CHAPTER FOUR

PRINCIPLES OF TEST CONSTRUCTION - CRITERIA OF A GOOD TEST

Principles of Test Construction

Test construction requires a systematic organized approach if positive results are to be expected. Firstly, the objective must be well defined. There are numerous points which are common to all types of tests and items which must be observed in constructing a test. Some of the more important are given below :

1. Avoid obvious, trivial, meaningless and ambiguous items;
2. Observe the rules of rhetoric, grammar and punctuation;
3. Avoid items that have no answer upon which all experts will agree;
4. Avoid trick, or catch items that are so phrased that the correct answer depends on a single obscure key word to which even good students are unlikely to give sufficient expression.
5. Avoid items which contain irrelevant clues.
6. Avoid items which furnish the answers to other items.
7. All the pupils are to take the same tests and permit no chance among items. Pupils cannot be compared with one another unless they all take the same tests.

Forms of New Type of Tests

No less than fifty different objective test techniques have been used in testing. For the purposes of the present test the following types have been used :

1. Multiple choice
2. Matching
3. Free response (analogy type)
4. Miscellaneous

It will be, therefore not out of place to know some details of the above types used in the test.

Multiple Choice Type

There are several forms of multiple choice items. The one used in the present test consists of an incomplete statement (including the description of situation if necessary) or stem followed by four possiblecompletions one of which is correct or definitely syperior to the others. Such items are comparatively less open to guessing than others. They are adaptable to a wide variety of materials. They could also be well adapted to measuring, understanding, discrimination and judgment. When well constructed the multiple choice items are probably the best, if not the best of all the objective tests.

Disadvantages

It is difficult to construct these items in such a way that the several responses appear plausible. They are more easily constructed to measure verbal memory rather than understanding. They are space consuming and time consuming. They are difficult to avoid including clues, which enable pupils to respond correctly on a superficial basis. It is difficult to avoid making the responses so plausible, that superior students at a given level choose them thus causing the item to discriminate negatively. The following rules should therefore be observed in constructing the multiple choice items :

1. The item should be practical and realistic.
2. The stated problem should be specific, clear and as brief as possible.
3. Obviously wrong responses should not be included. ?
4. Four to five choices are to be included.
5. The choices should be placed at the end of the incomplete statements.
6. Each choice should be tested on a separate line.
7. There should be no order in the placement of the correct response.
8. They should not be used, if a simple type is sufficient.
9. The choices should be homogeneous as far as possible since tests are designed for higher level of understanding.
10. Simple method of indicating response should be used.

The Matching Type

This type of test requires the matching of items placed in two or more columns. The item may consist of questions, incomplete statements, descriptive phrases, definitions, events, personages, vocabulary, dates, diagrams or other subject matter to prevent guessing, extra items may be placed in the response column.

The matching items are used for testing various outcomes. They are especially applicable for measuring pupil's ability to recognise relationships and make associations and for naming and identifying things learned. They are relatively easy to construct. They can be made quite objective and can be scored quickly. When properly constructed they are reasonably free from the guessing factor.

The following rules should be observed in constructing matching items :

1. There should be at least four and not more than 12 responses in each matching exercise.
2. Only homogeneous or related materials should be used in any one exercise.
3. All the items of a single matching exercise should be on the same page otherwise matching becomes difficult.
4. The basis on which matching is to be done should be clearly indicated.
5. Diagrams may also be used for matching.

On scoring matching type items no correction is necessary.

Free Response Items

There are a variety of forms of this type of test exercises. (a) A direct question is allowed by a blank space in which the pupil responds with a word, phrase, numerical answer, formula or other response, sometimes a name or a date or several different answers may be correct. (b) An incomplete statement contains a blank space at the end in which the pupil writes a short response completing the statement and making it true.

Analogies may also be grouped under the category of the "free response items" where the students are to study carefully the relationships between two words or diagrams out of the three given, then they are to suggest a fourth word or diagram as the second one bears with the first word or diagram. This type of items help in testing the ability to inter-relate in situations, things or events.

Criteria for a Good Test

The following seven points are taken as criteria of a good test.[1]

1. A good test must possess a very high validity.

[1] Sandford, Peter. The Educational Psychology on Objective Study (Longmans), 1939 Edition, p.318.

2. It must have a high reliability.
3. It must be very objective in nature.
4. It must pick out the good students from the poor i.e. it must possess high discriminating power.
5. It must be very comprehensive.
6. It must be easy to use. Its administration and scoring must be easy and there must be economy of time and effort.
7. Norms established on the basis of its results must be satisfactory.

Validity

The validity of a test is usually defined as that aspect of a test which ensures that it will actually measure what it claims to measure. To discover whether or not a test ✓ actually function^rs^ in this way two sets of measures are obviously needed, those of the test itself and of the thing it is stacked against, the measuring rod as it were. The latter is known as the criterion. Obviously, reputable measures both of the test and the criterion are needed. Hence the National Association of Directors of Educational Research has defined the validity of a test as " the correspondence between the ability measured by the test and ability otherwise objectively defined and measured." The degree of correspondence is shown by the coefficient of correlation (known as validity coefficient) obtained from accurate measures of the test and criterion. It finally boils down to the simple statement that the validity of a test is the degree of accuracy with which it measures what it purports to measures Thus Monroe writes " under the head of validity, we inquire into the degree of constancy of the functional relation existing between the scores yielded by the test and the abilities specified as being measures in the statement of its function."[1] and according to Barthelmess, " The term validity as applied to an intelligence test battery may be defined as the amount of agreement between the test's differentiation among individuals and the actual differentiation in intelligence among these individuals. The same definition applies to a sub-test or to a simple element. "[2]

Validity, therefore, is capable of a statistical interpretation by means of a coefficient of correlation. For a sound interpretation, two series of scores are needed, those of the tests and of the approved criterion. The farther the criterion falls short of theoretical perfection the more unsound the interpretation becomes.

In the case of intelligence tests it is obvious that there can be no curricular analysis to secure the pristine

[1] Monroe, W.S. An Introduction to the Theory of Educational Measurement (Boston: Houghton Miffin), 1923.

[2] Barthelmess, H.M. The Validity of Intelligence Test Elements-Contributions to Education No.505 (New York: Bureau of Publication; Teachers College, Columbia University), 1931.

validation of such tests. Practically all the tests, now on the market were secured by empirical means and selected because they worked, that is, correlated directly or indirectly with some criterion (usually teacher's judgments) which was accepted as a true measure of the function concerned. However Spearman early saw that the correlation technique might prove an instrument for the avoidance of this difficulty. Thurston's and Hotelling's factor analysis techniques attempt to generalize Spearman's techniques for any number of factors. If we accept Spearman's analysis of mental abilities and his two factor theory we are in a position first to select the best kind of test and afterwards to determine its validity. The latter which refers to "g" as a criterion may be used either with the test as a whole or with an individual item.

With intelligence test the earliest criterion used was that of the judgments of teachers ahd others who knew the pupil's ability. In general, however, it may be said with truth that teacher's judgments constituted the main criterion. Now Spearman's analysis is widely used and only which are highly saturated with "g" are selected.

Reliability

Any discussion of validity would be incomplete without some mention of so intimately related a concept as reliability. While validity is the degree of accuracy with which a test measures, what it claims to measure, reliability is the degree of consistency with which it measures whatever it actually does measure. The statistical measure of the validity of a test is the coefficient of correlation between test scores and an accepted criterion. The measure of its reliability is the coefficient of correlation between scores made when the test is administered to the same set of candidates on two separate occasions or between scores made on two equivalent forms. Validity involves a relationship between a test and an outside criterion; reliability is a property residing wholly within the test itself. A test has validity only with reference to this or that purpose, but its reliability is wholly independent of purpose. If the reliability of a test is 0.87, then it possesses that reliability whether it be used as a measure of intelligence or school achievement or humour or material wealth or fear of snakes.

But the relationship between validity and reliability must not be considered wholly in terms of contrast. In the main, they are characterized by intimate positive inter-associations. The validity of a test is conditioned by its reliability. It is obvious that unless a test measures consistently whatever it does measure it cannot hope to measure accurately what it claims to measure and while we cannot in strict truthfulness say that the validity of a test cannot exceed its reliability, yet in actual practice validity coefficients do tend to be lower. Certain it is that in order for a test to be perfectly valid it must be perfectly reliable. The reverse of course is not true. A test may measure consistently and yet yield a hopelessly inaccurate picture of the function it is attempting to evaluate. It may be perfectly reliable and utterly invalid.

Thus a highly reliable test should yield the same score when administered twice to the same students.

An objection may be raised that the students might have learnt something and it will affect their scoring. The answer to this is that all the students learn in this gap will be in accordance with their intelligence. Hence although they learn something that goes to improve their score, their relative position will remain unchanged so that a student who stood first on the first occasion, will also stand first on the second occasion and a student who stood second on the first occasion will stand second on the second occasion and so on.

No test is perfectly reliable. The student taking the test is a variable factor. He may become fatigued or bored or his health may vary. Thus there is the problem or variable physical and emotional relations from one test to another. For all these reasons, the test or re-test is rarely used for finding reliability.

In validity the emphasis is on a test's agreement with the objective, in reliability upon agreement with itself. In general the reliability of a test is increased by making it more objective, that is by reducing the personal equation of the examiner by increasing it in length and especially by increasing the number of items of the test by using final units in the score by which the test is scored and by keeping as constant as possible the conditions, under which the test is taken.

Objectivity

Objectivity of a test is an important factor for it affects both the validity and reliability of the test.

There are two aspects of objectivity which must be considered, while constructing a test. The first is concerned with the scoring of the test. The second is regarding the inter-pretation of the individual items of the test.

A test is perfectly objective when the examiners using the same test upon the same students give identical scores. The personal element of the examiner must not interfere in scoring. After the key has been prepared there should be no question as to whether an item is right or wrong, partly right or partly wrong.

The multiple choice, matching and some other forms of test may reach perfect objectivity but the recall and the completion test can never be made perfectly objective. Since objectivity reduces the sources of variability, it tends also to make a test more reliable.

The second aspect regarding objectivity is concerned with the interpretation of the items in the test. A given test item should mean the same to all students. This is difficult to achieve. However good a test item may appear in the beginning there may be some students who might interpret it wrongly. Naturally this will affect validity. The three criteria discussed so far, viz.,Validity, Reliability and Objectivity are inter-dependent and closely related.

Discrimination

A good test must possess the following properties :

1. It must measure small differences in achievement.
2. It must pick out the good students from the poor students.

In order to achieve these, three things are necessary.

(i) When the test is administered there should be a wide range of scores from the lowest to the highest.

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

(iii) Each item should discriminate between students who are high and those who are low in achievement. If the good students are just as likely to miss an item as the poor students the item does not measure in a positive direction. Sometimes poor students answer correctly certain items that the best students miss. That will be negative discrimination.

In such items, it means that the items are ambiguous and that they must be discarded.

After a test has been administered, item analysis should be made which will indicate the relative difficulty of each item and much more important than this, the extent to which each discriminates between good and poor students.

Comprehensiveness

The best battery should include enough items under each of the abilities so that it measures what it is supposed to measure.

Economy

The chief factors influencing the economy of an examination are the ease of its administration and scoring and the lesser cost of its printing. It should be so designed that a minimum of student's time will be consumed in answering each item. The test item should be so constructed as to enable students to score quickly and efficiently. The scoring costs could be considerably reduced by using easily scored forms in the construction of a test. In short, the test must be economic from the point of time, money and effort.

Norms

As usually calculated " A norm is the median or average of the present attainment of a given group." Norms are established by giving test to a sufficiently large cross section of students. The norms obtained in this way are used in interpreting the scores of individual students.

There are various types of norms such as grade norms, chronological age norms,mental age norms, etc. The grade norms are those that are obtained when students are grouped according to the school grade, chronological age norms and mental age norms are those that are worked out when the students are grouped according to their chronological age or mental age.

For getting satisfactory grade norms, it is necessary that the grade placing of students should be uniform and the standards of promotion in different schools should also be the same.

The Science Aptitude Test

The various test items constructed under the present 'Science Aptitude Test' are based on the abilities listed out. The test items constructed under each of them are designed to test that specific ability.

Sr. No.	Ability included	Test Items Constructed
1.	Numerical ability	29
2.	Spatial ability	25
3.	Reasoning ability	18
4.	Inter-relationship	19
5.	Mechanical ability	24
6.	Cause and effect relationship	24
7.	Infer from an experimental data	14
	Total	153

The whole test is divided into seven sections each of which intended to test a particular ability. In each of the sections, a model example is given wherever necessary for the guidance of the students. Some of the specimen examples included in the test under the various sections along with the directions to guide and help the pupils to answer the items in the booklet are given below :

SECTION I

Ability tested : Numerical Ability

(A) Directions:

After carefully studying the number series given on the left below tick off the correct succeeding number in the answer sheet, out of the numbers given on the right.

		a	b	c	d
Example :	1, 4, 9, 16, 25	32,	34,	36,	38

A study of the numbers given here on the left shows that they are all the squares of the natural numbers 1, 2, 3, 4, 5, the succeeding number should be the square of six (6). So the answer to be ticked off out of the given numbers on the right is 36 indicated under the letter 'c'.

(B) Directions :

In each of the following questions, we find two series of numbers. Some relationship exists between the corresponding numbers in the two series. The correct relationship is shown by one of the four alternatives given on the right hand side. Tick off the correct one.

Example :

If $A = 0, 1, 2, 3$ ...... a) $B = B + 1$

$B = -1, 0, 1, 2$ ..... b) $B = A + 2$

c) $B = A - 1$

d) $B = A - 2$

The numerical values of $A = 0, 1, 2, 3$ when substituted in the relation $B = A - 1$ give values $-1, 0, 1, 2$ for B. This is the only relation that satisfies the values of A & B and hence the right answer to be ticked off is B = A - 1.

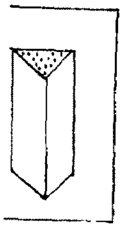
In this section two typical examples have been given for the benefit of students to understand the categories of items included in Section 1.

SECTION 2

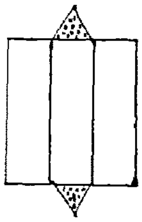
Ability tested : Spatial Ability

Directions :

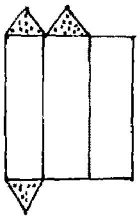
Paper models given on the left side, closed at both ends and hollow when opened assume one of the shapes shown on the right side. Tick off the correct shape from the figures given on the right.



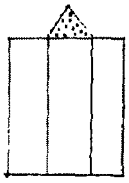
a



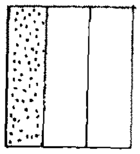
b



c



d



Out of the shapes given on the right, the one indicated under the letter 'a' is seen to be the correct shape and hence the student has to tick off the letter 'a' against this particular item in the answer sheet.

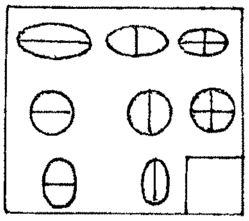
SECTION 3

Ability tested : Reasoning Ability

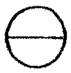
Directions :

In the matrices given below study carefully the columns and rows and tick off the correct number or symbol that should be written in the blank space from the set of given alternatives on the right

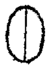
(A) Example :



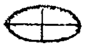
a



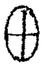
b

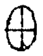


c



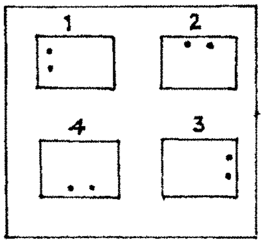
d



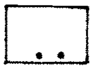
. A close study of the columns and rows in the example given above shows that the answer to be ticked off from the alternatives given ion the right is the symbol  given under 'd'.

Directions :

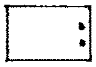
Carefully go through the figures given on the left side and tick off the figure that comes next in series after the fourth figure, from the set of alternatives given on the right side.



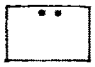
a



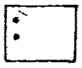
b

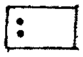


c



d



A close observation of the figures indicates that the figure that comes next in series after the fourth figure in the figures given on the left is the figure  given under 'd' on the right hand side.

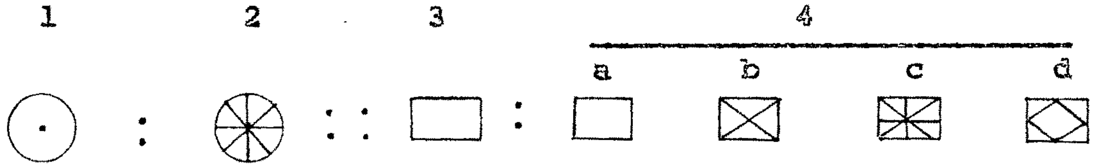
SECTION 4

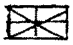
Ability tested: Ability to Inter-relate

Directions :

Closely observe and study the relationship that exists between the first two figures and suggest a fourth one. The relationship of 3 and 4 should be the same as the one that exists between 1 and 2.

Example :



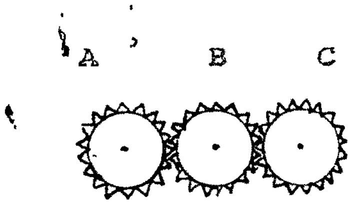
A close study of the figures given on the right reveals that the figure '' given under 'c' is the one that is to be ticked off as its relation with figure '3' is the same as the one existing between 1 and 2.

SECTION 5

Ability tested: Mechanical ability

Example :

When the toothed wheel 'A' rotates in the anti-clockwise direction the wheel 'C' rotates



a) in the direction as 'A'

b) opposite to the direction of 'A'

c) in the direction of 'B'

The toothed wheels A, B and C are inter-linked, as 'A' rotates in the anti-clockwise direction, B rotates in the clockwise and finally 'C' rotates in the anti-clockwise direction. So the wheel 'C' rotates in the direction of 'A' and hence the answer to be ticked off is the answer given under 'a'

SECTION 6

Ability tested: Ability to give the cause and effect Relationship

Example :

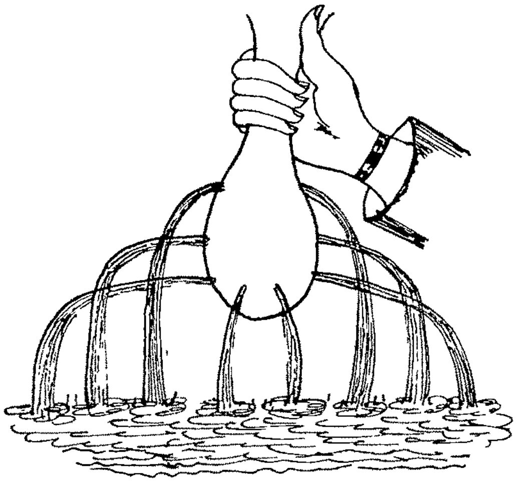
On a windy day there were gusty winds. The weather bulletin indicated the heavy inflow of winds in a particular area. It was due to the

(a) high pressure in the area
(b) temperature difference
(c) low pressure created.

The heavy inflow of winds is caused because of the low pressure created in the area and hence the alternative given under 'c' is to be ticked off in the answer sheet.

SECTION 7

Ability tested: Ability to Infer from an Experimental Data



Water is poured into a heavy paper bag. The bag is lifted carefully and four holes are punched with a needle at one inch difference at different heights on both sides as shown. Water from the holes near the bottom is seen to gush farther.

It shows that :

(a) sufficient air pressure is not exerted on the upper levels.

(b) limited space at the bottom levels compresses the water.

(c) pressure increases with depth.

Water near the bottom gushes farther because of the fact that the liquid pressure increases with depth. Hence the pupil has to tick off the answer given under 'c' in the answer sheet.

Thus the pilot test in all contains 7 sections with 153 items in total. The number of items included in each of the seven sections and the approximate time required by the pupils for answering the various items of different sections are given below :

Section No.	No.of Items included in the Pilot Test	Approximate Time Required
Numerical ability (Section 1) 1-29	29	20 minutes
Spatial ability (Section 2) 30-54	25	20 "
Reasoning ability (Section 3) 55-72	18	15 "
Ability to Inter-relate (Section 4) 73-91	19	15 "
Mechanical ability (Section 5) 92-115	24	20 "
Cause and effect Relationship (Section 6) 116-139	24	20 "
Infer from an Experimental data. (Section 7) 140-153	14	10 "
Total	153	120 Minutes

As the pilot test is to be administered in schools of Andhra Pradesh, it is necessary to translate the whole test into Telugu since education up to the secondary level is imparted in Telugu medium. It will help in having a better and wider sampling. The pilot test is translated into Telugu taking care to see that simple language is used in translating the items and to make them self explanatory.

The 'pilot test form' which is translated into 'Telugu' is finally got cyclostyled with neatly drawn diagrams in the form of a booklet. The answer sheets required for answering the test items have been separately given so that pupils can record their responses of the items separately without marking them on the test booklets.