage by multiplying it with lOO.

Step I: Sketch the normal curve and indicate the area (probability) to be determined on the graph.

Step 2: Convert x-value into z-score using the

formula $z = \frac{X - \mu}{\sigma}$



This process is called standardization.

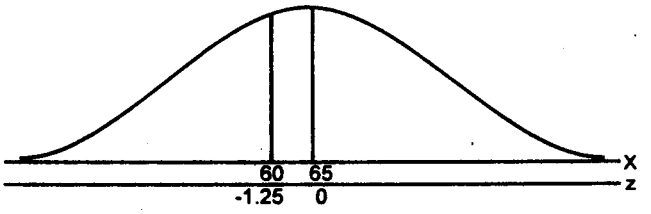
$$\text{For } x = 58 \; z = \frac{X - \mu}{\sigma} = \frac{58 - 65}{4} = 1.75$$

Step 3: Using the table 7.1 (Area under the standard normal curve from 0 to z). First we go down the left-hand column labeled Z to 1.7 .Then we go across this row until we are under .05 in the top row. The area in the body of the table there is 0.4599 which is the area - under the standard normal curve that lies between Z=O and z=l. 75 (or z=-I. 75). The required area (below 58) is 0.5-0.4599=0.0401.

Step 4: Percentage of the students below 58 inches =0.0401 x 100=4.01 %

(ii) above 60 inches.

z-score Area from 0 to z



$x = 60 \quad \rightarrow \quad z = \frac{60 - 65}{4} = -1.25 \qquad 0.3944$

The required area is 0.5+0.3944=0.8944

Percentage of the students above 60 inches =0.8944 x 100=89.44%

(iii) below 75 inches

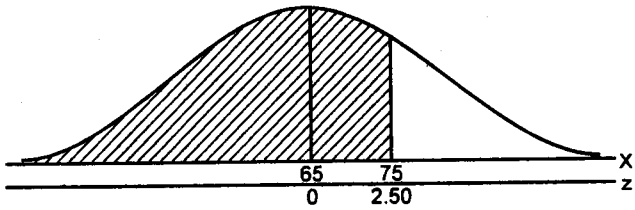
z-score Area from 0 to z

$x = 75 \rightarrow .\ z = \frac{75-65}{4} = 2.50$ 0.4938

The required area is 0.5+0.4938=0.9938

Percentage of the students below

75 inches = 0.9938 x 100=99.38%



(iv) above 75 inches

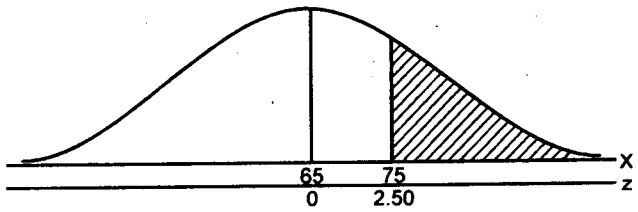
z-score Area from 0 to z

$x = 75 \quad z = \frac{75-65}{4} = 2.50$ 0.4938

The required area is 0.5-0.4938=0.0062

Percentage of the students above

75 inches = 0.0062 x 100 = 0.62 %



(v) between 60 and 75 inches.

z-score Area from 0 to z

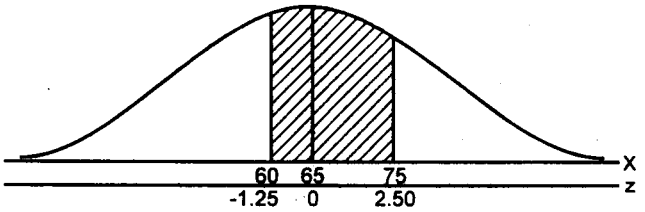
$x = 60 \rightarrow z = \frac{60-65}{4} = -1.25$ 0.3944

$x = 75 \rightarrow z = \frac{75-65}{4} = 2.50$ 0.4938

the required area is 0.3944 + 0.4938 = 0.8882

percentage of the students between

60 and 75 inches = 0.8882 x 100 = 88.82%

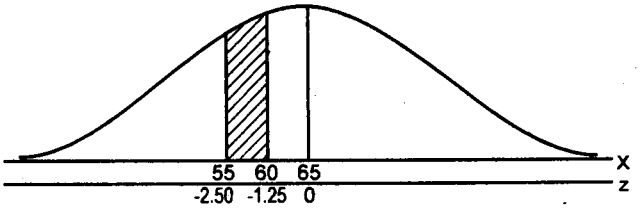


(vi) between 55 and 60 inches.

z-score Area from 0 to z

$X = 55 \rightarrow z = \frac{55-65}{4} = -2.50$ 0.4938

$X = 60 \rightarrow z = \frac{60-65}{4} = -1.25$ 0.3944



The required area is 0.4938 -0.3944=0.0994

Percentage of the students between

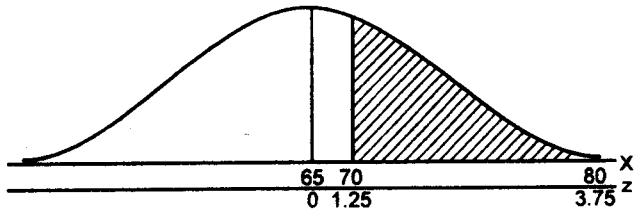
55 and 60 inches = 0.0994 x 100=9.94 %

(vii) between 70 and 80 inches.

z-score Area from 0 to z

$x = 70 \rightarrow z = \frac{70-65}{4} = 1.25$ 0.3944

$x = 80 \rightarrow z = \frac{80-65}{4} = 3.75$ 0.4999



The required area is 0.4999 -0.3944=0.1 055

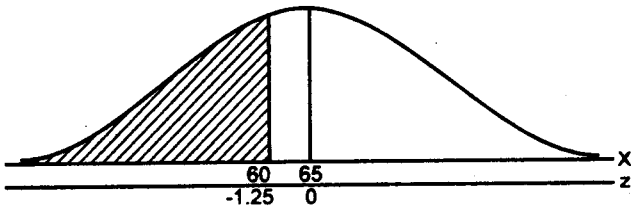
Percentage of the students between

70 and 80 inches = 0.1055 x 100=10.55 %

(viii) How many students are below 60 inches?

Z-score Area from 0 to z

$x=60 \rightarrow z = \frac{60-65}{4} = -1.25$ 0.3944 ~



The required area is 0.5-0.3944=0.0.1 056

Number of students having height below

60 inches=10000 x O.1056 =1056 students.

Finding x-value for a Specified Area

Using the formula of z-scores $z = \frac{x-\mu}{\sigma}$, we obtain the formula for converting z-score to x-values:

$x = \mu - z\sigma$

Example 24.

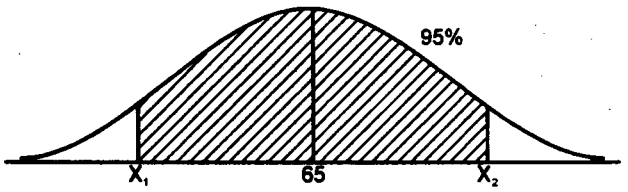
For the distribution given in Example 23 ,

(i) What limits will include the middle 95% of the cases ?

(ii) Find the height below which the heights of 80% of the students fall.

Solution

Step 1 Sketch the normal curve indicate the given area (probability) and the x-value to be determined on the graph.



Step 2 Find the z-score using table 7.1

Step 4 Obtain the required x-value using the formula $x = \mu + Z\sigma$.

(i) Area from 0 to Z Z-score

0.4750 1.96

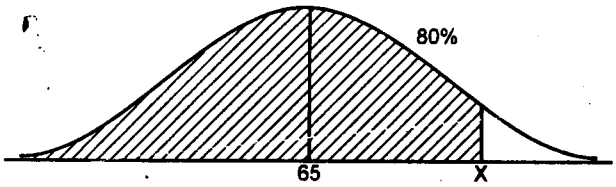
$x_1 = 65 + (-1.96)4 = 65 - 7.84 = 57.16$

$X_2 = 65 + (1.96)4 = 65 + 7.84 = 72.84$

(ii) Area from 0 to Z Z-score

0.3 0. 84 against 0.2996

$x = 65 + (0.84)4 = 65 + 3.36 = 68.36$



6.3- Self Assessment Questions 6

Read the following statements carefully and mark whether each statement is true or false.

(i) An event is a specified result that mayor may not occur when an experiment is performed.

(ii) The probability of an event is a measure of the likelihood of its occurrence.

(iii) The most important continuous probability distribution is the binomial distribution.

(iv) The binomial distribution is closely approximated by the normal distribution when n is small and neither p nor q is close to zero.

(v) There are two main types of random variables: discrete and continuous

Fill in the blanks

(i) The probability of an event is always between 0 and _____.

(ii) Random variables are usually denoted by _____.

(iii) A listing of the possible values and corresponding probabilities of a discrete random variable is called_____.

(iv) In a normal distribution, mean = 40, median = _____.

(v) The total area under the normal curve is equal to _____.

Answers

True/False

(i) T (ii) T (iii) F, It is the normal distribution. (iv) F (v) T

Fill in the blanks

(i) 1 (ii) X, Y, Z etc. (iii) Probability distribution. (iv) 40 (v) 1

6.4. Exercises

1. The scores made by candidates in a certain test are normally distributed with mean 500 and standard deviation 100. What percent of the candidates received scores (i) greater than 700 (ii) less than 400 (iii) between 400 and 600 (iv) which differ from mean by more than 150.

2. The heights of a large sample of men were found to be approximately normally distributed with mean 67.56 inches and standard deviation 2.57 inches. What height is exceeded by 5% of men?

Answers:

1. (i) 2.28% (ii) 15.87% (iii) 68.26% (iv) 13.36%

2. 71.79

7. BIBLIOGRAPHY

Garrett Henry E. and Woodworth R.S., (1966), Statistics in Psychology and education, Longmans, Green and Co. Ltd.

Ferguson George A., (1976), Statistical analysis in Psychology & Education, McGraw-Hill Kogakusha, Ltd.

Guilford J .P .and Fruchter Benjamin. (1981), Fundamental Statistics in Psychology and Education.

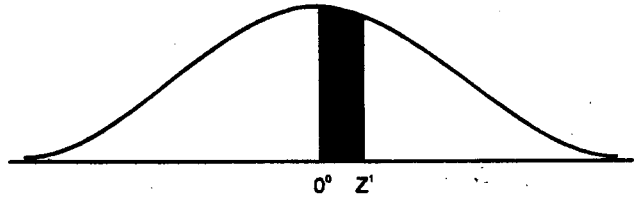
Weiss Neil A., (1996), Elementary Statistics, Addison- Wesley Publishing Company.

Daniel Wayne W., (1998), Biostatistics: A Foundation for Analysis in The Health Sciences, John Wiley & Sons (Asia) Pte Ltd. Singapore.

Games Paul A. and Klare George R., (1967), Elementary Statistics Data Analysis for the Behavioural Sciences, McGraw-Hill Book Company.

Table 7.1

Area under the Standard Normal Curve from 0 to Z



Z	0	1	2	3	4	5	6	7	8	9
0.0	0.0000	0.0040	0.0080	0.0120	0.0160	0.0199	0.0239	0.0279	0.0319	0.0359
0.1	0.0398	0.0438	0.0478	0.0517	0.0557	0.0596	0.0636	0.0675	0.0714	0.0754
0.2	0.0793	0.0832	0.0871	0.0910	0.0948	0.0987	0.1026	0.1064	0.1103	0.1141
0.3	0.1179	0.1217	0.1255	0.1293	0.1331	0.1368	0.1406	0.1443	0.1480	0.1517
0.4	0.1554	0.1591	0.1628	0.1664	0.1700	0.1736	0.1772	0.1808	0.1844	0.1879
0.5	0.1915	0.1950	0.1985	0.2019	0.2054	0.2088	0.2123	0.2157	0.2190	0.2224
0.6	0.2258	0.2291	0.2324	0.2357	0.2389	0.2422	0.2454	0.2486	0.2518	0.2549
0.7	0.2580	0.2612	0.2642	0.2673	0.2704	0.2734	0.2764	0.2794	0.2823	0.2852
0.8	0.2881	0.2910	0.2939	0.2967	0.2996	0.3023	0.3051	0.3078	0.3106	0.3133
0.9	0.3159	0.3186	0.3212	0.3238	0.3264	0.3289	0.3315	0.3340	0.3365	0.3389
1.0	0.3413	0.3438	0.3461	0.3485	0.3508	0.3531	0.3554	0.3577	0.3599	0.3621
1.1	0.3643	0.3665	0.3686	0.3708	0.3729	0.3749	0.3770	0.3790	0.3810	0.3830
1.2	0.3849	0.3869	0.3888	0.3907	0.3925	0.3944	0.3962	0.3980	0.3997	0.4015
1.3	0.4032	0.4049	0.4066	0.4082	0.4099	0.4115	0.4131	0.4147	0.4162	0.4177
1.4	0.4192	0.4207	0.4222	0.4236	0.4251	0.4265	0.4279	0.4292	0.4306	0.4319
1.5	0.4332	0.4345	0.4357	0.4370	0.4382	0.4394	0.4406	0.4418	0.4429	0.4441
1.6	0.4452	0.4463	0.4474	0.4484	0.4495	0.4505	0.4551	0.4525	0.4535	0.4545
1.7	0.4554	0.4564	0.4573	0.4582	0.4591	0.4599	0.4608	0.4616	0.4625	0.4633
1.8	0.4641	0.4649	0.4656	0.4664	0.4671	0.4678	0.4686	0.4693	0.4699	0.4706
1.9	0.4713	0.4719	0.4726	0.4732	0.4738	0.4744	0.4750	0.4756	0.4761	0.4767
2.0	0.4772	0.4778	0.4783	0.4788	0.4793	0.4798	0.4803	0.4808	0.4812	0.4817
2.1	0.4821	0.4826	0.4830	0.4834	0.4838	0.4842	0.4846	0.4850	0.4854	0.4857
2.2	0.4861	0.4864	0.4868	0.4871	0.4875	0.4878	0.4881	0.4884	0.4887	0.4890
2.3	0.4896	0.4896	0.4898	0.4901	0.4904	0.4906	0.4909	0.4911	0.4913	0.4916
2.4	0.4918	0.4920	0.4922	0.4925	0.4927	0.4929	0.4931	0.4932	0.4934	0.4936
2.5	0.4938	0.4940	0.4941	0.4943	0.4945	0.4946	0.4948	0.4949	0.4951	0.4952
2.6	0.4953	0.4955	0.4956	0.4957	0.4959	0.4960	0.4961	0.4962	0.4963	0.4964
2.7	0.4965	0.4966	0.4967	0.4968	0.4969	0.4970	0.4971	0.4972	0.4973	0.4974
2.8	0.4974	0.4975	0.4976	0.4977	0.4977	0.4978	0.4979	0.4979	0.4980	0.4981
2.9	0.4981	0.4982	0.4982	0.4983	0.4984	0.4984	0.4985	0.4985	0.4986	0.4986
3.0	0.4987	0.4987	0.4987	0.4988	0.4988	0.4989	0.4989	0.4989	0.4990	0.4990
3.1	0.4990	0.4991	0.4991	0.4991	0.4992	0.4992	0.4992	0.4992	0.4993	0.4993
3.2	0.4993	0.4993	0.4994	0.4994	0.4994	0.4994	0.4994	0.4995	0.4995	0.4995
3.3	0.4995	0.4995	0.4995	0.4996	0.4996	0.4996	0.4996	0.4996	0.4996	0.4997
3.4	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4997	0.4998
3.5	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998	0.4998
3.6	0.4998	0.4998	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.7	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.8	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999	0.4999
3.9	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000	0.5000