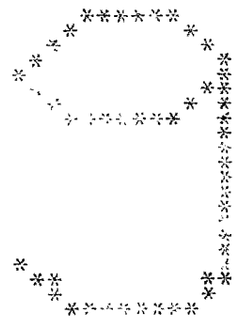
CHAPTER  VALIDITY AND RELIABILITY

Having evolved a composite index of creativity (the CCI) consisting of eight creativity measures, which were selected on the basis of their significant relationship (as described in Chapter 8) with criterion measures, next step in the efforts of the investigator would be to give correct estimates of validity and reliability of the said index. This has been described in this chapter. A brief review of the established statistical methods, conventions and procedures in estimating validity and reliability of the test battery (or the CCI) has been done before presenting the data and results.

9.1 Test Validity:

A psychological test is valid to the degree to which it measures a trait or an aspect of human behaviour and reliable to the extent it gives consistent results. The consistency may be in terms of time to time or alternate measurements.

Validity is the proportion that is common factor variance of the total variance of a measure.

$$\text{Thus Validity} = \frac{V_{co}}{V_t}$$

In certain respects, the definition given above seems to be more specific than telling in a traditional way that a test is valid if it measures what it is supposed to measure. Guilford (113, p. 461) recognises that the second definition is but one step better than saying a test is valid if it measures the truth.

Approaches to establishing validity seem to be from three directions, often recognised as 'type of validity'; they are:-

1. Content Validity
2. Criterion - related Validity
3. Construct and Factorial Validity.

In an article titled -- 'Do Creativity tests really measure Creativity?' Yamamoto (293) wrote - 'In general', three aims of test usage are assessment, prediction and explanation.' These three aspects of validity, are respectively content validity, criterion related validity, and construct validity.

9.2 Content Validity: "The content validity of a set of operations refers to the degree to which those operations measure the traits which we wish to measure as judged from the characteristics or content of those operations". (Ghiselli, 100, p.341). The judgement is subjective or 'professional'.

There are, as Ghiselli puts it, "two judgements involved: the extent to which the entire set of elements or items represents all aspects of the trait" (p. 342).

On the whole, it can be said that content validity is mainly a matter dependent upon the aspects of the test maker, his knowledge about the traits he wants to identify the amount of literature and tests available to him at the time of making the test and the skill with which he synthesises these in his tests.

Much has been said about the nature of the items and the nature of the hypothesised factor scores earlier while giving a rationale and developing scoring hypotheses. Hence there is little doubt about the content validity of the items

as well as the entire set of traits that they are supposed to measure. That each of the hypothesised factor scores contribute to the criterion measure is another point worth considering.

Another point needs to be mentioned in this context. Whenever the investigator used to meet the heads of schools, he used to give a description of salient features of creativity in general. And then he used to present his tests to them and ask whether the tests measure something different from school-tests, whether the tests really call for creative operations etc. Answers always used to be similar and overwhelmingly positive.

Anastasi's (4) opinion is more or less similar to that of Ghiselli (100). Content validity has also been called as definitional, logical or sampling validity by Yamamoto (293).

9.3 Criterion - Related - Validity:

Criterion related validity always refers to the communality between the test and some external criterion measures, such as supervisory ratings, job requirements, measures in a future performance. Under item-validity it has been stated that a test which contains items having significant correlation with criterion measures, on the whole will be valid. While constructing if care is taken to see that items are valid with respect to a criterion, the whole test constituted by such items will be valid.

Nunnally (1967 pp. 245-47) summarises Criterion - oriented approach to test construction in the following steps ---

" 1. Compose a large group of items.

2. Administer them to a large sample of individuals in the situation where the test will be used.

3. Correlate each item with the criterion.

4. Fashion a test out of those items that correlate most highly with the criterion.

When items have low correlations with one another and each correlates positively with the criterion, each item adds information to that provided by the other items; and when scores are summed over items, a relatively high correlation with the criterion will be found . . . " further he has recognised the advantage of developing hypothesis about the whole test-that might be predictive of a criterion.

Criterion - Related Validity "is demonstrated by comparing the test scores with one or more external variables considered to provide a direct measure of the characteristic of behaviour in question" (293).[1]

---

[1] Joint Committee on Test Standards of the American Psychological Association and National Council on Measurement in Education 'Standards for Educational and Psychological Tests and Manuals'. Washington, Amer. Psy. Assoc., 1966, p. 13. As cited in Yamamoto, (293).

Dealing under 'Predictive Validity' Ghiselli (100, p. 338) writes that it describes the accuracy with which we can estimate from the extent to which an individual manifests or possesses one property now the extent to which some other property will be manifested or possessed by him in the future, is now manifested or possessed by him, or was manifested of possessed by him in the past. All these three types of predictive situations involve the relationship between predictor and criterion scores. They differ only in terms of the time of occurrence of the criterion scores..."

Ghiselli (100) terms the second type as concurrent validity for example, proficiency in repairing automobiles at a time would be predicted by paper-pencil test of knowledge of automotive mechanics and repair. Such a test will be known to have concurrent validity.

Validity of the Present Test Battery:

(i) Validity of the Composite Creativity Index as decided by the Teacher Ratings as criterion:

Using the formula 4.50 given in Guilford (113) for the correlation of sums with a criterion, validity coefficient of the standard scores, sum of the hypothesised factor scores or composite creativity index has been calculated

against Teacher's Ratings as Criterion. Again help of the relevant intercorrelation matrix has been taken.

Hence we have -

$$r_{pI} = \frac{\Sigma r_{pi} \sigma_i}{\sqrt{\Sigma \sigma_i^2 + 2 \Sigma r_{ij} \sigma_i \sigma_j}}$$

where p = teacher ratings

i = Any variable from A . . . . . K.

j = Another variable from A . . . K.

$\Sigma r_{ij} \sigma_i \sigma_j$ = represents coveriance of all possible pairs of components from A to K.

Substituting relevant values from correlation matrix (Table 11) we get

$$r_{pI} = .332$$

.332 is significant at $p < .01$ for N=230. We can see that the composite index is highly valid.

(Note that as previously stated, Means and S.ds of all the eight individual measures are around 50 and 10 respectively. These have not been reported to avoid confusion. This is true for all the computation that succeeds wherever individual S.d's have been employed. And combined S.D's as required in the formula has been completed and used wherever necessary).

(ii) Validity of the composite creativity index as decided by circles test as criterion:

To calculate validity coefficient against Torrance's circles test as a criterion, same formula (ref. above) has been used. First correlation between composite creativity index and each of the Torrance's circle's test scores has been calculated. Finally using these correlation coefficients and the relevant intercorrelations among Torrance's circles factor scores, correlation between composite creativity index with a composite of Torrance's circle test scores has been calculated.

$r_{LI} = .146$ $r_{FI} = .183$ $r_{NI} = .093$ $r_{OI} = .190$

Where i = variables from A to K.

Finally

$r_{Ic} = .178$

where C = L, F, N and O.

.178 is significant at $p < .01$ for N=230.

As a matter of interest, correlation between composite of circles test scores with Teacher Ratings has been calculated using the same formula (ref. above).

$r_{pc} = .165$

.165 is significant at $p' < .05$ for N=230 and not at $p < .01$. As we already know correlation

of CCT with Teacher's Ratings as .332, we can say that it is better than the index given by Circles Test. And CCT has as much correlation (.178) with TCT as TCT has with Teacher Ratings (.165).

However one difference between the TCT measure and CCT should be made clear, the C.C.I. is a composite of eight measures derived from four tests. TCT - index is composite of four measures derived from single test. In Torrance's non-verbal form there are three different tests one of which is circles. Hence this is suggestive of the possibility that all three test measures (of TT CT) combined together would give a correlationship with Teachers Ratings say, relatively equal to that of CCT index with Teacher Ratings (.332).

(iii) Validity of Composite Creativity Index with a combination of TCT measures and Teacher Ratings:

It is informative to find out the correlation between composite creativity index with a composite of TCT measures and Teacher Ratings. Same formula (ref. above) has been used.

$$r_{I(c+p)} = .252$$

where C = is a combination of 4 measures L, M, N and O.

P = Teacher Ratings.

I = Combination of eight measures or C.C.I.

(iv) Validation by Domination Technique:

At this stage, along with Torrance's Circle Test and Teacher Ratings, Activities Checklist scores have been taken into account. Scoring activities checklist has been described earlier. From the 230 Ss under study, two groups - high creatives and low creatives - have been identified in the following way:

(a) For an S to be included in the high-creativity group, he or she should stand among the first 60 from the top in atleast two of the three criteria.

(b) For an S to be included in the low creative group he should stand among the first 60 chosen from below in atleast two of the three criteria.

A biserial correlation coefficient as given by Peters formula (Guilford: 112, p. 408-9) has been calculated. The formula requires the statistics given in Table 12 about the CCI, for the low, middle and high groups.

TABLE 12

Strength of the criterion groups and relevant statistics required by Peters formula

	No. of Ss in the group (n)	$p = \frac{n}{N}$	Y Ordinate	Mean (M)	
Low	62	.270	.3306	374.68	
Middle	117	(Not necessary)		402.51	45.817
High	51	.220	.2961	426.22	

As the statistic r_b using the said formula is based on the assumption of normality of the distribution of composite index a useful check for the assumption would be to see whether the mean of the middle group approximately divides the difference between means of high and the low groups at the midpoint. (Guilford: 112, p.408.9). It is seen from the means presented above that the condition has been fulfilled.

Thus,

$$r_b = \frac{(M_h - M_l)\, p_h p_l}{(p_l Y_h + p_h Y_l)\sigma_i}$$

where subscripts h and l stand for high and low groups.

Substituting the values as required we get,

$$r_b = .438.$$

A formula for standard error of the above statistic has also been provided (112)

$$\sigma_{r_b} = \frac{\sqrt{p_l p_h}}{(p_l Y_h + p_h Y_l)\sqrt{N}} \sqrt{p_h + p_l}$$

$$= .074.$$

True correlation (Pearson r) which would fall within $\pm$ 2.58 x .074 of .438 still exceeded the limits at $p < .01$ level for Pearson rand hence is highly significant.

The evidences provided hitherto impels one to accept that the composite creativity index of CPM I - IV is highly valid.

9.3 Construct Validity:

Yamamoto (293) quoting the Joint Committee (1) regarded criterion related validity, 'unless it is established in the context of same theory' as yielding 'no information about why correlation is high or low or about how one might improve the measures'. Hence he regarded construct Validity as 'the touch stone of scientific success'.

Following Cronbach and Meehl (48), Yamamoto (295) set forth the following steps in establishing construct validity.

' (1) Setting forth the proposition that a particular test measures creativity, (2) inserting this proposition to some theoretical formulations about creativity (3) deriving hypotheses concerning the behaviour characteristics correlated with test scores and also those which should show no relation to test scores if the test truly measures creativity. (4) collecting data to confirm or reject hypotheses'.

If the hypothesised relationships stand, it adds to the validity of constructs and measure. Otherwise, a failure may indicate one of these (1) inadequacy of the measure (2) inadequacy of the construct (3) inadequacy of the design.

Ghiselli (100, p. 350) writes -

> though largely based upon objective and quantitative data, construct validity is determined and evaluated by a subjective process of judgement; and the degree of validity cannot be expressed by any single quantitative index such as a validity coefficient but must be given in verbal terms.

Guilford has done much in establishing construct validity of creativity tests. It has often been pointed

out in the course of this thesis that constructs used here have been directly derived from the work of years of research by Guilford Group and there is much to feel indebted as it is so in the case of other researchers in the field too, to its proponent leader Guilford, for providing a systematic theoretical foundation.

Dr. Guilford writes (117, p.262) -

> Much has been said about the worrisome criterion problem, for example, I might say, incidentally, that our way of using factor analysis has one great advantage in that we do not have to worry about the criterion problem. Factor analysis provides its own criteria. You may not agree with me on this but I am quite ready to defend the idea that the best kind of construct validity is found through factor analytic approach . . . .

Cattell (1969) once compared the tool of factor analysis in psychological research to 'microscope' in biological research.

Commenting on the philosophical aspect of factor - constructs Baggley (12) writes:

> All these constructs represent man-made attempts to bring order into the natural phenomena observed by behavioural scientists. Their justification is ultimately pragmatic, do they or do they not aid in the development of scientific laws? Factors that represent effects of training and experience may be just as useful as factors representing genetic, endocrinological or central nervous system effects.

What types of variables are to be subjected to factor analysis? Anastasi opines that

> the set of variables analysed can, of course, include test and non-test data. Ratings and other criterion measures can thus be utilised, along with other tests, to explore the factorial validity of a particular test and to define the common traits it measures (p.148).

> Factor analysis as applied to psychological data is an act that is some what dependent on the skill and intention of the investigator. It is not a series of precisely defined procedures yielding a rigorously objective result.

In line with the opinions expressed by the writers (Yamamoto, Guilford, Anastasi and others) quoted above a factor analysis of the intercorrelation matrix given in the previous chapter has been undertaken. Eventhough validity of the composite creativity index has been established to a reasonable extent, it is felt that all variables considered previous to selection and rejection of scores on the basis of speculation, should be retained for factor analysis. Hence the matrix reported in Chapter 8 has been taken as it is with, however, correlation coefficients rounded upto two decimal places only. As the purpose of factorising is more interpretative than predictive, it is felt that taking coefficients upto two places is sufficient.

Of the various methods of factor analysis, Centroid method (Thurstone, 1947) has been selected for the purpose. The investigator has tried to depend upon opinion of experienced researchers in the field (Guilford, 112 p. 478) suggesting that centroid method would be suitable to the present investigation. Further, following considerations influenced selection of said method of factor analysis:

1. Previous to this, factors of creative ability have been identified by centroid method with sufficient number of replications; and constructs used in this study and the scoring methods, thus, already have considerable psychological standing.

2. For the first time large number of scores hypothesised for different factors of creativity have been derived using multiple scoring method and are being factor analysed. The purpose of factor analysis, here is more interpretative Mathematically more rigorous and more pragmatic and suitable methods can be employed as later research. Centroid method serves the purpose of interpretation.

Rationale underlying factor analysis has been given in detail in the books devoted to the subject (Thurstone: 155; Guilford: 112; Fruchter: 8; Harmon: 127).

The investigator has simply depended upon the opinion of experienced researchers, he has quoted already, in taking decision about the method. Rather his knowledge of mathematical basis of statistical tools he has used is scanty and insufficient.

Factor analysis of the sixteen variable intercorrelation matrix:

Extraction of factor: The sixteen variable matrix with intercorrelations upto two places has been factor analysed by centroid method. Communality estimates have been highest coefficients in the respective column or row of the matrix. Communality residuals have been reestimated after extraction of each successive factor. The principle of reflecting the column with highest negative total has been adopted. Method of computation has been exactly same as described by Fruchter (shown in Appendix B). In all eight factors (see Table 13) have been extracted. The following paragraph explains why the analysis was continued for eight factors.

Criteria for sufficient number of factors:

Number of factors extracted has been decided by three criteria: Tacker's ∅, Humplerey's Rule and Coomb's criterion.

Sample (N = 230) on which the factor analysis data are based, though not too large, is sufficient (N=200; see Guilford:112). Majority of the coefficients are positive. Remaining are low negative values. Hence the conditions for applying the latter two criteria are fulfilled. Table 14 summarises the statistics computed to decide whether all of the common factor - variance has been extracted from the correlation matrix. Tucker's phi reaches a level just near and below the level required by the criterion for the eighth factor. Product of the two highest loadings, as required by Humplerey's Rule, falls below twice the standard error for the fourth factor. More decisive is Coomb's criterion which shows that the number of negatives after reflection for the eighth factor falls within the value C (as obtained from Fruchter 87, p. 85) as required by Coombs (45). Thus two of the three criteria suggest that at the eighth factor necessary common factor variance has been extracted. Residuals in the eighth residual correlation matrix (see Appendix E) range between .000 and .045. Eighth factor loadings for all variables are within $\pm$ .20. Over-determination of number of factors and elimination while rotatig is our established practice (Fruchter). Thus all the eight factors have been retained for rotation.

TABLE 13

Factor Matrix

Test Score (Variable)	I	II	III	IV	V	VI	VII	VIII	h^2	$1-h^2$
A	68	49	-08	-30	31	-07	-07	-08	91	09
B	66	41	08	-24	36	-14	-21	-10	87	13
C	53	34	07	25	11	-24	05	15	56	44
D	55	39	24	33	09	22	15	08	71	29
E	62	55	-21	-25	-33	08	14	-08	93	07
F	65	55	-22	-25	-29	07	13	-01	94	06
G	55	31	-07	13	-15	-27	-04	16	54	46
H	56	36	18	30	-06	33	-13	-17	72	28
I	38	-33	21	-25	13	11	29	-12	49	51
J	37	-43	56	-33	-24	06	-09	15	84	16
K	27	-43	55	-29	-29	-03	-20	10	78	22
L	50	-58	-48	03	19	05	-05	-08	87	13
M	53	-62	-47	-02	17	12	08	12	95	05
N	32	-35	-34	39	-06	-12	02	02	51	49
O	48	-50	-38	29	15	18	18	-05	80	20
P	33	-18	30	21	-05	-14	19	-19	37	63
a_j^2	4.23	3.11	1.67	1.09	.72	.42	.34	.21	11.79	4.21
% Var	26.44	19.44	10.44	6.81	4.50	2.62	2.13	1.31	73.69	26.31

Note: Decimal points omitted.

TABLE 14

Criteria for sufficient number of factors

Factor	Tucker's Phi	Humphrey's Rule	Coombs' Criterion
I	.746	.4488	32
II	.534	.3596	32
III	.709	.3080	70
IV	.636	.1287	53
V	.594	.1188	56
VI	.816	.0891	76
VII	.676	.0609	55
VIII	.836	.0323	88
Criterion Value	.882	.132	92 ± 8

The Factor Matrix: Factor matrix with original loadings and recalculated communalities and percentage variance explained by each factor has been given on Table 13.

Checks: Correlation coefficients in the cells just above the principal diagonal of the matrix and the one of right corner have been tallied with those computed using factor loadings and residuals. This checks to the accuracy of computations (Table 15). Excepting the fact that addition and multiplication machines have been employed, computation has been done by hand by author himself.

Rotation of Axes: Graphic orthogonal rotations have been done to meet the criterion of simple structure, positive manifold and psychological meaningfulness. In all fourteen rotations have been done. First few rotations have been done to meet the requirements of simple structure and positive manifold. Afterwards rotations have been aimed to achieve maximum possible separation and psychological meaningfulness. Rotational Graphs have been presented in the Appendix B. The final rotated factor matrix has been presented in Table 16. The table also contains communality obtained before and after rotation and percentage variance explained by each factor. As suggested by Fruchter, correlation coefficients for the cells above the diagonal and right corner computed from rotated loadings have been compared with original coefficients to check the accuracy of rotation. Table 17 presents this comparison.

TABLE 15

Check Values

Variables	Sum of cross products	Residual Correlation after the last factor	Total sign changes for the two tests	Sum of Cross products +residual with adjusted sign	Correlation from original Correlation Matrix
A,B	.8594	-.005	Even	.8544	.85
B,C	.4825	-.002	Odd	.4845	.48
C,D	.5000	.013	Even	.5130	.52
D,E	.4251	.006	Odd	.4191	.42
E,F	.9325	-.010	Even	.9245	.92
F,G	.5287	-.004	Odd	.5327	.54
G,H	.3439	-.004	Odd	.3479	.35
H,I	-.0046	-.025	Even	-.0290	-.03
I,J	.4007	-.007	Odd	.4077	.41
J,K	.7893	.010	Even	.7993	.79
K,L	.0745	-.025	Odd	.0995	.10
L,M	.8755	-.001	Odd	.8745	.87
M,N	.5180	-.011	Even	.5070	.50
N,O	.5357	-.014	Odd	.5497	.55
O,P	-.1379	-.007	Even	.1309	.13
P,A	.0454	-.025	Even	.0204	.02

TABLE 16

Rotated Factor Matrix

Test scores (Variables)	I	II	III	IV	V	VI	VII	VIII	h^2	$1-h^2$	h^2 (from original loadings)
A	87	-16	-12	21	18	16	09	-05	91	09	91
B	85	-02	-21	27	13	12	-07	-07	88	12	87
C	34	-05	-15	57	11	25	08	18	58	42	56
D	27	06	-13	34	26	65	06	15	73	27	71
E	58	-10	-05	12	11	28	69	-12	94	06	93
F	60	-11	-04	15	09	28	69	-06	96	04	94
G	35	02	-09	53	-05	14	34	12	56	44	54
H	33	19	-06	25	04	70	09	-18	74	26	72
I	24	31	18	06	52	-21	-04	03	51	49	49
J	19	84	00	-12	19	-12	06	21	85	15	84
K	11	83	-07	-09	08	-20	03	12	78	22	78
L	33	10	84	17	02	-15	-11	-03	89	11	87
M	28	11	88	14	06	-12	-05	19	94	06	95
N	-07	06	54	46	-06	01	02	02	52	48	51
O	19	16	76	29	-13	14	-20	-03	80	20	80
P	-03	33	04	35	36	07	-04	-09	38	62	37
a_j^2	2.93	1.74	2.50	1.43	.62	1.35	1.16	.24	11.97	4.03	
% Var	18.31	10.88	15.63	8.94	3.88	8.44	7.25	1.50	74.83	25.17	

Note: Decimal points omitted.

TABLE 17

Check Values based on rotated factors

Test scores (Variables)	Sum of cross products	Residual correlation after the last factor	Total sign changes	Sum of the cross products +residual with adjusted sign	Correlation from original Matrix
A,B	.8644	-.005	Even	.86	.85
B,C	.5015	-.002	Odd	.50	.48
C,D	.5250	.013	Even	.53	.52
D,E	.4319	.006	Odd	.43	.42
E,F	.9506	-.010	Even	.94	.92
F,G	.5481	-.004	Odd	.55	.54
G,H	.3622	-.004	Odd	.37	.35
H,I	.0071	-.025	Even	-.02	-.03
I,J	.4279	-.007	Odd	.43	.41
J,K	.7951	.010	Even	.80	.79
K,L	.0699	-.025	Odd	.10	.10
L,M	.8794	-.001	Odd	.88	.87
M,N	.5246	-.011	Even	.51	.50
N,O	.5447	-.014	Odd	.56	.55
O,P	.1527	-.007	Even	.14	.13
P,A	.0663	-.025	Even	.04	.02

Slight differences have been tolerated.

Interpretation of factors:

In the following paragraphs, the rotated factors will be treated in turn. The hypothesised factor - scores having significant loadings are listed under each factor. A loading of .30 or more has been taken as significant in practice. If there are also significant loadings on factors other than the one under consideration, such loadings are mentioned inparentheses. Interpretation is based mostly on the concurrence of two similarly hypothesised scores, nature of the scores as understood from established ways of scoring procedures, content and possible reasoning on the loadings on other factors not under consideration. In mentioning scores name of the test from which it is derived and the alphabetical letters already used will be given.

Factor I

Test Scores	Loadings
CRM I A	.87
CRM I B	.85
CRM I C	.34 (.57 IV)
CRM II E	.58 (.69 VII)
CRM II F	.60 (.69 VII)
CRM II G	.35 (.53 IV; .34 VII)
CRM II H	.33 (.70 VI)
TCT L	.33 (.84 III)

It is obvious from the high loadings of scores A, B, E and F that the factor stands for fluency. The earlier guess that B does not differ from A has become true. Scores hypothesised for originality have just significant loadings on this factor. This is clear from the method of derivation of scores. Fluency score is parent for originality score too. In one of the studies (quoted earlier) such a relationship between originality and fluency scores derived from the same responses has been anticipated and has been attributed to 'method' variance. The present case is no exception to it. Further that TCT fluency score (L) has significant loading on this factor strengthens our naming of this factor as "fluency". Low (but significant) loading of fluency score L on this factor seems reasonable when we consider the restriction on 'S' in responding to 'circles'. He has to draw. Nature of score derived from CRM II seems to be a matter for discussion in that the score has high loadings on factor VII possibly referring to different 'content' of the responses.

As viewed from the fact that scores derived from symbolic tests (I, J and K) and scores derived from circles Test (M, N and O) have loadings which are not significant, the factor can be further termed to be 'ideational'.

Even 'H', a hypothesised elaboration score has low significant loading on this factor. This again can be accounted by a 'method' variance. That in other factors we do not find such significant loadings on different sets of scores derived from the same set of responses indicates that sufficient method variance has gone with 'fluency' score. It is meaningful to note that fluency score refers to all suitable responses and includes those used either for originality or elaboration.

Factor II

Test Scores	Loadings
CRM III - I	.31 (.52 V)
CRM IV - J	.84
CRM IV - K	.83
Teacher's Ratings - P	.33 (.35 IV .36 V)

Score J does not seem to differ from K even though they were derived from different methods, of course, from the same set of responses. There is a great deal of difference between the method of derivation of J and K and the method of derivation of A and C. In the latter case C (originality score) forms part of A (fluency score) and has simply a reduced range. J and K differ from each other in the sense that J refers to

number of different sets of relations (or arithmetic rules). Each provides different information about the same factor by virtue of the method by which they are derived. Besides each contributesto the criterion (loading on P is .33); their retention in the composite index seems to be justified. In contrast with this, there is no direct evidence that E which was rejected contributed to the criterion. Further A and E both referred to number of verbal responses. Hence preference for A which had significant correlation with P and L is justified. Thus J and K might be two different scores standing for the same factor, 'possibly' symbolic elaboration'.

Identification of factor II as symbolic elaboration has been done with some skepticism owing to the lack of other marker tests. It is definitely symbolic and refers to other than 'fluency for symbolic relations (I)' which can be identified with factor V. Further J and K both involve manipulation of numbers and relations resulting in 'implications'. This operation characterises Guilford's symbolic relations tests. This point has been made clear while framing the hypothesis. That teachers do generally regard a boy intellectually high if he achieves better in mathematics' tells us why significantly loaded these

symbolic test scores are with teacher ratings (P).

Factor III

Circles L	.84 (.33 I)
Circles M	.88
Circles N	.54 (.46 IV)
Circles O	.76

On the basis of the loadings on factor III nothing can be speculated. The reason is that all the four scores derived from cricles stand high on that factor. Scores L and N do contribute to ideational fluency (factor I) and originality (factor IV) as hypothesised. In the absence of tests from figural content category, very little can be guessed on the nature of this factor. The correlationship that circles test has with P in explained by originality factor (IV) having significant loadings on P and N.

Factor IV

CRM I - C	.57 (.34 I)
CRM I - D	.34 (.65 VI)
CRM II - G	.53 (.34 VII)
Circles - H_N	.46 (.54 III)
Teacher Ratings P	.35 (.33 II .36 V).

This factor is definitely 'originality' factor. Three scores C, & and N which have been hypothesised for originality have high loadings on this factor. Besides, Teacher Ratings has significant loadings on this factor. Teachers seemed to regard originality. The only interference is by D, supposed to be an elaboration score, which has low loading on this factor. This is reasonable. Sufficient variance of D will be explained by factor VI which should be regarded as elaboration factor. As elaboration too consists of supplementary ideas, a loading of .34 on originality is reasonable. Hence factor IV is definitely originality factor.

Factor V

CRM III - I	.52 (.31 II)
Teacher Ratings (P)	.36 (.33 II, .35 IV)

In Guilford's analysis a test known as alternate additions was designed to measure factor 'Divergent Production of symbolic Relations'. Though in the three analysis (Guilford and Hoepfner: 120, Guilford et al: 118) the test was a leading representative of DSR, However, in one analysis it went with Numerical facility (Gershon et al: 1963). Gershon et al (93) on the basis of results, stressed the need for new and better DSR tests. Guilford (123) stressed the need for appropriate letter tests of DSR. In CRM II the suggestions have been

met with. Instead of numbers individuals are given letters. Obviously numerical facility variance has been reduced by introducing letters. Secondly CRM IV which has been recognised as # symbolic implications might require memory for symbolic implications (MSI). In harmony with three analyses which brought CRM III type test as a DSR representative, it is speculated that CRM III - I is a DSR score. In the absence of marker tests it is difficult to account how much numerical facility variance is carried by CRM IV test scores (Factor II).

Factor VI

CRM I - D	.65 (.34 IV)
CRM II - H	.70 (.33 I)

Factor VI is definitely elaboration factor. Here two scores from two different tests of similar nature have gone together and stand differently from fluency and Originality scores of their respective parent test. D has low loading on originality and H has low loading on fluency. This might be due to the nature of elaboration being a set of ideas used to describe a central idea. Such low loadings either on originality or on fluency can thus be regarded as within resonable limits.

Factor VII

CRM II - E	.69 (.58 I)
CRM II - F	.69 (.60 I)
CRM II - G	.34 (.35 I .53 IV)

This factor is difficult to interpret. Even after conceding the failure of flexibility scores B and F, that E has sufficiently high loadings on another factor besides ideational fluency (factor I) suggests the need for marker tests different in content. On the basis of available information any speculation on the nature of this factor is out of question. Factor VIII consists of loadings mostly between+.20 and -.20.

In the interpretation of factors just done, there was been no factor which could be identified with teacher ratings. That is to say variance accounted by teacher ratings has been shared by those factors which have been regarded as factors contributing to creativity. This is reasonable because ratings as taken through CPCRS, simply identify characteristics important to creativity and do not enquire into general intellectual status of children.

Notably enough majority of the variance accounted by non-verbal-test scores (circles) has stood seperately. The semantic aspect of performance in response to circles

has been accounted by factor I (ideational fluency) and factor IV (originality).

In summary, we shall look to the construct validity of the individual scores of the Composite creativity index. A concurs with L and $\overset{E}{R}$ to suggest that it stands for ideational fluency. C and G concur with N and P to suggest that they stand for originality. D and N concur to tell us that they are elaboration scores. I, J and K, which are undoubtedly DP factors of symbolic content category have been significantly accounted by teacher ratings. Thus scores chosen as components of CCI have sufficient construct validity.

9.4 Factor Validity:

When considered factor analytically each score in C.C.I. has support of atleast one criterion measure. There is no single score in the C.C.I. which stands independent of criterion measures.

Loadings of the scores A, C, D, G, H, I, J and K on their respective factors range between .52 and .87. Thus validity of each scoring procedure in relation to factor which it is supposed to represent is sufficient and high. The test battery can be regarded as having factorial validity.

9.5 Reliability:

Test theory is centred around the assumption that a test score is a combination of true score and random error component. "By definition, random error is completely uncorrelated with the true score in the test and with random error in any set of measurement" (60; p. 388). It is possible to show total variance (V_x) as the sum of the true variance (V_a) and the error variance (V_e).

$$\text{i.e., } V_x = V_a + V_e$$

Or dividing by V_x we get,

$$\frac{V_a}{V_x} + \frac{V_e}{V_x} = 1.$$

in terms of proportions of variances.

Test Reliability: Considering that the two tests are equal in that time parts of each test measure exactly the same function and are perfectly correlated and contain a certain proportion of random error, it is possible to show that the reliability of self correlation of the test (Dubois: 60).

$$r_{xx'} = \frac{xx'}{NS_x S_{x'}} = \frac{V_a}{V_x} = 1 - \frac{V_c}{V_x}$$

where x and x' stand for deviation and S_x and $S_{x'}$ for standard deviations. Therefore, the reliability of a test, $r_{xx'}$, is $\frac{V_a}{V_x}$, the proportion of the total variance, or $1 - \frac{V_e}{V_x}$ that is one less the proportion of error variance (Dubois: 60).

Reliability can be estimated by four different methods:

1. Test-Retest Method.
2. 'Split-Half' Method.
3. Alternate Forms.
4. Rational Equivalence.

Test Retest Method: In this method a group of individuals is given the same test on two different occasions. Product - moment r between scores obtained by the group in the two occasions is used as the estimate of reliability. A test-retest r indicates the stability (Guilford: 112) with which the test identifies the individuals in different occasions.

Split-half Method: In this method the test is divided into two halves which are judged to be equivalent. Equating the two halves is some times done on the basis of item difficulties or sometimes by odd-even grouping of items. The scores of the two halves correlated, the product moment $r_{aa'}$ gives estimated reliability of one half of the test. Theoretically assuming the split halves are exactly equal it can be proved that

$$r_{xx'} = \frac{2r_{aa'}}{1 + r_{aa'}}$$

Thus reliability of a variable when doubled in length may be estimated by the reliability of the half ($r_{aa'}$) test. The above equation is a particular case of Spearman - Brown's Prophecy formula. Split-half reliability indicates how for a test is internally consistent (Guilford: 112).

Alternate forms:

In this method two alternative forms considered to be equivalent are administered to the same group and the results are correlated in order to get an estimate of reliability. Alternate forms reliability stands in between as an indicator for stability as well as internal consistency (Guilford: 112).

Estimating reliability by test retest or by alternate forms is applicable even to 'speeded tests'. Split - half method is not applicable to speeded tests. In case, administration of a test twice either by test-retest or using alternate forms is not possible, Anastasi (4) suggests a split-half method based on seperately timed parts or items grouped, into two halves, for estimating reliability of speeded tests or of those tests where effect of speed is present to a considerable extent.

The method of Rational Equivalence:

In this method estimate of reliability is the correlation

between the test with its hypothetical equivalent. "It assumes that variance of the existing test is identical with the variance of the hypothetical test, and that the sum of the item covariances within the existing test is proportional to the sum of the between test item - covariances". (Dubois: 60, p.394). Kuder - Richardson formula K.R. - 20 will be utilised for the purpose.

One form of Kuder - Richardson formula - 20 which uses - n, the number of items, ΣV_i the sum of the items' variances, and V_x, the total variance - is as follows:

$$V_{xx'} = \frac{n}{n-1}\left(1 - \frac{\Sigma V_i}{V_x}\right)$$

(Dubois: 60, p.397).

"Its use is not restricted to tests composed of dichotomous items; rather, the scoring system applied to the items may have any range as long as the total score is the simple sum of the item scores" (Dubois: 60, p.397). If applied to heterogeneous test, developed to predict external criterion, the resultant coefficient got using KR - 20 would be an underestimation of the true reliability (Dubois: 60, p. 396).

9.6 Reliability of the Present Test Battery:

As seen from the correlation matrix presented in the previous section, if highest correlation coefficients (either from a column or row) are to be regarded as communality estimates, they are sufficiently high. The lowest is .41 for score I and highest is .85 for score A. Communalities can be taken as lower-bound estimates of reliability of tests. This indicates that actual reliability of C.C.I. (made of eight scores) would be sufficiently high.

Various methods of finding reliability has been briefly given above. Three types of reliability have been estimated for the C.C.I. They are:-

1. Split-half version of Alternate forms.
2. Kuder Richardson - Reliability.
3. Inter-Scorer Reliability.

Split - half - cum - Alternate Forms Reliability:

A split half version of alternate forms technique has been suggested by Anastasi (4). For this division of the test into two halves should be on the basis of time rather than in terms of items. That is, the half-scores must be based on seperately timed parts of the test.

"If it is not feasible to administer the two half-tests separately, an alternative procedure is to divide the test to find a score each of the four quarters . . . . . " Scores from the "first and the fourth quarters can then be combined to represent one half-score, while those in the second and third quarters can be combined to yield the other half score. Such a combination of quarters tends to balance out the cumulative effects of practice fatigue and other factors. This method is especially satisfactory when the items are not steeply graded in difficulty level . . ."

In the case of CRM, items have been seperately timed. Hence reliability has been calculated using the technique just described.

Procedure: On the whole CRM I - IV contains 12 items. Each item has been seperately timed. Divided into two it yields halves of 6 items each. Each half gets 2 items each from CRM I and CRM II and 1 item each from CRM III and CRM IV. Table 18 gives how the division has been effected.

As scoring has been doneitemwise it is easy to total up for the items in the first and second halves seperately and get the scores in a manner similar to the scores from the whole test. Thus scores A, C, D from CRM I, G, H from CRM II, I from CRM III, and J and K from CRM IV have been

TABLE 18

Division of CRM into two halves

Test	Items*	I or II half
CRM I	H_5	I
	F_2	II
	D_5	II
	E_1	I
CRM II	C_4	I
	A_8	II
	C_3	II
	A_5	I
CRM III	1	I
	2	II
CRM IV	1	II
	2	I

* Order of the items is according to their positions in the test.

seperately got for each of the halves. Method followed to get a composite from each of the halves is same as the method followed to get C.C.I. from the whole test. That is each score has been converted in to standard scores with a distribution of mean of 50 and S.D. of 10. Table 19 gives the means and s.ds for raw scores of the eight componentsof C.C.I. as got for each of halves the test battery.

The standard score sums or the C.C.Is got from each of the halves have been correlated using the raw score version of the difference formula. (Garret: 90).

Necessary data for the two halves as required for computation of r have been given below:

First Half	Second Half
Mean M_1 = 400.46	Mean M_2 = 399.28
Standard Deviation (σ_1) = 36.96	Standard Deviation (σ_2) = 46.39
ΣX_1^2 = 37199165	ΣX_2^2 = 37162644

$$d^2 = \Sigma(x_1 - x_2)^2 = 212030$$

$$r = .758.$$

Using the Spearman - Brown Prophecy formula reliability of the whole test has been calculated.

TABLE 19

Means and standard deviations of the 8 creativity measures (N = 230) of the CCI as given by two halves of the test battery

Test scores	FIRST HALF		SECOND HALF	
	Means	S.Ds	Means	S.Ds
A	10.1304	3.4028	8.8304	3.6363
C	1.6522	1.4085	1.9213	1.6994
D	11.7391	6.3621	10.9130	7.0268
G	2.1826	1.6097	1.4826	1.3483
H	5.6174	3.7121	7.0261	4.1263
I	2.7261	1.3430	3.4783	1.2209
J	1.9304	1.6763	1.4609	1.2231
K	1.2913	.9795	1.1000	.9034

$$r_{xx'} = \frac{2r_{aa'}}{1 + r_{aa'}}$$

$$= .862$$

$$= .86.$$

Thus reliability of the present test battery is high enough to conclude that composite creativity index, derived from it is stable and consistent. Both aspects of efficiency of a test are approximately given by alternate forms reliability (Guilford: 112).

The Index of Reliability: The index of reliability of a test $r_{1\infty}$ is the square root of the reliability of the whole test.

Thus $r_{1\infty} = \sqrt{r_{xx'}}$

$= \sqrt{.862}$

$= .928$

$= .93$ which is the correlation of the test battery with a similar test of infinite length.

Standard Error of Measurement: A standard error of measurement for the obtained score can be given

$$SE_{meas} = \sigma_{sc} = S_x \sqrt{1 - r_{xx'}}$$

We know $S_x = 45.82$ and $r_{xx'} = .86$.

Hence $SE_{meas} = \sigma_{sc} = 17.15$,

As is true, true scores are hypothetical but ideal measures. Standard error gives us possible limits of discrepancy between obtained score and true score.

The Method of Rational Equivalence:

Limitations of this method when applied to a test - score which is a composite of heterogeneous measures have been given earlier. As variance of individual measures and total (i.e. of C.C.I.) are known, it will be additional information, though an underestimation, if we calculate reliability using formula KR - 20.

By this method the reliability,

$$r_{xx'} = \frac{n}{n-1}\left(1 - \frac{\Sigma V_i}{V_x}\right)$$

where n = number of items

V_i = variance of individual items

V_x = variance of the total.

Regarding individual scores as derived from individual items, we have 8 items. However regardig s.d of each of the individual standard scores viz., A, C, D, G, H, I, J and K as 10 and s.d of their total as 46 (See page 208 for σ = 45.82) will be convenient as the computation becomes simple without the loss of accuracy.

Hence we have

$$r_{xx} = \frac{n}{n-1}\left(1 - \frac{\Sigma V_i}{V_x}\right)$$

$$= \frac{8}{8-1}\left(1 - \frac{8\sigma^2}{\sigma^2_x}\right)$$

$$= \frac{8}{7}\left(1 - \frac{8 \times 10^2}{46^2}\right)$$

$$= \frac{8}{7}\left(1 - \frac{800}{2116}\right)$$

$$= \frac{8}{7} \times \frac{1316}{2116} = .71$$

Considering the limitation laid upon the method of rational equivalence, the reliability coefficient of .71 is high enough and real value must be higher than .71. This suggests that the test battery under consideration has least discrepency with its hypothetical counterpart.

Inter Scorer Reliability: Inter scorer agreement is necessary for any scoring procedure to be objective. Fluency, originality and elaboration scores (viz., A, C, D, G and H) of CRM I and II secured to involve subjective decision of the scorer to a limited extent. Unless this is within tolerable limits or in other words, there is high agreement between two scores, composite index based on such scoring procedure cannot be regarded as reliable.

Procedure: 50 test booklets of CRM I and II were selected randomlyout of the 230 available ones. Scorer x was investigator himself. Scorer Y was a person doing his post-graduation in Engineering and had thorough knowledge of nature of the investigator's work. Besides he was given adequate knowledge of the scoring procedures by the investigator.

Table 20 gives the means and S.Ds of 5 scores (A, C, E, G, H) the group of 50 Ss as given by two scores X and Y. Each score has been transformed into standard scores of mean 50 and s.d. = 10 and a composite of the scores (by adding the standard scores) has been obtained. Thus the two sets of scores to be correlated are the composite as obtained for each of the two scores X and Y.

Product - Moment r has been computed using scatter - diagram method. Appendix F presents the scatter of Ss and relevant computational steps.

A coefficient of correlation .87 indicates very high agreement between two scores. As the scores I, J and K are objective, the agreement between X and Y is complete and r would be equal to 1.00. Hence if r is to be calculated

TABLE 20

Means and standard deviations of 5 creativity measures (N = 50) of CRM I and II as given by two scorers X and Y

Test scores	Scoring by Y		Scoring by X	
	Means	S.D.	Means	S.D.
A	18.42	7.75	17.70	6.69
C	1.96	1.79	3.30	2.30
D	14.02	8.75	18.84	9.43
G	2.36	1.07	3.34	2.03
H	8.16	5.19	10.76	5.07

on the basis of composite of all 8 scores, the correlation coefficient would be definitely higher than .87, the correlation between the composites based only on 5 scores. Hence it can be concluded that the scoring is highly objective.

Considering all the three indices of reliability viz., split-half cum-alternate forms, rational - equivalence and inter-scorer agreement, it is concluded that the composite creativity index is a highly dependable measure. In other words, the measure is stable, trustworthy and objective.