Accepted 117/150

Sept. 23/46

AN ANALYSIS OF THE CALIFORNIA TEST OF MENTAL MATURITY; ADVANCED BATTERY

by

Louis Checov

A Thesis submitted in Partial
Fulfilment of The Requirements
for the Degree of
MASTER OF ARTS
in the Department
of
PHILOSOPHY AND PSYCHOLOGY

The University of British Columbia
October, 1946.

TABLE OF CONTENTS

TABLES AND CHARTS

CHAPTER I

STATEMENT OF PROBLEMS

The modern practice of mental measurement is dependent upon many basic principles or concepts - validation, standardization, objectivity, reliability, discriminating ability, norms, and ease of administration and scoring. Consequently, an intelligence test is to be judged by the extent to which it adheres to these principles. Unfortunately, the production of tests, and their adoption, has far outpaced proper critical evaluation, even after their use has become relatively widespread. As Stuit puts it, "Careful studies of validity and reliability coefficients and norms presented by test authors are all too rare." (23)

The California Test of Mental Maturity (Advanced Battery) consists of a total of 253 items, requiring two periods of about 45 minutes each to administer. It includes 3 pre-tests designed to detect gross defects of sight, hearing, and motor coordination, and 13 subtests arranged as power tests. The subtests numbering 4 to 16 are grouped into sections: A(memory, tests 4 and 5 - 53 items), B(spatial relationships, tests 6, 7, and 8 - 45 items), C(Reasoning, tests 9 to 15 - 105 items), and D(Vocabulary, test 16-50 items). The memory group is designed to test immediate and delayed recall. The tests on spatial relationships are intended "to reveal orientation in and ability to utilize spatial relationships"; those on reasoning are designed to "reveal evidences of the higher games of mental activity"; and the vocabulary test is to reveal "maturity of the apperceptive processes with references to ideas and growth of meaning." From the battery of subtests, three IQ's are obtained: Total Mental Factors (total scores), Language Factors (scores on subtest 5, 14, 15 and 16), and Non-Language Factors (tests 4 and 6 to 13 inclusive). The tests results are grouped to show a diagnostic profile (or the individual testee, and the manual of directions provides tables for the conversion of raw scores into IQ's.

The CTMM has received considerable praise in the few years since its publication, and whereas its use apparently has been reasonably widespread, such analyses as Stuit recommends have not been carried out. The only investigation dealing with reliability and validity that has appeared inthe literature since its publication has been that of Traxler (29). He attempts to investigate the reliability of the CTMM, the relation between language and non-language IQ's, the correlation of the CTMM with other tests of mental ability, and the relation of each type of IQ to reading ability. Although his sample is a small one, part of Traxler's findings on reliability point to some disagreement with the figures given by the test authors.

Apparently, the usual evaluation of a test is expressed by most authors after a study of the make-up of the test and the accompanying manual of directions, and any other pertinent information immediately available, and some recourse to statements of users of the test. Thus Harris claims that the California Test of Mental Maturity is probably the best test available above the five-year level. (8), and Kuhlman writes, "... we believe that the unabbreviated batteries are to be classed among the very best on the market for determining general levels of mental maturity." (15)

According to the Manual of Directions, the particular value of the CTMM (Advanced Battery) designed for use with students in Grade 7 to the college sophomore level, is in its diagnostic possibilities. Some of its significant features, as cited in the manual, are the following:

1. It purports to make a diagnostic evaluation, for each student, of those mental abilities which are related to success in various types of school activity- "in order that the teacher may utilize this information directly in aiding students who have learning difficulties".

2. It provides a diagnostic profile showing the extent to which the student possesses these abilities, "thus enabling the teacher to see at a glance the probable source of difficulty or success."

3. Being primarily diagnostic, the test yields not one

but three mental ages (MA's) and intelligence quotients (IQ's) - total, language, and non-language.

4. Again according to the manual, "the number and variety of test situations assures a high reliability."

5. The test correlates approximately .88 with the Stanford-Binet. The manual does not indicate the number of cases in the sample from which this particular coefficient was calculated.

6. The norms are adjudged comparable to those regularly obtained by use of individual psychological examinations and well-standardized group tests.

Educational Bulletin No. 14 adds further information about the value of the CTMM as an aid to diagnosis in the schoolroom, and also notes that its value is enhanced by the inclusion of pretests designed to detect gross defects of hearing, vision, and motor coordination. The Bulletin also quotes certain testimonials received in praise of the test, of which the following is illustrative: "we believe the California Test of Mental Maturity to be the finest intelligence test available, and we use it from the first to the twelfth grade."

Purpose and Method of Present Study

It is apparent that a certain service will be rendered by an analysis of this test. For this purpose answersxare sought to the following questions:

I.(a) What is the level and range of difficulty of individual items?

(b) Are the items arranged in order of difficulty?

(c) How difficult are the various subtests?

(d) To what extent do the items differentiate between superior and inferior students?

II.(a) How reliable are the scores for total, language, and non-language factors?

(b) How reliable are the various subtests?

III.(a) How well do Otis IQ's correlate with IQ's obtained from the CTMM?

(b) How well do CTMM IQ's correlate with academic subjects?

(c) What is the correlation between CTMM IQ's and certain technical school shop subjects?

IV. To what extent do the tests fall into the patterns suggested by the test makers?

CHAPTER II

A DESCRIPTION OF THE TESTS AND THE SUBJECTS

The subjects used in the present study were boys in Grades 10, 11 and 12 of the Vancouver Technical School, Vancouver, British Columbia. Approximately 195 students took the Advanced Battery of the California Test of Mental Maturity in November 1942; 180 of these had previously taken the Otis Quick-Scoring Mental Ability Tests (Gamma Test, Form Am). Table I shows certain statistics for the CTMM and the Otis test.

TABLE I

Means, Standard Deviations, and Ranges of IQ's on the CTMM and Otis Tests

	Otis IQ	CTMM Total IQ	CTMM Language IQ	CTMM Non-Language IQ
N	180	180	180	180
Mean	110.02	111.53	111.70	108.52
S. D.	11.92	9.89	10.58	10.49
Range	86-133	83-143	85-136	78-144

From the Table it is seen that the Means and Standard Deviations from the CTMM and the Otis are very comparable. It is also to be observed that the average IQ for all Vancouver students in Grades 10 to 12 is in the neighborhood of 111 or 112 (26) thus the sample in the present study appears to be representative of the Vancouver secondary school population.

Traxler, in his study, found that the median IQ based on language factors was much larger than the median IQ based on non-language factors. He assumed that this great difference was occasioned by the cultural background of his subjects, which favored language development. It would be expected, then, that the median IQ of non-language factors would, in the present sample of technical school boys, be at least as large as the median IQ for language factors. This is not the case; the difference between the two is not great, but the language factors remains the larger. In addition, Traxler finds that the variability (Q) for language factors is greater than for the non-language factors. This too is not the case in the present study.

Students' final marks in various technical subjects were obtained from the Technical School. Academic subjects included English, Science, and Mathematics; shopwork included woodwork and machine shop.

CHAPTER III

PRINCIPLES OF TEST CONSTRUCTION

The two main requisites for a measuring instrument are reliability and validity, and indeed, "the prime consideration in the construction and administration of tests is validity - that is, representation of the influence of factors that the test is to represent." (2) Besides this primary consideration of validity, other problems are mainly of an administrative sort; the test must not take too long to give, it should be reasonably easy to administer etc. A test which is highly valid (for a specific purpose) will represent individual variations in the character it is supposed to measure with great fidelity. A test which has a law degree of validity secures responses which represent strongly the influence of a number of other factors, so that the character we desire to measure is somewhat lost among the many present. What is done in a test, then, is to attempt to obtain test items which will stimulate responses of a given kind, and further, we try to get enough of these variance types of items so that the undesired factors will tend to neutralize each other and average out. Unfortunately, as one writer points out in this connection, "Investigators so far have attained only moderate success in these efforts." (3)

In this chapter it is proposed to deal with certain of the generally-accepted principles of test construction, all fundamentally aimed at developing a scientific measuring instrument.

1. Item Difficulty

The heart of the item analysis problem lies in the diagnostic value of items. An item must be able to distinguish between individuals who have more or less of the trait that the test attempts to judge. "No item which is answered correctly by all pupils in a given group can be of any functional value in a general achievement test for that group, nor can any item which is answered correctly by none of that group." (9) One consideration that could be admitted as an exception to this rule is that several very easy "shock absorbers" may be introduced at the beginning of a test so as not to discourage completely the testee.

Although it is generally believed that maximum discrimination among testees is obtained by items that about one-half the individuals can pass, test authorities are not in agreement upon what is the best form of distribution of item difficulty. According to some, the spread of scores on a test should extend from near zero to near the highest possible score (that there should be a range of success from about 5 to 20 percent to 80 to 95 percent) to ensure a maximum reliability. On the other hand, Symonds (24) and Thurstone (28) have shown that a test consisting of items of 50 percent difficulty value measures an individual most accurately. The former viewpoint conforms to what is generally known as the definition of a power test, and the latter is a feature of the speed test.

From the sample of the present study, as indicated in Table 2, few of the items qualify according to either of the principles or standards expressed above[3].

TABLE 2
Item Difficulty - Percentage Passing Each Item

	Subtests												
Item	4	5	6	7	8	9	10	11	12	13	14	15	16
1	91	96	94	20	100	34	97	59	87	93	94	98	99
2	78	95	80	63	99	71	93	41	55	92	73	96	92
3	78	58	91	22	99	24	92	65	14	82	94	93	65
4	43	81	82	43	95	79	76	7	39	82	91	33	80
5	77	83	95	27	84	34	85	9	61	81	81	93	41
6	90	89	91	58	41	79	67	12	63	90	78	95	61
7	97	81	62	35	16	16	47	86	14	85	30	80	67
8	95	88	70	31	6	18	87	61	18	68	86	90	41
9	95	91	97	37	3	20	74	35	15	59	28	85	77
10	96	89	88	17	2	25	13	78	30	45	54	88	66
11	96	80	91	23		80	85	29	17	32	55	91	47
12	93	86	91	38		79	57	12	7	17	43	82	26
13	96	95	95	25		28	22	31	2	16	21	76	28
14	98	67	61	26		57	52	21	0	11	16	54	34
15	93	90	59	15		68	45	11	0	2	5	17	46
16	92	72	41										27
17	96	41	72										42
18	90	94	78										33
19	90	79	93										30
20	93	59	94										16
21	78												53
22	92												17
23	88												77
24	75												18
25	94												9
26	85												12
27	86												34
28	70												22
29	92												23
30	56												8
31	90												44
32	71												20
33	81												16
34													45
35													12
36													8
37													11
38													15
39													10
40													25
41													67
42													16
43													10
44													13
45													10
46													7
47													8
48													2
49													5
50													7
M%	88.7	80.7	81.3	32.0	54.5	47.5	66.1	37.1	28.1	57.0	56.6	77.4	34.9

M% = Mean of the Percentages Passing Each Item in Each Subtest.

A quick glance at the Table reveals that there is a great range of difficulty of items of the whole test, varying from items that all students pass, to those that all fail. In addition, there are great discrepancies between the difficulty values of the items of the various subtests. For example, not a single item in subtest 7 was passed by over 80 percent of the subjects, whereas two-thirds of the items of subtests 4, 5, 6 and 15 were passed by over 80 percent of the students. Thus test 4 is composed primarily of items of a rather easy nature, whereas most of the items of subtest 7 are too difficult. Nearly 67 percent of the items of subtest 4 were passed by over 80 percent of the students; in subtest 7 only 13 percent of items were passed by more than 40 percent, and none of the items was passed by more than 60 percent. In subtest 12, over 50 percent of the items were passed by less than 20 percent of the subjects; in subtest 9, 6 percent of the items were passed by more than 80 percent of the group, and 20 percent of the items were passed by less than 20 percent of the group.

Table 3 summarizes the data of Tables 2 in order to illustrate more clearly the differences in the difficulty values discussed above.

TABLE 3.

Summary of Data on Item Difficulty

Sub-test	No. of Items	No. of Items passed by over 80% of the group	No. of Items passed by less than 20%	No. of Items passed by between 40 & 59% of pupils
4	33	24	0	2
5	20	14	0	3
6	20	13	0	2
7	15	0	2	2
8	10	5	4	1
9	15	1	3	1
10	15	6	1	4
11	15	1	5	2
12	15	1	8	1
13	15	7	2	2
14	15	5	2	3
15	15	10	1	1
16	50	3	21	8

It will be noted, also, that in subtest 12 more than 50 percent of the items were passed by less than 20 percent of the group, and only 6 percent of the items were passed by between 40 and 59 percent of the subjects. Test 8 includes 50 percent of the items passed by over 80 percent, and 40 percent passed by less than 20 percent. Tests 9, 13 and 14 are similar to test 8.

In the complete test, out of a total of 253 items, over 35 percent were passed by more than 80 precent of the group, and 20 percent were passed by less than 20 percent of the subjects. In none of the subtests is a majority of the items passed by between 40 and 59 percent of the group, The range recommended by Symonds (28), nor, with the exceptions of test 8, do any of the subtests exhibit a range of difficulty of from near zero to near 100 percent.

From these data it is apparent that few of the items fall within the range of difficulty generally. accepted as the most reliable, and the diagnostic value of many of the items is to be questioned.

2. Difficulty of the Subtests

Table 4 shows the means and standard deviations for subtests 4 to 16 inclusive as obtained in the present study. They are compared to similar data reported by the Los Angeles County Superintendent of Schools (16)

TABLE 4.

Means and Sigmas of Subtests

Sub-test	No. of Items	Obtained Scores		Los Angeles Scores	
		M.	S.D.	M.	S.D.
4	33	29.30	2.98	27.83	4.82
5	20	16.07	2.82	14.08	4.15
6	20	16.47	2.72	16.07	3.44
7	15	5.03	2.05	4.09	2.05
8	10	5.47	1.40	6.11	2.05
9	15	7.31	1.74	7.12	2.16
10	15	10.02	1.78	10.29	1.98
11	15	5.59	1.91	4.64	1.93
12	15	4.36	1.89	3.56	1.93
13	15	8.62	2.72	6.58	3.41
14	15	8.48	2.35	3.14	2.20
15	15	11.64	1.67	10.99	2.79
16	50	17.43	6.52	10.75	6.66

The mean difficulty of the subtests (as indicated in Table 2) varies from that of test 12 (28.1 percent of the items passed) to that of test 4 (88.7 percent of the items passed).

Of the battery of tests comprising the whole, only tests 8, 9, 13 and 14 approach the mean difficulty level of 50 percent generally considered the most reliable. Tests 4, 5 and 6 are apparently too easy for the sample of the present study, and tests 7, 12 and 16 appear too difficult. It might be noted, also, that for tests 7, 8, 11 and 12, the scores obtained are only slightly better than would be expected on a chance basis.

The mean scores of the present sample and for the Los Angeles study are roughly comparable, with, however, some exceptions, especially in subtests 14 and 16. The present sample appears to be less variable than that reported in the Los Angeles Study, especially for subtests 4, 5, 6, 8, 13 and 14. No attempt was made to determine the significance of these differences.

3. Order of Difficulty

Although time limits are provided, the test is, according to the publisher, a power rather than a speed test. If, as is claimed, the test is of the former variety, there is good reason to assume that there will be a range of scores of from near zero percent to near 100 percent, with the items in increasing order of difficulty. This claim was appraised by computing the rank-order correlating between obtained order of difficulty and test order in each of the subtests. The results appear below in Table 5.

TABLE 5.

Rho Between Obtained Order of Difficulty and Test Order

Subtest	Rho
4	.26
5	.37
6	.16
7	.32
8	1.00
9	- .114
10	.75
11	.27
12	.86
13	.95
14	.85
15	.67
16	.78

The values of these correlations indicate that for 6 of the 13 subtests, the items are not arranged even in approximate order of difficulty. Some of the items for this sample appear to be seriously misplaced. These results may be compared with those of Hovland and Wonderlic, who report rank order correlations between test difficulty and obtained order of .46 to .75 in various form of the Otis Self-Administering Test, Advanced Form (13), and with those of Tyler (31), whose reports rank order correlations of .69 to .91 in form D of the Terman-McNemar Test.

It was assumed for purposes of this study that a student attempted all items down to the last one he marked.

Table 6 shows the percentages of students attempting all the items in each subtest, i.e. the percentages who marked the last item in each subtest.

TABLE 6.

Percentages of Stuaents Attempting All Items

Subtest	Percentage
4	95
5	98
6	98
7	55
8	.03
9	95
10	87
11	69
12	7.6
13	7.6
14	17
15	87
16	49

Evidently, the test is not a speed test, with the exceptions of subtests 7, 8, 11, 12, 13, 14, 16. Also, the fact that the items on all subtests do not vary from very easy to very difficult suggests that this is not essentially a power test.

4. Item Validity

According to Hawkes, Lindquist and Mann (11), "an item may be said to have zero discriminating power when there is no systematic difference between the general achievement of the pupils who succeed on that item, and those who fail." For purposes of determining the discriminating power of the item, subtest 9 and 10 were selected (both of low reliability), and the phi coefficient determined for every item in both subtests. ( ) The subgroups used were the upper and lower 50 students, and the coefficients were obtained from the abac given in Guilford (4). The results appear in Table 7.

TABLE 7.

Phi Coefficients of Items in Subtest 9 and 10

	Subtest 9			Subtest 10		
	% Passing			% Passing		
Item	[x]L.group	[x]U. group	ϕ	L.group	U. group	ϕ
1	20	34	.14	98	98	0
2	72	72	0	90	90	0
3	36	32	-.08	88	96	.13
4	78	86	.10	80	82	0
5	32	42	.10	72	92	.26
6	88	94	.10	62	76	.18
7	12	28	.20	32	50	.19
8	20	20	0	82	94	.20
9	16	22	.10	64	80	.17
10	20	36	.16	14	10	-.08
11	76	82	.08	82	86	.01
12	64	82	.28	48	74	.28
13	32	30	-.03	16	18	0
14	48	72	.27	42	60	.19
15	66	74	.10	38	44	.02

[x]L. group - lower 50 of sample.

[x]U. group - upper 50 of sample.

It will be seen from these results that individual items fail to discriminate satisfactorily between the upper and lower groups of the present sample. Indeed, in some cases, the index of discrimination is negative. The results are not surprising in view of the low coefficients of reliability for these two subtests 9 and 10, which are .01 and .06 respectively.

CHAPTER IV

RELIABILITY OF THE TEST AND SUBTESTS

According to Warren (33) reliability refers to "a) either the degree of accuracy of a report, b) the self-consistency of a test or other measuring device, or c) the reciprocal of the variability of a series of measurements from s me chosen standards. (coefficient of reliability: the correlation among a set of measurement or between similar measurements)"

The term was originally introduced by Spearman (20) in 1904, who defined it as follows: "the (correlation) coefficient betwean one-half and the other half of several measurements of the same thing." In a later work he defined it thus: "reliability .... this means the amount of correlation between two or more ratings of the same kind." (21) For Ferguson and Jackson (14), these definitions give rise to the confusion evident in present-day considerations of reliability. According to Ferguson and Jackson the difficulty arises as a result of the connection, by Spearman, of reliability and coefficient, and since the interpretation of correlation coefficients is rather difficult, the connection has not been a happy one, and has tended to confuse rather than to clarify the issue. The correlation coefficient is a measure of the degree of relationship between two variables; reliability is a measure of departure from a perfect relationship. The possibility of misinterpretation is indicated by considering the definition of reliability by Thurstone (27): "a test that is subject to relatively small chance factors in its score is said to be reliable, while a test with considerable variation from one occasion to another is said to be unreliable." Here the difficulty is to determine what is meant by "relatively small" and "considerable."

Three These methods are generally employed for computing reliability. These include: 1) Test-retest Method: This is possibly the easiest method to use in determining the accuracy of a mental measuring instrument. The difficulty here, however, is the obvious one that, unlike the repetition of physical measurements, one can hardly be sure that the subject has not been changed. "With rare exceptions, what the testee learns during the first experience with the test is likely to carry over to the second trial." (5) Since the subjects are alive, they, as the objects measured, react to the process of measurement, and in any case, change over a period of time at different rates of change. This is an additional obstacle peculiar to measurements involving the living, and it is one of the main reasons why the repetition of a test is not used in all experiments concerned with the determination of reliability.

2) Alternate Forms Method: This method requires two or more equivalent form of a test. Here one is faced with somewhat the same problems as in the test-retest method. Although the items may not be identical in the two forms of the test, the more comparable they are, the more opportunity there will be for direct transfer between the two forms. However, the use of equivalent forms is usually satisfactoryxas long as a sufficient time interval is allowed to elapse in order to preclude the operation of a transfer-effect aided by memory and practice.

3) Split-half Method: This third general method is employed when it is feasible neither to repeat the test, nor to construct a parallel form. This method consists simply in giving the test once, and having divided the test into strictly comparable halves, two scores are obtained for the same individual. These are then correlated to give the coefficient of reliability. According to Guilford (6), "the split-half method is generally accepted as the best of the traditional procedures and it.... tells us of the accuracy of the scores at the time at which the individual was measured." The chief difficulty with the split-half method is that the subject is tested on only one occasion, and relatively temporary influences (feelings, attitudes, etc.) which would probably differ at other times and cancel out, affect the scores on both halves of the test the same way. Another flaw is that the reliability

of a test varies with its length. However, by means of the Spearman-Brown formula, one can estimate what the reliability of the full test would be if the two halves were really comparable.

Several other techniques have been developed in recent years which are arrived at eliminating the faults inherent in the methods described above. Of importance among these are the computation of reliability coefficients based on Rational Equivalence as developed by Richardson and Kuder (17), a method which is a variation of the split-half technique developed by Rulon (18), and a method termed analysis of variance largely developed by Fisher (1). None of these three methods lends itself to a concise description, but it can be noted that in the formulae developed by Richardson and Kuder a reliability coefficient is computed from the results of a single application of a test; that by Rulon is determined primarily from split-half scores with an additional formula utilizing differences between pairs of scores for individuals; the last mentioned discards the traditional reliability coefficients and attempts to analyse the measure of influence of components which are assumed to make up a score of an individual on a test.

In the present study, reliability coefficients were computed by the split-half method, and corrected by the Spearman-Brown formula, for the subtests as well as for total mental factors, language factors, non-language factors, memory, spatial relationships, reasoning, and vocabulary. Raw scores were used throughout. The results appear below in Table 8 where they are compared with the figures given in the manual of directions, and with those obtained by Traxler in his study.

TABLE 8.

Reliability Coefficients

Variables	Present Study	Manual	Traxler
Total mental factors	.82	.95	.92
Language factors	.86	.94	.91
Non-language factors	.74	.93	.86
A. Memory	.60	.92	.81
B. Spatial Rel.	.66	.89	.65
C. Reasoning	.68	.92	.83
D. Vocabulary	.93	.93	.91

The correlations obtained in the present study are lower than those reported in the manual of directions with the one exception of part D (Vocabulary). The difference in results is especially marked in the section entitled Spatial Relations. It is offered, then, that, except for the voacbulary section, the groupings of subtests are of value only in group prediction, and, although they do not appear highly reliable, can be considered to have rather limited value in diagnosis. With Traxler, this study raises the question about the usefulness of the diagnostic profile, for it may reasonably be assumed that the subtests within each section, being still shorter, will be even less reliable. Table 9 presents the reliability coefficients of the subtests.

TABLE 9.

Reliability of Subtests

Subtests	Reliability
4	.69
5	.60
6	.33
7	.23
8	.83
9	.01
10	.06
11	.50
12	.53
13	.80
14	.62
15	.35
16	.93

The reliability of a test is affected by a wide variety of causes in addition to the expected psychological influences (e.g. feelings, attitudes and the like). Symonds (25) has listed 25 factors which influence the reliability of tests. Of these factors several have been subject to specific study including the following: 1) the difficulty of the test items, 2) the number of responses in the items of the multiple-response type, 3) practice effect, 4) function fluctuation; 5) variability of the group tested.

In the present study, several of the factors noted above may have affected the results; certain of these will be discussed.

1. Influence of Variability of the Group tested: It is a well-accepted fact that the correlations between two variables are smaller when based upon scores obtained from homogeneous groups. Thus Vernon (32) notes in this connection, "one important reason why intelligence tests appear to be of much less value for predicting scholastic aptitude among secondary than among primary school pupils, and to be poorer still among university students, is that secondary pupils are more homogeneous, or more highly selected than primary ..... correlations necessarily sink as we pass from the unselected children to the primary school, from the primary to the secondary, and from the secondary to the university level."

2. Item Difficulty: Symonds (24) and Thurstone (28) have presented convincing arguments to show that a test made up of items of .5 difficulty value measured an individual most accurately, and that the best test was made up of items that could be answered with 50 percent accuracy by the average individual. Also, the diagnostic value of a test, and its reliability, are a maximum when the items are about 50 percent difficulty. The diagnostic value was found to decrease when the items departed from this 50 percent level.

3. Rānge of Scores: Closely connected to the factor of homogeneity is that of the effective range of scores on each subtest.

Table 10 indicates the ranges of scores within which more than 80 percent of the group fell.

TABLE 10.

Range of Scores containing over 80% of Group

Subtest	r	Range	Effective Range
4	.69	16-33	24-33
5	.60	7-20	12-20
6	.33	7-20	13-20
7	.23	1-11	3-9
8	.83	1-10	3-9
9	.01	3-12	5-10
10	.06	4-14	8-13
11	.50	0-11	3-10
12	.53	1-9	2-7
13	.80	0-15	5-13
14	.62	2-15	5-12
15	.35	7-15	10-14
16	.93	3-33	8-28

The distribution of scores is such that the range in each subtest is materially reduced, thus facilitating the activity of chance factors in seriously reducing the size of the reliability coefficients.

4. Puzzle - Nature of Items: It is apparent in some parts of the test that the items are of such a nature as to make the chance factor the major one operative. Particularly in the case of subtests of low reliability, the correct choice of answers can be disputed by a 16 year old subject. In many cases an incorrect answer can be justified on the basis of his experiences and thought processes. It can be noted also in this connection that for subtests 7, 8, 11 and 12 the mean scores obtained are only slightly better than those

to be expected by chance, thus indicating that the problems ~~passed~~ posED were not all of a type lending themselves to logical solution.

CHAPTER V.

VALIDITY

"The most important, the all-important characteristic of any test is its validity." (12). The validity of a test depends upon the effectiveness with which it measures that which it is intended to measure, or upon the effectiveness with which it accomplishes the purpose it is intended to accomplish. Although one may not be able to define with confidence just what it does measure, in the case of a valid test, one can be sure that it measures something indicative of success or failure. An intelligence test claims to measure intelligence, and therefore it should measure this factor or combination of factors, and the authenticity of a test depends, in turn, not on any theoretical analysis, but upon its relationship to a criterion.

One of the most difficult of all aspects of the validity problem is that of obtaining adequate criteria of what we are measuring, and, if we are dealing with intelligence tests, the intention is almost invariably success in the performance of intellectual tasks as exemplified by achievement in school. Hence the test measures simply in that it predicts performance at such tasks, and from a quantitative viewpoint, a high correlation between a test and its criterion may be taken as evidence of validity, provided both the test and its criterion are reliable. As Guilford points out "From a practical standpoint, the validity of a test is its forecasting efficiency in any measurable aspect of our daily living." (7)

In the present study, scores on the CTMM were correlated with scores obtained by the group in various academic and technical subjects. These correlations are presented in Table 11 below.

TABLE 11.

Correlations Between Test Scores and Academic Scores

	Non Language	English	Science	Math.	Machine Shop	Woodwork
Total IQ		*.036 .49	*.042 .35	*.042 .36		
Language IQ	*.042 .34	*.034 .52	*.042 .36	*.041 .37	*.039 .43	*.023 .74
Non-Language IQ	[illegible]	*.040 .39	*.043 .30	*.044 .28	*.044 .26	*.022 .73
Reasoning			*.032 .56	*.041 .37		
Vocabulary		*.043 .32	*.049 .04	*.047 .15		

$$*\text{P.E.} = .6745 \; \frac{1 - r^2}{\sqrt{N}}$$

Analysis of Table 11 gives some evidence of the validity of the CTMM and certain of its parts as an instrument for predicting success in school subjects -

1. The findings indicate that the language and non-language IQ's are correlated positively. The results in this case are in agreement with those reported by Traxler who found the two correlating about .6. These correlations are lower than those found ordinarily between two mental tests. On the other hand, language and non-language IQ's both correlate quite high with woodwork - .74 and .73 respectively - indicating apparently that either IQ predicts success in woodwork equally well. The language IQ correlates appreciably higher than the non-language IQ with English marks, but, again, it correlates higher also with machine shop scores.
2. In general, the CTMM total IQ's, language IQ's and non-language IQ's are not quite as predictive of success in English, Science and Mathematics as the majority of intelligence tests. Segel(19)/reports a correlation of .63 between average High School marks and scores on the American Council Examination. "The modal correlation between school marks and intelligence test scores are between .4 and .5." (22)
3. The language IQ alone predicts success in English, Science, Mathematics, Machine Shop and Woodwork, as well as either the total or the non-language IQ's.

3. This has important implications, especially in view of the fact that the reliability coefficient of the language scores tends to be higher than for either the total or the non-language factors. It would appear that the test could be reduced to subtests 5, 14, 15 and 16 with no loss in reliability or predictability. This would be of great value to the Busy-School-teacher and administrator, for auch a test would require but 40 minutes instead of the present 90 minutes of testing time. This reduction also seems permissible in view of the low reliability for many of the non-language subtests, and the profiles obtained from the scores probably have little diagnostic significance.

4. The subtests grouped under reasoning appear fairly valid when correlated with Science and Mathematics. This is not unexpected in view of the fairly high reliability coefficients of several of the subtests within this group. The correlation with Science is here appreciably higher than with mathematics.

5. The language IQ where correlated with English is one of the higher correlations obtained, thus indicating the language factors to be among the more valid of the subtests.

In general, with exception of the shopwork subjects, the CTMM cannot be considered to be among the more suitable tests for prediction of scholastic success for the sample cited in the present study.

CHAPTER VI.

CLUSTER ANALYSIS

According to the Manual of Directions, this test "samples the maturity of memory (delayed and immediate); of apperceptive processes; of spatial relationships; and of the logical and mathematical aspects of reasoning." The test is thus constructed to group certain of the subtests according to the particular factor which they are designed to test. Each of the subtests should accordingly conform to the particular pattern of the group. To investigate this possibility, intercorrelations were compiled for the 13 subtests. These appear below in Table 12 along with those reported by the Los Angeles study.

It will be noted that the present coefficients are roughly comparable to those reported by the Los Angeles study, but various exceptions are in evidence particularly where subtests 9 and 10 appear in the table. It will be seen, also, that the majority of the coefficients are fairly low. Of the total 78 intercorrelations 16 were over .30 as compared to the Los Angeles study which had 21 coefficients of .30 and over. In the present study 24 of the coefficients were lower than .10; in the Los Angeles study only one of the coefficients was lower than .10. Generally, then, the coefficients of correlation of the present study can be considered appreciably lower than those of the Los Angeles report.

TABLE 12.

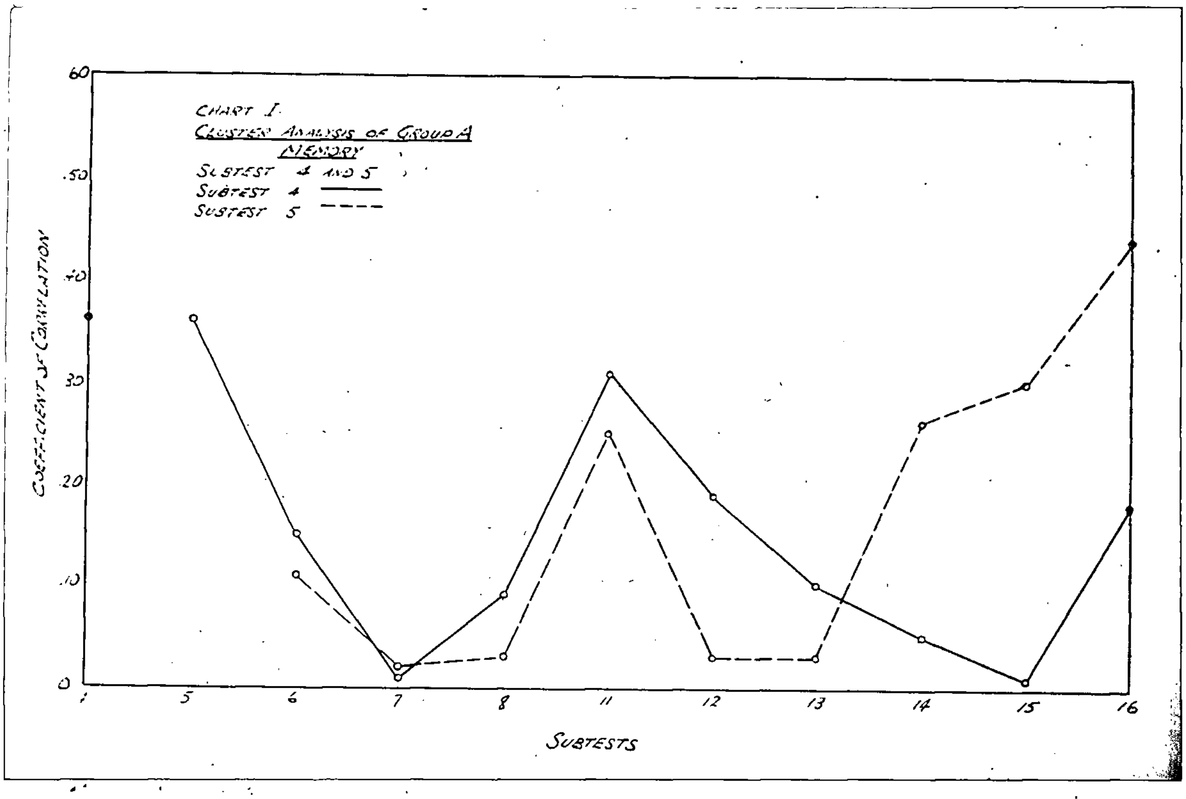
Intercorrelations of Subtests compared with Los Angeles Report

	4	5	6	7	8	9	10	11	12	13	14	15	16
4	**	*.37 .36	.22 .15	.20 .01	.15 .09	.14 .08	.22 .06	.28 .31	.25 .49	.35 .10	.32 .05	.34 .01	.17 .18
5	*** .041		.15 .11	.15 .02	.09 .03	.23 .15	.18 .09	.32 .25	.20 .03	.37 .03	.43 .26	.47 .30	.30 .44
6	.046	.047		.20 .06	.22 .18	.12 .11	.16 .09	.21 .30	.22 .12	.29 .17	.26 .17	.23 .30	.17 .11
7	.047	.047	.047		.21 .14	.23 .13	.18 .08	.25 .18	.17 .04	.26 .13	.26 .12	.23 .15	.16 .06
8	.047	.047	.046	.046		.15 .04	.11 .02	.15 .12	.15 .14	.27 .24	.17 .14	.16 .27	.15 .03
9	.047	.046	.047	.046	.047		.16 .03	.19 .21	.13 .07	.23 .09	.27 .17	.25 .22	.22 .06
10	.047	.047	.047	.047	.041	.046		.18 .21	.18 .32	.21 .14	.22 .11	.29 .57	.13 .16
11	.043	.044	.043	.046	.047	.046	.046		.29 .46	.31 .32	.41 .34	.34 .26	.30 .32
12	.046	.047	.047	.047	.046	.047	.043	.037		.36 .28	.36 .25	.34 .33	.22 .20
13	.047	.047	.046	.047	.044	.047	.046	.043	,044		.56 .42	.43 .34	.33 .09
14	.047	.044	.046	.047	.046	.046	.047	.042	.044	.039		.46 .26	.41 .31
15	.047	.043	.043	.046	.044	.046	.032	.044	.042	.042	.044		.33 .13
16	.046	.036	.047	.047	.047	.047	.046	.043	.046	.042	.043	,047	

* - Los Angeles Study r's
** - Present Study r's
*** - P. E. of obtained r's.

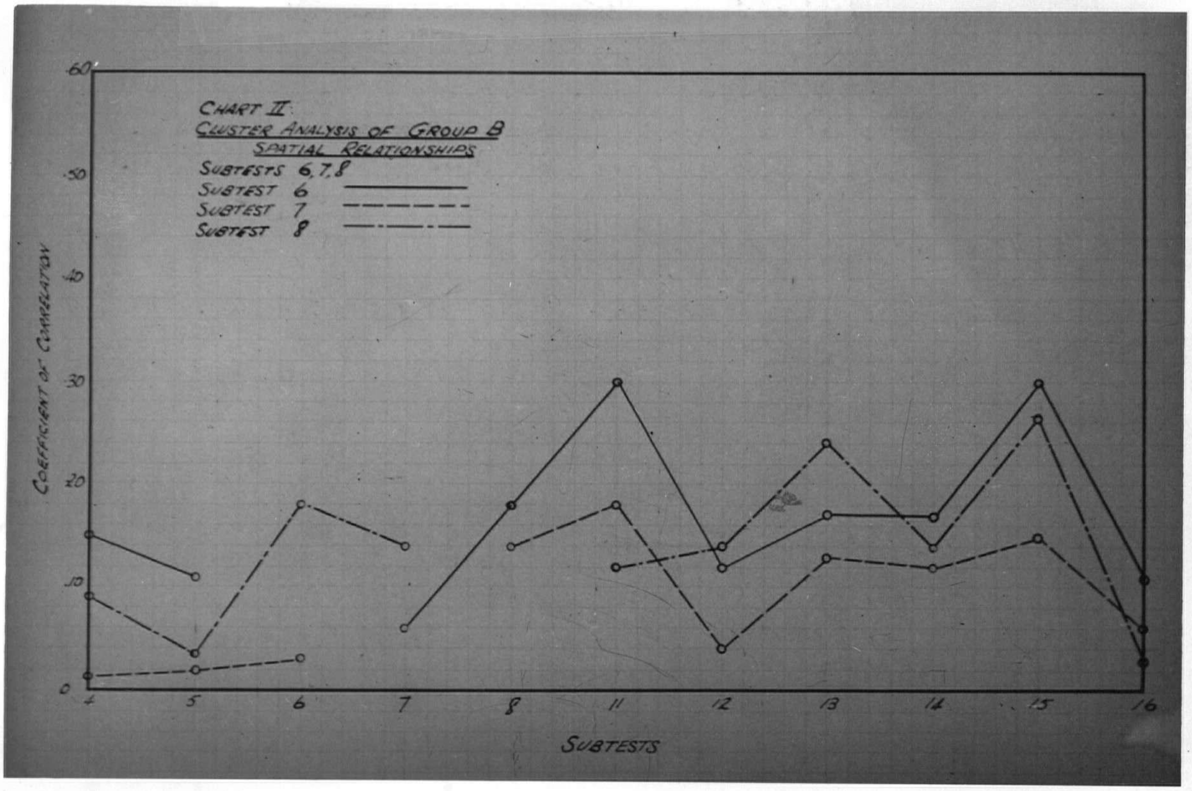
In order to determine the existence of patterns (or families) of factors appearing in the subtests, the intercorrelations were graphed in the group in which they are placed by the test-makers, according to the plan of Tryon (30). The results as grouped appear in Charts 1 to 4 inclusive. In these charts the correlations with subtests 9 and 10 are not included.

CHART 1.



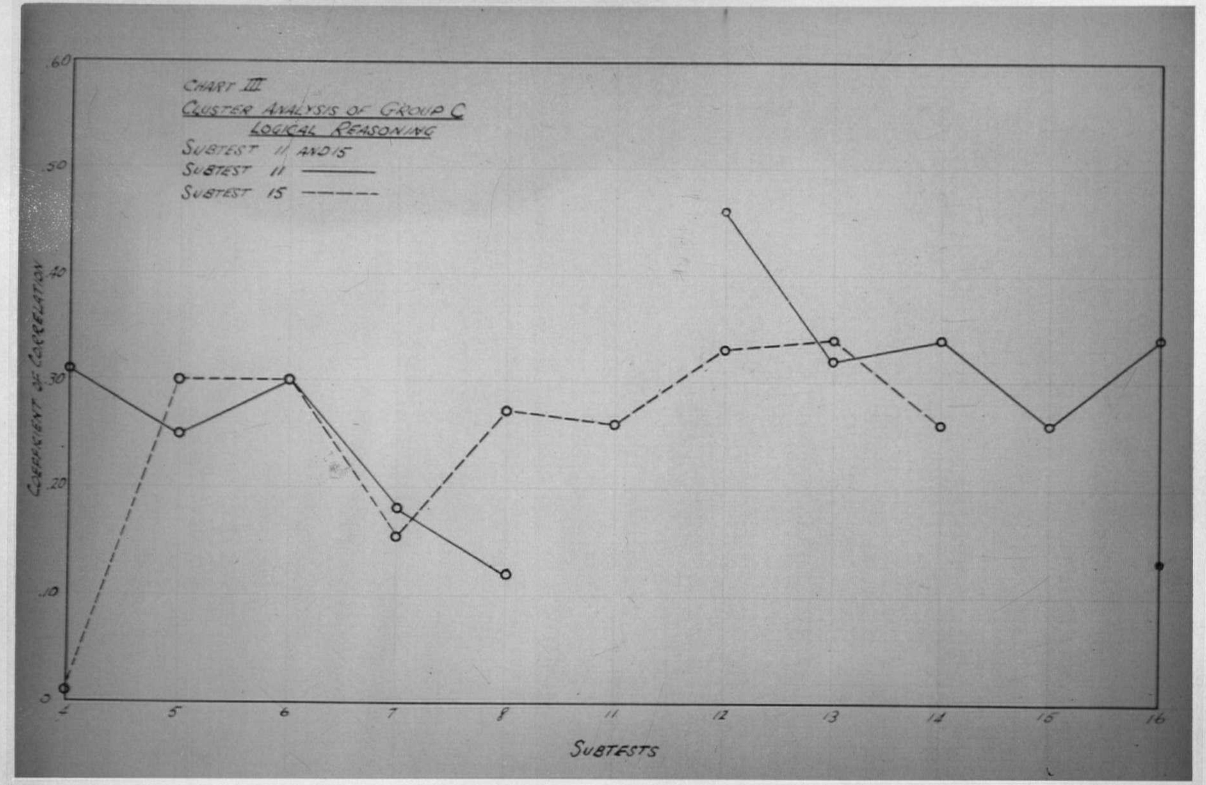
1. Group A (Memory) consists of subtests 4 and 5. It will be seen that the two subtests tend to a similar pattern, particularly when they are correlated with subtests 6, 7, 8, 11, 12 and 13. On the basis of the pattern elicited, the two subtests forming this group can be considered as distinctly related with respect to the above-mentioned tests, but show little tendency to coalesce where correlated with subtests 14, 15 and 16. In the latter case, the tendency is opposite to that exhibited in the first 5 intercorrelations. This pattern was also found when correlations were corrected for alterations.

CHART 2.



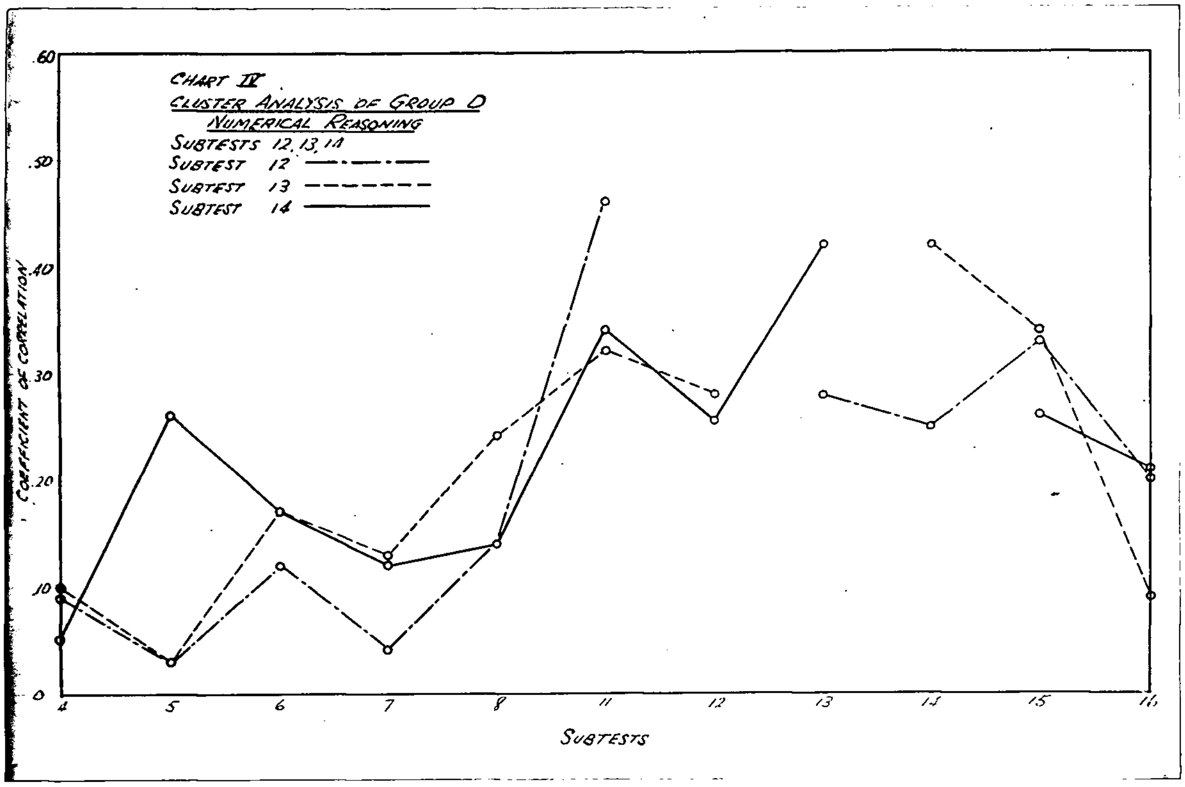
2. Group B (Spatial Relationships) consists of subtests 6, 7 and 8. The three subtests of this group evidently tend to a pattern indicating a certain "family" relationship. On the whole, however, the chart indicates a grouping of the three subtests into a pattern, with some emphasis on homogeneity of content of the three subtests. Analysis of the contents of thesexthree subtests suggests that the factor being measured by this subtest is reasonably entitled Spatial Relationships.

CHART 3.



3. Group C is composed of tests 9, 10, 11 and 15 and is labelled logical reasoning. In the present study, the low reliability coefficients of subtests 9 and 10 did not justify their inclusion in this cluster analysis, and only subtests 11 and 15 are presented in Chart 3. It will be seen that these two subtests do not present an immediately obvious patterning. However, when correlated with subtests 5, 6 and 7, a relationship is observed - a much closer and more evident pattern than that with 4, 12, 14 and 16. Evidently, the pattern is elicited only in subtests 5, 6 and 7.

CHART 4.



4. Group D consists of subtests 12, 13 and 14 and is termed numerical reasoning. A certain homogeneity is not unexpected, and Chart 4 indicates that a "family" relationship exists for the three subtests when they are correlated with subtests 4, 6, 7, 8, 9, 15 and 16. A pattern is not in evidence when the remaining intercorrelations are considered.

5. The profile of correlations of subtest 16 does not fit into any of the above clusters, and so is to be considered a residual, measuring some factors other than those included in the above families.

CHAPTER VII.

SUMMARY AND CONCLUSIONS

Modern practice in respect to mental measurement is characterized by a growing allegiance to the implements of the science coupled with an unfortunate tendency to neglect the proper investigation of the implements in use. Thus a test is published, utilized by many, and criticised without the empirical investigations necessary for its proper assessment. The present study investigates the California Test of Mental Maturity, Advanced Battery, in an attempt at such an assessment.

The test was given to 195 students of the Vancouver Technical School with a view to determining its value for predicting scholastic achievement. The study was divided into analysis of reliability, validity, and principles underlying test construction. These three main divisions were further subdivided in order to answer questions relating to 1) the level and range of difficulty of the individual items of the tests, the order of difficulty of items, the difficulty of the various subtests, the extent to which the individual items differentiated between inferior and superior students; 2) the reliability of the scores for the subtests and for the language and non-language factors; 3) the validity of various divisions of the tests when correlated with academic and Technical shopwork subjects, and 4) the extent to which the test falls into the patterns suggested by the test-makers.

The results as obtained in the study indicate that, for the sample employed, the level of difficulty of the items was not that generally considered as conducive to maximum reliability and diagnosis. The same was true for the general level of difficulty of the subtests. On the basis of correlations computed between obtained order and test order of difficulty, and on the basis of percentages of students attempting all items in a subtest, it was adduced that the test was neither a power nor a speed test, but a composite of the two. The ability of the items to discriminate between inferior and superior groups of students was left seriously in doubt as a result of the computation of an index of discrimination for two of the subtests of low reliability.

Reliability coefficients were computed for the subtests. These were far lower than those recorded in the manual of directions, and, except in the case of the Vocabulary section of the test, were not even roughly comparable. Certain factors influencing test reliability, such as the variability of the group tested, the item difficulty, the narrow range of scores, and the puzzle nature of some of the items were considered as contributing constituting to the low coefficients obtained.

Several of the test scores were correlated with academic and technical school marks. The results indicate that the part of the test measuring language factors was more valid and reliable than either the test as a whole or the non-language factors for predicting success in academic and shopwork subjects.

The groups of subtests comprising 'memory', 'spatial relationships', and 'numerical reasoning' appeared to fall into three clusters: The subtests contained in the 'logical reasoning' group do not illustrate any particular tendency to fall into a pattern.

On the basis of the analysis carried out in the present study, major discrepancies appear between the data here reported and similar data reported in the manual of directions. The manual is deficient in not reporting the results of any item analysis, any statement on level of difficulty of items and nor are the reliability coefficients included in the data. In terms of the recommendation made by Ferguson and Jackson (14), there are many gaps in the manual directions for the CTMM. In general, it appears that the test is not particularly suitable in the Technical High School. Further analysis is necessary to determine its value elsewhere.

BIBLIOGRAPHY

1. Fisher, R.A., Statistical Methods for Research Workers, 7th ed. Edinburgh: Oliver & Boyd: 1938. p. xv +356.

2. Good, C.V., Barr, A.S., and Scates, E.D., The Methodology of Educational Research, New York: D. Appleton-Century: 1938. p. 399.

3. ____________, Ibid., p. 402.

4. Guilford, J.P., Fundamentals of Statistics in Psychology and Education, New York: McGraw-Hill: 1942. p. 289.

5. ____________, Ibid., p. 274.

6. ____________, Ibid., p. 275.

7. ____________, Ibid., p. 285.

8. Harris, W.T., How to Increase Reading Ability, New York: Longmans, Green: 1941. p. 131.

9. Hawkes, H.E., and others, The Construction and Use of Achievement Examinations, Boston: Houghton Mifflin: 1936. P. 31.

10. __________, Ibid., p. 33.

11. __________, Ibid., p. 34.

12. __________, Ibid., p. 21.

13. Hovland, C.I., and Wonderlie, E.F., "Critical Analysis of the Otis Self-Administering Test of Mental Ability High Form." Journal of Applied Psychology, 23, 1939. pp. 367-387.

14. Jackson, R.W.B., and Ferguson, G.A., Studies on Reliability of Tests. Toronto: Department of Educational Research, Bulletin No.12, 1941.

15. Kuhlman, F. (in) Mental Measurements Year Book, New Jersey: Highland Park. 1941. p. 1385.

16. Report of Los Angeles County Superintendent of Schools.

BIBLIOGRAPHY (Continued)

17. Richardson, M.W. and Kuder, D.F., "The Calculation of Test Reliability Coefficients based on the Method of Rational Equivalence". Journal of Educational Psychology, 30, 1939. pp. 681-687.

18. Rulon, P.J. "A Simplified Procedure for Determining the Reliability of a Test by Split-halves". Harvard Educational Review, IX, 1939. pp. 99-103.

19. Segel, David, and others, U.S. Office of Education, Bulletin 15, 1934. p. 74.

20. Spearman, C. "The Proof and Measurement of Association Between Two Things." American Journal of Psychology, 15, 1904. pp. 72-101.

21. ____________ The Abilities of Man: Their Nature and Measurement, London: Macmillan and Co.: 1932.

22. Stroud, J.B., Psychology in Education, Longmans Green: 1946. p. 339.

23. Stuit, P.B., "Current Construction and Evaluation of Intelligence Tests". Review of Educational Research, 11, 1941. pp. 9-24.

24. Symonds, P.M., "Choice of Items for a Test on the Basis of Difficulty". Journal of Educational Psychology, 20, 1929. pp. 481-493.

25. ____________ "Factors Influencing Test Reliability". Journal of Educational Psychology, 19, 1928. pp.73-87.

26. Thirty-seventh and Thirty-eighth Annual Reports of the Vancouver City Schools for the Year ending December 31, 1939, and December 31, 1940, Vancouver, B.C.

27. Thurstone, L.L., The Reliability and Validity of Tests, Ann Arbor: Edwards Bros.: 1933. p. 113.

28. Thurstone, T.G., "The Difficulty of a Test and its Diagnostic Value". Journal of Educational Psychology. 23, 1932. pp. 335-343.

BIBLIOGRAPHY (Continued)

29. Traxler, A.E., "Study of the California Test of Mental Maturity: Advanced Battery". Journal of Educational Research, 32, 1939. pp. 329 -335.

30. Tryon, R.C., Cluster Analysis. Berkeley, California: Associated Students Store: University of California: 1939.

31. Tyler, F.T., "Analysis of the Terman-McNemar Tests of Mental Ability". Educational and Psychological Measurement, 5, 1945. pp. 49-58.

32. Vernon, P.J., The Measurement of Abilities, London: University of London Press: 1940. p. 141.

33. Warren, H.C., Dictionary of Psychology.