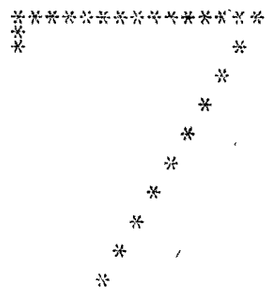
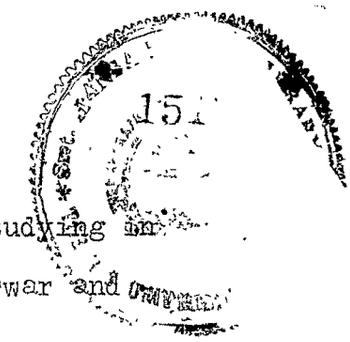


CHAPTER PILOT TESTING

7.1 OBJECTIVES OF PILOT TESTING

Pilot Testing was conducted with the main objective of shortening the test originally constructed. Shortening the test involved removal of unsuitable items. Deciding the difficulty level and item-validity coefficients was an 'additional' objective.

Finding out difficulty level and item-validity coefficients is 'additional' because it is known that in creativity tests speed is an influential component (274). In such cases, it has been held that traditional methods of item analysis fail to give correct picture of consistency of a test (124, 112, 4). On the other hand the test cannot be completely speeded one. Hence it was thought that it would be better to be skeptical about the results of the analysis.



7.2 Sample: Sample consisted of 426 children studying in X Std., drawn from six secondary schools of Dharwar and Udipi Taluks (Mysore State).

The two places, namely Dharwar and Udipi represent somewhat two distinct accents of spoken Kannada.

Children leave the school after passing the X std., Public Examination conducted by the Secondary Education Board and obtain their School Leaving Certificate. According to the present system of Secondary Education in the State of Mysore, a boy or girl would be at 15 + by the time he or she passes X Std., Extremes being less the age range of the group being studied may be taken to be 14-16 years.

7.3 Administration:

Testing was conducted for batches of 40-50 children at a time. Instructions were given orally. Steps followed in administering the tests have been given earlier; for giving examples black board was used. Complete testing required two 2½ hour sessions. Wherever parallel - batches were to be handled, help of the teachers was taken. In such cases, however, instructions were given by the author himself. This was made possible by starting parallel batches at a difference of 20 minutes. After the giving of instructions is over, a teacher used to take charge.

His work involved only asking the Ss to move to the next item at the end of specified interval of time. It may be mentioned once again here that each item in all the four subtests viz., CRM I - IV is seperately timed. The investigator (or the author) made himself available whenever necessary.

Children showed interest in the test. There was silence during testing except when anybody got doubts. In such cases they were asked to put up their hands. The examiner used to go near the boy or girl and solve the difficulty. Even such difficulties were very few because each test was made clear through examples.

7.4 Scoring:

Scoring was done only for fluency for the first two subtests and for number of triplets completed and number of correct arithmetical computations for third and fourth subtests respectively. At the item analysis stage, it was thought those scores would be sufficient to indicate the consistency of the test to be finalised. It may be mentioned here that other scores would be derived from the same set of responses used in the item analysis.

Fluency is the number of suitable responses given to a figure. In CRM I, fluency is the number of descriptive titles given to any stimulus figure. In CRM II fluency is the number of resembling objects, animals and things including those given in the form of titles as responses to a stimulus figure. In CRM III a response is a triplet which is correct and hence number of correct triplets written in response to a letter duplet was taken as the score. In CRM IV a response is a positive whole number brought at the end of the each correct set of arithmetical operations required in the test. Number of such sets formed the score for a set of numbers given as stimulus.

As already mentioned in the instructions, repetitions, geometrical responses and figures and totally irrelevant responses were discounted. Those responses in which the subject referred to only a part of the figures were also discounted. However, wherever there was a possibility of combining part responses in order to yield a whole responses, it was done so. This is possible only where S writes responses in two or more successive answer spaces combining which becomes a whole response. Such cases were very few.

7.5 A STUDY OF ITEMS

Items of CRM I and II were subjected to scrutiny based on author's understanding of the figures and the nature of responses to figures. Only those figures (or items) which met certain levels of ambiguity (complexity) and were structurally different in the sense that each elicited a broad class of responses different from its peer figure were retained for item analysis.

The figures did not have direct meaning or same nearness to real life situations and objects. In other words, the distance of a stimulus figure from a real life situation varied along the ambiguity continuum. Hence the purpose would be to see that the figures (in the CRM I especially) touch distinct points along the continuum so that the figures selected represent the population of possible figures.

Another point considered for retaining items for the final test is the structure and control of figures. For example, there are line drawings made of straight lines falling one upon another showing a set like structure. There are others full of curved lines showing either movement or wave like action or surface.

There may be dots scattered often in patterns. Each evokes different kind of responses. There may even be ink-blots. However ink-blots have not been included in this study. But some of thefigures which had streations circular or otherwise brought similar effects.

'Ambiguity' simply refers to 'expression capable of more than one meaning' (80). An ambiguous stimulus meant 'one that can mean more than one thing' (68). Ambiguous figure stood for 'any of a large category of pictures which when steadily regarded, are subject to changed interpretation without any actual change in picture' (68). Stimulus ambiguity has been defined in terms of response variability. TAT cards which elicit a variety of responses, that is, different themes or stories have been considered to be ambiguous and those which elicit stereotyped responses have been considered to be unambiguous (Kenny and Bijou, 166).

A direct application of the meaning inferred from the above definitions, gives number of possible responses as a measure of ambiguity. Highly ambiguous figure yields more number of responses than the one which is less ambiguous. Hence 'mean fluency score' was taken to indicate the position of a figure along the ambiguity non-ambiguity continuum. Mean fluency score was obtained by deciding the sum of

fluency scores for all Ss by the total number of Ss. It simply indicated the number of ideas an average individual can associate with a figure.

The question of ambiguity comes only in the case of CRM I. In CRM II, as the figures have the same components, to a larger extent complexity is controlled. Even ambiguity is not much varied. However even figures of CRM II were studied for ambiguity.

The question of ambiguity of a stimulus may be much more complex than what the simple mean fluency score indicates. The purpose behind the present scrutiny of figures is to see that relative position of each is decided. It is not to probe into the very question of ambiguity. Hence the measures of mean fluency arbitrarily used to fix the position.

Out of twentyfive figures of CRM I, one figure was completely dropped, for it did not evoke whole responses at all. Almost all responses were made to part of the figure. In CRM II, deciding whether a response is a part response or whole response was a slightly difficult job. Only those responses which specifically refer to part figures were discounted. Others, even though they were less specific, were taken as whole responses.

Mean fluency scores of all figures of CRM I and CRM II have been given in Tables 1 and 2 respectively. When converted into a frequency distribution the frequency polygons were approximately bell shaped along the ambiguity - non-ambiguity continuum. Tables 3 and 4 give the relevant details (means and s.ds.) about the distribution of stimulus figures of CRM I and CRM II respectively. That the curve is approximately bell-shaped is only of peripheral interest. Figures 1 and 2 give the respective frequency polygons.

Out of twentyfour figures (as one figure was not considered hence, not twentyfive) of CRM I eight figures from alternative class intervals (including the one containing the mean of the mean fluency scores and avoiding the ones at the extreme ends of distribution) were selected and retained for item analysis.

By choosing figures from alternative class intervals, sufficient distance in terms of mean fluency score was maintained between groups of figures. Further while choosing figures, author's experience with the nature of responses was taken into consideration. If the figure has evoked more responses under discountable categories (see scoring), it was rejected. These figures which on the whole evoked responses were retained at this stage for item analysis. Codes of figures selected have been underlined in the Table 3.

TABLE 1

Mean Fluency Scores of Stimulus items of CRM I.

S.No.	Figure	Mean fluency score
1	D_1	5.63
2	D_2	5.04
3	D_3	5.89
4	D_4	4.80
5	D_5	5.24
6	E_1	4.57
7	E_2	4.65
8	E_3	4.77
9	E_4	4.21
10	E_5	4.87
11	F_1	4.27
12	F_2	5.42
13	F_3	5.61
14	F_4	5.60
15	F_5	4.42
16	G_1	3.21
17	G_2	6.59
18	G_3	5.64
19	G_4	5.01
20	G_5	5.37
21	H_1	5.18
22	H_2	--
23	H_3	5.66
24	H_4	4.49
25	H_5	6.07

TABLE 2

Mean fluency scores of stimulus items of CRM II.

S.No.	Figure	Mean fluency scores
1	A_1	4.79
2	A_2	3.45
3	A_3	2.72
4	A_4	4.38
5	A_5	3.97
6	A_6	3.43
7	A_7	4.19
8	A_8	4.44
9	B_1	4.54
10	B_2	4.31
11	B_3	3.53
12	B_4	3.69
13	B_5	3.57
14	B_6	4.28
15	B_7	3.41
16	B_8	3.16
17	C_1	3.85
18	C_2	3.83
19	C_3	4.23
20	C_4	5.01
21	C_5	3.14
22	C_6	2.73
23	C_7	3.84
24	C_8	3.89

TABLE 3

Distribution of CRM I-figures according to mean fluency scores.

S. No.	Class Interval	Distribution of Figures*	f	M_X.P	fX	x	x^2	fx^2
1	6.3-6.6	G_2	1	6.45	6.45	1.37	1.8769	1.8769
2	5.9-6.2	$\underline{D_3}$ $\underline{H_5}$	2	6.05	12.10	.97	.9409	1.8818
3	5.5-5.8	D_1 F_3 F_4 G_3 H_3	5	5.65	28.25	.57	.3249	1.6245
4	5.1-5.4	$\underline{D_5}$ $\underline{E_2}$ $\underline{G_5}$ $\underline{H_1}$	4	5.25	21.00	.17	.0289	.1156
5	4.7-5.0	D_2 D_4 E_3 $\underline{E_5}$ G_4 E_2	6	4.85	29.10	.22	.0484	.2904
6	4.3-4.6	$\underline{E_1}$ F_1 F_5 $\underline{H_4}$	4	4.45	17.80	.62	.3844	1.5376
7	3.9-4.2	E_4	1	4.05	4.05	1.02	1.0404	1.0404
8	3.5-3.8	-	0	3.65	.00	1.42	2.0164	.0000
9	3.1-3.4	G_1	1	3.25	3.25	1.82	3.3124	3.3124
			24		122.00			11.6796

$$\text{Mean} = \frac{\Sigma fX}{N}$$

$$= \frac{122}{24}$$

$$= 5.08$$

$$\text{Standard Deviation} = \sqrt{\frac{fx^2}{N}}$$

$$= \sqrt{\frac{11.6796}{24}}$$

$$= .70$$

*Note: Figures underlined were selected for item analysis.

In CRM II, out of the twentyfour figures ten figures were selected and retained for item analysis. The ten figures were selected above with one difference. All the ten figures were above the mean of mean fluency scores for the twentyfour figures. Thus ambiguity-spread is further reduced in the case of CRM II. This is apart from the fact that all the figures have controlled parts O, V and I. Regarding the nature of the figures in evoking different kinds of responses, the principle remained same as in CRM I. That is, each figure retained should evoke different class of responses in relation to its peers. As the figures are to be studied by item analysis, the subjective conditions mentioned above are not to be mistaken as rigid rules. Codes of figures selectedhave been underlined in Table 4.

As mentioned earlier, the question of ambiguity or of operational differences does not arise in the case of remaining two subtests namely, CRM III and IV. Hence all the items in these were subjected to item analysis.

7.6 Item Analysis

Before attempting to analyse the test items, a brief description of established procedures of item analysis will be presented at this juncture.

TABLE 4

Distribution of CRM II figures according to mean fluency scores

S. No.	Class Interval	Distribution of Figures*	f	x'	fx'	fx'^{2}
1	4.9-5.2	$\underline{C}_4$	1	+3	+3	+9
2	4.5-4.8	$\underline{A}_1$ $\underline{B}_1$	2	+2	+4	+8
3	4.1-4.4	$\underline{A}_4$ A_7 $\underline{A}_8$ $\underline{B}_2$ B_6 $\underline{C}_3$	6	+1	+6	+6
4	3.7-4.0	$\underline{A}_5$ B_4 $\underline{C}_1$ C_2 $\underline{C}_7$ C_8	6	0	0	0
5	3.3-3.6	A_2 A_6 B_3 B_5 B_7	5	-1	-5	+5
6	2.9-3.2	B_8 C_5	2	-2	-4	+8
7	2.5-2.8	A_3 C_6	2	-3	-6	+18
					-15	
					-2	+54

$$\text{Mean} = AM + ci$$
$$= 3.85 - .08 \times .4$$
$$= 3.82$$

$$c = \frac{\Sigma fx'}{N}$$
$$= \frac{-2}{24}$$
$$= -.08$$
$$c^2 = .0064$$

$$\text{Standard Deviation} = i\sqrt{\frac{fx'^2}{N} - c^2}$$
$$= .4\sqrt{\frac{54}{24} - .0064}$$
$$= .60$$

*Note: Figures underlined were selected for item-analysis.

Item analysis refers to computation of statistics describing items that constitute a test. By item analysis we come to know how far the items are adequate to constitute the test.

7.7 Difficulty or Popularity Analysis

Simplest form of item analysis involves difficulty or popularity analysis. Item difficulty, usually denoted by p, refers to the proportion of persons answering an item correctly or passing in it. Usually an item variable is dichotomous taking values of either 0 or 1 according as an individual fails or passes. However this is not always the case. In certain tests where each item is continuous, some pass - fail cut-off point such as the median or even theoretical midpoint of the continuum has been preferred (Bhogiwal: 30).

In deciding p it has become rather customary to take extreme 27% groups defined by the performance on the total s score. That too, when the size of the sample is such that the extreme groups are of a strength of 100 each, averaging the proportion of persons passing in the item yields relatively same p values as the p calculated on the basis of whole group. A group of 370, 27% of which is 100 is taken usually for the item analysis study.

Thus $p = \frac{U + L}{2}$ Where U and L are proportions of persons passing in the item from upper and lower 27% groups.

Dubois (60) mentions two uses of p. It helps in determining the order of items in a test or a subtest

> A second use of the p values is in the development of speeded tests. Speed in a function may be measured by a collection of items of uniform difficulty, of which p is an accepted measure. The number of items answered in a set amount of time can then be taken as a measure of speed...

He further writes ___

With items on structural interest and personality tests, p is a measure of popularity rather than difficulty. Again, however items with P value of .00 or 1.00 do not make discrimination among the respondents....

It has been possible to show that the maximum possible seperation of individuals according to their performance on the itam is possible when P is around .50 (Garret; 90; p.363).

7.8 Internal Consistency

Internal consistency is indicated by high inter-correlations among items. Instead of calculating all the intercorrelations only item-total correlations are calculated. A high internal consistency refers to homogeneous set of items.

Total of item scores will be a continuous variable. Item total correlation is generally a point - biserial or biserial r, which is also an algebraic variant of product moment r, Item-total correlation is usually denoted by r_{it}. r_{it} is referred as item validity index or discriminating index.

As already pointed, in the case of difficulty analysis, extreme groups method has been the one mostly preferred. To facilitate quick estimation of biserial coefficients, Flanagan prepared a table of normalised biserial coefficients (Flanagan; 76) to be read directly when the percent of successes on the item in the upper and lower 27% groups defined by the total score is known. Davis (53) translated values the non-linear r into linear scale of discrimination values using Flanagan's tables. He also calculated difficulty indices as linear transformation of the percent of passing. Fan (72) found that p values by Davis method are affected by level of r_{it}. He (73) estimated p values for various levels of r_{it}. Harper and Gupta (129) translated Davis discrimination indices and Fan's difficulty p's into a single abac. Using the abac the discrimination index and difficulty value (p) for an item could be read directly if the P_1, percentage of passes in the lower

27% of the group and P_h, percentage of passes in the upper 27% of the group are known. The procedure given by Harper and Gupta has been employed in the present study.

Sample: Initial sample, as it has been stated earlier, consisted of 426 pupils. It has been suggested by different researchers that it would be convenient to take 370Ss for item analysis and thus discard others at random. (Guilford, 112; Dasgupta and Harper 129).

As each of the four tests (CRM I - IV) has been constructed under the hypothesis that they would measure at least one independent factor, item analysis was conducted seperately for each subtests. Only 370 answer sheets for each subtest were randomly and asavailable retained and others were discarded.

However as tesing was done in two Sessions some children who were present in the first session were not present in the second session. Some gave blank sheets. Had only those children who had appeared for all the four subtests in two sessions and who had taken the tests correctly been accounted for item analysis 340 students would have been available. As in each subtest there were more than 370 children, and the aim was to conduct seperate item analysis for each subtest, the former course was taken. Hence 370 children in each subtest was taken into account for item analysis.

Procedure: For each of the subtests scores obtained by 370 Ss were tabulated under the respective items selected for item analysis. In CRM I and II number of items were eight and ten each and in CRM III and IV . . six . .. and ..five .. respectively. Total of scores for all items under each subtest for each of the 370 Ss was calculated and tabulated accordingly. It should be noted that the procedure described in this section, unless otherwise mentioned, applies for each of the four subtests and stands for item analysis of each subtest seperately.

Highest and lowest totals were marked. That is the S who stood highest on the total of scores on all items of a subtest and also the S who stood lowest were marked. From the topmost S (or total) 100 Ss (i.e., upper 27%) in the decreasing order were marked. From the lowermost S (or total), 100 Ss (i.e., lower 27%) in the increasing order were marked.

In some cases where each item is a continuum and not dichotomous (such as right or wrong), median has been successfully used as the cut-off point (Bhogiwal: 30). This case arises where item scores are continuous. In the present case too, median score for each item was calculated. Those who have scored above the median in any particular item were assumed to have passed in that item.

Thus number of persons passing in each item among the upper 27% (or 100) Ss, and also among the lower 27% (or 100) Ss were noted. Davis difficulty and discriminating indices were calculated using the abac prepared for the purpose by DasGupta and Edwin Harper (1962). Using the conversion scale given by Dasgupta and Edwin Harper, Flanagan's discriminating (r's) were also noted. Relevant information regarding computation of difficulty and discriminating indices for each item has been presented in Tables 5, 6, 7 and 8, successively for each of the four subtests CRM I - IV.

It is observed that difficulty values are around .50 and discrimination indices are very high for all the items. This would mean almost all items which were retained for item-analysis are suitable to be included in the final test. In the first two sub-tests, screening done by the author on his experience over the nature of items may be the reason for the high discrimination indices. All inconsistent items did not find a way into item analysis. Secondly the test scores are partly speeded. In a similar study (Wallach and Kogan, 274) item-total rs for number score (fluency in the present case) were very high.

Had the item analysis been conducted for all items (i.e., twentyfive in CRM I and twentyfour in CRM II) there would have been inconsistent items having low discrimination indices. Screening done at the Ist stagehas been eventual

TABLE 5

The Davis difficulty and discriminating Indices for CRM I stimulus figures

Sl. No.	Figure*	P_h Percentage of successes in the upper 27 %	P_l Percentage of successes in the lower 27 %	Davis Difficulty Index**	Davis Validity Index**
1	D_3	90	17	53(55%)	55(.72)
2	$\underline{D_5}$	79	11	45(41%)	50(.68)
3	$\underline{E_1}$	72	19	47(45%)	35(.53)
4	$\underline{F_2}$	81	13	48(47%)	50(.68)
5	G_5	81	14	48(47%)	50(.68)
6	H_1	72	10	43(37%)	47(.65)
7	$\underline{H_4}$	78	19	44(39%)	52(.69)
8	$\underline{H_5}$	87	21	53(56%)	47(.65)

* Figures retained for the final test are underlined.

** Values given in the parenthesis are percent difficulty and biserial coefficients (Flanagan's) determined using the conversion scale (Harper and Gupta, 1962).

TABLE 6

The Davis Difficulty and Discriminating Indices for CRM II stimulus figures

Sl. No.	Figure*	P_h Percentage of successes in the upper 27 %	P_l Percentage of successes in the lower 27 %	Davis Difficulty Index**	Davis Validity Index**
1	A_1	88	17	52(53%)	54(.71)
2	A_4	86	8	47(45%)	60(.76)
3	$\underline{A_5}$	89	11	50(50%)	60(.76)
4	$\underline{A_8}$	82	9	45(41%)	55(.72)
5	B_1	84	7	45(41%)	60(.76)
6	B_2	86	9	48(47%)	59(.76)
7	C_1	85	11	49(48%)	56(.73)
8	$\underline{C_3}$	85	2	42(35%)	74(.84)
9	$\underline{C_4}$	91	15	53(56%)	59(.76)
10	C_8	90	14	52(53%)	58(.75)

* Figures retained for the final test are underlined.

** Values given in the parenthesis are percent difficulty and biserial coefficients (Flanagan's) determined using the conversion scale. (Harper and Gupta, 1962).

TABLE 7

The Davis Difficulty and Discriminating Indices for CRM III Stimulus Items

Item No.	P_h Percentage of successes in the upper 27 %	P_l Percentage of successes in the lower 27 %	Davis Difficulty Index**	Davis Validity Index**
1	85	10	48(47%)	57(.74)
2*	88	1	42(35%)	80(.87)
3	86	10	48(47%)	58(.75)
4*	75	7	42(35%)	53(.71)
5	90	13	51(52%)	59(.76)
6	84	10	48(47%)	56(.73)

* Items retained for final test.

** Values given in the parenthesis are percent difficulty and biserial (Flanagan's) coefficients determined using the conversion scale (Herper and Gupta, 1962).

TABLE 8

The Davis Difficulty and Discriminating Indices of CRM IV items

Item No.	P_h Percentage of successes in the upper 27 %	P_l Percentage of successes in the lower 27 %	Davis Difficulty Index**	Davis Validity Index**
1*	92	9	50(50%)	68(.81)
2	92	3	46(42%)	78(.86)
3*	94	15	54(57%)	65(.79)
4	90	3	45(41%)	75(.85)
5	98	3	51(52%)	90(.90)

* Items retained for the final test.

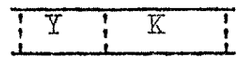
** Values given in the parenthesis are percent difficulty and biserial coefficients determined using conversion scale.

in eliminating such items. Hence the items chosen are consistent enough among themselves in discriminating individuals.

In the third and fourth subtests there has been very little reason to believe that the items differ. Each item has the same content followed by the same operation of responding. All the difference that can be thought off should be due to the individuals affinity towards a particular number of letter. As the numbers and letters used in the tests are those of daily use, very little is seen why the items should differ from each other. Hence the items only measure difference between individuals. The items are consistent among themselves in doing this. Hence the items have high discriminating indices.

7.8 Selection of items: Finally, selection of items for the final test was a matter of need and experience.

In CRM I H_5, F_2, D_5 and E_1 and one practice item F_4 were retained for the final test. The figures have been mentioned above, in the order of their mean-fluency scores. That is, meanfluency scores of H_5, F_2, D_5 and E_1 are 607, 5.42, 5.24 and 4.57 respectively (Table 1).

These values have been obtained from Table 1. In CRM II, $\underline{C}_4$, $\underline{A}_8$, $\underline{C}_3$ and $\underline{A}_5$ were retained. No practice item was added as operation remains same as in CRM I. Their respective mean fluency scores are 5.01, 4.44, 4.23 and 3.97. These values have been obtained from Table 2. In CRM III the duplets  ,  and the one practice item were retained. In CRM IV number sets 3,5,6,7, 9, and 2,3,5,6,8 and one practice set were retained. Practice items for CRM III and IV have not been printed on the test sheet. They are to be worked out on the black board for one response only allowing others to come from Ss. Formats of the above subtests remained same. Tests in the final form can be seen in Appendix C. Except that the number of items are less in the final forms of the subtests, procedure of administration too (including instructions) remains unchanged.