at a single sitting, the results are relatively uniformly affected by the examinees's physical condition and attitudes, and prevailing environmental conditions. When there is a time interval, the retest results will be affected by the normally expected fluctuations in individual performances and by changes in environmental conditions.

4. The effect of practice and learning : Such effect depends upon the content of the test, the length of the interval and upon the examinee's experiences during the interval.
5. Reliability of subtests : Other factors being equal, the reliability of test increases with increase in lengths although not in direct proportion.
6. Consistency of scores : Some tests are not entiraly objective in scoring, because the examiner at times finds it necessary to judge the correctness or quality of responses.[14]

6.3 <u>Methods of Estimating Reliability</u> :

Various methods of estimating test reliability are known. The following four methods will be considered, in view of their suitability or otherwise, so far as the present test is considered :

1. Item-interrelationship method
2. Equivalent form method
3. Subdivided test method
4. Retest method

6.3.1 Item Interrelationship Method :

These procedures have been developed with a view to avoiding the subdivision of total test into two halves on arbitary choices. They are based on homogeneity or internal consistency of a test. This type is analogous to the split-half methods and like them, open to the objection that they do not provide any indication of the fluctuation in scores over time: besides like the split-half methods, they are not applicable to speeded tests. But a more serious objection to the procedure is the doubt whether this type of estimate can be more properly regarded as a separate property of tests, distinct from traditional concepts of either reliability or validity.[15] Guilford gives his views on this method as,

> Since so much of the statistical thinking concerning reliability is put in terms of variances, it is not surprising that the estimation of reliability can be made by a more conventional analysis-of-variance approch. Several investigators have proposed this kind of approch, among whom are Jackson, Hoyt, and Alexander.[16]

Anastasi futher states :

> Although homogeneous tests are to be prefered because their scores permit relatively unambiguous interpretation, a single homogeneous test is obviously not an adequate predictor of a highly heterogeneous criterion. Moreover, in the prediction of a heterogeneous criterion, the heterogeneity os test items would not necessarily represent error variance. Traditional intelligence tests provide a good example of highly heterogeneous tests designed to predict a highly heterogeneous criterion.[17]

Hence from this it follows that these procedures are are not applicable to traditional intelligence tests. In spite of the limitations, the suitability of their application to the present test was considered. One of the requisites for their application is that the scores on the test should be the number of correct answers and that no coreection is applied. In the present test scoring is done on the bases of time and the raw scores are weighted; hence this method is not applicable to the present tests.

6.3.2. Equivalent From Method :-

In this method the reliability estimates are not based on a single trial like the subdivided test method The effect of memory on the reliability coefficient is little less than in the case of retest method. Thus one way of avoiding the difficulties encountered in retest reliability is through the use of equivalent forms of the test. The subjects can then be tested with one form on the first occasion and with another, comparable form on the second. The correlation between the scores obtained on the two forms represents the reliability coefficient of the test and this coefficient is a measure of both temporal stability and equivalence or adequacy of item sampling. Such a coefficient thus reflects two aspects of test raliability. Hence though this method is not absolutely free from limitations it is considered to be the best, if properly applied, among the existing ones. But this method is rarely used because it is not an easy job to prepare two equivalent forms of tests.

The preparation and administration of equivalent test forms, though quite satisfactory as a procedure for estimating reliability presents certain practical difficulties. These center around the problem of the time and labour involved both in the construction and the administration of two complete test forms. If only a single form of a test is needed for the research or practical use to which the test is to be put, it often seems unduly burdensome to prepare two separate tests merely in order to obtain an estimate of reliability. Furthermore, when a test is developed and administered asapart of research project, time for the administration of an equivalent form of the test is often not conveniently available.

In most of the usual types of tests of ability or achievement, preparing equivalent form does not present undue difficulty. These are some situations, however, in which equivalence will be difficult to achieve. This is true when either (a) the test is essentially unique or (b) a single exposure changes the individual to such an extent thet he is really a different individual at the second exposure.

Moreover as Anastasi puts,

> Parallel forms of a test should be independently constructed tests designed to meet the same specifications. The tests should contain the same number of items, and such items should be expressed in the same form and should cover the same type of content. The range and level of difficulty of items should likewise be equal. Instructions, time limits, illustrative examples, format, and all other aspects of the test need to be checked for comparabillity. Only when the two forms are actually equivalent can the differences in scores from one form to the other be considered as error variance.[18]

For the present work, it was quite difficult, rather impossible, to prepare a parallel form of the test. To prepare the performance test material is a very difficult and expensive job. It becomes quits impossible to prepare such material for two parallel performance tests which completely obeys the rules layed down by Anastasi as seen in the above discussion. It was for these reasons that the idea of finding test reliability by this method could not be put into practice.

6.3.3 Subdivided Test Method : This method is also known as split-half method. By this method from a single administration of one form of a test it is possible to arrive at a measure of test reliability by various split-half procedures. Thus two scores are obtained for each individual by dividing the test scores into two comparable halves. The testitems are split up in various ways. These are :

1. Alternate items as a basis for splitting test
2. Alternate groups of items as basis for splitting test
3. First Vs second half as basis for splitting test

The tests may be divided into different number of parts and then any two parts can be correlated and coefficient of correlation can be found out. As the coefficient is affected by the test length, the Spearman-Brown prophency formula is applied for the correction of test length. Freeman gives the reason for using this formula as, "The scores of the whole test, being based upon a larger numbef of items, is a more adequate sampling of traits 6r functions and hence reduces the possible effects of chance solutions and accidental errors"[19]

In the present investigation Spearman-Brown formula was used to apply the correction for the test length. So coefficient of correlation was corrected by applying this formula.

This method of estimating reliability has been most frequently abused by the practical psychologists. The statistical principles, underlying all split-half methods, prohibit its application to 'purely' or 'highly' speeded tests. Cronbach remarks,

> Only tests which nearly all pupils finish can be studied by this method.[20]

Besides, Freeman warns the investigators as,

> Items in most tests are grouped together according to type and are graduated according to type and are graduated according to difficulty, from easiest to hardest. Thus when this systematic arrangement is employed, the odd-even procedure yields very close approximations to equivalent half-scores, because each half score is based upon the same types of items and the same number of each type; and each half score is based upon items which progress in difficulty in approximately the same degree.[21]

The items in PPTI are arranged according to the difficulty value and liberal time limits are fixed for all test items and hence this method was found suitable to find the reliability of the test.

6.3.4 4 Retest Method : This method is quite a simple one in which the test is administered twice to the same sample on two different occasions. The reliability coefficient in this case is simply the correlation of the scores obtained by the same subjects on the two administrations of the test.

But it is in reality a special class of reliability coefficient and should more accurately be termed as retest coefficient. There are some limitations of this method. Kuder and Richardson say,

> The retest coefficient on the same form gives, in general, estimates that are too high, because of material remembered on the second application of the test. This memory factor cannot be eliminated by increasing the length of the time between two applications because of variable growth in the function tested within the population of individuals. These difficultics are so werious that the method is rarely used.[22]

But Bhatia thinks that the use of this method is unavoidable. He says,

> The difficulty in regard to reliability is due to the particular type of tests which we are dealing with. In tests of our type the repetition of the scale on the same group after an interval of time. as has been pointed out by Alexander, is the only practicable method for establishing reliability, although this method too is considered by many to be unsatisfactory.[23]

Anastasi discusses the drawbacks of this method as :

Although apparently simple and straight forward, this technique presents serious difficulties when applied to most psychological tests. Practice will probably produce varying amounts of improvement in the retest scores of different individuals. Moreover, if the interval between retests is fairly short, the subjects may recall many of their former responses.[24]

Thorndike opines about this as :

> Repeating the same test form holds the sampling of items constant so that this factor is treated as systematic rather than error variance. Reliability coefficients calculated from a repetition of the same test may be expected to be higher than those based on parallel, equivalent forms by an amount that is equal to this variance associated with sampling of items. A second possible difficulty with repeition of the same test is actual memory of particular items and of the previous response to them.[25]

In this method the practice of the first administration affects the results of second administration. This depends upon the time interval of both the administration. If the interval is small the practice and memory effect is more. Hence the coefficient of correlation is spuriously high. If the interval is very big the subject's growth takes place which naturally affects the results. Psychologists think that the interval between two administrations should be neither too much nor too less. In this connection Cattell says,

> In many instances one wishes to retest a child's intelligence after the lapse of some months or years. When more than a year elapses it is quite safe to use the same test, for the test items are almost invariably forgotten, and in any case the child's growing intelligence ecounters the critical questions within a new region of the scale.[26]

But Mursell objects to the long interval and he says, "In a short time interval there is very likely to be some specific practice effect carried over from the first to the second testing.. If the time interval betwe n testing is long, the obtained correlation will probably reflect the effect of growth, of learning and of environmental influence generally quite as much as it does the reliability of the instrument.[27]

Anastasi gives her views about the time interval as :

> Thus in checking this type of test reliability, an effort is made to keep the interval as short as feasible. In testing young children, the period should be even shorter than in the testing of older subjects, since at an early age progressive developmental changes are discernibke over a period of a month or less. For any type of subjects, the interval between retests should rarely exceed six months.[28]

The question of practice effect troubles the reseqrcher a lot. The immediate retesting of test is objected only because of this. Cattell gives his views about the practice effect and says that, "Since practice does not increase intelligence itself, the better the intelligence test - i.e. the more it is saturated with 'g' - the less it is susceptible to practice effects. Experiment shows that in this field we must distinguish between a practice effect and what P.E. Vernon has called test sophistication. By test sophistication we mean getting familiar with the type of questions asked, with the timing arrangements and with other features which are strange and sometimes disturbing when one first meets an intelligence test. By practice we mean the improvement in the actual intelligence operations,.....Actually the evidence indicates that there is extremely little practice effect in intelligence tests in the sense just described".[29]

In view of the foregoing discussion, as the age range of the testees was from 16+ to 22+, it was thought proper to keep the interval of eight months between two testings which would be suitable for both low and high age ranges.

6.3.5 Correction for Range :

As it has been pointed out, no method of estimating reliability is perfect, Each has its own limitations. The reliability coefficient of a test administered to a group of wide range of talent cannot be compared directly with the reliability coefficient of a test administered to a group of relatively narrow spread, a single grade, for example psychometrists have objected to the application of reliability estimates obtained from heterogeneous groups, to groups which are less diverse and narrower in range of ability.

If the reliability coefficient of a test in a wide range is known, the reliability coefficient of the same test in a group of narrow range can be found out, provided the test is equally effective throughout both ranges. The formula is -

$$\frac{\sigma n}{\sigma w} = \frac{\sqrt{n - y_{ww}}}{\sqrt{1 - y_{nn}}}$$

Where n and w = the 's of the test scores in the narrow and wide ranges respectively. Ynn and Yww = the reliability Coefficients in the narrow and wide ranges.[30]

The reliability Coefficients, obtained were cofrected for range, by the use of this formula, where 14 and 25 were accepted as the population SDs of Scores, used in calculating the deviation IQs for the age groups 16+ to 22+ respectively.

6.4 The reliability of the present test :

It should be noted that no one type of measure of test reliability is universally preferable. The choice depends upon the use to which the test scores are to be put. Hence the selection of the method to be used for finding the test reliability depends upon the types of the tests as well as on the uses of the results.

The reliability of the present test was estimated by the Test-Retest method and split-half method as the other two methods were not applicable as discussed in the foregoing paragraphs. The implementations of the methods used are described and the results are recorded.

1. The retest method : For obtaining the retest reliability estimate, 50 pupils were retested at the interval of about 8 months. As it has already been discussed before, in the opinion of the present researcher, the period of 8 months was sufficient enough to minimize, as far as possible, the effect of memory, practice and familiarity. However, he was not unaware of the fact of growth in mental maturity, attained by the testees during the period. That perhaps accounts for an increase in the mean score during the second testing.

Table 6.1 gives the analysis of the sample, retested and tables 6.3 to 6.8 give testwise distributions of Scores on the testings as shown in the Scale- tergrams.

TABLE- 6.1 : SAMPLE FOR RETEST RELIABILITY ESTIMATES

Ages	Grades							Total
	XI	XII	FY	SY	TY	P.G.I	P.G.II	
16+	2	2						4
17+	3	3	1					7
18+		3	3	2				8
19+			2	3	3			8
20+				3	2	2		7
21+				4	3	1	1	9
22+						3	4	7
Total	5	8	6	12	8	6	5	50

TABLE-6.2 : RELIABILITY ESTIMATE BY THE RETEST METHOD - FULL SCORE SCATTERGRAM

Score in the first testing

Scores in the Second Testing	96-105	106-115	116-125	126-135	136-145	146-155,	156-165	166-175	176-185	186-195	196-205	205-215	216-225	To tal
216-225													1	1
206-215											1	1		2
196-205										1	1			2
186-195									1	2				3
176-185									5					5
166-175								7						7
156-165							8	4						12
146-155					2	5								7
136-145					3									3
126-135				2										2
116-125			1	1										2
106-115			1											1
96-105	1	1	1											3
Total	1	1	3	3	5	5	8	11	6	3	2	1	1	50

Statistics :

	First Testing	Second testing
$\bar{X}$	161.3	160.5
SD	25.38	27.627

r 0.0703

SC ± 4.3739

TABLE 6.3 : RELIABILITY ESTIMATE BY THE RETEST METHOD

TEST 1 FORNBOARDS

SCATTERGRAM

Scores in the Second Testing	Scores in the first testing										
	6-10	11-25	16-20	21-25	26-30	31-35	36-40	41-45	46-50	51-55	56-60	Total
56-60										1		1
51-55									1		1	2
46-50							1		1			2
41-45							3	3				6
36-40						3	2	2				7
31-35					4	5						9
26-30					3	4						7
21-25			1	4								5
16-20			1									1
11-15	1	5	3									9
6-10	1											1
Total	2	5	5	4	7	12	6	5	2	1	1	50

Statistics :

	First testing	Second testing
$\bar{x}$	30.1	30.07
SD	11.65	12.38
r	0.973	
se	±11.9142	

TABLE 6.4 : RELIABILITY ESTIMATE BY THE RETEST METHOD

RETEST METHOD

- Test -2 BLOCK DESIGNS - SCATTERGRAM

Scores in the Second Testing	Scores in the first testing									Total
	1- 8	9- 16	17- 24	25- 32	33- 40	41- 48	39- 56	57- 64	65- 72	
65-72									2	2
57-64								3		3
49-56						2	3			5
41-48						8				8
33-40					12					12
25-32			1	9						10
17-24		1	4							5
9-16		3								3
1-8	2									2
Total	2	4	5	9	12	10	3	3	2	50

Statistics :	First testing	Second testing
$\bar{x}$	35.54	36.18
SD	15.216	15.158
r	0.9806	
Se	± 1.8003	

TABLE 6.5 : RELIABILITY ESTIMACE BY THE RETEST METHOD

- Test -3 . PICTURE ARRANGEMENT-SCATTERGRAM

Scores in the Second Testing	Scores in the first testing									Total
	1-3	4-6	7-9	10-12	13-15	16-18	19-21	22-24	25-27	
25-27								1	1	2
22-24								2		2
19-21						1	5			6
16-18				1		7				8
13-15					14	1				15
10-12				6		1				7
7-9			4							4
4-6		4								4
1-3	2									2
Total	2	4	4	7	14	10	5	3	1	50

Statistics :	First testing	Second testing
$\bar{x}$	13.88	13.94
SD	5.49	5.64
r	0.9745	
Se	± 1.087	

TABLE-6.6 : RELIABILITY ESTIMATE BY THE RETEST METHOD

- Test-4, BLOCK BUILDING - SCATTERGRAM

Scores in the second Testing	Scores in the first testing									Total
	1- 6	7- 12	13- 18	19- 24	25- 30	31- 36	37- 42	43- 48	49- 54	
49-54								1		1
43-48								1	1	2
37-42						1	4	1		6
31-36					1	5	1			7
25-30					8	1				9
19-24			1	7	2					10
13-18		2	6	1						9
7-12		2	1							3
1-6	3									3
Total	3	4	8	8	11	7	5	3	1	50

Statistics :	First testing	Second testing
$\bar{x}$	25.34	25.1
SD	11.74	11.43
r	0.9472	
Se	± 2.7012	

TABLE-6.7 : RELIABILYTY ESTIMATE BY THE RETEST METHOD

- Test-5. MAZES- SCATTERGRAM

Scores in the Second Testing	Scores in the first testing					Total
	1-4	5-8	9-12	13-16	17-20	
17-20				1	1	2
13-16			9	3	1	13
9-12		3	11	4	1	19
5-8	3	6	1			10
1-4	2	4				6
Total	5	13	21	8	3	50

Statistics :	First testing	Second testing
$\bar{x}$	9.78	10.10
SD	4.10	4.22
r	0.7271	
Se	± 2.1472	

TABLE-6.8 : RELIABILITY ESTIMATE BY THE RETEST METHOD

-Test-6. - PICTURE COMPLETION-SCATTERGRAM

Scores in the Second Testing	Scores in the first testing							Total
	1-2	3-4	5-6	7-8	9-10	11-12	13-14	
13-14						2	2	4
11-12				2	2	2		6
9-10				3	4	2		9
7-8			4	7	4			15
5-6		3	6					9
3-4	1	2	1					4
1-2	2	1						3
Total	3	6	11	12	10	6	2	50

Statistics :	First testing	Second testing
$\bar{x}$	7.34	7.78
SD	3.04	3.13
r	0.8914	
Se	± 1.0018	

TABLE-6.9 : RETEST RELIABILITY COEFFICIENTS

Sr. No.	Tests	r	SEmeas	Index of REIABILITY
1.	Form Boards	0.9732	± 1.9142	0.9865
2.	Block Designs	0.9806	± 1.8003	0.9902
3.	Picture Arrangement	0.9745	± 1.0872	0.9871
4.	Block Building	0.9472	± 2.7012	0.9732
5.	Mazes	0.7271	± 2.1472	0.8527
6.	Picture Completion	0.8914	± 1.0018	0.9441
7.	Full Scale	0.9703	± 4.3739	0.9850

As can be seen grom the observations of table 6.9 the reliability estimates, that is Product-moment rs vary from 0.73 in case of Mazes test to 0.98 in case of Block Designs test. The reliability coefficient fro the full Scale Score is 0.97.

6.4.1 Standard Error of Measurement :

Another way of estimating reliability is the standard error of measurement. The effects of variable or change errors in producing divergences of test scores from their true values is given by the formula

$$\sigma Sc = \sigma 1 \sqrt{1-r11}$$

(Standard error of an obtained Score) in which

Sc = the SE of an obtained Score

(also called the SE of measurement)

1 = the Standard deviation of test Score (test-1)

= the reliability Coefficient of test 1

The Subscript Sc indicates that this SE is a measure of the error made in taking an obtained Score as an estimate of its true Score.[31]

It will be observed that the error of measurement is independent on the test units. Hence, SEmeas of one test cannot be compared with SEmeans of the other test. All the reliability coefficients were used in callulating the testwise error of measurement. The SE meas. obtained for each test is given in table 6.9.

6.4.2 Index of Reliability :

The reliability coefficient is an estimate of relative reliability. It measures the dependability of test scores by showing how well obtained scores agree with their theoretically true values. It gives the maximum correlation which the given test is capable of Yielding in its present form. This is true because the highest correlation which can be obtained between a test and a second measure is between the test scores and their corresponding true scores. The true score of an individual, on a test has been defined as the mean of a very large number of determinations made of the same person on the same test or on Parallel forms of test administration under standard conditions. The correlation between a set of obtained scores and their corresponding true counterparts is given by the formula.

$$r_{1\propto} = \sqrt{r_{11}}$$

where r_1 = the correlation of the obtained and true scores

r_{11} = the reliability coerricient of the test 1

The symbol $\propto$ (infinity) designates true scores. The coefficient $r_1\propto$ is called the index of relability.[32]

The index of reliability of all the tests were found out. The results are recorded in table 6.9.

There are different views about the answer to the question "How much reliability must a test Possess ?" No definite answer can be given to this question. As discussed before, the reliability estimate depends upon the amount of heterogeneity of the trait measured, in the sample used, as well as on the method employed. Thus while thinking about reliability all these factors should be taken into account. Nunnally opines as,

> No definite rule can be stated as to how high the reliability coefficient should be for a test, but in general one suspects a test that has a coefficient less than 0.80 some of the better standardized instruments have reliability coefficient over 0.90.[33]

The present tests are individual tests and hence are considered to be more reliable. The age range in the Present retesting is from 16 to 22; hence the group can be called a heterogeneous one. In construction and standardization of performance tests of intelligence for Pupils of grades II to XI in Gujarat, Leelaben Patel got the retest reliability coefficients of the performance tests as follows :

TABLE :6.10 : TEST RETEST RELIABILITY COEFFICIENTS OF PATEL PERFORMANCE TEST OF INTELLIGENCE FOR PUPILS OF GRADE II TO XI IN GUJARAT[34]

Test	Reliability coefficient
Form Boards	0.8814
Block Designs	0.9181
Picture Arrangement	0.8888
Block Building	0.8619
Mazes	0.9491
Picture Completion	0.8426

She had retested the sample of grades II to XI and hence it was just as heterogeneous as in the Present test.

6.4.2 Split-Half Method :

To find out the coefficient of correlation by this method all the 420 record blanks were used. The composition of the sample was the same as that of the population tested in the present scale given in table 4.20 and hence it is not given again. In these 420 record blanks each subtest was divided into two halves, one containing the odd intems and the other, the even items. The score on each subtest obtained was added so as to obtain the two halves of the whole test. Agewise correlation coefficients were found out between scores of these halves. The prophecy formula was used to correct for the reduced length. That is from the reliability of the half test, the self-correlation of the whole test is estimated by spearman-Brown prophecy formula. The formula used is as follows :

$$r1I = \frac{2r\tfrac{1}{2} - \frac{I}{II}}{1 + r\tfrac{1}{2} - \frac{I}{II}}$$

Where r1I = reliability coefficient of the whole test.

$r\tfrac{1}{2}\ \frac{I}{II}$ = reliability coefficient of the half test found experimentally.[35]

The SEmeas and Index of reliability where calculated for each r corrected for length. The r for ages 16 to 22 were calculated by transforming the r's in to Fisher's Z function and taking the arithmetic mean of the Z's. This mean Z is then converted in to an eauivalent r.

The following table 6.11 gives the agewise correlations, SEmeans, and Index of reliability for the age group of 16 to 22 boys and girls. In each age group 30 pupils were taken.

TABLE : 6.11 : SPLIT - HALF RELIABILITY COEFFICIENTS

Sex	Age-Range	X	X Corrected for Length	Odd Score $\bar{X}$	Even Score $\bar{Y}$	Odd Score 6X	Even Score 6Y	SE meas	Index of Reliability
Boys	16 +	0.60	0.69	69.53	67.90	10.42	11.20	± 4.89	0.77
Girls	16 +	0.89	0.92	69.66	69.63	10.65	11.91	± 4.99	0.94
Boys	17 +	0.97	0.98	69.86	73.73	10.60	11.49	± 5.30	0.98
Girls	17 +	0.95	0.98	71.67	74.70	11.65	11.95	± 5.83	0.97
Boys	18 +	0.94	0.95	71.63	72.40	11.12	10.60	± 5.79	0.96
Girls	18 +	0.89	0.91	72.10	73.20	11.78	10.54	± 6.12	0.94
Boys	19 +	0.72	0.78	84.2	83.43	7.17	7.66	± 4.42	0.85
Girls	19 +	0.74	0.79	84.4	85.40	7.82	9.15	± 4.82	0.86
Boys	20 +	0.63	0.68	89.06	91.27	7.29	7.83	± 4.55	0.79
Girls	20 +	0.66	0.70	88.20	90.03	6.36	7.38	± 3.97	0.81
Boys	21 +	0.80	0.82	93.63	92.30	8.52	9.43	± 4.89	0.89
Girls	21 +	0.81	0.82	94.71	95.20	8.02	8.30	± 4.61	0.90
Boys	22 +	0.89	0.91	97.77	10.79	10.79	10.14	± 6.10	0.94
Girls	22 +	0.75	0.81	101.26	100.36	8.17	6.77	± 4.62	0.87

It can be observed that the X corrected for length ranges from 0.61 to 0.78, and the X is 0.60 to 0.97.

Bhatia's sample reliability estimates was selected at random from the entire population used for standardization. His tests are standardized for the age range 11^+ to 16^+ and hence this sample can also be assumed to be eaually heterogeneous. The split-half coerricient for the literate group is 0.851 and for the lilliterate group it is 0.841. Bhatia does not correct the coefficient for reduced length by applying the Prophecy formula. Yet it can be compared with the one for the present test.[36]

Thus the results of reliability estimate can be considered as satistactory. Reliability estimate of each subtest was found out by test retest method and the agewise coefficient of correlation was found out by the split-half method. Thus an attempt has been made to get the finer estimates,

In the next section, the problem of validity of the test result is dealt with.

6.5 Validity :-

The reliability of any measuring instrument is the necessary condition but it is not a sufficient condition. The test may measure something consistently but may not measure what it purports to measure. If an instrument measures accrately and consistently what it purports to measure, it is called a reliable and valid instrument. The validity of a test, or of any measuring instrument depends upon the fidelity with which it measures. "A home made yardstick is entirely valid when measurements made by it are accurete in terms of a standard measuring rod. And a test is valid when the performances which it measures correspond to the same performances as otherwise independently measured or objectively defined"[37]

To find out the validity of a test, one must compare the reality of what it does measure with some ideal conception of what it ought to measure. Cureton says, "Validity is therefore defined in terms of the correlation between the actual test scores and the 'true' criterion scores"[38]

Ross defines validity as follows :

> One kind of validity concerns the degree to which the test or other measuring instrument measures what it claims to. In a word, validity measures truthfulness.[39]

Gullikson defines it in a more particular form when he says,

> ..... the validity of a test is the correlation of the test with some criterion.[40]

Guilford says,

> In a very general sense, a test is valid for anything with which it correlates.[41]

Guilford explains the meaning of validity in statistical terms as follows :

> ..... what a test measures, in common with other tests and other measures, is in the form of common factor. Common factor variance, then, is the basis for validity. ..... the correlation of a test with each common for measuring that factor. A test may have a validity (factor loading) of .50 for measuring the factor of numerical facility and a validity of .60 for measuring reasoning.[42]

In a very general sense, a measuring instrument is valid if it does what it is intended to do. Proper performance of some instruments is rather easily verified. e.g., of the yard stick as a measure of length. It takes very little "research" with this instrument to find that resulting measurement -

1. fit in perfectly with axiomatic concepts of the nature of length and
2. relate to many other variables.

The validity of the physical instruments can be obtained very easily and accuretely. But it is very difficult to get independent standards in mental measurement; the validity of a mental test can never be estimated very accuretaly. The validity of the psychological test is a relative term. A test is valid for a particular purpose or in a particular situation; it is not generally valid.

Validity is a matter of degree rather than an all-or-none property, and validation is an unending process. Now evidence may suggest modifications of an existing measure or the development of new and better approch to measuring the attribute in question, e.g. anxiety, intelligence or the :emparature of stars.

The validity of a test is generally found out by finding the correlation between the test and some independent criterion.

6.5.1 Kinds of Validity :-

Different kinds of validity are identified as it is not an absolute characteristic. According to a report prepared by a joint committee of the American Psychological Association, American Educational Research Association and National Council on Measurement used in Education, four types of validity have been distinguished, namely content validity, concurrent validity, predictive validity and construct validity. Factorial validity may be added to this list. Anastasi classifies them as (1) Face validity (2) Content Validity (3) Factorial Validity (4) Empirical Validity[43]

According to Thorndike and Hagen, Validity may be broadly classified into (1) rational, and (2) Empirical or experimental. The former consists of two classes viz. Content and Construct Validity and the later of three namely, Congruent, Concurrent and Predictive Validity.[44]

6.5.1. 1. Content Validity :-

What has been called "Content vallidity" is employed in the selection of items in educational achievement tests. Standard educational achievement examinations represent the consensus of many educators as to what a child of a given age or grade should know about a particular subject. A test of a particualr subject is judged to be valid if its content consists of questions covering these areas.

Barr, Davis and Johnson say;

> Logical content validity is obtained when an investigator defines and describes the abilities, traits, concepts or skills that he expects to be measured by an instrument of research, analyses them to indentify the elements needed in measuring instrument and designs the instrument with the demands of the situation as his criteria.[45]

6.5.1 2. Construct Validity :-

It is not concerned with content or subject matter acted upon but with the "functions" or "processes" that are applied to some content.[46] Though this is rational it helps in deciding what is to be measured and hence in a way helps in predicting efficiency of a test. To have a construct validity the test items must be specific, concrete and precise. They must consist of definite limited tasks.. The problem of preparing a test that has construct validity is that of bridging the gap from borad general concept to specific language tasks or test items.[47]

6.5.1.3 Conqruent Validity :

To find out this type of validity a test is correlated with an existing similar measure of the the same function. The validity of the test used as the ccriterion should be testified. The second type of validity is based on evidence that is empirical or statistical, one that comes from the relationship of the instrument to some other measure or fact.[48]

6.5.1. 4. Concurrent Validity :-

Concurrent validity is concerned with the relation of test scores to an accepted contemporary criterion of performance on the variable which the test is intended to measure.

As a validation criterion, school marks obiviously leave a great deal to be desired. Even a composite make has considerable unrealiability. And as an average it is made up of component usually unspecified, and each with a weighting which is not reported..... The reason why the criterion is so widely used is chiefly that it is about the only readily available numberical rating to be obtained on large number of persons.[49]

A secondary criterion quite often employed is that furnished by teachers' ratings. It has been frequently used in connection with the validation of intelligence tests.[50]

6.5.1. 5. Pradictive Validity :-

Predictive validity refers to the relation between test scores and criterion scores which can be obtained after the laps of some time. Predictive validity is the simplest of the three types of validity to understand. Predictive validity is at issue when the purpose is to use an instrumentt to estimate some important form of behaviour, the latter being refered to as the crierion. Primarily it consists of correlating scores on the predictor test with scores on variable. The size of the correlation is a direct indication of the amount of validity. Predictive ability, however, is limited to only a comparatively small part of the total domain of uses of psychological measures and to that of prediction problems in applied situations. Such predictor instruments have proved very useful in school settings and usually to a lesser extent in industrial, clinical, governmental and millitary settings.

6.5.1. 6. Factorial Validity :-

The factorial validity of a test is the correlation between that test and the factor common to a group of tests or other measures of behaviour.[51] While discussing about validity she further gives the new concept of factor validity which is seen from her following views :

> Returning now to the concept of the factorial validity 6f a test, we may note that such validity is simply the "factor loading" of a particular factor in the test in question. Such a factor loading is also equivalent to the correlation of the test with factor.[52]

Thus validity of a given test is defined by its factor loading and these are given by the correlation of the test with each factor. This phase which can be treated as part of construct validity, Will be discussed later in this chapter.

6.6. Validity of the Present Test :-

The difficulty of validation lies in securing a suitable validation criterion against which the test may be validated. Various investigators have therfore employed various devices to establish the validity of correlation with any of them.[53] Bhatia then quotes Wechsler's two ststements which are as follows :

> The Bellevue scales were devised because of the belief that the Binet scales were not sufficiently "good" measures of intelligence for adults. Otherwise, indeed, we should not have gone to the trouble of devising our tests.[54]

The second statement runs as follows :

> No new test can be markedly out of line with established measures of intelligence and still claim to be "good" measure of it, because that would be tentamount to saying that all other tests were not reliable measures of it. But the degree to which any new test correlates with established test (e.g., the Binet) cannot in and of itself be accepted as a basic proof of the new tests validity.[55]

From the above statement it follows that a certain degree of correlation with the established measures is desirable but that a high degree is not essential.

The present adapted I.Q. measurement test scores have been validated against the performances of the pupils at their annual examination. For this the percentage marks of the pupils falling in particular age groups have been recorded and their levels of I.Q. as measured by the present adapted test have been juxtaposed. The means and standard deviations of the total percentage marks of the high and low I.Q. groups were computed and the C.R.S. were calculated to show the significance of the difficulte of marks between the high and low I.Q. groups. The relevant statistics have been given in the table 6.12 below.

TABLE 6.12 : STATISTICS FOR ESTABLISHMENT OF CROSS VALIDATION STUDY OF THE PRESENT TEST

Age groups	Total Percentage in Achievement of H I.Q.		L.I.Q.		Difference in	σM	C.R.
	$\bar{x}$	σ	$\bar{x}$	σ	$\bar{X}$		
16^+	54.31	11.09	37.68	9.72	16.63	2.69	6.182
17^+	61.72	12.40	39.12	13.42	22.60	3.34	6.77
18^+	58.16	15.13	36.44	12.73	21.72	3.61	6.02
19^+	52.92	8.09	35.76	10.02	17.16	2.45	7.01
20^+	49.79	10.91	32.22	8.71	17.57	2.55	6.89
21^+	54.33	8.97	41.06	10.09	13.27	2.46	5.38
22^+	59.12	11.52	41.16	13.12	17.96	3.19	5.63

All the CRs are significant at 0.01 level.

From the table, it is observed that the high I.Q. groups Performance inthe school subjects were better than those of the low I.Q. groups in all the age groups. This proves that the calegories of high and low intelligence obtained school achievement according to their level of intelligence. In fact this type of validaity is useful in prediction of school $uccess. Hence this is an example of predictive validity of the present adapted test -

This test Predicts the academic success so well that it could be used to predict the future academic achievement of the pupils of adult age.

The present test was also validated against the WAIS IQ. Though Weschler Adult Intelligence Scale is not adapted for Gujarati population and the results may not be reliable & valid But as thise test is also having performance type of test items it was thought for the sake of inquisitiveness to compare the scores on WAIS with those on PPTI adapted for adult groups. The WAIS IQs have been given in age-range of 16-17, 18-19 and 20-21 in the initial stage. the investigator had calibrated IQ against the raw scores age wise, i.e. IQ for ages 16, 17, 18, 19, 20, 21 & 22.

To suit the pattern of WAIS I.Q., grouped the I.Q. of the 16-17, 18-19 and 20-21. The remaining I.Q for age 22 had been dropped because the age group of 23 had not been covered in the sample by the present researcher.

The co-efficients of correlation of the above age-groups were computed by product-moment formula using scatterogram. The scatterogram and requisite statistics have been given in the tables below.

TABLE 6.13 : VALIDATION OF PRESENT ADAPTED TEST WITH WAIS FULL SCALE I.Q.s OF AGE GROUP 16-17 BOYS

WAIS SCORE Present Test Score	57- 66	67- 76	77- 86	87- 96	97- 106	107- 116	117- 126	127- 136	137- 146	147- 156	TOTAL
121-123											
118-120										2	02
115-117										3	03
112-114										4	04
109-111									3	3	06
106-108									6		06
103-105								4	1		05
100-102							3	2			05
97- 99							2				02
94- 96						4	1				05
91- 93				4	2	6					12
88-90			2	2							04
85-87			3								03
82- 84		2									02
79- 81	1										01
76- 78											
73- 75											
TOTAL	01	02	05	06	02	10	06	06	10	12	60

N = 60

Statistics:-

$\overline{x}$ = 100.38	Ex = 6023	x = 10.02
$\overline{Y}$ = 120.67	Ey = 7240	y = 26.62

r : 0·86

TABLE 6.14 : VALIDATION OF PRESENT ADAPTED TEST WITH WAIS FULL SCALE I.Q s OF AGE GROUP 18-19 BOYS

WAIS SCORE Present Test Scores	57- 66	67- 76	77- 86	87- 96	97- 106	107- 116	117- 126	127- 136	137- 146	147- 156	TOTAL
121-123											
118-120											
115-117										1	01
112-114										2	02
109-111										6	06
106-108								4	6		10
103-105							4	2			06
100-102							5				05
97- 99						3	1				04
94- 96						5					05
91- 93					5						05
88- 90				3	1						04
85- 87				4							04
82- 84			1	2							03
79-81		1	1								02
76- 78			2								02
73- 75			1								01
TOTAL		01	05	09	06	08	10	06	06	09	60

Statistics :-

N = 60

$\bar{x}$ = 97.95 Ex = 5877 x = 10.16

$\bar{y}$ = 117.31 Ey = 7039 y = 22.22

r = 0.87

TABLE 6.15 : VALIDATION OF PRESENT ADAPTED TEST WITH WAIS FULL SCALE I.Q.s OF AGE GROUP 20-21 BOYS

WAIS SCORE	57-	67-	77-	87-	97-	107-	117-	127-	137-	147-	total
Present Test Scores	66	76	86	96	106	116	126	136	146	156	
121-123											
118-120										1	01
115-117										1	01
112-114									3	2	05
109-111								2	33		05
106-108							4	5			09
103-105						4	2				06
100-102						4					04
97- 99				1	5						06
94- 96				6							06
91- 93			3	2							05
88- 90		5	2								07
85-87	2	1									03
82- 84	2										02
79- 81											
76- 78											
73- 75											
total	04	06	05	09	05	08	06	07	06	04	60

Statistics:-

$\bar{x}$ = 100.11   Ex = 6007   x= 9.16

$\bar{y}$ = 106.75   Ey = 6405   y = 26.62

r = 0.89

TABLE 6.16 : VALIDATION OF PRESENT ADAPTED TEST WITH WAIS FULL SCALE I.Q.s OF AGE GROUP 16-17 GIRLS

WAIS SCORE Present Test Scores	57- 66	67- 76	77- 86	87- 96	97- 106	107- 116	117- 126	127- 136	137- 146	147- 156	TOTAL
121-123											
118-120										2	02
115-117									2	2	04
112-114								3	2	2	07
109-111								2	3		05
106-108								4	1		05
103-105							3	2	1		06
100-102						3	2				05
97- 99							2	1			03
94- 96						2	1				03
91- 93					4	3					07
88- 90				1	2						03
85-87			2	2							04
82- 84			2	1							03
79- 81		1	2								03
76- 78											
73- 75											
TOTAL		01	06	04	06	08	08	12	09	06	60

Statistics :-

N = 60

$\overline{x}$ = 102.26 Ex = 6136 x = 12.17

$\overline{y}$ = 98.35 Ey = 5901 y = 20.21

r = 0.82

TABLE 6.17 : VALIDATION OF PRESENT ADAPTED TEST WITH WAIS FULL SCALE I.Q.s of AGE GROUP 18-19 GIRLS

WAIS SCORE	57-	67-	77-	87-	97-	107-	117-	127-	137-	147-	TOTAL
PRESENT TEST SCORE	66	76	86	96	106	116	126	136	146	156	
121-123											
118-120											
115-117									1	2	03
112-114								1	2	2	05
109-111								2	1		03
106-108							2	1	3		06
103-105							2	2			04
100-102						2	1				03
97- 99					2	1	3				06
94- 96						3	2				05
91- 93				2	2	3					07
88- 90				1	2						03
85-87			2	2							04
82-84			2	1	1						04
79- 81			1	2							03
76- 78			1	1							02
73- 75		1	1								02
TOTAL		01	07	09	07	09	10	06	07	04	60

Statistics :- N = 60

$\bar{x}$ = 103.32 Ex = 6199 x = 11.32

$\bar{y}$ = 112.26 Ey = 6736 y = 18.33

r = 0.88

TABLE 6.18 : VALIDATION OF PRESENT ADAPTED TEST WITH WAIS FULL SCALE I.Q.s OF AGE GROUP 20.21 GIRLS

WAIS SEORE PRESENT TEST SCORES	57- 66	67- 76	77- 86	87- 96	97- 206	107- 116	117- 126	127- 136	137- 146	147- 156	TOTAL
121-123											
118-120											
115-117										1	01
112-114								1	2		03
109-111							2	3	1		06
106-108							4	3	2		09
103-105						3	2	1			06
100-102					2	1	2	1			06
97- 99					1	2	4				07
94- 96				3	2	2					07
91- 93				4	2						06
88- 90				2	3						05
85- 87			1	1							02
82- 84			2								02
79- 81											
76- 78											
73- 75											
TOTAL			03	10	10	08	14	09	05	01	60

Statistics :-

N = 60

$\bar{X}$ = 116.30   Ex = 6978   x = 13.52

$\bar{Y}$ = 118.32   Ey = 7099   y = 17.28

r = 0.88

TABLE- 6.19 : CORRELATION BETWEEN PRESENT ADAPTED TEST AND WAIS

Present Adopted Test Age Group		coefficient of correlation	WAIS Age group	coefficient of correlation
16 - 17	Boys	0.86	18 - 19	0.84
	Girls	0.82		
18 - 19	Boys	0.82	25 - 34	0.86
	Girls	0.88		
20 - 21	Boys	0.89	45 - 54	0.89
	Girls			

It is observed that the rs in all the three age-groups far exceed the critical values. Hence it could be said that the present adapted test is a valid instrument for measuring the intelligence of the adult group having age range 16^{+}-22^{+}.

From the foregoing discussion one can easily conclude that the total Score of the present test is a good indicator of G.

To sumup, the results of Reliability & validity studies show the efficiency of the present test as the measure of intelligence.

-x-x-x-x-