CHAPTER X

THE ESTIMATION OF TEST RELIABILITY

THE MEANING OF RELIABILITY

If a test is applied a second time under similar conditions, and the testees' scores differ widely from those previously obtained, the test is obviously a poor one. It is said to be 'reliable' only if the two sets of scores correlate highly with one another.

The term reliability was first introduced by Spearman. He used it in his basic papers on test theory.

There is a distinction between reliability as defined in theory and reliability as estimated in test operations. There are several meanings attached to reliability. Common synonyms for reliability include: dependability, consistency, and stability. Each signifies something different as applied to measurements. Even the same term has slightly different meanings as applied to different measurement operations.

The original approach to the problem of reliability by Spearman was based on the correlation of parallel tests.

As reliability is a general term and refers to several types of evidence regarding the consistency of measurement,

different types of reliability coefficients answer different questions and permit different inferences regarding the evidence. Different types of reliability coefficients should,therefore,be found out to establish the reliability of any measuring tool. Each kind of reliability coefficient should be estimated from empirical data. Before these different kinds of reliability coefficients and methods to measure them are discussed, the different meanings given to test reliability should be studied in details.

Dr.Micheels and Dr.Karnes[1] say,

> A reliable test is a 'trustworthy' test. It is accurate. It is consistent. If the test measures in exactly the same manner each time it is administered, if the factors that affect the test scores affect them to the same extent every time the test is given, the test is said to be high in reliability.

Thus, according to them, a test which is trustworthy, accurate and consistent, is a reliable one.

Again as Ross[2] puts it,

> By reliability is meant the degree to which the test agrees with itself. To what extent can two or more forms of the test be relied upon to give the same results, or the same test to give the same results when repeated? If the scores on the test are stable under these conditions, the test is said to be reliable. In a word,reliability means consistency.

---

1 Micheels, W.J. and Karnes, M.R., "Measuring Educational Achievement", McGraw - Hill Book Company,Inc.,1950, p. 111.

2 Ross, C. C., "Measurement in To-day's Schools", Prentice-Hall, Inc., New York, Third Edition, 1955, p.121.

Reliability or consistency in a measurement procedure is a matter of degree and not an all-or-none matter. Whenever we measure anything, whether in the physical, the biological, or the social sciences, that measurement contains a certain amount of chance error. The amount of chance error may be large or small, but it always is present to some extent. If the chance errors are small in size, relative to the variation from person to person, the reliability or consistency of the measure is high. If the chance errors become large in proportion to the variation from person to person, the reliability of the measure is low.

Not only that a testees' score on a reliable test should be the same or nearly the same on his taking the same test for a second time but his scores should also tally much reasonably when he is administered and scored the same reliable test by different testers at different times. This fact suggests that the test's consistency is maintained only if it is administered in certain standardized procedures, and that it should be scored on a definite pattern, meaning, it should be objective.

Summarising, it can be said that a test, for being a reliable one, should be precise, trustworthy, consistent and objective.

IMPORTANCE OF RELIABILITY

Shakespeare said: "consistency, thou art a jewel",

and he was right. But a person can be consistently wrong and a test also can measure consistently something which is wrong. So it is not the greatest jewel. High reliability is no guarantee for the goodness of a test; but the low reliability is a definite proof of a poor test. What a test measures consistently, should be the one which the test is meant to measure. The ideal test tells the truth consistently.

Just as a person's consistency in behaviour should be ascertained before he is entrusted with any responsible work, so a test user would naturally, like to know first how much dependable the test is before he thinks of using it for some purpose. Clearly, any degree of unreliability in the score resulting from the application of a measuring device is distress -ing to the educator, guidance worker, industrial personnel officer, or other individual who must use that score as a basis for a practical decision. Unreliability introduces a question mark after the score, and means that any judgment based upon it must be tentative. The use of an instrument with low reliability will give an incorrect and deceiving estimate of a testee. It is for this reason that the test-users attach much more importance to high reliability of a test and select for use only the tests with high reliability.

Reliability becomes of critical importance in research studies at a number of points. Thorndike[1] says,

---

1 Lindquist, E. F., Op.Cit., p. 563.

> In any study of prediction and in any study of improvement resulting from training, some degree of reliability in the measure of the criterion being predicted or in the ability being trained is imperative if one is to achieve any prediction on the one hand or any evidence of improvement on the other.

To conclude this discussion, it must be accepted that validity is the first quality that is sought in a test, and granted that, reliability is a valuable auxiliary.

FACTORS INFLUENCING THE RELIABILITY OF A TEST

As it is discussed above, reliability is considered as the degree to which a true or perfect measurement of each testee is obtained when a measure is applied. It is impossible to obtain such a perfect measure in any field of science. The measure is always contaminated by chance factors which affect the accuracy of the measurement. Some possible sources of variation in psychological measurement are listed by Thorndike[1] under the following headings:

(1) Lasting and general characteristics of the individual.

(2) Lasting but specific characteristics of the individual.

(3) Temporary but general characteristics of the individual.

---

1 Lindquist, E. F., Op.Cit., p. 568.

(4) Temporary and specific characteristics of the individual.

(5) Systematic or chance factors affecting the administration of the test or the appraisal of test performance.

(6) Variance not otherwise accounted for (chance).

A similar analysis, differing in detail, has been formulated by Cronbach.[1]

The variance in score on a test is caused by one or more factors listed above. Thorndike[2] says,

> There will also be variance which is associated only with the one particular set of measurements, that is, which will not be reproduced another time. This may be designated "error" variance. The existence of this error variance corresponds to the fact of unreliability, and its amount relative to the total of all variance is a measure of the degree of unreliability.

The reliability coefficient is also affected by the length of the test. The importance of lengthening tests is that, with every item added, the sample of performance becomes more adequate. By lengthening the test, the reliability coefficient is increased proportionally. Longer tests also are less influenced by guessing.

---

1 Cronbach, L.J., "Essentials of Psychological Testing", Harper and Brothers, Publishers, New York, 1949, p. 60.

2 Lindquist, E. F., Op.Cit., p. 565.

The reliability coefficient, since it deals with the variance of item performance as a ratio to variance on total test performance, is closely related to the spread or range of scores in the group studied.

And lastly, a test may give reliable measures at one level of ability, and unreliable measures at another level.

DIFFERENT METHODS OF ESTIMATING TEST RELIABILITY

The discussion in the above paragraphs, concerning the various sources of true and error variance reveals that there is a large number of determiners of reliability coefficient. These sources sometimes increase the reliability coefficient and sometimes decrease it. Moreover, these sources of variance can be classified in different types. Each type of sources of variance gives rise to reliability coefficient. Three major types of reliability coefficients are generally used in describing consistency of measurement for psychological tests and techniques. These are:

(i) Coefficient of internal consistency,

(ii) Coefficient of equivalence,

and (iii) Coefficient of stability.

The choice of which type of reliability estimate to use can often be facilitated by deciding what sources we want to go into error variance and what sources into true variance.

The interpretation of a coefficient will also be more enlightened if we keep in mind all the things that might have contributed to its size. If the circumstances are favourable to collect different types of data needed to calculate different types of reliability coefficients, the present test author feels it necessary to calculate all types of reliability coefficients.

We shall first elaborate each type of reliability coefficient and then discuss methods used to calculate each type.

(1) Coefficient of internal consistency: This is the estimate obtained from the single administration of a test to a representative group of individuals. It indicates how consistently the test measures the individual's performance at a particular moment. The methods, generally, used to estimate this coefficient are: (1) split-half method, (2) Kuder-Richard-son method, and (3) Cyril Hoyt's method - based on 'Analysis of variance technique.

(2) coefficient of equivalance: It measures fluctuations from day-to-day in the individual and fluctuations in the sampling of content of the test. Or in other words, it indicates both equivalence of content and stability of performance. The preferred method of determining this coefficient, is the alternate or parallel-forms method.

(3) Coefficient of stability: This indicates the degree to which the scores on a particular test are stable over a given period of time. This coefficient tells us nothing

concerning the internal consistency of a test. In fact, the parts or even the items of a test might inter-correlate zero and yet the coefficient could be high. This coefficient is estimated by the test-retest method.

The table below summarises the above discussion.

TABLE NO. 46

SHOWING TYPES OF RELIABILITY COEFFICIENTS AND METHODS TO ESTIMATE THEM

Sr. No.	Type of reliability coefficient	Methods used to estimate it	Method depending upon
1	Coefficient of internal consistency	(1) split-half method	(1) Correlation
		(2) K-R method	(2) Item-analysis data
		(3) Cyril Hoyt's method	(3) Item-analysis data
2	Coefficient of equivalence	Alternate or parallel-forms method	Correlation
3	Coefficient of stability	Test-retest method	Correlation

Each method to estimate corresponding reliability coefficient is, now, discussed in brief.

SPLIT-HALF METHOD

It is generally agreed that "split-half" method of establishing reliability should be applied to power tests only.

Or, when only one form of a test is available, reliability is determined by a "split-half" method.

In the split-half method, the test is first divided into two equivalent "halves" and the correlation found for these half-tests. From the reliability of the half-test, the self-correlation of the whole test is then estimated by the Spearman-Brown prophecy formula.

The major problem in using half-tests for the purpose of estimating reliability is dividing the original test into two equivalent halves.

Harold Gulliksen[1] suggests following methods to divide a test into two equivalent halves.

(1) Successive halves or thirds.

(2) Odd versus even items or every nth item.

(3) Matched random sub-tests.

The odds-evens split is the one most commonly used. The main advantage in using this method is that all the data for computing reliability are obtained upon one occasion, thereby eliminating the variations brought about by differences between the two testing situations. On the other hand, a marked disadvantage is that chance errors may affect scores on the two

---

1 Gulliksen, Harold, "Theory of Mental Tests", John Wiley & Sons, Inc., New York, 1950, pp. 201-210.

halves of the test in the same way, thus tending to make the reliability coefficient too high.

TEST-RETEST METHOD

This is the simplest method of determining reliability coefficient. This method involves repetition of the test. The test is given and repeated on the same group after a due course of time. The reliability coefficient is found out by computing correlation between the first and the second set of scores.

This method should be used only when it is not possible to use other methods. It has the only merit of being simple and easy to apply, while there are a number of disadvantages. It is extremely difficult to control the conditions that influence scores on retest, and this makes its use inadvisable.

ALTERNATE OR PARALLEL-FORMS METHOD

The alternate-forms method bears resemblances to both the internal-consistency approach and the retest approach. The alternate-forms method is satisfactory when sufficient time has intervened between the administration of the two forms to weaken or eliminate memory and practice effect.

In drawing up alternate test forms, care must be exercised to match test materials for content, difficulty and form. The definition of the alternate-forms method has been sharpened somewhat in recent years.

Thorndike[1] speaks of equivalent forms by which he means tests having identical true variance and no overlap of error variances.

Gulliksen[2] speaks of parallel tests, which he defines statistically. Parallel tests have equal means, equal variances, and equal inter-correlations with one another. For the purpose of determining whether they are parallel in all these respects, he recommends the construction of at least three forms so that there can be three estimates of inter-correlation. He presents various statistical tests for determining whether these properties of parallel tests have been satisfied.

THE METHOD OF "RATIONAL EQUIVALENCE"
('K-R' - METHOD - FORMULA - 20)

Dissatisfied with the split-half method, Kuder and Richardson developed new procedures based on item statistics. They split a test into n parts of one item each.

This method stresses the inter-correlations of the items in the test and the correlations of the items with the test as a whole. The formula known as 'K-R' formula - 20, is the basic one used for computing reliability coefficient by this method.

Rational equivalence formula tends to under-estimate somewhat the reliability coefficient as found by other methods.

---

1 Lindquist, E. F., Op.Cit., p. 575.

2 Gulliksen, Harold, Op.Cit., pp. 174-180.

This formula provides an estimate of the internal consistency of the test and thus of the dependability of test scores.

HOYT'S 'ANALYSIS OF VARIANCE' TECHNIQUE

The problem of estimating test reliability from consistency of individual performance upon the items of a test has been attacked directly by analysis of variance techniques by Cyril Hoyt.[1] He assumes that the score of an individual on a test may be divided into four independent components, as follows:

(i) A component common to all individuals and to all items.

(ii) A component associated with the item.

(iii) A component associated with the individual.

(iv) An error component that is independent of (i), (ii) and (iii).

It is assumed further that the error component of each item is normally distributed, that the variance of the error component is the same for each item, and that the error components for any two distinct items are uncorrelated. When these conditions are fulfilled, it is possible to analyse the

---

1 Hoyt, Cyril, "Test Reliability obtained by Analysis of Variance", Psychometrica, 6, (1941), pp. 153-160.

variance in test scores into the variance contributed by each of the last three components. Reliability is, then, estimated from the following expression.

$$\text{Reliability} = 1 - \frac{\text{Error variance}}{\text{Variance among individuals}}$$

METHODS USED TO ESTIMATE RELIABILITY OF THE PRESENT TEST

The reliability of the present test is estimated by the following three methods.

(1) Split-half method,

(2) Hoyt's 'analysis - of - variance' technique,

and (3) K-R method - formula 20.

The 'test-retest' method is not applied, as it is not advisable to repeat the test itself.

The 'parallel-forms' method also is not applied here as it is not feasible to construct parallel forms of the test.

We shall now proceed to the discussion of calculation details involved in the application of each method to estimate reliability of the present test.

APPLICATION OF 'SPLIT-HALF' METHOD ('WHOLE-TEST' - 'TOTAL SAMPLE')

The test was divided into two equivalent halves. For dividing the test into two parts, the odd - even method was

followed. The scores on the two halves of the test were correlated and the product-moment coefficient of correlation was found out. This gave the half-test reliability. The self-correlation of the whole test was, then, calculated by applying the Spearman-Brown prophecy formula.

SAMPLE

A small sample of 200 testees out of the total sample of 530 testees was selected for the purpose of applying 'split-half' method to estimate reliability of the whole test. The testees selected for the 'reliability sample' were in order of 1, 4, 7, 10, 13 .....etc. The sample should be the similar one to the parent-sample. The mean, median and standard deviation of the new distribution were calculated. The results are shown in the following table.

TABLE NO. 47

DATA GROUPED FOR THE CALCULATION OF MEAN, MEDIAN AND STANDARD DEVIATION OF THE DISTRIBUTION OF THE 'RELIABILITY SAMPLE'

Scores	Mid.pts.	f	cum.f.	x'	fx'	fx'^2
100-104	102.0	5	200	+4	20	80
95- 99	97.0	14	195	+3	42	126
90- 94	92.0	21	181	+2	42	84
85- 89	87.0	41	160	+1	41	41
					+145	
80- 84	82.0	47	119	0	0	0
75- 79	77.0	34	72	-1	-34	34
70- 74	72.0	20	38	-2	-40	80
65- 69	67.0	10	18	-3	-30	90
60- 64	62.0	7	8	-4	-28	112
55- 59	57.0	0	1	-5	0	0
50- 54	52.0	1	1	-6	- 6	36
					-138	
		N = 200			$\Sigma fx' = +7$	$\Sigma fx'^2 = 683$

Mean = 82.175

Mdn. = 82.480

SD = 9.250

The mean score and standard deviation of this 'new' sample are, thus, 82.175 and 9.25 respectively. These measures very fairly tally with the corresponding measures obtained from the parent sample. These are 79.09 and 9.27 respectively.

The 'reliability sample' can, therefore, be said to be most satisfactorily representative of the total sample.

CALCULATION OF RELIABILITY COEFFICIENT

The total scores of all the 200 testees on both the 'halves' were found out and then Pearson's product-moment coefficient of correlation between them was computed. The scattergram of scores on both halves is shown in the following table.

TABLE NO. 48

SHOWING THE SCATTERGRAM OF SCORES USED IN 'SPLIT-HALF' METHOD

		'Even-items' scores (X-variable)						
		20-24	25-29	30-34	35-39	40-44	45-49	fy
'Odd-items' scores (y-variable)	50-54	-	-	-	2	-	-	2
	45-49	-	-	-	4	9	2	15
	40-44	-	-	15	28	9	-	52
	35-39	-	13	48	24	-	-	85
	30-34	6	13	12	-	-	-	31
	25-29	4	9	-	-	-	-	13
	20-24	1	1	-	-	-	-	2
	fx	11	36	75	58	18	2	200

The product-moment coefficient of correlation was, then, calculated from the above table in the usual procedure. It was found to be equal to 0.782. This coefficient of

correlation gives the reliability of a test of half the original test-length. It is necessary to find out the self-correlation of the whole test. The reliability of the whole test was computed by applying the Spearman-Brown prophecy formula. The formula[1] used, is given below.

$$r_{11} = \frac{2r_{\frac{1}{2}\,1/11}}{1 + r_{\frac{1}{2}\,1/11}}$$

Where r_{11} = reliability coefficient of the whole test,

and $r_{\frac{1}{2}\,1/11}$ = reliability coefficient of the half-test.

Substituting the value of each, in the above formula, we get,

$$r_{11} = \frac{2 \times 0.782}{1 + 0.782}$$

$$= \frac{1.564}{1.782}$$

$$= \underline{0.878}$$

Thus, the reliability-coefficient of the present aptitude-test, as calculated by the 'split-half' method is 0.878.

The P.E. of the 'r' (0.878) was found as under:

---

1 Garrett, H. E., Op.Cit., p. 339.

$$P.E._{'r'} = 0.6745 \times \frac{1 - r^2}{\sqrt{N}}$$

$$= 0.6745 \times \frac{1 - (0.878)^2}{\sqrt{200}}$$

$$= 0.6745 \times \frac{0.229}{14.142}$$

$$= 0.0109$$

$$= \underline{0.011}$$

APPLICATION OF HOYT'S METHOD
('WHOLE-TEST' - 'TOTAL-SAMPLE')

The procedure[1] was carried out in the following way:

(1) The total sample of 530 testees was used for this purpose. All the test-scripts were arranged according to their rank order.

(2) The 'Student-item' chart was constructed.

A specimen of the same is given on the next page.

---

1 Micheels, W. J. and Karnes, M.R., Op.Cit.,pp. 472-476.

TABLE NO. 49

A SPECIMEN OF 'STUDENT-ITEM' CHART

Testees	Items				.....	120 (n)	Each testee's total score 't'	t^2
	1	2	3	4				
1							t_1	t_1^2
2							t_2	t_2^2
3							t_3	t_3^2
4							t_4	t_4^2
.							.	.
.							.	.
.							.	.
.							.	.
530 (k)							t_{530}	t^2_{530}
Numbers of correct responses on each item: 'p'	p_1	p_2	p_3	p_4	.....	p_{120}	$\sum_{i=1}^{n} pi = \sum_{i=1}^{n} ti = 35059$	
p^2	p_1^2	p_2^2	p_3^2	p_4^2	.....	p^2_{120}		
$\sum t^2 = 238407$							$\sum p^2 = 11696132$	

In the table No. 49 on page No. 311,

n = number of items,

k = number of individuals,

p = number of correct responses on each item,

t = each testee's total score,

Σp = sum of all 120 p's,

Σt = sum of all 530 t's,

Σp^2 = sum of all p^2's,

Σt^2 = sum of all t^2's.

(3) Then the 'analysis of variance' table was construct-ed and necessary calculations were made to fill in the blanks in the table. The table, with blanks duly completed, is shown on the next page.

TABLE NO. 50

AN ANALYSIS-OF-VARIANCE TABLE WITH THE FORMULAS FOR COMPUTING THE VALUES UNDER EACH HEADING AND THE ACTUAL VALUES OBTAINED UNDER EACH HEADING

Source of variance	Degrees of freedom - df	Sum of squares	Mean of squares	
Between individuals	$k - 1 = 529$	$\frac{1}{n}\Sigma t^2 - \frac{(\Sigma t)^2}{nk}$ $= 530$	$\frac{\text{Sum of squares}}{\text{df}}$ $= 1.002$	(a)
Between items	$n - 1 = 119$	$\frac{1}{k}\Sigma p^2 - \frac{(\Sigma t)^2}{nk}$ $= 2740$	$\frac{\text{Sum of squares}}{\text{df}}$ $= 23.03$	(b)
Residual	$(n-1)(k-1) = 62951$	Total - (between individuals + between items) $= 12460$	$\frac{\text{Sum of squares}}{\text{df}}$ $= 0.198$	(c)
Total	$nk - 1 = 63599$	$\frac{(\Sigma t)(nk - \Sigma t)}{nk}$ $= 15730$	-	

DETAILS OF ABOVE CALCULATIONS

Degrees of freedom between individuals $= k - 1$

$= 530 - 1$

$= 529$

Degrees of freedom between items $= n - 1$

$= 120 - 1$

$= 119$

Residual $= (n-1)(k-1)$

$= (120-1)(530-1)$

$= (119)(529)$

$= 62951$

Total degrees of freedom $= nk - 1$

$= (530 \times 120) - 1$

$= 63599$

Sum of squares between individuals $= \frac{1}{n} \Sigma t^2 - \frac{(\Sigma t)^2}{nk}$

$= \frac{1}{120} \times 2384070 - 19340$

$= 19870 - 19340$

$= 530$

Sum of squares between items $= \frac{1}{k} \Sigma p^2 - \frac{(\Sigma t)^2}{nk}$

$= \frac{1}{530} \times 11696132 - \frac{(35059)^2}{530 \times 120}$

$= 22080 - 19340$

$= 2740$

Total sum of squares $= \frac{(\Sigma t)(nk - \Sigma t)}{nk}$

$= \frac{(35059)(63600 - 35059)}{63600}$

$= 15730$

Residual sum of squares = total - (between items + between individuals)

$= 15730 - (530 + 2740)$

$= 15730 - 3270$

$= 12460$

Mean of squares for individuals $= \frac{530}{529} = 1.002$

Mean of squares for items $= \frac{2740}{119} = 23.03$

Mean of squares for residual $= \frac{12460}{62951} = 0.198$

The reliability coefficient of the aptitude test was obtained by the following formula:

$$r_{tt} = \frac{a - c}{a}$$

$$= \frac{1.002 - 0.198}{1.002}$$

$$= 0.802$$

The reliability coefficient obtained by this method is, then, 0.802, which is a little less than that obtained by the 'split-half' method.

APPLICATION OF THE KUDER-RICHARDSON METHOD (FORMULA - 20) ('WHOLE TEST' -'TOTAL SAMPLE')

The K - R formula - 20, used here, is given below.[1]

$$r_{11} = \frac{n}{(n-1)} \times \frac{\sigma_t^2 - \Sigma pq}{\sigma_t^2}$$

Where r_{11} = reliability coefficient of the whole test,

n = number of items in the test,

σ_t = the SD of the test scores,

p = the proportion of the group answering a test item correctly,

q = $(1 - p)$ = the proportion of the group answering a test item incorrectly.

To apply this method also, the original sample of 530 testees was used. In the above formula n is, therefore, equal to 530 and σ_t is equal to 9.27. The proportion of the group answering a test item correctly, p, was found out for each of the 120 test-items. From these values of 'p', the values of corresponding q 's were also found out. In the following table, the value of 'pq ' for each item is given. The sum of all pq - values is equal to 17.5305.

---

1 Garrett, H. E., Op.Cit., p. 341.

TABLE NO. 51

SHOWING 'pq ' VALUES OF 120 TEST-ITEMS

Item No.	'pq '	Item No.	'pq '
1	0.1785	23	0.1240
2	0.1707	24	0.1783
3	0.1314	25	0.0539
4	0.1853	26	0.1400
5	0.1216	27	0.2072
6	0.1623	28	0.1706
7	0.1302	29	0.1008
8	0.1412	30	0.1179
9	0.2101	31	0.1347
10	0.1602	32	0.0412
11	0.1700	33	0.1287
12	0.1882	34	0.0928
13	0.1703	35	0.0473
14	0.2001	36	0.0432
15	0.1307	37	0.0666
16	0.1618	38	0.0808
17	0.0581	39	0.1130
18	0.2101	40	0.1252
19	0.1391	41	0.1371
20	0.1603	42	0.2000
21	0.1218	43	0.1743
22	0.1421	44	0.2257

Item No.	'pq '	Item No.	'pq '
45	0.0820	69	0.1170
46	0.0963	70	0.1003
47	0.0927	71	0.1072
48	0.1732	72	0.1275
49	0.1029	73	0.0898
50	0.1369	74	0.1337
51	0.1124	75	0.1305
52	0.1040	76	0.1721
53	0.1314	77	0.1782
54	0.1406	78	0.1957
55	0.0514	79	0.1543
56	0.1615	80	0.2176
57	0.1042	81	0.2074
58	0.0927	82	0.2109
59	0.0900	83	0.2065
60	0.0817	84	0.1759
61	0.0569	85	0.1887
62	0.1110	86	0.1150
63	0.1432	87	0.1071
64	0.1389	88	0.0899
65	0.1473	89	0.0667
66	0.1879	90	0.0144
67	0.2187	91	0.1380
68	0.1313	92	0.1477

Item No.	'pq '	Item No.	'pq '
93	0.1414	107	0.1794
94	0.1568	108	0.1935
95	0.1507	109	0.2039
96	0.1888	110	0.2089
97	0.2055	111	0.2144
98	0.2188	112	0.1910
99	0.2038	113	0.1683
100	0.2187	114	0.1773
101	0.2231	115	0.1512
102	0.2279	116	0.1727
103	0.2324	117	0.1163
104	0.2432	118	0.1084
105	0.2419	119	0.0761
106	0.2460	120	0.0415

$$\Sigma pq = 17.5305$$

The reliability coefficient was, then, calculated as shown below:

$$r_{11} = \frac{n}{(n-1)} \times \frac{\sigma_t^2 - \Sigma pq}{\sigma_t^2}$$

$$= \frac{120}{119} \times \frac{(9.27)^2 - 17.53}{(9.27)^2}$$

$$= \frac{120}{119} \times \frac{85.94 - 17.53}{85.94}$$

$$= \frac{120 \times 68.41}{119 \times 85.94}$$

$$= 0.803$$

The reliability coefficient of the present aptitude test as measured by K-R method is, therefore, 0.803.

The results obtained by the use of different methods are summarised in the following table:

TABLE NO. 52

SHOWING THE RELIABILITY COEFFICIENT OF THE PRESENT TEST AS OBTAINED BY DIFFERENT METHODS

Sr. No.	The method used	Reliability coefficient obtained	P.E.r
1	'Split-half' method	0.878	± 0.011
2	Hoyt's method	0.802	-
3	Kuder-Richardson method	0.803	-

It can be seen that the results obtained from K-R formula-20 and Hoyt's method, are identical. In fact, as shown in various text books, they should be so. The split-half method gives a little higher value of reliability coefficient. This discrepancy might be attributed to overestimation of reliability coefficient by the split-half method or to underestimation by the use of K-R formula-20 or Hoyt's method.

This may mean that two good methods confirm each other and that the best estimate that can be obtained of reliability for this test is about 0.80 or it may mean that the Hoyt's method, like the K-R method, gives biased results. At any rate, it is safe to say that the reliability of this test is probably not lower than 0.80.

COMPARISON OF METHODS

The factors, which influence the reliability of a test, are discussed in details, in the earlier pages of this chapter. Each method, applied to estimate the reliability, is associated with some sources of variation that cause errors in measurement. A table is given below, to enable a summary comparison of the different sources of variation represented in three different procedures applied for estimating reliability of the present test.

TABLE NO. 53

SOURCES OF VARIATION REPRESENTED IN DIFFERENT PROCEDURES FOR ESTIMATING RELIABILITY

Sr. No.	Sources of variation	Experimental procedure for estimating reliability		
		'Split-half' method	'K-R' method	Hoyt's method
1	Variations arising within the measurement procedure itself.	✓	✓	✓
2	Changes in the individual from day to day	-	-	-
3	Changes in the specific sample of tasks	✓	✓	✓
4	Changes in the individual's speed of work.	-	-	-

RELIABILITY IN TERMS OF TRUE SCORES AND MEASUREMENT ERRORS

For the purpose of interpretation and other uses, the reliability as obtained by K-R method or Hoyt's method, namely, 0.80, will be considered as the reliability of the present test.

A score on a mental test may be thought of as an index of the testee's "true ability" plus errors of measurement.

(A) THE RELIABILITY COEFFICIENT AS A MEASURE OF TRUE VARIANCE

The variance of the obtained scores can be divided into two parts: the variance of the true scores and the variance of chance errors. This is expressed mathematically as under:[1]

$$1 = \frac{\sigma_\infty^2}{\sigma_x^2} + \frac{\sigma_e^2}{\sigma_x^2}$$

Where $\frac{\sigma_\infty^2}{\sigma_x^2}$ = true score variance

and $\frac{\sigma_e^2}{\sigma_x^2}$ = error variance

Under certain reasonable assumptions, reliability coefficient becomes a measure of true score variance i.e.

$$r_{11} = \frac{\sigma_\infty^2}{\sigma_x^2}$$

and the above equation becomes:

---

1 Garrett, H. E., Op.Cit., pp. 345-346.

$$1 = r_{11} + \frac{\sigma_e^2}{\sigma_x^2}$$

$$\text{or} \quad r_{11} = 1 - \frac{\sigma_e^2}{\sigma_x^2}$$

This shows that if error variance is low the reliability coefficient is high and vice-versa. This formula also reveals the reliability coefficient to be an index of the precision of measurement.

The reliability of the present test is 0.80. Therefore, 80 per cent of the variance of test-scores is true-score variance and only 20 per cent error variance.

(B) ESTIMATING TRUE SCORES BY WAY OF THE REGRESSION EQUATION AND THE RELIABILITY COEFFICIENT

The regression equation[1] which estimates true score is given below:

$$\bar{X}_\infty = r_{11}X + (1 - r_{11})\ M$$

Where $\bar{X}_\infty$ = estimated true score on the test,

X = obtained score on the test,

M = mean of test distribution (79.09),

r_{11} = reliability coefficient of the test (0.80).

---

1 Garrett, H. E., Op.Cit., pp. 347-348.

The regression equation for estimating true score on the present test is worked out as under:

$$\bar{X}_\infty = 0.8\,X + (1 - 0.80) \times 79.09$$

$$= 0.8\,X + 0.2 \times 79.09$$

$$\bar{X}_\infty = 0.8\,X + 15.82$$

The standard error of an estimated true score is given by the following formula:[1]

$$SE_\infty = \sigma\sqrt{r_{11} - r^2_{11}}$$ Where $\sigma = 9.27$, $r_{11} = 0.80$

The SE_∞, of the true score on the present test is calculated below:

$$SE_\infty = 9.27 \times \sqrt{0.8 - (0.8)^2}$$

$$= 9.27 \times \sqrt{0.8 - 0.64}$$

$$= 9.27 \times \sqrt{0.16}$$

$$= 9.27 \times 0.4$$

$$= 3.71.$$

The 0.95 confidence level is $\bar{X}_\infty \pm 1.96 \times 3.71$, i.e. $\bar{X}_\infty \pm 7$.

(C) THE INDEX OF RELIABILITY

The correlation between a set of obtained scores and

---

1 Garrett, H. E., Op.Cit., pp. 347-348.

their corresponding true counterparts is given by the formula:[1]

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

Where $r_{1\infty}$ = the correlation between obtained and true scores,

r_{11} = the reliability coefficient of the test.

The coefficient $r_{1\infty}$ is called the index of reliability.

The index of reliability for the present test is:

$$r_{1\infty} = \sqrt{0.8}$$
$$= 0.894$$
$$\therefore \quad r_{1\infty} = 0.894$$

Thus, 0.894 is the maximum correlation which the test is capable of yielding in its present form.

ESTIMATION OF THE RELIABILITY OF EACH SUB-TEST

The reliability coefficient of each sub-test included in the present aptitude-test battery was also found out by applying split-half method. Test data obtained from the 'reliability sample', which was selected for applying the split-half method to estimate the reliability of the whole test, were used in all the five sub-tests here. The reliability coefficient

---

1 Garrett, H. E., Op.Cit., p. 349.

obtained for each sub-test is shown in the table below:

TABLE NO. 54

SHOWING THE RELIABILITY COEFFICIENT OF EACH SUB-TEST ALONG WITH ITS PROBABLE ERROR AND THE INDEX OF RELIABILITY

Sub-Test	Reliability coefficient	P.E.$_r$	$r_{1\infty}$
I	0.588	± 0.031	0.767
II	0.752	± 0.021	0.867
III	0.467	± 0.037	0.683
IV	0.325	± 0.043	0.570
V	0.432	± 0.039	0.657

The reliability coefficient as estimated from each unit sample data for the whole test scores is shown in the following table. The split-half method was applied to estimate reliability coefficient for each unit.

TABLE NO. 55

SHOWING r_{11} ALONG WITH ITS P.E.r AND $r_{1\infty}$ FOR THE WHOLE TEST WITH RESPECT TO EACH UNIT SAMPLE

Unit sample	N	Reliability coefficient	P.E.r	$r_{1\infty}$
Unit A	100	0.673	± 0.037	0.820
Unit B	74	0.673	± 0.043	0.820
Unit C	66	0.706	± 0.042	0.840
Unit D	94	0.720	± 0.033	0.848
Unit E	78	0.736	± 0.035	0.858
Unit F	118	0.795	± 0.023	0.892

IS THE RELIABILITY COEFFICIENT SATISFACTORY

There is no set formula to test the significance of the obtained reliability coefficient so that one can say immediately that it is satisfactory. But one can satisfy one's self by comparing the reliability of the present test with that of some other tests known to measure aptitude for teaching.

There are a number of tests measuring specific aptitudes. A detailed information of each test is also available. But there are very few tests which measure aptitudes for teaching. The investigator could not get detailed information even for these few tests. It is not possible, therefore, to compare the reliability of this test with that of any other test

measuring aptitude for teaching.

But Froehlich and Hoyt[1] suggest,

> As an arbitrary guide, the guidance worker should buy a test only if its reliability coefficient is 0.80 or higher. He can readily do this because most of the standardised tests on the market have coefficients which are in this range.

If one is to accept this suggestion, one can say that the reliability coefficient of the present test is satisfactory.

---

1 Froehlich, C. P. and Hoyt, K. B., "Guidance Testing", Science Research Associates, Inc., Chicago, 1959, p. 83.