PROPOSALS FOR THE EVALUATION OF MATHEMATICS ACHIEVEMENT IN SECONDARY SCHOOLS AFFILIATED TO THE BOARD OF SECONDARY EDUCATION, LAHORE.

By

ABDUL HAMID WAIN

A Thesis
Submitted in partial fulfilment of the requirements
for the degree of Master of Arts
in the Education Department of the
American University of Beirut
Beirut, Lebanon.
February, 1963

MATHEMATICS TESTING: PAKISTAN

WAIN

ACKNOWLEDGEMENTS

I am deeply indebted to Professor Frederick R. Korf, the chairman of my thesis committee, under whose stimulating guidance this study has been carried out. Professor Korf directed my thinking by giving invaluable suggestions and ideas. I am grateful to him in yet another way. I took two courses with him - one in educational statistics and one in educational measurement - which provided the basic background for this study. I take pleasure in saying that the present work owes a great deal to his sympathetic encouragement.

I am grateful to the members of my thesis committee, Professor Habib A. Kurani, Professor Louis P. Cajoleas and Professor J. Edgar Lowe, for their help and guidance. I studied philosophy of education with Professor Kurani. The rich experience I gained in that course was of tremendous help to me in the treatment and exposition of the problem. At the back of my mind, also, was an abundant and illuminating experience I acquired in a course on social change with Professor Cajoleas which immensely coloured the thinking which went into the last chapter and clarified in my mind ways to cause desirable changes in the existing system of evaluation.

Special thanks are due to Professor Faizeh Antippa for helping me in arranging the administration of the algebra test in the local schools.

I owe my thanks to Mrs. Wadad Cortas, Principal of the Ahliah Girls School, Miss Rida Jureidini, Principal of the National Evangelical School for Girls, Mr. Faiz Asaad, Director of International College, and Mr. Kamil Deeb, Principal of the National Protestant College, who very kindly permitted me to administer the algebra test in their institutions.

I also acknowledge my thanks and sincere appreciation to the students and the respondents who helped by cheerfully participating in the study.

I am extremely grateful to Mr. Anwer Sikander Khan, Principal Sadiq Public School Bahawalpur (West Pakistan), Mr. A. Zaffar Hanif, Mr. H. Jivanandham, Mr. S. M. Rafi, Mr. A. Majid Wain, Mr. A. Rashid Rana and Mr. Muhammad Abdullah Khawja for sending from Pakistan the material and information I needed in connection with this study.

I have to express my gratitude to my wife for helping me in analysing the test results; for copying portions of the manuscript for typing; and for keeping me buoyant all through.

February, 1963 Abdul Hamid Wain

ABSTRACT

The thesis addresses itself to some serious short-comings in the system of evaluating mathematics achievement in Pakistan. It is shown that the system of evaluation is not related to instruction in the sense that tests are not used to improve instruction. Tests, it is believed, should aid learning, but it is seldom, if ever, that tests in Pakistani schools are used to serve this end; instead the present tests encourage the pupils to memorise set methods and steps which they reproduce mechanically in the examination. The problem of improving the tests has been a concern of the investigator for quite some years. It grew in intensity and gained in significance when he came into living contact with the modern concepts of education during his study for the M.A. degree (Education) at the American University of Beirut. He, therefore, decided to probe it by analysing the defects of the present system of evaluation and suggesting some remedies consistent with modern theories.

The first chapter states the problem, which is to examine critically the present day achievement tests in mathematics and to acquaint the teachers with better techniques of evaluation.

The second chapter undertakes a critical appraisal of the system of evaluation prevailing in Pakistani schools. The defects in the existing system are pointed out by making use of internal and external evidence. Internal evidence is provided by a close study of

the mathematics examination papers of the Secondary Board of Education, Lahore and the teacher-made tests presently in use in Pakistani schools. The Board tests and the teacher-made tests, judged by the modern standards available in the literature on educational evaluation, have been found to be defective in many ways. It has been shown that the present tests (1) are poorly organised, (2) offer a liberal choice of questions, (3) encourage rote memorisation, (4) involve time-consuming calculations, (5) contain questions constructed around more than one objective, (6) lack in reliability due to poor representation of content coverage, (7) are difficult to score, (8) have low discriminating power, (9) employ a defective system of grading, and (10) do not attempt to use results to improve instruction. External evidence of the existence of these short-comings has been sought through a questionnaire sent to twelve eminent mathematicians in Pakistan, each of whom is associated with the evaluation of mathematics achievement in Secondary schools.

The third chapter sets forth modern concepts of evaluation. It stresses the need for keeping the objectives of the course vividly in mind when constructing tests, and discusses at some length the criteria for a good achievement test.

The fourth chapter goes into the problem of constructing objective achievement tests, and shows the process of test construction through the following three steps:

(i) Planning the test

(ii) Item writing

(iii) Organising, administring, scoring and interpreting.

The fifth chapter presents a practical example of constructing, administring and analysing a test in algebra. The test was designed by the investigator to measure achievement at the end of a first year course in algebra, and was administered in four schools in Beirut. The main object of the chapter is to illustrate for the guidance of the Pakistani teachers the entire process of test development in a specific situation.

The sixth chapter makes recommendations for the improvement of the present system of evaluation. It contains particular recommendations relating separately to the Board tests and to teacher-made tests. A plan has been suggested to help introduce better achievement tests within the limitations of the present system of education, and its anticipated outcomes are examined.

TABLE OF CONTENTS

Chapter Page

Chapter

CHAPTER I.

INTRODUCTION

The Problem and its Significance

The present study arises from a perception of defects in the present system of evaluating mathematics achievement in the secondary schools affiliated to the Secondary Board of Education, Lahore. The Commission on National Education voices dissatisfaction with the current evaluative techniques, as follows:

> The present system of evaluation is confined to the intellectual attainment and is based on written examinations conducted by the schools at the end of every term... These are hardly more than formalities, and are not taken very seriously by either the pupil or the teacher, and there is no attempt to base promotion through the school on an objective and comprehensive assessment of the work done throughout the year.[1]

The report also states:

> On the other hand the public examinations (conducted by the Education Board)[2] held at the end of classes X and XII are taken too seriously and absorb the entire attention of the teachers and the school authorities. Their short-comings are notorious: they consist solely of written tests in which success can be achieved through mere memorisation and practically no effort is made to test the pupils' intelligence...[3]

Whatever is generally stated in the above citations by way of criticism of the overall system of evaluation existing in Pakistani schools is particularly true in the special case of the evaluation of

[1] Government of Pakistan, Report of the Commission on National Education,(Karachi, Ministry of Education, 1960) pp 122-123

[2] The remark in parenthesis is from the writer.

[3] Ibid., P.123

mathematics achievement. The conventional essay test, which is the sole technique employed to evaluate mathematics achievement, measures primarily the ability of the examinees to perform mechanical tasks in the basic skills. The tests are made up of a few questions taken from the prescribed courses of study. They fail to test the total achievement or understanding, but, instead, encourage the pupils to reproduce memorised materials, which are not necessarily understood.

The present day system of evaluation does not exploit the use of testing to improve instruction or to shed light on individual differences in the classroom situation. The mathematics teachers continue to teach without gaining knowledge of individual difficulties. The result is that thousands of pupils fail for not receiving the particular assistance they need in school. It is seldom realised that the failure of pupils may be due to the failure of the teacher to diagnose, by means of tests, the causes of unsatisfactory progress.

The problem, then, is to examine critically the mathematics tests (with special emphasis[4] on the Board tests) current in Pakistani schools and to make proposals for their modification and improvement.

Objectives of the Study

The foregoing pages contain implicitly the purpose of the present study. For the sake of clarity, however, the various objectives may be explicitly enumerated as follows:

[4] Such an emphasis is justified on the ground that the Board tests set a pattern for classroom mathematics teachers in various schools.

(i) To describe modern concepts of evaluation in as simple and direct a manner as possible. This may serve as a guide for evolving better evaluative techniques.

(ii) To acquaint the mathematics teachers in Pakistan with objective techniques of evaluation.

(iii) To bring to light some glaring defects in the traditional method of evaluation, which is presently in use in Pakistan.

(iv) To illustrate in some detail the actual process of evaluation by constructing an algebra test, administering it, and following through with a detailed analysis of results.

(v) To explain the use of statistical methods employed in scoring and in interpreting results.

(vi) To help, indirectly, in evolving similarly improved evaluative techniques in fields other than mathematics.

Reasons for Selecting the Problem

The present study has been centred on the field of mathematics for the following reasons:

(1) It has been proposed to acquaint the teachers in Pakistan with objective techniques of testing. In the field of mathematics, this can be achieved with less difficulty because, according to Cliff, (i) the evaluation of mathematics achievement lends itself rather well to objective testing; (ii) there is nothing new in the objective type test, the oldest type being the completion type (e.g., 7x8 = ?) to which have been added true-false, matching, multiple-choice and other

types of items.[5]

(2) Training mathematics teacher in the use of statistical methods to interpret results has a greater chance of success because mathematics teachers are likely to learn statistical concepts more quickly than teachers of other subjects.

(3) The writer, being himself a mathematics teacher, is in a better position to handle this problem than a problem of evaluation of achievement in a field other than mathematics.

Method of Study

The methods of the present investigation consist of critical analysis of the data furnished by the following sources:

(1) Mathematics question papers of the Secondary Board Examination.

(2) Mathematics question papers of the internal examinations conducted by the schools.

(3) An objective test in algebra, constructed, administered scored, and analyzed by the investigator in Beirut schools.

(4) Literature pertaining to evaluation in education.

(5) A questionnaire sent to some eminent mathematicians presently associated with evaluating mathematics achievement in the

[5] Marian C. Cliff, "The Place of Evaluation in the Secondary School Program", The Mathematics Teacher, XLIX (April, 1956), p272

secondary schools of Pakistan.[6]

Limitations

(1) The investigator being in Beirut, the test in algebra could not be administered in Pakistan.

(2) The present study is a foundation and a beginning only. No standardised tests have yet been produced in Pakistan.

Definition of Terms:

(1) Board stands for the Board of Secondary Education, Lahore.

(2) Pakistani Schools should be taken to mean those Secondary Schools affiliated to the Board.

(3) Pakistani students, Pakistani teachers etc, should be interpreted to mean students and teachers in schools affiliated to the Board.

[6] Twelve mathematicians known to the investigator to be eminent in the field of evaluating mathematics achievement were asked to express their views on the present tests by answering a questionnaire. These mathematicians include Paper - Setters, Head Examiners, and Sub-Examiners of the Board, as well as mathematics teachers in institutions of repute. Seven of them responded. The questionnaire may be seen in Appendix A

CHAPTER II

CRITICAL APPRAISAL OF THE SYSTEM OF EVALUATION PREVAILING IN PAKISTANI SCHOOLS

This chapter points out some glaring defects in the present system of evaluating mathematics achievement in Pakistani schools.

Internal and external evidence of the existence of defects in the system are presented. Internal evidence is sought by studying the examination questions - teacher-made tests as well as Board tests. External evidence is collected through eliciting the opinions of some eminent mathematicians of Pakistan who have been associated with evaluating mathematics achievements.[1]

Similarities Between the Board Tests and Teacher-Made Tests

The Board tests and the teacher-made tests are similar in the following ways:

(1) Both are of the conventional essay type.

(2) The questions contained in the teacher-made tests are usually on the pattern of the Board questions. This is so because the classroom teachers are conscious that the hurdle which the students must eventually clear is the Board examination. The teacher-made tests very

[1] Opinions elicited by means of a questionnaire given in appendix A

often contain a number of questions directly lifted from the Board examination papers of the preceding few years. Classroom instruction is geared to serve the end of passing the Board Examinations.[2] The Pakistani teachers devote greater classroom time and attention to those topics which usually appear in the Board Examinations.

(3) The organization of teacher-made tests is similar to the Board tests in the following ways:

(i) Both include optional questions

(ii) The questions on both are arranged in the sequence of topics included in the syllabus (rather than in the order of difficulty level of the questions, or any other order).

(4) Both follow the same grading system i.e., the percentage system. The individual score according to this system, is interpreted in terms of comparison with the highest possible score on the test.

(5) The minimum pass mark for both the teacher-made and the Board, tests is 33 per cent.

(6) After the administration of test, no attempt is made in either case to analyse the test results so as to appraise the strengths or weaknesses in the test itself.

[2] Government of Pakistan, Report of the Commission on National Education, (Karachi, Ministry of Education, 1960), p.123

Thus both the Board tests and the teacher-made tests have a great deal in common and share common defects. Hence the analysis of the Board tests which follows is essentially applicable to the teacher-made tests as well.

The Board Examination

The Board Examination in mathematics consists of two written papers of three hours each as follows:-[3]

		Marks
Paper A	Arithmetic	50
	Algebra	50
Paper B	Geometry	100

The arithmetic section of Paper A generally contains six questions. The first question is in three parts, and each of the remaining in two parts. In all then the number of arithmetic problems is thirteen. The directions on the top of the paper read as follows:

> "Attempt any two parts from question No.1 and any six parts from the remaining five questions 2 to 6"[4]

In other words, the examinees have to do eight questions out of thirteen. They may omit one part of the first question and five parts from the remaining.

The algebra section usually contains six questions. The first question in this section (being in fact the seventh on the test) is

[3] Board of Secondary Education, Lahore, Secondary School Examination, 1961-62 (Syllabus), p.57

[4] This statement of direction is the same word by word on the question papers of the Board for the years 1961, and 1962.

divided into three parts and the rest of the questions into two parts each. The directions for this section run as follows:

> "Attempt any two parts from question 7 and any six parts from the remaining five questions 8 to 12" [5]

This section also contains thirteen problems and the examinee is expected to select eight of them to obtain the maximum score possible on the section.

The question paper on geometry usually contains ten questions. The directions generally read as follows:

> Attempt seven questions in all, including questions 1 and 2 which are compulsory.

The first question, which is compulsory, has an alternative; the second question is in five parts, out of which the candidate is required to do only three. The first and the second questions are invariably on practical geometry and deal with geometrical constructions. From the remaining eight questions, each of which is in two parts (a) and (b), the students are required to do five questions. The (a) parts of the questions from 3 to 10 are all statements of theorems lifted verbatim from geometry text books, and the students are required to prove them. The (b) parts of these questions are problems which can be solved by the application of the theorems in the (a) parts.

[5] The "direction" is taken from the question paper of the Board for the year 1962

Examples:[6]

6 (a) Prove that equal triangles on equal bases are of the same altitude.

6 (b) Two equal triangles have been drawn on the opposite sides of the same base, prove that the base bisects the line joining their vertices.

General Objections Relating to the Structure of the Tests

There are two relevant objections which can be raised about the structure of the tests.

(1) Choice of questions: The foregoing discussion clearly reveals that the test constructors offer a liberal choice of questions to be answered. Since the achievement of the objectives of a common program of study is to be evaluated through the examination, it is necessary that each examinee is made to answer the same questions. "Giving a choice of questions reduces the common base upon which different individuals may be compared."[7] The use of optional questions does not find favour with Stalnaker who writes:

> ...several studies have shown that optional questions complicate measurement and introduce factors of judgement which are extraneous to the ability being measured. For sound sampling, it is recommended that optional questions be avoided and that all

[6] Questions 6(a) and 6(b) are taken from the Board Examination paper for 1962.

[7] Robert L. Thorndike & Elizabeth Hagen, Measurement and Evaluation in Psychology and Education, (N. York, John Wiley and Sons Inc., 1961) p.55

examinees be asked to run the same race.[8]

Another drawback in offering choices of questions is that students fall into the habit of not preparing certain difficult portions of the course in the hope that they will be able to avoid questions on such portions by making use of the choice offered by the examiners. As the pass mark is 33 per cent, a student is tempted to ignore about two-thirds of the course material in the hope that he will pass via the choices offered. It is, in any event, a common practice to omit the study of hard parts of a course.

(2) Sequential organisation of the tests: A closer look at the layout of teacher-made tests as well as the Board tests indicates that the questions are usually organised in the sequence of the topics found in the syllabus and are not arranged in the order of increasing difficulty - i.e., from easy through moderately difficult to most difficult. The examinees may become discouraged and confused when confronted with more difficult questions in the beginning of the test and may not be able to exhibit their full achievement. While giving suggestions for improving the essay type examination, Remmers and Gage, express the view that "the questions should be arranged in order of difficulty."[9]

[8] J. M. Stalnaker, "The Essay Type Examination", Educational Measurement, ed. E.F. Lindquist. (Washington D.C., 1951), p.506.

[9] H. H. Remmers and N.L. Gage, Educational Measurement and Evaluation, (N. York, Harper & Brothers, 1955), p 184

Specific Objections Relating to Content and Type of Questions

(1) Rote learning is encouraged

The arithmetic section of the Board Examination contains a number of typical questions, the solution of which does not require real understanding. The examinees memorise the steps and mechanically arrive at the correct answer. Certain typical questions, having a set pattern, are repeated year after year in the Board Examination. The candidates are thus encouraged to learn the steps by rote, and the teachers frequently drill and coach for these specific questions.

Examples:

Arithmetic. Board Examination, 1961

(i) Find by practice the cost of 2 tons, 5 cwt. and 21 lbs of coal at the rate of Rs.4, annas 8 per cwt.
(1 ton = 20 cwt., 1 cwt = 112 lbs)

(ii) Find the day of the week on the 19th June, 1940

Questions exactly on the above pattern appeared, as if by clockwork, in the Board Examination paper for the year 1962. Examples:

(i) Find by practice the cost of 3 tons 6 cwt., and 24 lbs of coal at Rs 70 per ton (1 ton = 20 cwt, and 1 cwt = 112 lbs).

(ii) Find the day of the week on the 5th May, 1920.

The algebra section, like the arithmetic section, contains questions of fixed patterns repeated every year, thereby encouraging the candidates to memorise the process of their mechanical solution.

Examples:

Algebra. Board Examination 1961

(i) Find the H.C.F. and L.C.M of
$2a^4 + 3a^3 - 2a^2 - 2a + 1$ and
$2a^4 + a^3 - 4a^2 + 1$

Algebra. Board Examination 1962

(ii) Find the H.C.F. and L.C.M. of
$n^4 + n^3 + 2n - 4$ and
$n^3 + 3n^2 - 4$

The geometry test provides the students a still greater opportunity of cramming. The geometry question paper contains eight statements of theorems lifted verbatim from the text books. These theorems are offered as parts of questions, and the candidates are required to prove five of them. The theorems carry a weight of 35 per cent of the geometry part test. The examinees are, therefore, tempted to memorise all the theorems verbatim to ensure 35 per cent coverage. (only 33 per cent marks are needed to make the pass grade.)

It is interesting to point out that the Board in its rules for paper-setters explicitly lays down:

> Questions should aim at testing the ability of a candidate (a) to understand a topic (b) to apply his knowledge to solve practical problems and not merely testing his ability to reproduce the answers which have been given to him in notes or learnt by him from a book.[10]

But the paper-setters appear to pay no heed to this injunction and the examination, instead of motivating the students to develop an understanding of mathematical concepts, applications and process, encourages rote learning of set methods and steps which lend themselves to

[10] Board of Secondary Education, The Calendar 1959-61 (Lahore, 1958), p 123

unthinking, mechanical reproduction by students.

That the students reproduce crammed material in the examination is borne out by the opinions of some Examiners in mathematics:

> (i) "A candidate can easily get pass marks by cramming theorems and construction propositions."
>
> (ii) "The candidates mostly get through by cram work. The knowledge gained is mostly applied mechanically."
>
> (iii) "Candidates are not well-versed in mathematics. It appears that they learn much by rote."[11]

Judged by the criterion which is being presented below, the mathematics tests used in Pakistani schools can in no way be considered of high quality:

> Good testing..., should penalise rote learning rather than put a premium upon it... It should be the teacher's objective in test construction so to phrase or present the questions and responses that only a genuine understanding of the concept will enable the student to respond correctly.[12]

(2) Questions involve time-consuming calculations

Some questions in arithmetic require time-consuming calculations. Example:[13]

What sum of money increased by 27½ per cent becomes

₺1 1S 2d.

The aim of the evaluator is to test the understanding of the

[11] The statements are edited but substantially accurate extracts from the replies of Examiners to the questionnaire given in appendix A

[12] Herald E. Hawkes, E.F. Lindquist and C.R. Mann, The Construction and use of Achievement Examination. (Boston Miffin & Co., 1936), p.95

[13] The question is taken from a teacher-made test.

principle of percentage, but in working out the problem the examinee will have to go through a labyrinth of calculations. It is quite possible that even if the student understands the principle he might get lost in the maze of arithmetical calculations.

Another Example:[14]

$$\text{Solve: } \frac{1/3 + \frac{1}{4} + 1/5 + 1/6}{1/3 - \frac{1}{4} + 1/5 + 1/6} \times 4\ 2/3 \div \frac{1/15 + 1/20 + 1/24}{\frac{1}{2} - 1/3 + \frac{1}{4} - 1/5 - 1/6}$$

Here the objective of the examiner seems to be to test the ability of the candidates in carrying out fundamental operations with vulgar fractions. But the question framed is too involved. With this type of question, not only will the examiners vary in their judgements, if marks are to be awarded for steps, but they will also not be able to diagnose the specific weaknesses of individual students. The question could have been better split into several shorter questions which could be marked objectively and would, at the same time, have provided better opportunities for the teacher to diagnose individual deficiencies.

(3) Questions are constructed around more than one objective

"Each question in an essay test should be planned to measure one defined objective of instruction."[15] This rule is not always observed by the test-makers. Aiming, by means of one problem, to test

[14] The question is taken from a teacher-made test.

[15] J. W. Wrightson, Joseph Justman and Irving Robbins, Evaluation in Modern Education, (N. York, American Book Company, 1956) p.104.

the knowledge of a series of rules and formulas, the examiners, sometimes, create such a complex structure that the problems can only be said to have an artificial and confusing form. In their substance the problems are sometimes actually more elementary than problems which are briefly stated. The problem [16] below is constructed around more than one objective:

$$\text{If } x = 3 + \sqrt{8} \quad \text{find the value of } x^4 + \frac{1}{x^4}$$

This problem would require the student to find the value of $\frac{1}{x}$ by rationalising the denominator. Then x and $\frac{1}{x}$ are to be added to obtain the following equation:

$$x + \frac{1}{x} = 6$$

The next step will be to square the two sides, thus getting the result

$$x^2 + \frac{1}{x^2} + 2 = 36$$

And then by transposition, the result is to be stated as

$$x^2 + \frac{1}{x^2} = 34$$

Finally the student has to square this result to obtain the following result:

$$x^4 + \frac{1}{x^4} + 2 = 1156$$

and then by transposition to arrive at the desired answer:

[16] The problem is taken from the Board Examination Paper for 1958.

$$x^4 + \frac{1}{x^4} = 1154$$

The objectives around which this problem is constructed may be analysed as follows:

(i) Knowledge of rationalising the denominator

(ii) Ability to add surds

(iii) Knowledge of the principle of equation

(iv) Ability to use the formula $(a + b)^2 = a^2 + b^2 + 2\ ab$

Instead of stating the problem in a complicated form, the achievement of the above objectives could be tested directly and briefly by the use of short-answer completion items stated below:

(i) <u>Rationalising the denominator</u>:

when $x = 3 + \sqrt{8}$

$\frac{1}{x} =$ ________ (write the answer after rationalising the denominator)

(ii) <u>Adding surds</u>

If $x = 5 + \sqrt{6}$ and $\frac{1}{x} = 5 - \sqrt{6}$

$x + \frac{1}{x} =$ ________

(iii) <u>Use of formula</u>

Expanding $(x + \frac{1}{x})^2$ we get ________

(iv) <u>Principle of equation</u>

If $x^2 + \frac{1}{x^2} + 4 = 18$

$x^2 + \frac{1}{x^2} =$ ________

Another Example:[17]

> The cost of turfing a field at the rate of Re 1 8as per sq. yd. is Rs. 144. The length is twice the breadth of the field. Find the cost of fencing the field at the rate of Rs. 2 8 as. per yd.

To begin with, the problem contains a definite ambiguity, because it is not explicitly stated what shape the field has. Is it of the shape of a parallelogram or a rectangle? It is probably assumed by most candidates that by "field" the examiner means a rectangular field.

To arrive at the correct answer (after assuming a rectangular field) the examinee has to go through a number of steps which may be enumerated in order:

(i) Finding the area of the field by using the unitary method--the objective is the knowledge of the unitary method.

(ii) Finding the breadth of the field by working out the square root of half of the area and then doubling the breadth to find the length of the field--the objective is to test the ability of the student to find a square root.

(iii) Finding the sum of four sides--the objective is to test the knowledge of the concept of perimeter.

(iv) Finding the cost of fencing--the objective is to test the knowledge of the unitary method or proportion.

Suppose now that a student has a full grasp of the unitary method or proportion, and that he also has a clear understanding of

[17] The problem is taken from the Board Examination Paper for 1959.

the concept of perimeter, but he lacks the ability to find the square root. Such a student will surely fail to do this problem, and will receive no credit for his achievement of two of the main objectives which this problem sets forth to test. The scores of pupils on tests which contain problems scored as a unit, but which are constructed around more than one objective can hardly be considered valid.

(4) Reliability

A good test must measure accurately and consistently what it is supposed to measure. The consistency with which a test measures is called its reliability. The reliability of the tests presently in use in Pakistani schools can be investigated by getting answer books of a set of students scored by different examiners. Such an investigation, though worth while, could not be made by the investigator, because of his being outside Pakistan. It is, however, presumed (and it is believed that any later investigation will support the presumption) that the current tests are low in reliability for two reasons:

(i) Narrow sampling of content

(ii) Subjectivity in scoring.

A study of the Board tests and teacher-made tests readily reveals that the sampling of subject matter is narrow. Take, for example, the geometry question paper of the Board for 1962. The question paper contains eight theorems and their applications, out of which five theorems and their applications are to be done by the candidates. The geometry syllabus which the students are supposed to cover contains about forty three theorems.[18] Clearly the test is a

[18] Board of Secondary Education, Lahore, Secondary School Examination 1961 and 1962 (Syllabus), (Worldlight Press Dil Mohd Road, Lahore) pp 59-62

poor representative of the total population of questions which might be asked. In testing the students by means of such a narrow sample of subject matter (about one-eighth of the syllabus), there exists the real danger that chance variations in the learning-testing situations will have a maximum effect on test scores.

Again, the achievement tests, being of the conventional essay type, are liable to be scored subjectively, thereby diminishing reliability. Numerous studies have shown that teachers do not agree at all well as to the grades to be assigned to examination papers of the essay type. Studies have also shown that individual teachers cannot even agree well with their own prior judgments when they assign a second series of grades to the same papers. A scoring range, for example, has been found from 2 out of 20 to 19 out of 20 when a single answer on an essay question was scored by different examiners. A difference of 25 points (50 to 75) was found when the same answer was marked by the same teacher after an interval of two months.[19] Another example of the subjectivity in scoring in essay type test is reported by Traxler and others:

> ...one geometry instructor had reproduced a geometry test paper handed in by one of his pupils and sent it to many mathematics teachers with the request that they rate it on a scale of 100 points. The paper came back with grades ranging from 10 to 90.[20]

[19] R. L. Thorndike & Elizabeth Hagen, Measurement and Evaluation in Psychology and Education, (New York, John Wiley and Sons Inc., 1961) pp 45-46.

[20] Arthur E. Traxler el al, Introduction to Testing and the Use of Test Results in Public Schools, (New York, Harper Brothers, 1953), p.8

Yet another experiment is reported by Orleans and Sealy in which 37 teachers marked an arithmetic answer paper and the scores in percentage varied from 45 to 75. [21]

(5) Discrimination

As the present day achievement tests do not adequately cover the content, they are very likely to be lacking in discriminating power. The following statement[22] of an examiner helps to draw some important inferences relating the discriminating quality of the tests. He writes:

> In one of the annual examinations (the Board Examination), I was the Head-examiner in mathematics. I prepared the following list from the scores of about 6,000 answer books:
>
> Percent of students obtaining zero score 5%
>
> . . . .
>
> . . . .
>
> Percent of students obtaining perfect scores of 100) 4%

It is clear from the above that this particular test to which the figures relate failed to discriminate among the topmost and the bottom-most examinees. The perfect score recorded by the topmost 240 students (4% of 6,000) indicates that they have been denied the opportunity to show the full extent of the differences in their achievements. The zero score obtained by 300 students (5% of 6,000) shows that the test failed to reveal any differences of achievement among

[21] J. S. Orleans and Glenn A. Sealy, *Objective Tests,* (New York, World Book Co., 1928), p. 17-18.

[22] Extract from a statement received by the investigator in reply to the questionnaire in appendix *A*

the 300 poor pupils. One is tempted to wonder at the shape of a distribution in which 9% of the candidates are concentrated at the two extreme scores. There is almost unlimited evidence that achievement distributions if derived from accurate tests, seldom show end-concentration, and never show both-end concentration.

(6) Scoring

Under the present system the scoring of answer papers is tedious and takes a considerable time. To the question: "Do you consider the scoring of answer papers arduous?"[23] All the respondents replied in the affirmative. One of the respondents pointed out:

> I find it extremely hard to mark 400 answer papers in a limited time of 24 days especially so when I have to attend to the usual classwork as well. The present system needs improvement.[24]

Some Other Objections

(1) Grading system

The present system of grading is far from satisfactory. The Board and the classroom teachers interpret the scores of students as a percentage. The following regulation of the Board lays down the grading procedure:

> Successful candidates who gain sixty per cent of the aggregate number of marks or more shall be placed in the First Division, those who gain not less than forty five per cent

[23] See appendix A

[24] The time limit is imposed by rule 3.1 contained in Book of Instructions for Examiners, (Board of Secondary Education, Lahore, 1962), p.14

> in the Second Division and all below in the Third Division.[25] Pass mark in a subject is 33 per cent.[26]

The present practice of interpreting individual scores in terms of percentages is defective on various grounds:

(i) The score in the percentage form is not meaningful. A pupil's 100 per cent score can hardly indicate a state of perfection, and the score of zero per cent does not necessarily imply a complete lack of knowledge. The score in this form, on the other hand, may be misleading. The teacher instead of understanding what the pupil _did_ on the test may wrongly take it to mean what he _can_ do. For example, a teacher may erroneously infer that a student who gets 70 per cent marks on a test has mastered 70 per cent of the material covered in the course. In making such a decision the teacher ignores some important factors which exercise their effect on the score, namely the difficulty level of the test, the effect of conditions in which the test is administered, the degree of subjectivity that enters into constructing and scoring, and the like.

(ii) Some arbitrary percentage bands have been chosen--33 per cent as the passing grade, 45 per cent and above Second Division and so on--which are unalterable and bear no relation to the actual difficulty levels of the tests.

[25] Board of Secondary Education, Lahore, _Secondary School Examination 1961 & 1962 (Syllabus)_, p.5

[26] _Ibid_, p.11

Furthermore, these arbitrary bands have been left undefined. An unfortunate student who gets a score of 59 per cent falls in the arbitrary band of Second Division, but his lucky companion who manages to get a score of 60 per cent is placed in the First Division. Whether the pupil in the Second Division is inferior in achievement to the one in the First Division is doubtful. No present-day test is so accurate as to measure absolutely such small differences of achievement. Again, the tests being of the conventional essay type, the nominal difference could easily be the result of subjective judgment of the examiner. The arbitrary bands, thus, present a distorted picture of the achievement of the individuals.

(iii) Averaging per cents over two or more tests can reverse the actual and true relative standings. To see this problem clearly, consider the data in the following table:

Table Showing Relative Standing of Pupils

Tests	Mean	Standard Deviation		Scores for	
				Student Y	Student Z
A	55	5		50	65
B	50	12		74	44
			Total	124	109
			Average	62	54.5

It can be easily noticed from the table that under the present per cent grading practices, student Y will receive a First Division and will be considered superior to student Z who will be placed in the Second Division. But in terms of relative standing student Z is superior to student Y. This can be shown by changing the raw scores to standard scores,[27] in the following manner:

Students	Standard Score on Test A	Standard Score on Test B	Total
Y	$\frac{50 - 55}{5} = -1$	$\frac{74-50}{12} = +2$	+ 1
Z	$\frac{65-55}{5} = +2$	$\frac{44-50}{12} = -.5$	+ 1.5

The per cent grading system made use of by the classroom teachers and the Board, thus, gives a false impression of the relative position of students as it does not take into account the spread of scores on various tests. In effect, this oversight gives high weights to tests having much score variation, even though the computations appear to weigh the test equally.

(iv) It is extremely difficult to build into a test exactly the difficulty level desired. Suppose a student gets 70 per cent of the highest score possible on a test. It is highly improbable that the same student would get 70 per cent of the highest possible score on a second test constructed to measure the same objectives.

(v) The passing grade is 33 per cent. It has been shown in the foregoing pages that a passing grade can often be obtained

[27] For a discussion on standard scores please see chapter IV.

by the examinees by memorising set portions of the course and by attaining a mastery in some typical processes often repeatedly asked in the examination.

(2) Use of test results

Results of the Board tests or teacher-made tests, under the present system of evaluation, are not being used to further certain important functions like instructional facilitation, pupil guidance, improving the tests themselves etc. The National Commission on Education, having realised this weakness in the existing system, stresses the need of subjecting the test results to a critical analysis:

> It is essential that completed examinations should themselves be subjected to critical analysis so as to find out whether they have properly tested what the examinees have been taught on the basis of prescribed curriculum, how far they have tested memory and understanding, and what were the factors contributing to failure or success.[28]

The report of the National Commission on Education further adds:

> We consider that the Board of Secondary Education should make arrangements for evaluating their examinations and set up permanant sections in their offices for this purpose.[29]

Data Collected Through Questionaire

Answers to the questionnaire further support the evidence that the system of evaluating mathematics achievement followed in Pakistani

[28] Government of Pakistan, Report of the Commission on National Education, (Karachi, Ministry of Education, 1960) p.124.

[29] Ibid., p.124

schools is defective. The seven respondents agree that the mathematics achievement tests (i) are not well planned; (ii) do not contain a liberal sampling of all phases of subject matter, being confined to a small number of problems; (iii) encourage memorising of theorems in geometry; (iv) encourage rote learning of typical processes in arithmetic and algebra;(v) are likely to yield varying scores if marked by different examiners; (vi) are arduous to score (vii) contain some questions which involve long and time-consuming computations (viii) do not take into account individual differences; (ix) make it possible for some students to get a pass mark by mastering some set parts of the course by rote memorisation, without understanding.

While indicating the purposes of testing as they <u>exist</u> at the present day and <u>not</u> as they <u>ought to be</u>, none of the respondents indicated that tests were being used for "identifying trouble spots and difficulties of students," which leads to the conclusion that the most important object of testing-to guide learning and teaching--is being overlooked entirely.

All the respondents support the idea of using tests which have a large number of short items such as the widely successful multiple-choice forms, covering wider ground, but stressing understanding of method and principles, and giving less weight to detailed calculations. They suggest the introduction of the new types of tests which "include Yes or No type, filling up blanks, etc".[30]

[30] The words in the quotation form part of the answer of a respondent to querry 17 on the questionnaire given in appendix <u>A</u>

CHAPTER III

MODERN CONCEPTS OF EVALUATION

INTRODUCTION

The foregoing chapter discussed at some length the defects in the existing system of evaluation prevailing in Pakistan. To remove these defects and to put the evaluative process on a sound basis it is necessary that the teachers in Pakistan have a clear and vivid understanding of the modern concept of evaluation. To explain the concept three questions will be answered:

1. What is evaluation? 2. Why is evaluation needed?
3. How is evaluation done?

The answer to the first question requires discussion of the nature of evaluation, explaining the steps in the process and the mentioning of the various evaluative techniques, of which the present investigation will restrict itself to a detailed study of achievement test as an instrument of evaluation. This is being done because the problem is confined to achievement tests, though it is realised that other instruments of appraisal are equally important. To answer the second question the purposes of evaluation will be mentioned and the role of evaluation in education will be discussed. The answer to the third question will form part of the next chapter where the present study will go into the intricacies of constructing, administring and scoring an achievement test and shall as well discuss how the scores should be interpreted.

Nature of Evaluation

Evaluation has been described by various writers in various ways. In the words of Travers:

> The evaluation of education in terms of outcomes is not merely a process of determining what the actual outcomes are, but it also involves a judgment of the desirability of whatever outcomes are demonstrated to occur... The process of making an evaluation consists not in measuring the changes but in judging whether the change is or is not desirable.[1]

Remmers and others explain the concept by bringing out a difference between measurement and evaluation:

> Measurement refers to observations that can be expressed quantitatively and answers the question "how much". Evaluation goes beyond the statement of "how much" to concern itself with the question of "what value"... Evaluation, therefore, presupposes a definition of goals to be reached--objectives that have been set forth.[2]

To Victor H. Noll, the important feature of the evaluative process is to determine the progress of the learner with respect to a desired goal by using techniques that can be depended upon.[3]

[1] Robert M. W. Travers. Educational Measurement (New York: The Macmillan Co., 1955), p.7

[2] H. H. Remmers, N.L. Gage, and J. Francis Rummel, A Practical Introduction to Measurement and Evaluation, (New York: Harper & Brothers, 1960), pp 7-8

[3] Victor H. Noll, Introduction to Educational Measurement Boston: Houghton Mifflin Co., 1957) p.13.

Micheels and Karnes point out the comprehensive nature of the evaluative process.

According to them:

> ...you will be evaluating when you take into consideration many factors that are inherent in student growth--proper attitudes towards others, safety habits, manipulative skills, acquisition of knowledge, appreciation, understanding and the like.[4]

The extracts quoted and the views presented above cover almost all the aspects involved in the concept. As a useful step we may separate the compound statements into simple expressions as follows:

(1) Evaluation presupposes objectives. We evaluate when we try to find out how far the behaviour of a pupil conforms to certain objectives.

(2) Evaluation is a comprehensive process as it deals not merely with achievement, but with all aspects of the pupil--attitudes, appreciations, safety habits, understandings.

(3) Evaluation is qualitative in nature because it involves a value judgment - achievement is to be a desirable achievement and is to be judged against a norm.

(4) The process of evaluation is also quantitative in the sense that it measures as precisely as possible the trait or traits it is supposed to measure.

(5) Evaluation is closely knit with the aim of education. It forms part of the educational fabric. The purpose of education is to

[4] William J. Micheels and M. Ray Karnes, <u>Measuring Educational Achievement</u> (McGraw--Hill Book Company Inc., 1950), p.23.

change students from a given state of experience to a desired state beyond the given one by means of learning experiences. The purpose of evaluation is to measure and to determine whether any change has occured in the students.

Steps in Evaluation

Evaluation proceeds through a sequence of steps. Three definite steps can be pointed out:

(1) Adequate formulation of objectives for evaluation purposes.

(2) Selection and use of evaluative techniques.

(3) Recording data obtained through evaluative techniques and analysing and interpreting it.

Formulating Objectives

Objectives are the goals in the direction of which the curriculum seeks to change pupils. An objective, thus, is a normative concept carrying implications of the good and the desirable.

The formulation of objectives is, perhaps, the most important of all tasks in evaluation. Unfortunately many mathematics teachers in Pakistan do not much care about the objectives of the course when constructing a test. Some of the ideas which may prove of value in the formulation of objectives will be discussed below:

(i) Specifying content and behavior. While formulating objectives their two-dimensional aspect may be kept in mind. The most familiar dimension of a set of objectives **is** the specification of the subject matter. These objectives consist of items of content and technical terms or names of concepts such as exponents, square root, factors; or facts and theorems to be remembered; or formulas to be recalled. The second dimension is related to the behaviour, the student is expected to learn. The behaviour description uses such terms as recall, understand, apply, analyse and the like.

(ii) Establishing relationship between content and behaviour. The task of formulation of objectives involves not only the specification of content and behaviour elements, but also of establishing relations between the two elements. As an example consider the following objectives:

(i) understanding the use of formulas in algebra.

(ii) competence in the use of formulas in algebra.

Although the specification of content for both the objectives is the same, the behaviour to be tested will be different in each case.

(iii) Objectives in terms of pupil changes. Objectives should be worded in terms of changes expected in the pupil rather than as duties of a teacher, since attainment of objectives must in any case be evaluated in terms of pupil changes. The difference between the two ways of stating objectives may be seen from the following illustration.

Teacher's objectives: To teach the knowledge
skills, habits involving the multiplication
in arithmetic

Pupil's objectives:

Knowledge	-	To know the multiplication table
Skill	-	To be able to apply multiplication table
Habit	-	To use small numbers as multipliers.

(iv) As far as possible objectives should be specific. They should be put in terms of observable changes in the pupil between the beginning and end of his experience in a defined segment of the educative process. Objectives that are stated at a level of generality can be made specific by analysing behaviour associated with it. For example, in the objective, 'appreciation of mathematics', we can replace 'appreciate' by terms like 'wants more of' or 'wants to know more about'. Only when the objectives are specifically formulated can the teacher decide to teach for them and evaluate achievement of them.

(v) Each statement should contain one objective only, to prevent confusion and facilitate identification of the objective.

Selection of Techniques.

The next step in the evaluative process is the selection and use of a suitable technique or techniques to measure the desired behaviour. These techniques may be classified as follows:

i) Achievement tests

ii) Scholastic-aptitude tests (Intelligence Tests)

iii) Special-aptitude tests

iv) Interest Inventories

v) Character or personality instruments.

Besides these techniques, some other devices such as interview, observation, parents' reports and the like may be resorted to.

An evaluator's job is to select and use a measurement technique or techniques that are optimally effective. Up to the present time the last four of the five techniques mentioned above have not been developed in Pakistan, nor do the teachers make sufficient use of the devices of observation and interview. In these conditions, the teachers in secondary schools of Pakistan depend heavily upon achievement tests for purposes of evaluation. It is fully realised that an evaluation program, to be comprehensive, should employ a variety of techniques and devices. It is hoped that efforts to develop other techniques besides achievement tests will be made by educators in Pakistan. Meanwhile, an attempt is being made at improving the present day achievement tests. It seems worth-while to describe what an achievement test is under the modern concept of evaluation and to explain at some length the criteria for a good achievement test.

Achievement Test Defined. An achievement test is an instrument designed to measure the extent to which any specified objective has been attained. There are two main types--general achievement tests and diagonostic tests. There is no fine line of demarcation between the two. The general achievement test, as the name implies, samples the entire field of work being tested and yields a single score indicating a pupil's achievement. A diagonostic test is designed to reveal a person's strengths and weaknesses in one or more areas of the field being tested. It assists the teacher in determining where the learning or teaching has been successful or where it has failed.

Criteria for a Good Achievement Test. An achievement test, to be good and useful, has to conform to some criteria, which according to Micheels and Karnes, may be briefly stated as follows:

(i) A good test must actually measure what it is supposed to measure. (Validity)

(ii) It must do this accurately and consistently. (Reliability)

(iii) It must be fair to the students. (Scoring independent of the personality of the scorer)

(iv) It must pick out the good students from the poor. (Discrimination)

(v) It must be long enough to do the job. (comprehensiveness).

(vi) It should be easy to use (Ease of administration) 5

Reliability

Reliability refers to the consistency with which the test measures. To know whether a test is reliable such questions are raised: How accurate are the measurements? How stable are they? Will the results be approximately the same, if the test is administered immediately again? Statistical procedures to estimate the degree of reliability will be described in the next chapter.

It may, however, be pointed out here that reliability is closely connected with validity of a test. If a test is valid it must have a reasonable degree of reliability. The validity of an evaluative device, after all, is the degree to which it measures what it is intended to measure. Remmers and others break the definition of validity in two parts: (i) "the degree to which it measures" and (ii) "what it is intended to measure." The degree to which a test measures anything, and measures it accurately is the reliability of

[5] Michaels & Karnes, op.cit., pp.103-104

the test. What it is intended to measure is the criterion for relevance of the test.[6]

Thus it is to be remembered that a test might be highly reliable and still not be valid. A test in mathematics, for example, may be valid for a certain grade in arithmetic. In order to be valid it has to be reliable. This same test might also be highly reliable if given several times to a grade in woodwork. But no matter how reliable it might be in measuring the woodwork students, it would not be valid, because it would not be measuring achievement in woodwork. To repeat, a test must be reliable in order to be valid, but a test can be highly reliable without being valid.

There are three important considerations regarding reliability: (i) The reliability of a measuring device depends to a large extent upon the extensiveness of the sample of student experience. For example, a long test usually provides more reliable data than a short test. (ii) Whenever important decisions are to be made, sufficient testing time should be allowed to secure reliable results. (iii) The quality of questions can affect reliability. Ambiguous or vague questions result in lowering the reliability of the evaluative instrument.

Validity

A test is valid when it measures what it purpots to measure. The basic question in validity is how well a test does the job that it

[6] H.H. Remmers and N.L. Gage, Educational Measurement and Evaluation. (Harper & Brothers, N.York, 1955), pp 122-123.

is employed to do. Likewise, a single test item is valid when it does the job expected of it.

Suppose a teacher has devised a mathematics test to measure the ability of his students to apply certain principles of mathematics in solving problems in their daily living. If the actual test measures only the students' ability to recall and write down certain facts on paper, it would not be valid for the purpose specified. There is a tremendous difference between memorising facts and being able to apply them in real situations.

The validity of a test may be affected by other factors, for example: the student may be a poor reader; he may have difficulty in measuring; he may have a limited vocabulary. The total score of such students will not be a valid indication of their real achievement.

Thus it is a complex problem to determine validity in an accurate manner. In fact it is unwise to think about validity in terms of a single statement. "A test may be valid for one purpose but not another."[7] For example, we may develop an arithmetic test that does very well in measuring the achievement of our students. In other words it is a valid test for one purpose. But it does not mean that it will measure equally well the students of another teacher who is teaching in another town. This is because the objectives of the courses may be different. Thus validity of a test is in terms of definite, specified conditions.

[7] Marion Epstein and Sheldon Myers, "How a mathematics Test is Born," The Mathematics Teacher, (1 April, 1958), p.299.

Attempts to determine the validity of an achievement test might follow two methods. In the first instance one has to know beforehand that a certain test was highly valid for the purpose one has in mind. One will then try to develop a second test that is equally valid. The results of the students on the two tests will then be compared. This method is known as the correlation method and yields a correlation coefficient of validity. It is easy to understand this method but the difficulty lies in finding a first test, that is known to be doing a good job. And, if such a test is available, the question arises as to why a new one is needed for the same job. For practical purposes this method is seldom used by teachers. The second method is the one most often used, although in a much more subjective manner.

When making use of the second method the first thing to do is to clearly specify the objectives to be measured. On the basis of these objectives, a group of competent people well conversent with the content of the field being tested, may carefully study various parts of the test and give their opinion on the validity of the test. As a result various changes, corrections and additions may be made. This is, no doubt, a crude and subjective method, but if it is carried out thoroughly and conscientiously, it will very likely result in tests much more valid than those ordinarily used.

One important aspect in the validity discussion is the validity of individual items. It is true that the validity of the total test will be dependent upon how well each item does what it is supposed to do. The validity of a test item can be improved by a process of self-questioning.

For example, suppose that a teacher of mathematics is in the process of constructing an achievement test designed to measure the attainment of certain stated objectives. Suppose further that one of the objectives is to measure the skill of the student in computing with understanding, the operation of multiplication of decimal numbers.

He jots down a completion item in the following form:

$$\text{Item:}\quad \begin{array}{r} \text{Rs.}\quad .31 \\ \underline{\quad\quad \text{x } 24} \end{array}$$

Having jotted it down if he were to question himself, "Does this item really measure what it is supposed to? Is it valid?" Of course the answer will be 'No'. The item as it is written may be a good item to test skill in computing accurately and efficiently, but certainly it does not provide information on the understanding of the operation of multiplication. The teacher could then make an effort to revise the item. Maybe he rewrites the same item as shown below:

Item: Find the product of the numbers below. Then fill in the blanks to show the full meaning of the partial products.

$$\begin{array}{rl} \text{Rs.}\quad .31 & \\ \underline{\quad\quad \text{x } 24} & \\ & = \ \ 4 \text{ x Rs. } .31 = \text{---} \\ \underline{\quad\quad\quad\quad} & = \text{--- x Rs. } .31 = \text{---} \\ & = \text{--- x Rs. } .31 = \text{---} \end{array}$$

The revised item certainly measures whatever the original item measures, but it also provides the teacher with information that should enable him to distinguish between those who have learned by rote and those who understood the process.

Objectivity

Objectivity is an important factor that affects both the validity and reliability of an examination. A test is objective to the extent that competent persons agree on the scoring of answers. In other words, a test is said to be objective to the extent that the opinion or judgment of the scorer is eliminated from the scoring process. Objectivity is usually attained by (i) stating the questions precisely (ii) requiring precise short answers (iii) scoring the test by using a previously determined key.

Criticism of objective testing. Objective tests have often been severly criticised on the basis that they emphasise and measure only specific, unrelated bits of information instead of broader concepts and understanding. The critics claim that such tests encourage only the memorisation of miscellaneous facts. This is not true. In fact objective tests can be and have been devised to measure some of the broader outcomes of education. If objective tests on the whole have been based on the narrower goals of instruction it is largely because the makers of these tests have lacked the necessary skill, vision and perhaps the motivation to base their questions on larger concepts. If a teacher sincerely believes in the importance of the broader goals he can do much to encourage their attainment by emphasizing them in his own tests. Epstein and Myres strongly recommend the use of objective mathematics tests:

> Contrary to a prevailing misconception about objective mathematics achievement tests, such tests are not inherently limited to the measurement of mere recall

> of factual material or of mechanical manipulations of symbols. On the contrary, the full gamut of concepts and understandings of a good mathematics course can be sampled and tested by ingenious item writers.[8]

Discrimination

A test discriminates if it is constructed in such a way that it will detect or measure small differences in achievement--when it picks out the good from the poor--the 'cans' from the 'cannots'. This is an absolute essential if the test is to be reliably used for ranking students on the basis of achievement or for assigning marks. Three things will be true of a test that meets this standard:

(i) There will be a wide range of scores. Scores are likely to vary from nearly the lowest to nearly the highest possible scores.

(ii) The test will include items at all levels of difficulty. That is, the items will vary uniformly in difficulty from the most difficult one, which will be answered correctly by the best students, to an item so easy that practically all the students will answer it correctly.

(iii) Each item will discriminate between students who are high and those who are low in achievement. Each item will be missed more frequently by poor students than by good students.

[8] Marion Epstein and Sheldon Myres, "How a Mathematics Test is Born", The Mathematics Teacher, L I (April, 1958) p.300

As is true with validity, reliability and objectivity, the discriminating power of a test is increased by concentrating on and improving each individual item in the test. After a test has been administered a simple item analysis can be made which will help to indicate the relative difficulty of each item and the extent to which each discriminates between good and poor students. The technique of item analysis has been practically illustrated in chapter V.

Comprehensiveness

In constructing an achievement test it is important to sample liberally all phases of instruction which are supposed to be covered by the test. It would, perhaps, not be practical to test every point that is taught. How comprehensive should a test be? How much of sample should it include? These are the natural questions which crop up.

A test should be comprehensive enough to be valid. It should include enough points so that it measures what is is supposed to measure. This is an easy statement to make but somewhat more difficult to put into practice. There is no formula to indicate when a test meets the criterion of comprehensiveness. It is a matter of judgment. The best that a teacher can do is to put a question to himself: Is this test comprehensive enough to measure accurately and well what I expect to measure? Action taken by a test-maker after careful consideration of this question will have a strong effect on the comprehensiveness of the test.

Ease of administration

Consideration must be given to the features of the test which make it readily administered and scored. It should be so designated that a minimum of student time will be consumed in answering each item. The test items should also be constructed in such a manner that they can be scored quickly and efficiently.

The next chapter on the actual construction of an achievement test and specific type of items will contain some suggestions that will make for ease of administration and scoring.

Recording and Analysing Data

The third major step in the evaluation process is obtaining a record of what happens under test situations and subjecting such a record to thorough scrutiny and analysis. Several interpretations are made on the basis of analysis, and useful information obtained, which is applied to improve the effectiveness of instruction.

How the record is obtained and interpreted will be taken up in some detail in the next chapter.

Role of Evaluation

What is meant by Evaluation has been discussed in the foregoing pages. An answer to the second question, (why is evaluation needed?) is now sought. Here a discussion of the purposes of evaluation and its role in education seems relevant. The evaluators of mathematics achievement in Pakistani schools limit the purposes of evaluation to a few obvious ones, viz. assigning grades, promoting students as ends in themselves are not acceptable reasons in a sound evaluation program. No doubt one evaluates for these purposes too but the most important use of evaluation is in improving the quality of instruction--making the teaching and learning of mathematics more effective. To achieve this end a sound evaluation program has to play numerous roles:

(i) Determining levels of understanding. Evaluation can determine levels of understanding of students. Levels of understanding are associated with structures of mathematical relationships. One student may be able to recognise a special case of a general principle, another may be able to state the principle in words, the third may state it in symbols, the fourth may be able to apply it to a new problem. Every one of these four students is at a different level of understanding. Again a boy may sense that $25 \div 2\frac{1}{2}$ will produce the same result as $50 \div 5$ He has reached a higher level of understanding than the boy who would work out the problem in an ordinary way. Teachers of mathematics should be concerned with the evaluation of these levels of understanding. The techniques of evaluation in Pakistan need modification, to serve this important end.

(ii) Improving mathematics program. Evaluation can be useful in improving the mathematics program in terms of curriculum content and organisation; selection of learning material and methods of instruction. It can furnish data which should be used in making value judgments on questions such as : Is text book x better than test book Y as a tool for learning geometry? Is teaching device 'A' better than 'B' ? How well do the students progress towards the goal? Are the students able to apply their skill and knowledge in new situations? Are the topics of suitable difficulty? Many of these questions concerning instruction can be answered by good evaluative procedures.

(iii) Helping research work. The present mathematics curriculum is in a ferment. Decisions need to be made about the topics included, the sequence of topics and the immediate and remote goals. This calls for careful evaluation to furnish information upon which final decisions can be based. Data collected through the evaluation process can become a basis for educational research. Thus schools in Pakistan must evaluate as validly as possible and maintain good records to be used for research purposes.

(iv) Determining mental ability. A sound system of evaluation reveals a good deal about the student's mental ability, his maturity and his background of experience and learning. This information can be used in helping the student learn mathematics effectively. One would not under the existing mathematics structure, teach logarithms, for example, unless a student has sufficient background of equations, indices and the like. A first obligation of a teacher is to guide the student into study for which he is ready in terms of mental maturity and background of experience and learning.

(v) Discovering modes of learnings. The right kind of evaluation can discover the mode of learning employed by the students--whether their achievement is the outcome of the thinking process or a rote memorisation. Mode of learning in mathematics is as important as the contents, because a pattern of learning that features inquiry and thinking is extendible by the student and has, perhaps, greater transfer value than learning by rote. Thus the teacher must constantly be evaluating the modes of learning employed by students.

(vi) Guiding learning. Evaluation can guide teaching by furnishing specific strengths and weaknesses in pupils' achievements or capacities. "By studying the results of good tests (plus information obtained from other evaluative devices) you can obtain a fairly accurate picture of the way your students learn."[10] The teacher may then either seek to eliminate those weaknesses by using special teaching methods or by directing learning towards areas where the pupils' efforts will be more fruitful. A pupils' difficulties, for example, may be traced by means of tests to a specific inadequacy such as inability to deal with certain combinations of digits, or lack of correct techniques for carrying, Diagonostic testing may thus reveal the precise sources of pupils' trouble spots and guide the teacher to an optimum way of overcoming them. Teachers may discover what parts of a topic or unit need to be retaught, or taught differently. The pupils who are capable of doing exceptional work may be discovered by evaluation procedures and may be guided towards special tasks.

The role of evaluation is very succinctly summed up by Eugene D.Nichols in the following passage:

> It is truism to state that tests are an essential part of teaching. They have the function of judging the effectiveness of teaching done, as well as having implications for teaching to be done. Through testing, however, the teacher not only must

[10] Micheels and Karner, op. cit., p86

> obtain an estimate of the quality of his teaching and the quality of his student's learning, but he also must develop an insight into ways of judging the quality of the test itself. That is, he must develop the sense somewhat as the artist does, by which he will be able to interpret the results of the test so as to use them for improving the quality of his instruction, the amount of learning on the part of the students, and the effectiveness of the instrument with which he chose to test the two.[11]

[11] Eugene D. Nichols. "Testing the Understanding of the Concept of Equation," The Mathematics Teacher, L (May, 1957), p.400

CHAPTER IV

CONSTRUCTING THE ACHIEVEMENT TEST AND INTERPRETING THE SCORES

Process of Construction

An attempt was made to explain the theory of evaluation in the previous chapter, which dealt with the theoretical elements involving the determination of objectives, content and the technical qualities of the measuring instrument such as validity, reliability and discriminating power. The present chapter will be concerned with the procedure of constructing an achievement test and interpreting scores.

Written achievement tests used by teachers may be classified as essay examinations or short-answer examinations. The latter are often referred to as objective or 'new-type' tests. The conventional essay examination requires the formulation of an extended answer to a question. Imbedded in the question are steps which necessitate understanding of basic principles, but these are generally lost in the maze of manipulation. Short-answer tests on the other hand, consist of questions to which the pupil responds by the selection of one or more of several given alternatives or by filling in a word or phrase.

In recent years, the short-answer tests have gained widespread popularity. This chapter will explain the process of constructing a short-answer (objective) achievement test. At present the essay-type examination is in vogue in Pakistan. In the prevailing system "...there is no attempt to base promotion through the school on an objective and comprehensive assessment of the work done throughout the year."[1] It is the writer's opinion that the objective tests coupled with essay-type examination can put the evaluation system in Pakistani schools on a sounder basis. It is for this reason that the techniques of objective test construction and interpretation are being explained for the use of Pakistani teachers in general and mathematics teachers in particular.

The Process of constructing objective achievement test and interpreting scores may be viewed in terms of three steps:

(1) Planning the test.

(2) Item writing.

(3) Organising, administering, scoring and interpreting.

Planning the Test

The Planning of an objective test generally involves the following considerations and decisions to be made by the mathematics teacher:

(1) The objectives of the course must first be established. The individual teacher can do this in collaboration with the mathematics department.

(ii) The objectives of the course are then to be defined and analysed in terms of student behaviour. Items of subject matter are to be related to specific student outcomes.

(iii) The number of items is to be decided by the teacher.

[1] Government of Pakistan, Report of the Commission on National Education, (Karachi; Ministry of Education, 1960), p.123.

The amount of time given to the administration of the test is a factor in determining the number of items. If the time is limited the number of objectives to be covered may also have to be limited. There is nothing wrong with using a test that reliably samples one or two objectives.

(iv) The teacher has also to decide what type of items, he will use on the test. The common types of items which are used in constructing a test are: Completion, true-false, multiple-choice, and matching. Their use is discussed below:

Completion Items. The completion item requires the pupil to complete the thought of a sentence by providing a single word or words or number or phrase which is omitted or it directs him to respond to a question by writing his answer in the blank space provided. Completion items are convenient for testing recall of concepts or computational skills.[2] A completion item can generally be scored with a good deal of accuracy. Computation problems such as 'find the L.C.M. of 4,12,16' where only the answer is of concern, may be classified as completion items. For more examples of this type, attention is invited to the Elementary Algebra Test contained in chapter 5.

True-False Items. True-false items require the pupil to express his judgment of a given statement by indicating True or False, Yes or No, or some similar response. True-false items are adapted to

[2] H. H. Remmers and N.L. Gage, op.cit., p.77

the testing of simple facts, ideas and concepts. They usually take little time to answer and can therefore cover a relatively large amount of subject matter in a given length of time. They should be based on statements that are unequivocally true or false such as 'A pentagon is a plane figure bounded by five straight lines.' True-false items provide many opportunities for the student to get the right answer by guessing and a correction for chance success is usually applied. These items can be scored quickly and objectively.

Multiple-Choice Items. Multiple choice items require the pupils to recognise which of the several suggested responses is the best or the correct way to answer a question. While the completion item requires the pupil to produce the correct response without suggestion, the multiple-choice item calls for recognition only. It is adapted to the testing of complex ideas and interpretations. The multiple-choice item is superior to the true-false, which presents only two alternatives, in that it reduces the opportunity for guessing the correct answer. The difficulty level of a multiple-choice item is related to the stem (the part before the choices) as well as to the choice of options presented along with the right answer. Like the true-false items, the multiple-choice items can be scored quickly with a high degree of objectivity. An example of this type is:

The simple interest on Rs. 1000 for 8 years is

A Rs. 320
B Rs. 3200
C Rs. 32
D Rs. 3.20

The student is required to indicate the correct answer from the given choices. Further examples of this type are contained in the algebra test given in appendix B.

Matching Items. Matching items consist of two parallel columns of words, phrases or sentences. The pupil is required to match or associate each item of one column with the item which corresponds to it in the other column. Each matched pair is scored separately. Actually, however, the pairs are interdependent because an incorrect response may make an item unavailable for correct pairing. (This is not true, however, when the instructions permit the use of the same choice more than once.) For this reason one of the two columns should contain more items than the other. Also it is better to have two short matching exercises than one long one. For examples of matching items, a reference is invited to the algebra test contained in appendix B

(v) A decision having been made about the number of items and types of item to be included in the test, the evaluator should prepare a table of specifications. This table is to be the blueprint for constructing test items. It will serve as a guide. The specification table will represent the number of items to be used for each objective. Such a table is provided in chapter V.

Item Writing

After the establishment of a detailed set of test specification based on the rationale (including content coverage, number and type of questions, timing) the next step is item writing.

An item may be defined as a scoring unit. An objective item is one that can be scored by mechanical devices. The job of any single item is to separate those students who have in large degree attained the objective sampled, from those who have not, to any extent, reached it. Each item marked right or wrong divides the students in two sections. Anything about the structure and presentation of the item that leads more of the upper section (high scorers) than of the lower section (low scorers) to get it wrong will lead to unreliable information.

Item writing directed to the purpose explained above is, therefore, to be done with care and caution. It is an art and requires an uncommon combination of special abilities:

First, the item writer must have a thorough mastery of the subject matter being tested. Not only must the item writer be acquainted with the facts and principles of the field, but he must fully understand their implications. He should also be aware of the popular fallacies and misconceptions in the field.

Second, and of utmost importance, the writer must possess a developed set of educational values (aims and objectives), which ought to permeate his thinking. It is difficult for one whose sense of values is inadequate to produce good achievement items consistently.

Third, the item writer must understand, psychologically and educationally the individuals for whom the test is intended. He must be familiar enough with their probable levels of educational development to adjust the difficulty of the items appropriately. He must also know what would constitute plausible distractors for multiple-choice items.

Fourth, the item writer must have a certain degree of mastery in verbal communication. He should be able to use words with precise meaning and should have sufficient skill to arrange them so that they communicate the desired meaning and no other.

Fifth, the item writer must be skilled in handling the special techniques of item writing. Obviously he needs to be familiar with the types and varieties of test items and with their possible limitations.

As an aid to minimise the number of errors in item writing, an evaluator may be guided by a set of established maxims. The number of such maxims is fairly large. For the sake of brevity, only a few basic maxims will be discussed here. For a fuller statement see appendix D

Basic Maxims[3]

(i) Make every effort to avoid ambiguities. The objective test item must be clear in and of itself. Lack of clarity may arise from inappropriate choice and awkward arrangement of words. It may

[3] The present discussion of item writing is directed primarily towards objective items used in paper and pencil tests of achievement in mathematics. However many of the suggestions made here will apply to other types of test in other fields of education.

also arise from lack of clarity in the thinking of the person who writes it. Before emerging in final form items need critical examination and revision. An example of an ambiguous item is given below.

> T-F The area of four walls of a room is equal
> to length and breadth multiplied by
> twice the height of the room.

The item as stated above will leave even the student who knows the answer in doubt. If he interprets the item as saying that the item means length + breadth x 2 height, he will mark it false. But if he interprets it as (length + breadth) x 2 height, he will mark it true. A little rewording avoids the difficulty:

> T-F The area of four walls of a room is equal to
> the sum of length and breadth multiplied by
> the height of the room.

(ii) Avoid the use of non-functional material in the item. Do not put any more in the item than is necessary unless ability to recognise irrelevant material is being tested.

For example:

> Saeed has Rs. 2000/- out of which he
> divides Rs.900/- between A and B
> in the ratio 1:2. How much did A
> receive? ______

This is a completion item. The item has been unnecessarily lengthened by the use of irrelevant words which have no bearing in the working out of the item.

An improved statement after deleting the extraneous material would be simply:

> Saeed divides Rs. 900/- between A and B in the ratio of 1:2. How much did A receive? ________

(iii) In computational problems specify the degree of precision expected, or better still arrange the problems to come out even except where ability to handle fractions and decimals is one of the points being tested. (Degree of accuracy may be specified in the item or in a series of similar items.) Consider the following example:

> A store that can hold 700 bags of rice, has now 300 bags in it; what part of the store is vacant? _____

Here some students may give the exact answer 4/7 and others may give decimal answers to various degrees of accuracy. If the evaluator wants this item to the nearest hundredth, it may be written:

> A store that can hold 700 bags of rice, has now 300 bags in it; to the nearest hundredth what part of the store is vacant?_____

If the correctness of a numerical response depends upon stating the unit of measurement, make this fact clear. If not it is best to include the unit of measurement in the statement of the question as for example:-

> The volume of a cube nine feet on an edge is _______ cubic feet.

(iv) Avoid giving irrelevant clues. Unless it is carefully constructed a test item may in itself, furnish an indication of the correct or expected answer. For example:

In the expression 3^n, the n will be called an

1. root
2. base
3. radical
4. exponent.

Here the right answer is the only alternative that correctly follows the article an. An improvement would be:

In the expression 3^n, the 'n' would be called

1. a root
2. a base
3. a radical
4. an exponent.

(v) Avoid irrelevant sources of difficulty. Quite frequently reasoning problems in mathematics are answered incorrectly by examinees who have reasoned correctly but who have slipped in their computations. Such problems are often met in the question papers meant for secondary school students of Pakistan. Consider the following example designed to measure the principle of price discount:

> Mr. Rashid was given a 12½% discount
> when he bought a table whose list
> price was Rs. 97.75
> How much did he pay for the table?

A number of examinees who understand the principle will miss the item because of errors in multiplications and in the placement of decimal points. If the item is revised as follows:

> Mr. Rashid was given a 10% discount
> when he bought a table whose list
> price was Rs 100/- How much did he
> have to pay for the desk?

The computational difficulty is removed so that the principle alone can be tested. Test constructors may differ concerning the advisability of eliminating computational difficulty from certain mathematical problems, but when complex and time-consuming calculations are included in the item, the item writer should recognise that he is chiefly testing computational skill rather than understanding of a mathematical principle.

(vi) Avoid giving clues to one item in the statement of another. Consider the following two items which might appear on the same test, though not necessarily following one another:

> Which of the following is a special type of quadrilateral?
> 1- rhombus
> 2- pentagon
> 3- regular hexagon
> 4- triangle.
>
> Which of the following quadrilaterals is by definition equangular?
> 1- trapezium
> 2- rhombus
> 3- square
> 4- parallelogram.

Most of the students would expect that all the alternatives given for the second item are quadrilaterals. Rhombus would appear to be a reasonable selection for the first item based on its appearance in the second. This difficulty could be avoided by deleting or replacing rhombus as an alternative in the second item.

(vii) In a multiple-choice item make all distractors plausible and attractive to examinees. The following test item illustrates this point:

> The ratio of 25 paisas to five rupees is
> 1. 1/20
> 2. 1/5
> 3. 5/1
> 4. 20/1

The examinee who carelessly overlooks the distinction between paisas and rupees or inverts the ratio, will arrive at one of the distractors rather than at the answer. Sometimes it is helpful for item writers to first present multiple-choice stems as free response items, and then to use incorrect responses of some examinees as distractors.

(viii) Avoid weighting of items. Each single response should be numbered and should count one point. To a test maker certain objectives may be relatively more important than others, but investigations on the subject have revealed that little is gained by spending time on assigning various weights to various items.[4]

(ix) Finally, all items are to be carefully checked for incorrect mathematics, incorrect grammar, triviality, or trick questions that have a twist unimportant to the objective.

The essentials of item writing have been well summed up by Epstein and Sheldon:

> In the writing of questions, there are certain aims that are always kept in mind. Each question should be designed to test a single, well-defined concept, and if it is written to test the concept, it should not be workable by an easier, alternative technique. Answer choices should be based on plausible, but incorrect methods and misconceptions common among students, not on arithmetic errors unless skill in arithmetic process is the thing being tested. Time consuming calculations not germane to the concept being tested are to be avoided, and fairness to the candidate should be given constant consideration by the test constructor. Problems whose solution depends on a trick are always excluded and care is extended to measure only those concepts which are considered appropriate for the group being tested, according to the original test plan and specifications.[5]

[4] Micheels and Karnes, op. cit, p.147

[5] Epstein and Sheldon, op.cit., p.300

Organising, Administring and Scoring the Test and Interpreting Data

After items are written they are to be organised into a test. There are two major considerations in setting up the test: (i) the ease of the students to understand what they are required to do, and (ii) the ease with which the teacher will be able to locate and score answers. The following general rules are helpful:

Grouping items. If there are more than one type of items, they are to be put in sections under the headings multiple-choice questions, matching questions and so on. All the items are to be numbered consecutively from the first item to the last.

The various sections are to be arranged so that the easier items come before the more difficult ones.

Putting part of an item on the bottom of a page and the remainder on the top of the next page is a bad practice and should be avoided.

Directions. Each group of items is to be preceded with a simple and clear statement telling how and where the student is to indicate his answers. For computational items the directions might be, "For each problem below write your answer to the nearest hundreth on the line in front of the problem number."

Directions for computational items requiring answers in denominate numbers should state that the denomination is required for full credit.

Directions for multiple-choice items should tell the student whether he is to seek the right answer or the best answer.

Directions for true-false and multiple-choice items should inform the student if any correction for guessing will be used in scoring. A good statement is as follows: "A portion of your wrong answers will be substracted from your correct answers, so it is best to guess only on those items where you are reasonably sure of the answer."

If the method of arriving at the answers will be used in evaluating performance, a statement indicating that work should be shown and will be considered in the scoring should be included.

Recording of answers. It is generally desirable to provide answer spaces that correspond with the placement of the items on the page. But separate answer sheets may also be used. The answer sheets are to be divided in columns with spaces for writing answers, one column for each page of items. The student may then place the answer sheet under the test booklet and line the answer spaces up with the items on each page. The use of separate answer sheets reduces the labour of scoring. If students answer the problem on the test booklet, the answer spaces may be placed down the side of the page. This lessens the labour of scoring.

Spaces for computational work may be provided below problems which are to be scored for computation.

Administration. Each student should work from a separate copy of the test, rather than from a test written on the blackboard. Best possible physical surroundings should be provided with adequate light, ventilation and desk space. If speed is not being tested, enough time should be allowed so that each student has an opportunity to attempt all items.

Scoring the Test. A key is to be prepared containing all answers that are to be given credit. If the items are carefully prepared, there may be only one acceptable answer for each item. If the ambiguity of a statement makes two or more answers acceptable, they should be included on the key. The list of acceptable answers is to be prepared in such a way that it can be placed beside the answer spaces used by the students.

Correction for guessing. One of the decisions which must be made about the scoring is concerned with whether the score is to be the number of right or whether a formula to correct for guessing (chance success) is to be employed. The generalised formula for correcting for guessing is:

$$S = R - \frac{W}{N-1}$$

where S = score
R = the number of right responses
W = the number of wrong responses
N = the number of choices offered by each item

When this formula is used for true-false items it reduces to $S = R - W$. If used for multiple-choice items with four suggested answers as in the Algebra Test in appendix B, it would be reduced to

$$S = R - \frac{W}{3}$$

Analysis of Test Scores for Interpretation

It has been mentioned in chapter II that in the Pakistani schools, the mathematics teachers practically make no attempt to interpret the scores of pupils on various tests. If an attempt is made at all it is limited to interpreting a students' score in terms of a comparison with the highest possible score on the test. This method usually expresses the score as a percentage. Some of the characteristics of the score make it very unsatisfactory.[6] Test results as has been pointed out earlier are not used to yield information about the strengths and weaknesses of individual students or the class as a whole nor are they used to provide information about the test itself. It is, therefore, considered essential to provide a procedure which the mathematics teachers in Pakistan should adopt for analysing and interpreting scores on achievement tests.

The writer is of the view that the use of statistical methods in the analysis of test results is directly in line with good scientific techniques. The most important statistical techniques from the standpoint of the frequency of their use in the classroom are abilities to (i) classify and tabulate data (ii), determine the common measures of central t endency (iii) determine the common measure of variability or spread (iv) determine the relationship of two groups of data (v) utilize simple graphical methods in the presentation of facts (vii) secure derived scores and use them in the interpretation of results.

[6] See chapter II

It is not proposed to elaborate upon these techniques here for such an elaboration would require a separate treatise in itself. It is hoped that the mathematics teachers would make themselves sufficiently versed in the use of these techniques to be better able to analyse and interpret test scores. The procedure of analysis and interpretation of scores is, however, being explained.

The set of scores from an achievement test can be tabulated, summarised, and analysed to provide many different types of information. The three types of information which can be obtained from test scores will be considered here: (i) information about the achievement of the class, (ii) information about the achievement of individual students, (iii) information about the test.

Information About the achievement of the Class

A good teacher is much concerned with the level of performance of the class and the spread of scores. The level of performance may be summarised in a measure of central tendency, such as the mean, and the spread of scores is indicated by a measure of variability, such as the standard deviation. These two measures may be used as a basis of interpreting the score of an individual student or for comparing the scores of two or more groups.

Measures of central tendency indicate the performance of the group as a whole whereas the measure of variability indicates the amount of spread of the scores.[7]

[7] There are three measures of central tendency--mean, median & mode-- of which the first two are often used. Mean and median are used when the distribution of scores is symmetrical; median is used when the distribution is asymmetrical or skew & it is desirable to minimise the contribution of the extreme scores. But if the distribution is asymmetrical & the contribution of each score is to be in direct relation to its size, the mean is used.

The means or medians of different distributions may be the same, and yet the measure of variability be different for different distributions. For purposes of interpretation of scores the standard deviation is used as a measure of variability because it takes into account the deviation of each score from the mean.[8] In addition to its usefulness as a measure of variability, it may be used in describing the test performance of the individual students as shall be seen in the next section.

The teacher's information about the standard deviation of a set of scores will give him a general idea of the ability level of the group and he can, therefore, construct a test that should yield roughly the distribution results he is seeking. If he wants to identify students needing remedial work, he will give a relatively easy test that would spread the low ability students and will bunch the rest at nearly perfect scores. But if he wants to measure achievement of all of his students, he will need a distribution with enough spread to show differences among them.

Information about the Achievement of Individual Students

One of the aims of an evaluator in mathematics is to acquire information about the individuals through tests. A raw score or a score in percentage form is not of any substantial help for the teacher. A score gains meaning when compared to the score of other members of the

[8] The range which is defined as the difference between the highest and lowest scores plus one is an unsatisfactory measure of variability because it tells nothing about the scatter of scores between the extremes.

groups called norm groups.

The following normative groups would be among those to be considered for a particular mathematics test:

(i) The mathematics class of which a student is a member.

(ii) All mathematics students of a particular grade or age level in a school in Pakistan.

(iii) All mathematics students or a representative random sample of a particular grade or age level in a city or district in Pakistan.

(iv) All mathematics students or a representative random sample of a particular grade or age level in a broad geographical region say West Pakistan or East Pakistan.

(v) All mathematics students or a representative random sample of a particular grade or age level in the whole country--both East and West Pakistan included.

In order to use norms intelligently we need a clear definition and description of the norm group and a clear definition and description of the group or individual to be compared with the norm. Any conclusion must then relate to the characteristics of both groups.

The practice in Secondary Schools of Pakistan to express the score on a test as a percentage is for reasons explained earlier[9] is not satisfactory. To increase the amount of information obtained from a test score, we may interpret a students' performance on a test through the use of derived scores. A derived score indicates a students'

[9] See chapter II.

performance in terms of the performance of other students. Three types of derived scores are being considered here: Ranks, Percentile Ranks, and Standard Scores.

Ranks. These are readily obtained by putting the papers in order according to the raw scores and numbering them from the best to the poorest. The person with the highest score has a rank of 1 and next highest a rank of 2, and so forth. Ranks hide the size of the differences in the raw scores. For example, on a test in Algebra suppose the three top students obtained scores of 69, 63, 62 their ranks being 1, 2 and 3 respectively even though six points separate the top two and only one point separates the second and the third.

Percentile Ranks. Percentile ranks also provide ordering of students in a group. Thus if Rashid has a raw score of 68 and 38 percent of his classmates have scores below 68, Rashids' percentile rank is 38. Percentile ranks also like the ranks hide the size of the differences in the raw scores. Further they cannot be added or averaged.

Standard score. A type of derived score that retains more than the ordinal information of ranks and percentiles is the standard score. A students' standard score tells how many standard deviations his score is above or below the mean. It is obtained by subtracting the mean from his score and dividing the difference by the standard deviation. Carrying out this information for each student's score gives a set of standard scores with a mean of zero and a standard deviation of one. These are called Z-scores.

Examples: Suppose the mean of a set of scores on a geometry test is 65.5 and the standard deviation is 5. If one student gets a raw score of 70 and another 55, their Z-scores will be

$$\frac{70 - 65.5}{5} = + .9 \quad \text{and} \quad \frac{55-65.5}{5} = - 2.1$$

The first student is .9 standard deviation above the mean and the second 2.1 standard deviation below the mean.

As can be noticed, Z-scores involve the use of negative values and decimals. Other sets of standard scores that avoid these two characteristics can be established. A set of standard scores with any desired mean and standard deviation can be obtained by the linear transformation $xz + y$ where y is the desired mean and x is the desired standard deviation. It is often convenient to establish standard scores with a mean of 50 and a standard deviation of 10. Thus the students of the last example will have the following standard scores:

$$\text{(i)} \left(10 \times .9\right) + 50 = 59$$

$$\text{(ii)} \left[10 \times (-2.1)\right] + 50 = 29$$

Advantages. There are two basic advantages of standard scores on percentiles. First they preserve the shape of the distribution of raw scores. If the raw scores are asymmetrical, the standard scores will be asymmetrical too. Second, standard scores are additive. If a student has a series of standard scores that are based on the same class or group of students, these scores may be weighted in any way or averaged, to obtain an overall summary of his performance on the different tests.

Ranks, percentiles, and raw scores provide information concerning what a student did on a test. Standard scores, on the other hand, not only show what a student did on a test but also indicate where he stands among others taking the test and thus lend to a value judgment. The use of standard scores would, therefore, signify that we are evaluating achievement rather than just measuring it.

Remedial Information

Besides interpreting raw scores in terms of standard scores, a teacher can locate trouble spots and difficulties of individual pupils by charting the responses of pupils. The chart may be made in the following form:

Names Items.	A.K.	R.S.	B.T.	M.N.	R.T.	R.W.	P.Q.	A.B.	E.R.	S.P.	Number Right	percent Right
1			✓		✓						8	80
2		✓			✓						8	80
3		C			B						8	80
4	B	C	B	D	C	B	B	C	D	C	X	Zero.

Note: checks indicate wrong answers on completion or true-false items whereas letters show the choice of wrong alternatives on multiple-choice questions.

The teacher by looking at the chart can diagonose the weaknesses of students in different areas of learning. He would, for example, know that both R.S. and R.T. have not understood the objective sampled by the multiple-choice item 3. Again by looking at item 4, he will find that none of the students has answered the item correctly.

He may draw some inferences on this basis. He may, for example, think that the item has not been properly worded or he may suppose that he did not perhaps teach well for the objective he was testing in item 4, and should, therefore, take it up with a new approach.

Information About the Test

Test results are analysed to determine the effectiveness of the measuring techniques. For the purpose, a test maker has to evaluate the merit of the individual items as well as the merit of the total test. To determine the merit of any test item, test results have to be subjected to an item analysis whereas the merit of the total test is determined largely by a study of its validity and reliability.

Item Analysis. The purpose of item analysis is to determine the discrimination or the extent to which an item differentiates between the 'cans' and 'cannots'. If more 'cans' than 'cannots' answer the item correctly, the item is working in the desired direction and is called a positive discriminator. But if more 'cannots' than 'cans' get the item right, the item is working against the test and is called a negative discriminator.

The item analysis data also tells us something about the difficulty of the items. The difficulty index of an item is defined as the percentage of a total group getting the item right. To determine the level of achievement of students in a whole class, it is desirable to have some relatively easy and also some relatively difficult items, with most items having difficulty indexes that centre on the .40 to .60 range.

The item analysis data also gives us information concerning the effectiveness of the distractors in multiple-choice items. Good distractors are those which are selected to a greater extent by the members of the low scoring group, the 'cannots'. When reversals are found, the test maker has to consider the distractor, and if he still wishes to use the item, he revises the distractor, trying to make it less appealing to the better students. The analysis also reveals some distractors which are not chosen by any one of the group. Such distractors should be revised or eliminated by the test maker.

The whole procedure of item analysis will be fully illustrated in the next chapter by practically carrying it out in all its aspects.

Validity. The meaning and the methods of appraisal of validity have been explained in chapter III. Two observations are to be made here: (i) valid items add to the validity of the test (ii) the validity of the individual items largely depends on their discriminating power.

Reliability. Results of a test provide information about its reability. The information on reliability is usually summarised in a reliability coefficient. Many procedures have been proposed for obtaining reliability coefficients--the test-retest, equivalent forms and split-half methods. There is also a shorter method (known as kuder-Richardson procedure) which is accurate enough for classroom tests and other ordinary situations. It requires the calculation of only two measures, the mean and the standard deviation of the distribution of scores. The formula is stated below:

$$r_{11} = \frac{n}{n-1} \left\{ 1 - \frac{M - \frac{M}{n}}{S^2} \right\}$$

where r_{11} = reliability of the test, n = number of items on the test, S = standard deviation of the scores, and M = the mean of the scores.

After a reliability coefficient for a test is obtained there remains the problem of interpretation. It may be said in general that the higher the reliability coefficient the better. Usually the reliability coefficients fall between the limits 0.00 and 1.00. The more reliable a test, the closer the reliability coefficient will be to 1.00

Standard Error of Measurement. Using a test with low reliability coefficient has its effect on the reliability of scores obtained by pupils. The probability of error in a particular score can be estimated by the following formula:

Standard error of measurement = $S\sqrt{1 - r_{11}}$ where S is the standard deviation and r_{11} is the reliability coefficient of the test. This statistic tells the range within which scores on the same test would be expected to fall two-thirds of the time.

Suppose that a certain test has a standard deviation of 6 and a reliability coefficient of .75. Then the standard error of measurement is $6\sqrt{1 - .75} = 3$. If Azam has a score of 55 on this test there is 68% probability that Azam's true score is included in the range of ± 3 or 52 to 58.

Concluding Statement

There are many refinements of test analysis which have not been mentioned here. The average mathematics teacher in the Secondary Schools of Pakistan should be able to follow the procedures presented here.

Certainly if he does use such procedures, he will achieve steady improvement in the construction of his tests and in his whole evaluation program and will thus experience a sense of genuine satisfaction in this important part of his job.

CHAPTER V

ACTUAL CONSTRUCTION, ADMINISTRATION AND ANALYSIS OF ALGEBRA TEST

The object of this chapter is to illustrate, for the mathematics teachers in the Secondary Schools of Pakistan, the principles of test construction and evaluation by means of a practical, detailed account of the actual construction and administration of a test in algebra. The test has been designed to measure achievement at the end of the first year course in Algebra. The test which consists of forty-three items is included as appendix B

The whole process of construction, administration, scoring and interpreting is explained below.

Planning the Test: Before the actual writing of items decisions were made about the objectives to be tested, the percentage of items to test each objective, the types of the items and the length of the test. Decision about the objectives was made on the basis of:

(i) a careful study of the first year algebra syllabus in use in some local schools in Beirut,

(ii) an analysis of standard algebra text books.[1]

This study made it possible for the test constructor to arrive at some common objectives.

The second problem facing the test maker was to decide upon the percentage of all items which should be assigned to test each of the objectives. The decision in this matter was taken on the basis of the judgment and the experience of the test constructor. It was, for example, considered necessary to put relatively more emphasis on the understanding of algebraic concepts and relatively less emphasis on mere recall of facts and relations. The test constructor, therefore, decided to include eighteen items (42%) to test whether the pupils have a sound understanding of concepts. The objectives around which the test was constructed as well as the number of items pertaining to each objective are as follows:

Objectives	Number of Items	Percentage of Test content
(i) To understand the terminology of algebra taught and to have an understanding of algebraic concepts. (Items 1-18)	18	42%

[1] The following books were consulted in preparing the test: (i) Panel of the School Mathematics Study Group, *First Course in Algebra Part I & II* (New Haven, Yale University Press, 1961). (ii) Mansur Hanna Jurdak, *School Algebra Book One,* (American Press, Beirut, 1957),(iii) A.M. Welchons & W.R. Krickenberger, *Algebra Book Two,* (Ginn & Co., N. York, 1949) . (iv) D.J. Aiken, K. B. Henderson & R.E. Pingry, *Algebra Book One*,(New York: Mcgraw-Hill Book C., Inc., 1960) (v) W. G. Shute and others, *Elementary Algebra*, (New York: American Book, Co., 1956).

Objectives	Number of Items	Percentage of Test content
(ii) To be able to evaluate algebraic expressions.	2 (Items 19, 20)	5%
(iii) To perform fundamental operations (addition, substraction, division & multiplication) that have been taught.	10 (Items 21-30)	23%
(iv) To solve simple equations	5 (Items 31-35)	12%
(v) To be able to find factors of algebraic expressions.	4 (Items 36-39)	9% .
(vi) To perform fundamental process with fractions	4 (Items 40-43)	9%
	Total 43 items	100%

It should be evident that the two-dimensional aspect of objectives has not been lost sight of. The objectives have been stated in a manner which establishes a relation between the content elements and the behaviour elements.

The statement of objectives as presented above served as a sort of a table of specifications or blue-print for the test constructor.

It was decided to use three types of items--matching, completion and multiplê-choice. This was done with a view to illustrate their use to mathematics teachers in Secondary Schools of Pakistan. It is to be mentioned, however, that "...among the various forms of so called objective tests the more objective, such as the multiple-choice are to be preferred to the less objective such as the completion."[2]

The length of the test was determined by the amount of time likely to be available for the administration of the test. The test constructor estimated that the number of items feasible for a class period of fifty minutes would be about forty three. This would allow the examinees roughly one minute for each item and give them seven minutes to go through the directions. It was noticed that the bulk of students finished the test in the allotted time. Those who could not finish--they were about fifteen in a group of 99--were given additional time through the courtesy of the different schools where the test was administered.[3]

Item Writing. Having decided upon the objectives, the number of items and the types of items, the actual process of item writing was begun. Every care was taken not to violate the maxims of item writing stated in chapter IV.

[2] K.W. Vaughn, "Planning the Objective Test," Educational Measurement, ed. E.F. Lindquist, (American Council on Education, Washington D.C., 1951) p.172

[3] Additional time was allowed because of the experimental nature of the test.

The test items having been prepared were organised in three groups under the headings, Matching Questions, Completion Questions, and Multiple-Choice Questions.[4] Care was taken that within each group the easier items were placed before the more difficult ones.[5] Also it was decided that the matching items be placed in the first group, the completion items should follow and the multiple-choice items be placed in the last group. How far these subjective decisions were correctly made could be checked through item analysis after the administration of the test. It will be indicated in the later part of this chapter that the organization of this particular test was defective in several ways and could be improved.[4]

Administration of the Test:

The test was administered to a total of 99 students in four Beirut schools, namely the National Evangelical School for Girls, the International College, the National Protestant College and the Ahliah Girls School.[6]

The following administrative steps were carefully followed:

(i) Seating was arranged such that there was minimum opportunity for copying, and so that each student had a proper writing surface.

[4] For reasons explained in Chapter IV

[5] For reasons explained in Chapter II

[6] The four schools will be identified throughout this study as schools A, B, C and D, their order being different from that shown above so as to avoid invidious comparisons on the basis of the scanty data available.

(ii) To induce maximum motiviation, the examinees were told that the grades on the test will form part of their final grade in algebra. This was agreed to by the teachers.

(iii) Before distributing the copies of the test it was made sure that each student possessed writing material.

(iv) Copies of the test were distributed and the students were told to fill in the identification data accurately, but they were warned not to turn over the first page until instructed to do so.

(v) The directions on pages 1 and 2 of the test were read and explained. The examinees read them silently while the examiner read them aloud and explained them with the help of the illustrative examples given on the first page of the test booklet.

(vi) The examinees were encouraged to ask any questions they had concerning the directions.

Scoring the Test:

For ease of scoring a key was prepared in such a way that it could be placed beside the answer spaces used by the students. During the course of scoring it was detected that items 13 and 17 could have two correct answers. The scoring key was amended accordingly. These two items with possible answers are stated below:

Item		Answers
13 The difference between m and 3 is	1.	m - 3
	2.	3 - m
17 In $6y^3$, the exponent is	1.	3
	2.	Cube

It was made clear in the directions that the score on multiple-choice and matching questions will be computed by subtracting a portion of wrong answers from correct answers. To obtain the score on these items the "correction for chance success formula" was used.[7] For each correct answer in completion section one point was awarded. All the items on the test were equally weighted for reasons explained in Chapter IV.

Three actual examples of how the 'correction for chance success formula' was applied are presented below:-

Example 1	Number Correct	Number Wrong	Number Omitted
	13	6	4
Score = $13 - \frac{6}{3} = 11$			
Example 2.	15	8	0
Score = $15 - \frac{8}{3} = 12$			
Example 3.	10	4	9
Score = $10 - \frac{4}{3} = 9$			

Analysis of Test Scores for Interpretation:

The test scores were scrutinised to yield information about the groups tested, about individual achievement, and about the test itself.

Information about the Groups

The scores of the four groups tested were separately tabulated and the mean and the standard deviations were obtained in each case.

[7] See Chapter IV, p62

The following table presents information on the general level of performance of each group:

Table Showing the Performance Level of Individual Groups.

Names of Schools	Highest score obtained	. Lowest score obtained	. Mean	. Standard Deviation
School A	23	4	15.31	6.87
School B	39	13	24.9	6.3
School C	21	3	9.32	4.13
School D	33	3	10.85	5.85

This information is a help in making a comparative study of the different groups. The schools which are clearly low in the level of performance are thus given impetus to seek ways and means to improve.

Information about Individuals:

It is important to realise that a single raw score has little meaning. Some basis for comparison is always necessary for the interpretation of a single raw score with a view to relate the performance of pupils to one another. The scores of all the 99 students were, therefore, arranged in a frequency distribution and the mean and standard deviation were worked out. These were found to be 16.5 and 10.1 respectively. By means of these measures individual scores can be interpreted in relation to the large group comprising all the students tested.

For example, a student in school A actually obtained a score of 21. This score can be interpreted in this way:

In relation to large group (using mean and standard deviation of the large group), the score of 21 is

$$\frac{21 - 16.75}{10.1} = .4 \text{ (nearly) standard deviation above the mean,}$$

but in relation to its own group, (using mean and standard deviation of A - School), it is

$$\frac{21 - 15.31}{6.87} = .6 \text{ (nearly) standard deviation above the mean.}$$

Deriving mean and standard deviation of these sets of scores makes it possible to transform the raw scores to standard scores (say with a mean of 50 and standard deviation 10). Making use of the transformation formula[8], the mathematics teacher of school A can record the score of 21 as a standard score of 56. Since standard scores are equal units of measurement they can be manipulated mathematically. The standard scores of any group of students on this test can be added to standard scores on some other algebra test or tests and averaged to give an overall picture of a students' progress, without the improper weighting which is so commonly (and unwittingly) done when scores from distributions having different standard deviations are averaged.

Another way of interpreting the individual scores is by the use of percentile norms. A percentile indicates the percent of students who made lower scores. A percentile of 01 indicates the raw score to be in the lowest one percent. While a 99th percentile indicates the score in the top one percent. Percentile norms on this test were worked out for each score. A complete percentile report is added as

[8] See Chapter IV

appendix E A Summary view of the report is being presented below:

Percentile Norms for The Algebra Test Based on the 99 Examinees in the Four Schools:

Highest (Perfect) score possible on the Test	Possible Score	43
	Percentile	Raw Score
Superior Performance	99th	39
	95th	35
	90th	33
	80th	29
High average performance	75th	24
	70th	22
Average performance	60th	19
	50th	14
	40th	11
Low average performance	30th	9
	25th	8
Poor performance	20th	7
	10th	5
	5th	4
	1st	3

With the help of this table a students' raw score can be compared with the scores of other students on the test. Percentile norms, by providing a basis for comparison, make it possible for the evaluator to interpret the scores of individual pupils.

How to interpret scores (An example)

Name of Student	Raw Score	Percentile
M.W.	14	50
R.M.	39	99
S.D.	24	75
S.A.	9	30
M.A.	4	5

M.W's score of 14 places him at the 50th percentile or exactly in the middle of the group of all other students who had had algebra for about a year. R.M.'s score of 39 being the highest recorded on the test places him at the 99th percentile--not the 100th--there is no such percentile as the 100th, because percentile means percent of the group falling below his score--and since he is part of the group, 100% of it cannot be below him. S.D's score of 24 places her at the 75th percentile. She falls among the students whose achievement would be termed of high average level. S.A with a score of 9 is at the 30th percentile and thus falls among the lot of students whose achievement is of low average level. M.A who obtains a score of four is placed at the fourth percentile and compared to other students his achievement may be termed poor.

Information about the Test:

Test scores if properly analysed can yield useful information about the test itself. The test maker obtained the following information about the test:

(i) Difficulty index and discrimination power of each item through a detailed process of item analysis.

(ii) The reliability coefficient of the test.

(iii) Validity of the test in relation to teacher's grades.

Item Analysis--Simplified Procedure

The 99 test papers were first arranged in order of total score starting with the highest score of 39 and ending with the lowest score of 3. The top 27 papers (27% of all the test papers) were selected to represent the top group (score range 22 to 39) and the bottom 27 papers (27% of all the test papers) were selected to represent the bottom group (score range 3 to 9).[9] The responses made to each item by each individual in the top and bottom group were tallied and counted to give frequency of choosing each option. Examples:

Item-wise tables showing frequencies of choosing each option

Item 3

	. A	. B	. C	. D	. E	. F	. G	. H	.OMIT.	.
Top	1	1	-	-	22	1	-	-	2	27 responses
Bottom	2	3	2	-	-	5	-	1	14	27 responses

[9] "...it is customary to compare the responses in the top 27 percent papers with those in the lowest 27 percent of the papers."--M. M. Downieand R. W. Heath, Basic Statistical Methods, (Harper & Brothers, New York 1959), p205. For more details, the reader is referred to Flanagan, J.C., "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from the Data at the Tails of the Distribution", J. Educ. Psychol., 30:674-680, (1939)

Item 19

	Right	Wrong	Omit	
Top	11	14	2	27 responses
Bottom	1	18	8	27 responses

Item 41

	A	B	C	D	Omit	
Top	-	17	8	-	2	27 responses
Bottom	1	3	12	5	6	27 responses

The frequencies of choosing each option were then shown as percentages of 27. Examples:

Analysis charts showing in percentages
the Frequencies of choosing each option.

Item 3

	A	B	C	D	E	F	G	H	Omit
Top	4	4	0	0	81	4	0	0	7
Bottom	7	11	7	0	0	19	0	4	52

Item 19

	Right	Wrong	Omit
Top	41	52	7
Bottom	4	66	30

Item 41

	A	B	C	D	Omit.
Top	0	63	30	0	7
Bottom	4	11	44	19	22

To indicate correct response to each item the designation of the correct response was underlined.

A complete statement of item analysis is added as appendix F.

The difficulty index of each item is simply the proportion of those attempting it who answer it correctly. It is easily determined by adding the proportion of correct responses of top and bottom groups and dividing the sum by 2. For example the difficulty indices of items 3, 19, 41 were determined to be .41, .22, and .37 respectively. The discrimination of each item is a measure of the extent to which it is measuring what the test as a whole is measuring. In a nutshell, it tells how well that particular item is contributing to consistent measurement of what the test as a whole is measuring. This may be found by using John C. Flanagan's table of the values of the Product Moment Coefficient of Correlation in a Normal Bivariate Population corresponding to given proportions of success.[10] The difficulty indices and the discrimination indices were recorded in a table[11] and plotted on a graph.[12]

[10] Available in R.L. Thorndike, Personnel Selection, (New York: John Wiley & Sons, Inc., 1959) pp.347-351. An abridged version as adapted by Professor Korf is shown in appendix C.

[11] For table see appendix G

[12] For graph see appendix H

Inferences Based on Item Analysis

Some useful inferences may be drawn on the basis of item analysis.

Matching Questions.

In this section there are twelve items. The items on the whole are hard. The difficulty indices of all the items in this section except items 5 and 8 are lower than .50. Two conclusions can be drawn:

(i) The students have not well mastered the objective around which the items have been constructed

(ii) The items are, perhaps, poorly written.

There doesn't seem to be something seriously wrong with the writing of these items, but maybe the difficulty could be slightly reduced by revising some of them. The evaluator noticed that many students, while taking the test, wanted to know the meanings of 'represent', 'indicated' and 'denote'--the words which are used in the items. To keep the conditions of the test uniform for all pupils, the test administrator did not explain these words to the pupils. It is, therefore, presumed that a slightly larger number of students would have responded to these items, if they had not to face the language difficulty. These items, if the test were to be administered again, could be revised as follows:

Item 1 ---- $\frac{a}{b}$ is

Item 3 ---- a-b is

Item 5 ---- 3x, 4x, 10x are

Item 7 ---- The expression $x^2 + 2xy + y^2$ is

Item 9 ---- $A = L \times B$ is a

Item 11 --- In the given root $\sqrt{P + q}$, the

symbol $\sqrt{}$ is a

Item 12 --- If p, q, r be three numbers

pqr is a

The items in this section are fairly discriminating (the Flanagan r ranges between .47 and .81). It therefore seems appropriate to keep them on the revised test. The decision as to whether to retain an item may be made roughly by computing $2 \left(\frac{1}{\sqrt{N-1}} \right)$ where N is the number of cases in the top and bottom 27% combined. An index above this value is probably a real non-chance value. Lower indices could be the results of pure chance, and items having such indices should be discarded or revised if possible.

A very useful conclusion which a teacher can draw from the difficulty indices of the first twelve items is that the pupils, particularly in the bottom group, are not well grounded in and have not a satisfactory understanding of the objective which was intended to be measured by these items. The mathematics teachers of the four groups ought to direct instruction toward the attainment of this objective i.e., to understand the terminology of algebra and to have an understanding of algebraic concepts. It is in this sense that "test results provide a diagonostic technique for studying the learning of the class and the failures to learn and for guiding further teaching and study."[13]

The items in the section may be looked at more closely one by one. Example:

Item 5

78 percent of the top group 23 percent of the bottom group got this item right.[14] The difficulty index and the discrimination index are .50 and + .55 respectively.[15] The item is of average difficulty and is a positive discriminator. It is working basically, the way it should. On the basis of this information it would be worth using it for a similar group.

Completion Questions:

This section contains eight items. (13-20). Except items 13 and 16 the rest of them have a difficulty index varying between .35 and .22. These items were thus among the hardest for the pupils, but all of them are definitely positive discriminators.[16] Some of them may be looked at closely.

Item 15:

Item:....Five times N diminished by 9 is -----

This item was done correctly by 57 percent from the top group

[13] Robert L. Thorndike & Elizabeth Hagen, Measurement and Evaluation in Psychology & Education, (New York: John Wiley & Sons, 1961) p.90

[14] See appendix F

[15] See appendix G

[16] See appendix G

and 7 percent from the bottom group. The item is a positive discriminator (discrimination index being 166). But 11 percent from the top and 48 from the bottom group omitted the item. Possibly if the difficult word "diminished" is replaced by a more familiar word 'decreased', some of the examinees who omitted the item would have tried it. (It sometimes happens, however, that such revisions actually change the discrimination in an undesired direction. If so the original form should be reinstated). This item is worth using for a similar group in the following revised form:

Five times N decreased by 9 is -----

Item 19: The difficulty index is .22 and the discrimination index is .56. Of the top group 52 percent got it wrong. The item as structured does not seem to be defective. The teacher should impart more effective instruction to his pupils on "evaluating algebraic expressions" on which the item is based.

Use of completion items: Completion items can in some cases be used to find out usual wrong responses of examinees. These wrong responses can be offered as attractive distractors in an effective multiple-choice item. A study of the wrong responses to items 14 and 15 was made. The two items can be presented in multiple-choice form in the following way:

Item 14: Khalid is 14 x years old, his age four years from now will be

A 14 x + 4   B 10 x   C 18 x   D 18

Item 15: Five times N diminished by 9 is

A $5N - 9$ B $\frac{5N}{9}$ C $5N + 9$ D $- 4N$

(The distractors are the common wrong responses of the students which the evaluator came across during scoring).

Multiple-Choice Questions. The section contained twenty two items (21 to 43). Compared to the first twenty items, the questions in this section were relatively easy; the difficulty index varying between .24 and .76. Some of them are analysed below.

Item 21: Subtracting 2b-4a from 5a-3b

gives A a-5b B 9a-b

C -a+5b D 9a-5b

		A	B	C	D	Omit
Analysis	Top	15	19	11	55	0
Chart	Bottom	26	22	26	19	7

Interpretation:

(i) The item is hard, difficulty index being .37

(ii) It is positively discriminating the discrimination index being .39

(iii) The several distractors seem to be functionning reasonably well.

(iv) For this particular test and this particular group this item is performing well.

(v) It would, perhaps, have served better if placed towards the end of this section.

Item 28: K - K is equal to

A -2K B 2K C 1 D 0

Analysis		A	B	C	D	Omit
Chart	Top	0	0	15	81	4
	Bottom	11	7	0	71	11

Interpretation:

(i) The item is fairly easy, difficulty index being .76

(ii) The item is not significantly discriminating, the discriminating index being + .15 and is a good example of an item to be thrown away because of low r.[17]

(iii) The several distractors appear to be functioning reasonably well, except C which is negatively functioning.

(iv) About 4 percent from the top group and 11 percent from the bottom group omitted the item. It is probable that the omissions were caused by a lack of knowledge on the part of the students and not because of ambiguity of the item.

(v) Being an easy item it ought to have come earlier in this section.

[17] An item to be discriminating on this al gebra test should have a discriminating index of +.27 or above. This has been worked out by using the formula $2 \left(\frac{1}{\sqrt{N-1}}\right)$ explained earlier in this chapter.

The vaue .27 would be lower, of course, if there had been more persons tested, giving proportionally more in the top and to bottom 27%.

Item 30: The product of $x^4 . x^5 . x^2$ is

A x^{40} B x^{10} C x^{11} D x^8

		. A	. B	. C .	D	.Omit.
Analysis	Top	0	4	96	0	0
Chart	Bottom	52	4	37	0	7

Interpretation:

(i) The item is relatively easy, difficulty index being .66

(ii) It is positively discriminating, the discrimination index being .67

(iii) The distractor 'D' is not functioning. It is a dead weight in the item. Also detractor 'B' **is not discriminating.** However detractor 'A' attracted 52 percent of the lower group. The evaluator may infer that in the process of multiplication the students in the lower group multiply the powers instead of adding them. Instruction may be directed to drive out this erroneous notion from the minds of students. The item might also be offered to an experimental group as a completion item to find out some other plausible and functioning detractors besides 'A'.

Item 32: In $3/2\ n + \frac{1}{4}\ n - n = 9$, n equals

A 12 B 3 C -12 D 4

		. A .	B .	C .	D .	Omit.
Analysis	Top	22	15	11	4	48
Chart	Bottom	26	22	15	11	26

Interpretation.

(i) This item just did not work. It was keyed 'A' and more from the bottom group than from the top marked this alternative. The item thus is discriminating negatively, the discriminating index being - .06

(ii) The item is among the hardest items, the difficulty index being .24

(iii) 48 percent of the top and 26 percent of the bottom groups omitted the item. The large number of omissions is perhaps on account of the defective structuring of the item.

(iv) This item ought to be revised or dropped.

(v) Perhaps the item could function better if the denominator and the numerator of the fraction $\frac{3}{2}$ would have been placed one above the other like $\frac{3}{2}$ instead of placing them side by side with a slanting line dividing the two.

Organisation of the Test:

The subjective organisation of items can now be viewed objectively in the light of item analysis. By looking at each item as it is plotted on the graph[18] in relation to difficulty level and discrimination index the test constructor can improve the organisation of this test. It is, for example, clear that in general the items in the completion and matching sections were relatively more difficult than the

[18] See appendix H

items in the multiple-choice section. Hence the test ought to have started with the multiple-choice section. Again, the multiple-choice section can be graded in difficulty by placing the easier items like 24, 28, 40 in the beginning and the hardest items like 28, 35, 21 can be placed last of all. A similar process can be applied to items in the other two sections.

Again, a look at the graph may reveal that items which are low in discrimination and are too hard or too easy may be dropped, rewritten or revised. The test constructor, for example, may ponder over items such as 32, 35, 17, 16, 28 and decide whether they are worth trying again or not.

Validity of the Test As a Predictor of Grades Earned:

The validity of the test was measured by correlating the algebra grades of the students given by the teachers and the grades earned by them on the test. The following correlations were found.

Name of School	Number of Examinees	Correlation coefficient Algebra Test Vs. Teachers grades.
School A	16	.69
School B	30	.41
School C	19	.43
School D	34	.53

Reliability of the Test

The reliability coefficient of the test was found by the Kuder-Richardson formula explained in Chapter IV. It was possible to use this formula because the test was not speeded and each item was weighted one point. The reliability coefficient was calculated to be .92. The error of measurement was also calculated by the formula explained earlier.[19] The error of measurement was discovered to be nearly 3. There is thus 68% probability that an obtained score of say, 21, represents a true score, between 18 and 24.

The test constructor is of the view that the validity coefficients and the reliability coefficient found are surprisingly high for a test in the process of development. Both may be expected to increase, of course, when more reliable items are substituted for those discarded.

Concluding Statement.

The project presented in this chapter stresses careful work on test construction and interpretation of results. It is hoped that the mathematics teachers in Secondary Schools of Pakistan would follow as far as possible the process of test development as explained in the foregoing pages.

[19] See chapter IV.

CHAPTER VI

RECOMMENDATIONS AND CONCLUSIONS

The defects of the present system of evaluating mathematics achievement in Pakistani schools have been pointed out. It has been shown with sufficient vividness that the current mathematics tests encourage memorisation, are not well-structured, contain questions involving time-consuming calculations, lack reliability, have a narrow sampling of content, consist of questions which are constructed around more than one objective, do not have sufficient discriminating power and are tedious to score. Further, the grading system as it exists today presents a distorted picture of the progress of pupils, and the results of tests are not used to improve instruction or to make decisions. To improve the situation, the use of objective tests has been suggested and its advantages fully explored. The suggestion, however, does not mean that a complete replacement of essay type tests by objective tests is desirable. Such a complete change-over, it may be asserted, is hardly the object of the present study. The author, on the contrary, shares the view with thorndike that, "Neither the essay test nor the objective test is satisfactory as the sole type of test to measure academic achievement."[1]

It is held that each type of test should be used in "those

[1] Robert L. Thorndike & Elizabeth Hagan, Measurement & Evaluation in Psychology & Education, (New York, John Wiley & Sons Inc., 1961) p.41

situations[2] where its advantages are maximised and its weaknesses minimised."[3] To be brief, what is being said is that the _exclusive_ use of essay tests in their present state is not serving the needs of pupils,teachers and schools. The mathematics teachers in Pakistani schools, therefore, ought to know what objective tests are, how to construct them and how to make use of test results. In other words, the aim has been to introduce to the mathematics teachers the technique of objective testing and its effective use in evaluation. This problem, it is believed, has been squarely dealt with by explaining important phases of the technique in chapters III and IV and later by applying it in a specific practical situation which forms the theme of Chapter V.

The problem of how to initiate changes in the traditional system of evaluation, which is deeply rooted in the present day educational fabric, remains. Any attempt to thrust the change upon the teachers is likely to end in a fiasco. Well thought-out steps will have to be taken if any changes from traditional to more modern methods of evaluation is to succeed. To serve this end a plan is suggested below which, it is hoped, will prove helpful in facilitating and guiding the course of change in the existing educational system. Before describing the plan, the following recommendations based on the study carried out in the foregoing chapters are made:

Recommendations Relating to Teacher-made Tests

(i) The Pakistani teachers should make a frequent use of objective

[2] For situations when to use essay examination see Ibid., pp.50-53

[3] _Ibid._, p.41

tests to evaluate achievement in mathematics.

(2) The objective tests should be constructed with due regard to the principles of test-construction explained in this thesis.

(3) The classroom teachers should make use of test results to collect information diagnosing particular strengths and weaknesses of individual pupils in the way explained in Chapter IV. It may be stressed again, as has already been done, in chapter IV, that the teacher may learn much by examining the individual papers closely, and studying the pattern of class responses to individual items. If the teacher studies the pattern of right and wrong answers,[4] he may discover classwide misconceptions and lines of necessary class work for the future.

(4) In schools where the mathematics staff is large enough, it is suggested that the more important tests are constructed by a committee instead of a single teacher. In addition to the improvement in item writing, the discussion of which items to be included and rejected is likely to require discussion of the objectives being tested and thus lead to better definition of objectives.

(5) The system of percentage grading in schools may be abolished and in its place the classroom teachers should make use of standard scores. A system such as this (or any other similar system suited to needs of schools) based on standard scores may be used:

[4] To study the pattern of right and wrong answers, a remedial chart has been suggested in Chapter IV.

A for + 1.5 S[5] or greater
B for + .5 s upto 1.5 s
C for - .5 s upto + .5 s
D for - .5 s down to - 1.5 s
E for - 1.5 S or below.

Such a system automatically preserves the shape of distribution while assigning a grade to a student strictly in terms of his standing relative to the rest of his class. There is, however, nothing particularly sacred about the boundaries and "S" intervals; they should, in fact, be adjusted in situations where for example, it is clear to the teacher that no real failures have occurred or where the top grade is inappropriate for any class member.

Recommendations Relating to Board Tests

(1) The Board tests should be mainly objective, but a portion of the tests, should contain essay type questions which may later be studied for their validity in predicting future academic success, along with the objective questions. Such studies can help in the decision as to whether more or fewer essay-type questions should be adopted in the future. In making use of essay type questions some basic principles should be kept in mind,[6] e.g., each question should measure one defined objective of instruction; pupils should not be allowed to make a choice among several questions; a standard answer should be formulated in which a specific number of credits is allotted to each significant point; the questions should not be too lengthy,

[5] S Stands for standard deviation.

[6] For detailed treatment of these principles, the reader is referred to R.L. Thorndike and Elizabeth Hagen, Op. Cit., PP.53-56

involving time-consuming calculations; the questions should test the understanding of the principles of mathematics and not merely require a reproduction of memorised material, such as the reproduction of theorems in geometry in the present board tests.

(2) The Board should obtain the services of interested and experienced mathematicians to serve on a committee responsible for test-construction. Intensive work by the committee should precede the final assembly of any test. Each year new kinds of test questions should be explored. The specialists should write large numbers of questions and later review them for possible misinterpretations, their suitability to the general purpose of the test, and their relative level of difficulty. After the questions that are considered satisfactory by the specialists have been selected, they should be tried out on groups of students comparable to those who will be taking the test. Before giving a final form to the test a number of additional questions may have to be discarded on the basis of analysis of the responses of students.

(3) The Board tests every year should contain a number of recently-devised but experimental mathematical questions. The students should not know which questions are experimental. The responses to experimental items should not be used in arriving at test scores. A separate analysis of the experimental questions should be made by the Board every year, and those questions which function well should be offered as regular items, in the following year's examination.

(4) The Board should discard the use of arbitrary grade bands, namely, First Division, Second Division and Third Division, and in its place the performance of the pupils be judged on the basis of percentile norms.[7] Instead of having the achievement of 33 per cent of total marks as the perenially fixed pass mark, the Board should decide about the pass mark every year, basing the decision on the difficulty level of the test found after analysis of results, the general performance of the examinees and related factors.

General Recommendations

(1) The Board invariably should analyse the test results to yield information about the test by employing the method of item analysis explained in Chapter IV and practically illustrated in Chapter V. Classroom teachers will produce better classroom tests if, they too, undertake such analysis.

(2) The possibilities of making city-wide tests, district-wide tests, division-wide tests, and developing grade norms for such standardised tests should be fully explored.

The Plan

(1) It is necessary first of all that the mathematics teachers should feel a need for objective tests and should realise that for certain purposes such tests are the best devices available to them. Usually teachers gladly do a great deal of work and willingly spend a great deal of time and effort in carrying out a project if convinced of its utility

[7] The use of percentile norms is explained in Chapters IV and V.

in the school system. If they are not convinced of the desirability of any project they may remain lukewarm or even resentful. It is, therefore, essential that teachers understand the weaknesses of the traditional tests and see the advantages of the new type tests. One way of making the mathematics teachers appreciate the need for objective tests is to use Mathematics Teachers' Conferences for demonstrating a series of simple experiments aiming to answer questions similar to the following:

(i) To what extent do teachers vary when each is required to work on identical answer paper to one of the conventional types of mathematics tests in use at the present time ?

(ii) To what extent do teachers vary in their opinions of the order of difficulty of a number of essay-type questions and of the difficulty of any one question ?

(iii) To what extent do teachers' estimates of the difficulty of a mathematical problem differ from its actual difficulty determined by a survey of pupils' reactions ?

(iv) To what extent is the total range of the content of mathematics represented in a test of the conventional type, which is designed to evaluate achievement over a period of six months to one year ?

(v) What are the advantages of expressing test scores[8] as derived scores rather than percentages?[9]

[8] For the meaning of derived scores see chapter IV

[9] For the disadvantages of per cent system of grading see Chapter II.

It is perhaps unnecessary to add that data actually collected by research in Pakistan would be the most convincing data to present to answer these questions at these conferences.

Carrying out experiments and presenting research data related to the questions stated above will, it is hoped, serve as a practical and concrete basis for a gradual and teacher-supported change toward more objective-type examination items. Those connected with the work will, it is believed, feel convinced of the utility of objective tests.

(2) The next step is to ensure adequate preparation for the work of making tests and treating test results. This preparation may be begun by a careful study of this thesis, which contains necessary material to guide in test construction and treatment of scores, and also provides a sample objective test[10] as well as some sample objective questions in Chapter IV.

A more practical way to train teachers in making objective tests is, however, to give instruction and practice in summer sessions for mathematics teachers, extended over periods of about six weeks. To begin with, such a session may be convened at Lahore or some other suitable place under the auspices of the Education Directorate and in collaboration with the Board of Secondary Education, Lahore. The Director of Public Instruction, Lahore, may ask each district inspector of schools who is within the

[10] See appendix B.

limits[11] of the Education Board to send a specified number[12] of capable mathematics teachers (say five teachers from each district making a total of 95), to attend the session.

Meanwhile, the Director of Public Instruction should also obtain the services of a suitable number of experts[13] in the field of objective testing for the purpose of training the teachers during the session. The group of teachers may be split into smaller groups or committees, each committee to be headed by a leader from among the experts. One of the experts should be selected to work as general coordinator. Each committee may be assigned[14] to construct test items for a particular class. One committee, for example, might construct a test in Geometry for the eighth class, another making a test in Algebra for the nineth class and so on. The items constructed by the various committees should be assembled by the committee leaders. The committees then may discuss, under the supervision of the committee leaders, how some items could be improved, which items are to be selected for the test, and which of them are to be discarded. After the selection of items for each test has been made, each committee should prepare a final draft of the test, to be printed or mimeographed

[11] Nineteen districts fall within the territorial jurisdiction of the Board. These districts are, Lahore, Sheikhupura, Gujranwala, Sialkot, Multan, Jhang, Lyallpur, Montgomery, Bahawalpur, Bahawalnagar, Rahim Yar Khan, Dera Ghazi Khan, Muzzafargarh, Campbellpur, Mianwali, Rawalpindi, Jhelum, Gujrat, Shahpur.

[12] The number will depend on local conditions and the availability of the resources of training in the sessions.

[13] It should not be difficult for the Director to acquire the services of some capable professors from the various Teachers' Training Colleges in the country.

[14] The assignments to be the results of discussions among the members of the committees and the leaders.

after having been edited by the committee leader. Scoring keys will also be prepared by the various committees and the experts will explain the methods of administring and scoring of the tests to the members. The session at this stage may end. The coordinator should announce the date for the next session and request the participants to administer the tests in their schools. The teachers may also be asked to bring statements of raw scores of the examinees to the next meeting.

The sedond session of these teachers should be held on the scheduled dates. In this session the committee leaders will provide sufficient understanding of the elementary statistical concepts involved in testing to prevent the teachers from misusing and misunderstanding the test results. Only a very little in the way of statistics is necessary,[15] but this small amount should be thoroughly learned. The committee leaders will acquaint each teacher with such fundamental concepts as median, standard deviation, percentiles, standard scores etc and with some basic facts such as the existence of a probable error of measurement in every test score (even when the scoring is completely accurate). The committee leaders, while interpreting the scores, would demonstrate by practical examples how the scores could be used in organising the class or school, in deciding how and what to teach each pupil or group of pupils, how to determine where a particular teacher needs to improve his teaching, and how the tests can be further improved by the process of item analysis.

(3) At the end of such a program each district will have a number of teachers trained in the technique of objective testing.

[15] See chapter IV.

Now the District Inspectors of Schools can convene conferences of mathematics teachers in which the teachers of the District already trained in the technique can teach the procedure to those who attend such conferences.

(4) The convening of conferences will not have to be done year after year, for once the majority of mathematics teachers starts to make use of objective tests, the new teachers will continue to learn the technique from the experienced hands, and thus follow suit.

(5) It is necessary to mention here that the pupils in the schools, at present, are entirely unaccustomed to the objective types of test questions. It would, therefore be advisable to give them practice on the types of questions used, otherwise the newness of the experience may interfere with their performance on the actual tests.

The Outcomes of the Plan

Once the mathematics teachers are convinced of the desirability of objective tests and a majority of them are trained in the technique in accordance with the plan set forth, it will be possible for the Board, the education officers, and the classroom teachers to use such tests in evaluating achievement in mathematics.

The Board can appoint a committee consisting of some capable teachers to develop objective tests for the Secondary School Examination. The committee in course of time would be able to develop norms and standardised tests for the region within the territorial limits of the Board.

The District Inspectors of Schools and the Divisional Inspectors of Schools will be able to develop any of the following tests according to specific needs and demands of varying situations:

(i) City-wide tests.

(ii) District-wide tests

(iii) Division-wide tests.[16]

Such testing instruments will be helpful in achieving greater uniformity in grading. The scores interpreted in the light of percentiles on these tests could be used, along with other information, in arriving at the pupils' final grade, Such tests may also provide a good motivation, especially for review, as the students may feel interested in their own scores and how, as individuals and as a class, they compared with city percentiles, district percentiles, etc. The use of such tests might also cause the teachers to adhere more closly to the prescribed course of study.

Classroom teachers will be able to make frequent use of short objective tests to locate the weaknesses of pupils. The results of a short objective test, based on a single unit of subject matter, if administered before the unit is presented, can tell the teacher how much the pupils already know. By administring the test again after testing the unit, the teacher can form an idea of the growth or progress of individual pupils, can tell how well the unit has been learned by the class; what particular points he should

[16] District-wide and Division-wide tests may have to be further categorised as urban, suburban and rural.

again emphasise and in what way he should improve his method of teaching to make the pupils learn effectively.

The training of classroom teachers in the use of statistical methods will help them in interpreting scores, in improving tests and in the use of standard scores to replace the easily misused system of per cent grading.[17]

Concluding Statement and a Caution

The thesis has advocated the use of objective tests to evaluate mathematics achievement. Let it, however, be borne in mind that it has not been merely a case of better tests and a better method of stating test results which has been advocated, but also of finding out where pupils and teachers need improvement and how to effect these improvements. The effort to use objective tests will succeed to the extent that the pupils are discouraged from learning to pass tests in which, as the Report of the Pakistan Commission on National Education has stated, "...success can be achieved through mere memorisation."[18] In other words tests are not to be considered as ends in themselves but as means to ends.

It is further to be stated that while the writer has presented many suggestions, these can by no means be considered to constitute an ideal program to fit any situation without modifications. Whatever may be ideal for one situation may be far from ideal for another.

[17] The Short-comings of per cent grading system were shown in Chapter II.

[18] Government of Pakistan, Ministry of Education, Report of the Commission on National Education, (Karachi: Manager of Publications, 1960) p.123.

Every school, therefore, has to design the evaluation program according to its particular needs. But the thesis, it is sincerely hoped, may serve as a guide and reference in the solution of many practical problems that teachers and schools may face in devising improved evaluation programs.

APPENDIX A

COVERING LETTER AND QUESTIONNAIRE

P.O.Box 1294,
American University of Beirut,
Beirut, Lebanon.

December 3, 1962.

Dear Sir,

I am studying at the American University of Beirut for M.A. (Education). By way of partial requirement of the course, I am writing a thesis on "Proposals for the Evaluation of Mathematics Achievement in Secondary schools affiliated to the Punjab Education Board." I understand that you have been associated with the evaluation of mathematics achievement at the Secondary School level. Your views on the problem will be of great value to me in my work. A questionnaire is being enclosed to elicit your opinion. Kindly express your views on the questions in the spaces provided for the answers. The questions are not restricted to the examination conducted by the Secondary Board of Education but also cover the overall system of evaluation followed by classroom teachers of mathematics.

You may skip questions which you prefer not to answer. If instead of answering the questionnaire, you prefer to send me your critical appraisal of the present system of evaluation, please do so.

I shall deem it a great favour and will feel highly obliged if you could kindly reply as early as possible because I am very much hard pressed for time.

Sincerely yours,

A. H. Wain.

(By "Tests" is meant the present day mathematics Tests.)

1. Are the tests in your opinion, well planned in the sense that the objectives of the course are clearly envisaged before preparing them?
....

2. Do you think the tests to be sufficiently comprehensive? In other words, is there a liberal sampling of all phases of subject matter?
....

3. Would you like the tests to be more comprehensive? ....

4. Do the tests, in your view, measure small differences in achievement?
....

5. Do the tests encourage memorising of theorems in Geometry? ....

6. Do the tests encourage rote learning of some processes of typical Arithmetic and Algebra problems (Viz problems on Calendar, Practice, Etc.)
....

7. Do you think that the mathematics answer papers if marked by different examiners will yield varying scores? ....

8. Do you consider the scoring of answer papers in the present system arduous and should therefore, be made easier?.............................

9. Are you in favour of reducing the duration of examination time (say from 3 hours to 2 or less or no time limit? ....
....

10. Do you think that some questions on Arithmetic tests are time-consuming as they involve long computations? ....
....

11. Do you think it could be possible for some students to get a pass mark by mastering some set parts of the course? ....
....

12. Are you satisfied with the present marking system? If not what system would you recommend? ....
....
....

13. Do you think it would be of greater value if tests are constructed by a team of experts than by a single examiner? ....

14. Do the classroom tests (Teacher grade) take in view the individual differences of the pupils? ....
....

15. In the existing system, what are generally held to be the purposes of testing? (Please do not mention what you think ought to be the purpose, but only indicate what they are in the existing system.) Please, put tally marks against as many of the following statements as answer the query.

Tests are used for

a. reporting progress to parents
b. promotion to a higher grade
c. testing computational skills
d. testing application to life problems to some extent
e. appropriate placement of students
f. guiding learning
g. identifying trouble spots and difficulties of students.
h. discovering the mode of learning employed by students - whether the achievement is the outcome of thinking process or rote learning
i. improving the mathematics programme at school in terms of curriculum content
k. deciding about the sequence of topics and setting up of goals of instruction
l. determining levels of understanding of pupils
m. assigning marks
n. testing understanding of pupils
o. any other purposes not mentioned above (please specify)
....
....
....

16. Do you support the idea of tests, having a large number of short items, covering wider ground, but stressing understanding of method and principles, and giving less weight to detailed calculations?......

17. What kind and type of tests, in your opinion should replace the present day tests? Please express your opinion freely and frankly.

APPENDIX B

Time 45 minutes

Elementary Algebra Test

__________________ Name   __________________ School   ______________ Class

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.

This test contains the following types of questions:

Matching Questions

Directions: For each statement in column I, choose a word belonging to it from column II and write in the blank space provided before each statement, the capital letter mentioned before the word you choose. (One is done for you, do the second yourself)

Sample 1 :	Column I	Column II
B	The number by which we divide is called	A a difference
___	The result of addition is called	B a divisor
		C a sum

Completion Questions

Directions : Complete the following statement by writing answers in the blanks provided. (One is done here, do the second yourself.)

Sample 2 : Subtracting 8 from 15 gives ___7___

Sample 3 : If we multiply 9 by 4, we get __________

Multiple Choice Questions

Directions : Make a circle around the letter in front of the correct answer.

Sample 4 : 9 - 4 + 3 equals

A 2
B 15
[✓ C] 8
D 10

(Note that C has been encircled, because 8 is the correct answer. Now do sample 5 yourself.)

Sample 5 : If 2x = 4, then x equals

A 2
B 1
C 4
D 8

Directions about Guessing

Don't make a wild guess on multiple-choice questions and matching questions because your score will be computed by subtracting a portion of your wrong answers from your correct answers.

If a question is difficult skip it and go on, returning to it later if you have time. No one is expected to answer all the questions in the time allowed.

Scratch work You may do calculations or extra work on the blank sides of the question paper.

Matching Section

For each statement in column I choose a word belonging to it from column II and write in the blank space provided before each statement the capital letter mentioned before the word you choose. You may use the same letter as many times as you need it.

Column I

1. ______ $\frac{a}{b}$ represents

2. ______ the expression $x + Y - a + b$ is

3. ______ $a - b$ represents

4. ______ the term $8Y^2$ is of the second

5. ______ $3x$, $4x$, $10x$ represent

6. ______ In ab^2c the symbols are

Column II

A a polynomial

B factors

C similar terms

D a root

E a binomial

F a formula

G degree

H a ratio

Column I

7. _____ the expression $x^2 + 2x + y^2$ represents

8. _____ In the fraction $\frac{p+q}{r}$ the symbol r is

9. _____ $A = 1 \times b$ represents a

10. _____ In $8p^2$ the number 8 is a

11. _____ In the indicated root $\sqrt{p + q}$, the symbol $\sqrt{\ }$ is a

12. _____ If p, q, r, denote three numbers pqr is a

Column II

A denominator

B remainder

C trinomial

D radical sign

E formula

F quotient

Completion Section

Complete the following statements by writing answers in the blank spaces provided.

13. The difference between m and 3 is ________

14. Khalid is 14x years old, his age four years from now will be ________

15. Five times N diminished by 9 is ________

16. m increased by n gives ________

17. In $6y^3$, the exponent is ________

18. Stating in symbols p is 3 less than q will be ________

19. If $a = -4$, $b = 2$, $c = 0$ the value of $6a - (3b^2 - 6c)$ is ________

20. If $x = 5$, $y = 3$ the value of $x^2 - 5xy + y^2$ is ________

Multiple - Choice Section

Each problem below is followed by four choices, only one of which is the correct answer. Make a circle around the letter in front of the correct answer.

21. Subtracting $2b - 4a$ from $5a - 3b$ gives

A $a - 5b$

B $9a - b$

C $-a + 5b$

D $9a - 5b$

22. Adding $2x^2 - 5x - 3$ and $x - 1 + 3x^2$ gives

A $5x^2 - 4x - 4$

B $5x^2 - 6x - 4$

C $5x^2 - 4x + 4$

D $6x^2 - 4x - 4$

23. Subtracting $- m^2 + 6m$ from $3m^2 - m + 4$ equals

A $2m^2 - 5m + 4$

B $4m^2 - 5m + 4$

C $4m^2 - 7m + 4$

D $2m^2 - 7m + 4$

24. Multiplying $(5mn^2)$ by $(-8mn)$ equals

A $40m^2n^3$

B $-40m^2n^3$

C $13 m^2n^3$

D $-40mn^2$

25. Dividing $(-28abc^4)$ by $(-4ac)$ equals

A $7bc^3$

B $-7bc^3$

C $7abc^5$

D $7a^2bc^5$

26. Multiplying $(5x-6y)$ by $(-3y -2x)$ equals

A $18y^2 - 3xy - 10x^2$

B $-18y^2 - 27xy - 10x^2$

C $18y^2 -27xy -10x^2$

D $-18y^2 -3xy + 10x^2$

27. Dividing $(3m^2 - 7m - 6)$ by $(m-3)$ equals

A $3m^2 - 2$

B $3m -2$

C $3m + 2$

D $3m^2 + 2$

28. $k - k$ is equal to

A $-2k$

B $2k$

C 1

D 0

29. $\frac{4x + y}{4}$ is equal to

A $x + y$

B $4x + 4y$

C $x + \frac{y}{4}$

D $\frac{x + y}{4}$

30. The product of $x^4. x^5. x^2$ is

A x^{40}

B x^{10}

C x^{11}

D x^8

31. If $3b + 3 = 45 - 4b$, b equals

A -3

B 4

C 6

D -2

32. If $3/2n + \frac{1}{4}n - n = 9$, n equals

A 12

B 3

C -12

D 4

33. If $5x - 3(4x - 5) = 1$, x equals

A 2

B 0

C -1

D 5

34. If $4m - 7c = c$, m equals

A 2

B $3c$

C $2c$

D $-2c$

35. If $\frac{2k + 3}{2} - \frac{3k + 1}{4} = 1$ k equals

A 2

B 3

C 4

D -1

36. The factors of $25x^2 - 1$ are

A $(25x - 1)(25x + 1)$

B $(5x - 1)(5x + 1)$

C $(5x^2 + 1)(5x^2 - 1)$

D $(5x + 1)(5x^2 - 1)$

37. The factors of $x^2 - 9x + 18$ are

A $(x-3)(x-6)$

B $(x-9)(x+2)$

C $(x-3)(x-9)$

D $(x-6)(x-9)$

38. The factors of $100x^2 + 25$ are

A $25(2x + 1)(2x - 1)$

B $(10x + 5)(10x - 5)$

C $(10x + 5)(10x + 5)$

D $25(4x^2 + 1)$

39. The factors of $x^2 - 5x - 6$ are

A $(x -3)\ (x -2)$

B $(x - 3)\ (x + 2)$

C $(x - 6)\ (x + 1)$

D $(x + 3)\ (x + 2)$

40. The simplest form of

$\frac{25a^3b}{-100ab^2}$ is

A $\frac{a^4b^3}{-4}$

B $-\frac{a^2}{4b}$

C $\frac{a^2b^3}{-4}$

D $\frac{a^2b^3}{4}$

41. Changing $\frac{x^2-y^2}{x^5}$ X $\frac{x^2}{x-y}$

to simplest form gives

A $\frac{-yx}{x^3}$

B $\frac{x + y}{x^3}$

C $\frac{x - y}{x^3}$

D $(x^3 - y^3)\ (x^{10})$

42. When reduced to lowest terms

$\frac{36a^2b^3c}{4a^2bc}$ becomes

A $\frac{9b^4}{c^5}$

B $\frac{9b^2}{c^3}$

C $\frac{9}{b^4\ c^5}$

D $9b^3c^4$

43. After simplification

$\frac{x^2 -9}{x^2+5x+6}$ reduces to

A $\frac{x + 3}{x + 2}$

B $\frac{1}{5x -3}$

C $\frac{-9}{5x + 6}$

D $\frac{x - 3}{x + 2}$

APPENDIX C

PROPORTION OF SUCCESSES IN THE 27% SCORING HIGHEST ON THE CONTINUOUS VARIABLE

02	06	10	14	18	22	26	30	34	38	42	46	50	54	58	62	66	70	74	78	82	86	90	94	98	PROPORTION OF SUCCESSES IN THE 27% SCORING LOWEST ON THE CONTINUOUS VARIABLE
0	19	30	37	43	48	51	55	58	61	63	66	68	70	72	73	75	77	79	80	82	84	86	88	91	02
	0	11	19	26	31	36	40	44	47	50	53	56	59	61	64	66	68	71	73	76	78	81	84	88	06
		0	08	15	21	26	30	34	38	41	45	48	51	54	57	60	63	65	68	71	74	77	81	86	10
			0	07	12	18	22	27	31	34	38	42	45	48	51	54	57	60	63	67	70	74	78	84	14
				0	06	11	16	20	25	28	32	36	39	43	47	49	53	56	60	63	67	71	76	82	18
					0	06	10	15	19	23	27	31	34	38	42	45	49	52	56	60	63	68	73	80	22
						0	05	09	14	18	22	26	30	33	37	41	44	48	52	56	60	65	71	79	26
							0	04	09	13	17	21	25	29	33	37	40	44	49	53	57	63	68	77	30
								0	04	09	13	17	21	25	29	33	37	41	45	49	54	60	66	75	34
									0	04	08	13	16	20	25	29	33	37	42	47	51	57	64	73	38
										0	04	08	12	16	20	25	29	33	38	43	48	54	61	72	42
											0	04	08	12	16	21	25	30	34	39	45	51	59	70	46
												0	04	08	13	17	21	26	31	36	42	48	56	68	50
													0	04	08	13	17	22	27	32	38	45	53	66	54
														0	04	09	13	18	23	28	34	41	50	63	58
															0	04	09	14	19	25	31	38	47	61	62
																0	04	09	15	20	27	34	44	58	66
																	0	05	10	16	22	30	40	55	70
																		0	06	11	18	26	36	51	74
																			0	06	12	21	31	48	78
																				0	07	15	26	43	82
																					0	08	19	37	86
																						0	11	30	90
																							0	19	94
																								0	98

A Table of the Values of the Product -Moment Coefficient of Correlation in a Normal Bivariate Population Corresponding to Given Proportions of Success.

(Adapted from the Original Table computed by John C. Flanagan)

DECIMALS ARE OMITTED FOR GREATER CLARITY

Note: Negative values are obtained by reading "HIGHEST" for "LOWEST" and vice-versa

APPENDIX D *

A Summary of Maxims for Item Writing

From Chapter 4 of Thorndike and Hagen, Measurement and Evaluation in Psychology and Education. Numbers in parentheses refer to the page of the second edition on which the supporting text material is found; the number following the dash refers to the number of the maxim on that page.

General Maxims

(61-1) Keep the reading difficulty of test items low.
(61-2) Do not lift a statement verbatim from the textbook.
(62-3) If an item is based on opinion or authority, indicate whose opinion or what authority.
(62-4) In planning a set of items for a test, take care that one item does not provide cues to the answer of another item.
(63-5) Avoid the use of interlocking or interdependent items.
(63-6) In a set of items, let the occurrence of correct responses follow essentially a random pattern.
(63-7) Avoid trick and catch questions.
(63-8) Try to avoid ambiguity of statement and meaning.
(65-9) Beware of items dealing with trivia.

Maxims for True-False Items

(66-1) Be sure that the item as written can be unequivocally classified as either true or false.
(67-2) Beware of "Specific Determiners".
(67-3) Beware of ambiguous and indefinite terms of degree or amount.
(68-4) Beware of negative statements and particularly of double negatives.
(68-5) Beware of items that include more than one idea in the statement, especially if one is true and the other is false.
(68-6) Beware of items where the correct answer depends upon one insignificant word, phrase, or letter.
(69-7) Beware of giving cues to the correct answer by the length of the item.

Maxims for Short-Answer and Completions Items

(69-1) Beware of indefinite or "Open" completion items.
(70-2) Omit only key words.
(70-3) Don't leave too many blanks in a statement.
(70-4) Blanks are better put near the end of a statement rather than at the beginning.
(70-5) If the problem requires a numerical answer, indicate the units in which it is to be expressed.

* Adapted for use in a course in Tests and Measurement, American University of Beirut.

Maxims for Multiple-Choice Items

(73-1) The stem of a multiple-choice item should clearly formulate a problem.
(73-2) Include as much of the item as possible in the stem.
(74-3) Don't load the stem down with irrelevant material.
(74-4) Be sure that there is one and only one correct or clearly best answer.
(75-5) Items designed to measure understandings, insights, or ability to apply principles should be presented in novel terms.
(75-6) Beware of clang associations.
(76-7) Beware of irrelevant grammatical cues.
(76-8) Beware of the use of one pair of opposites among the options if one of the pair is the correct or best answer.
(77-9) Beware of the use of "None of these", "None of the above", "All of these", and "All of the above" as options.
(78-10) Use the negative only sparingly in the stem of an item.

Maxims for Matching Items

(80-1) When writing matching items, the items in a set should be homogeneous.
(80-2) The number of answer choices should be greater than the number of problems presented. (Unless using master list -- see pp. 80-81)
(80-4) Response options should be arranged in a logical order, if one exists.
(80-5) The directions should specify the basis for matching and should indicate whether an answer choice may be used more than once.
(80-3) The set of items should be relatively short.

APPENDIX E

Percentile Norms on Elementary Algebra Test

Maximum Possible Score 43

Raw Score	Percentile Rank
39	99
38	98
37	98
36	97
35	95
34	93
33	90
32	87
31	85
30	83
29	81
28	78
27	77
26	76
25	76
24	75
23	73
22	71
21	68
20	64
19	61
18	59
17	57
16	54
15	52
14	50
13	47
12	44
11	36
10	33
9	29
8	21
7	18
6	13
5	8
4	4
3	1
Number of pupils tested	99
Highest score reported	39
Lowest score reported	3

APPENDIX F

Elementary Algebra Test - Item Analysis

(Numbers are percentages of top and bottom groups--these groups being the top 27% and bottom 27% of examinees. The correct response for each item has been indicated by underlining.

Matching Questions

Item No.	A	B	C	D	E	F	G	H	Omit
1	4	18	0	4	0	0	0	56	18
	7	37	0	11	0	0	0	4	41
2	71	7	0	0	4	11	0	0	7
	0	4	15	7	4	26	0	7	37
3	4	4	0	0	81	4	0	0	7
	7	11	7	0	0	19	0	4	52
4	0	0	0	4	4	0	89	0	4
	7	4	0	19	0	11	7	0	52
5	0	0	78	0	0	4	0	7	11
	7	4	23	7	0	7	7	7	38
6	0	44	4	7	0	15	0	4	26
	0	0	7	4	19	0	4	4	62
7	0	0	74	0	11	0	4	0	11
	4	0	7	0	7	4	15	7	56
8	100	0	0	0	0	0	0	0	0
	22	15	0	11	4	4	4	0	40
9	0	0	0	0	62	0	19	0	19
	7	0	4	0	19	11	11	7	41
10	0	0	0	0	0	0	0	81	19
	4	4	11	0	0	15	7	4	55
11	0	0	0	63	11	7	0	0	19
	0	0	4	19	4	7	0	4	62
12	0	0	7	7	4	7	70	0	4
	7	4	15	0	4	7	0	7	56

Completion Questions

Item No.	Right	Wrong	Omit	Item No.	Right	Wrong	Omit
13	96	0	4	17	37	22	41
	63	15	22		11	37	52
14	63	33	4	18	59	37	4
	11	70	19		4	63	33
15	67	22	11	19	41	52	7
	7	45	48		4	66	30
16	74	22	4	20	63	30	7
	56	37	7		7	71	22

Multiple-Choice Questions

Item No.	A	B	C	D	Omit	Item No.	A	B	C	D	Omit
21	15	19	11	55	0	25	89	11	0	0	0
	26	22	26	19	7		44	30	0	7	19
22	96	4	0	0	0	26	92	4	4	0	0
	36	19	30	0	15		30	18	11	11	30
23	11	0	63	15	11	27	7	7	80	0	4
	40	15	11	15	19		7	15	26	22	30
24	0	96	4	0	0	28	0	0	15	81	4
	22	49	0	22	7		11	7	0	71	11
29	19	4	55	7	15	37	96	0	0	0	4
	51	4	11	19	15		26	19	15	7	33
30	0	4	96	0	0	38	0	11	4	81	4
	52	4	37	0	7		11	4	19	22	44
31	0	0	93	0	7	39	15	7	78	0	0
	15	11	55	0	19		15	7	15	7	56
32	22	15	11	4	48	40	0	93	0	0	7
	26	22	15	11	26		4	59	18	4	15
33	85	0	4	0	11	41	0	63	30	0	7
	44	7	19	11	19		4	11	44	19	22
34	7	0	86	0	7	42	0	96	0	0	4
	45	22	19	7	7		11	48	4	15	22
35	4	15	7	37	37	43	0	0	7	86	7
	22	19	11	29	19		4	19	36	15	26
36	4	92	0	0	4						
	29	26	19	7	19						

APPENDIX G

Table showing Difficulty Index and Discrimination Index of Each Item on the Algebra Test

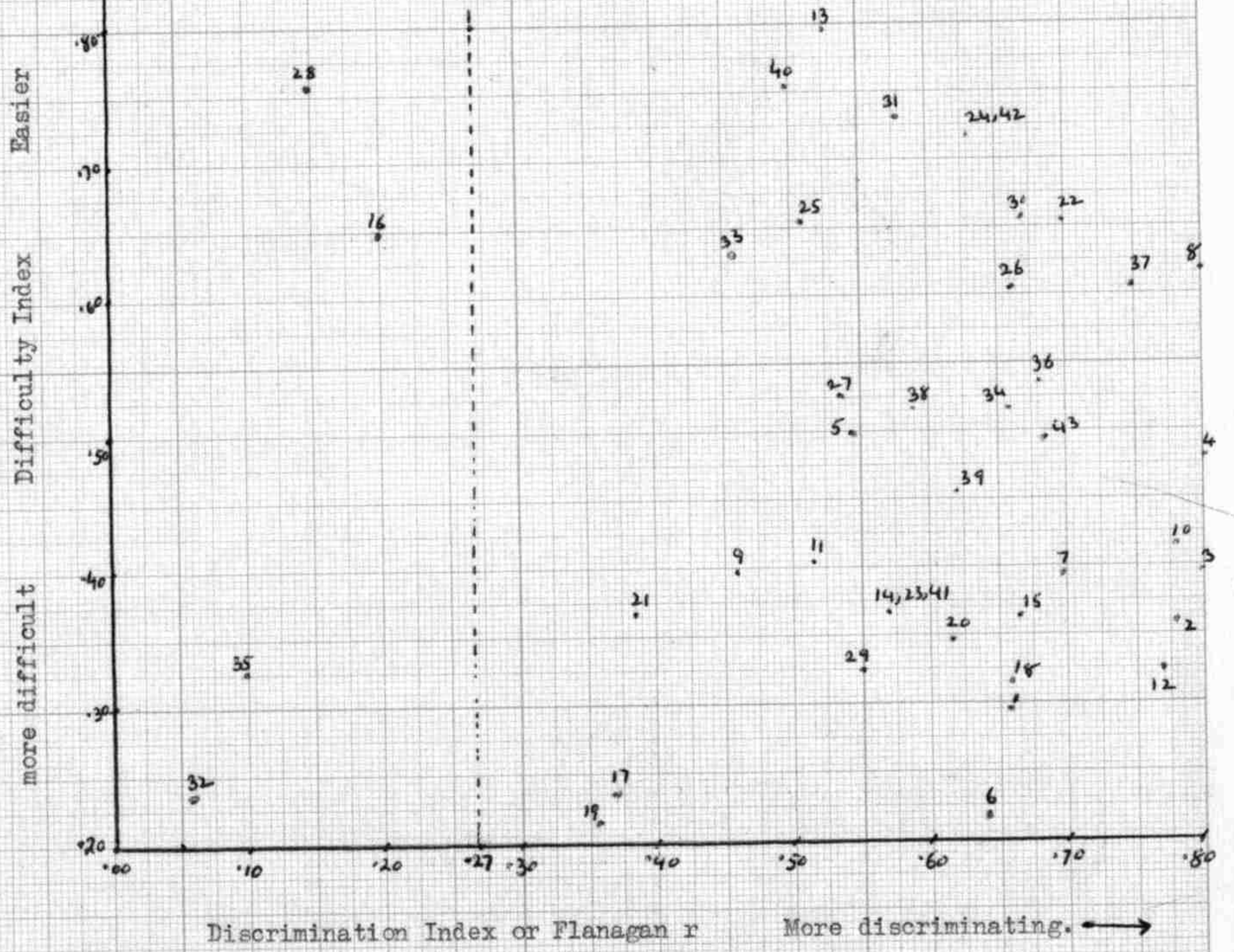
Item	Difficulty		Item	Difficulty		Item	Difficulty	
No.	Index	Flanagan r	No.	Index	Flanagan r	No.	Index	Flanagan r
1	.30	.66	20	.35	.62	39	.46	.62
2	.36	.78	21	.37	.39	40	.76	.50
3	.40	.80	22	.66	.70	41	.37	.57
4	.48	.80	23	.37	.57	42	.72	.63
5	.50	.55	24	.72	.63	43	.50	.69
6	.22	.64	25	.66	.51			
7	.40	.70	26	.61	.66			
8	.62	.80	27	.53	.54			
9	.40	.46	28	.76	.15			
10	.42	.78	29	.33	.50			
11	.41	.47	30	.66	.67			
12	.35	.77	31	.74	.53			
13	.80	.53	32	.24	-.06			
14	.37	.57	33	.64	.46			
15	.37	.66	34	.52	.66			
16	.65	.20	35	.33	.10			
17	.24	.37	36	.54	.68			
18	.32	.66	37	.61	.75			
19	.22	.56	38	.52	.59			

APPENDIX H

Graph Showing Difficulty Index and Discriminating Index

of Each Item on the Algebra Test

(Numbers are Item Numbers)



The broken line is drawn at .27 r . It shows that items on the left of the line are not sufficiently discriminating and the test-constructor has to make a decision if these items are to be included in the revised test or not. (N.B. the value .27 is not fixed, but varies from test to test, depending on the number of papers used in obtaining data).

BIBLIOGRAPHY

Aiken, D. J. and others. Algebra Book One.
New York: McGraw-Hill Book Co., Inc., 1960.

Bean, Kenneth L. Construction of Educational and Personal Tests.
New York: McGraw-Hill Book Co., Inc., 1953

Bradfield, James M. and H. Steward Moredoek - Measurement and
Evaluation in Education. New York: The Macmillan Co., 1957

Baron, Dennis and Harold W. Bernard. Evaluation Techniques For
Classroom Teachers. New York: McGraw-Hill Book Co., Inc., 1958.

Board of Secondary Education, Lahore. Secondary School Examination,
1961-62 (Syllabus)

______________________. The Calendar 1959-61.Lahore, 1958

______________________. Book of Instruction for Examiners.
Lahore: 1962

Blommers, Paul and E.F. Lindquist. Elementary Statistical Methods.
Boston: Houghton Mifflin Co., 1960

Cliff, Marian C. "The Place of Evaluation in the Secondary School
Program." The Mathematics Teacher
49: 270-73, (April, 1956)

Downie, M. M. and R.W. Heath. Basic Statistical Methods.
New York: Harper & Brothers, 1959.

Epstein, Marion and Sheldon Myres. "How a Mathematics Test is Born."
The Mathematics Teacher. 51: 299-302, (April, 1958).

Flanagan, J.C. "General Considerations in the Selection of Test Items
and a Short Method of Estimating the Product-Moment Coefficient."
Journal of Educational Psychology. 30: 674-680, 1939

Greene, Harry A. and others. Measurement and Evaluation in the
Secondary School. N. York: Longmans Green & Co., 1943

Government of Pakistan, Ministry of Education. Report of the
Commission on National Education. Karachi: 1960.

Hawkes, Herald E. and others. The Construction and Use of Achievement
Examinations. Boston: Mifflin & Co., 1936.

Jurdak, Mansur Hanna. School Algebra Book One. Beirut: American Press, 1957.

Lipitt, Ronald and others. The Dynamics of Planned Change. New York: Harcourt Brace and World Inc. 1958.

Miel, Alice. Changing the Curriculum. New York: Appleton. Century-Crofts, Inc., 1946.

Micheels, William J. and M. Ray Karnes. Measuring Educational Achievement. New York: McGraw-Hill Book Co., Inc. 1950

Noll, Victor H. Introduction to Educational Measurement. Boston: Houghton Mifflin Co., 1957.

Nichols, Eugene D. "Testing the Understanding of the concept of Equation." The Mathematics Teacher, 50:400-02, (May, 1957)

Nunnally, Jum Co. Tests and Measurements. New York: McGraw-Hill Book Co., Inc., 1959.

Orleans, Jacob S. and Glenn A. Sealy. Objective Tests. New York: World Book Co., 1928.

Panel of the School Mathematics Study Group. First Course in Algebra Part I and II. New Haven: Yale University Press, 1961

Ross, C. C. and Julian C. Stanley. Measurement in Today's Schools. Englewood Cliffs: N.J. Prentice-Hall Inc., 1954.

Remmers, H. H. and N.L. Gage. Educational Measurement and Evaluation. New York: Harper and Brothers Publishers, 1955.

Remmers, H. H. and others. A Practical Introduction To Measurement and Evaluation. New York: Harper and Brothers Publishers, 1960.

Shute, W. G. and others. Elementary Algebra. New York: American Book Co., 1956.

Stalnaker, J.M. "The Essay Type of Examination." Educational Measurement, ed. Lindquist. Washington D.C., 1951, pp.495-530

Thorndike, R.L. Personnel Selection. New York: John Wiley & Sons, 1959.

Thorndike, R.L. and Elizabeth Hagen. Measurement and Evaluation in Psychology. New York: John Wiley & Sons Inc., 1961

Traxler, Arther E. and others. Introduction to Testing and the Use of Test Results in Public Schools. New York: Harper & Brothers, Publishers, 1953.

Travers, Robert M.W. Educational Measurement. New York: The MacMillan Co., 1955.

Vaughn, K. W. "Planning the Objective Test." Educational Measurement. E.F. Lindquist. Washington D.C., American Council on Education, 1951, pp. 159-184.

Wrightstone, J. Wayne and others. Evaluation in Modern Education. New York: American Book Co., 1956.

Welchons, A.M. and W.R. Krickenberger. Algebra Book Two. New York: Ginn and Co., 1949.