The reliability and validity of a number of intelligence tests when applied to approved school boys

Mills, Leslie F.

THE RELIABILITY AND VALIDITY OF A NUMBER OF INTELLIGENCE TESTS WHEN APPLIED TO APPROVED SCHOOL BOYS.

by

LESLIE F. MILLS, B.Sc. (Durham), B.Ed. (Edinburgh).

--oOo--

Thesis presented in fulfilment of the requirements for the degree of Doctor of Philosophy in the University of Durham.

31st. March 1952.

N O T E.

Permission of the Managers of Aycliffe Home Office Approved School, and also of the Home Office Children's Department must be obtained, through the writer, before any part of this thesis may be quoted in print.

ACKNOWLEDGEMENTS.

The writer wishes to make grateful acknowledgement to the Trustees of the Leverhulme Fellowship Research Fund for the financial assistance which made this research possible.

Thanks are also due to the Managers of the Aycliffe Home Office Approved School, County Durham where the investigation was carried out; to the Principal and Staff of Aycliffe School for their support and co-operation throughout the two years during which time the writer was in residence at the School, and to Professor Sir Godfrey H. Thomson who advised on the original planning of the investigation.

In planning and carrying out the research, the writer wishes to acknowledge a special debt of gratitude to Professor E. A. Peel for his continual guidance and inspiration.

—oOo—

C O N T E N T S.

	Page
Introduction.....	viii
I. THE AIM OF THE INVESTIGATION.....	1
II THE TESTS AND ASSESSMENTS USED IN THE INVESTIGATION.....	8
III THE TEST RELIABILITIES.....	19
1. The Reliability Coefficient.....	19
2. The Kuder Richardson Method of Estimating Reliability.....	21
3. The Application of the Technique of Analysis of Variance to the Problem of Estimating Test Reliability.....	25
4. Factors Influencing Test Reliability.....	29
5. The Methods Adopted for Calculating the Reliability of the Tests in the Battery.....	32
6. Comments on the Reliability Coefficients obtained...	37
IV THE VALIDITIES OF THE TESTS.....	43
1. The Meaning of Test Validity.....	43
2. The Importance of General Intelligence.....	44
3. Intelligence as Measured by Tests.....	46
4. Comments on the Tests in the Battery.....	47
5. The Validation of the Tests in the Battery by Factor Analysis.....	50

Contents. continued.

6. The Validities by the Criteria..... 59

(i) The assessment of general intelligence..... 59

(ii) The assessment of practical ability..... 60

V THE READING ABILITY NECESSARY TO PRODUCE A VALID SCORE ON A VERBAL GROUP INTELLIGENCE TEST 68

1. The Findings of Schonell and Mellone..... 68

2. The Present Investigation..... 70

3. An Enquiry into the Structure of Essential Form A... 76

VI THE REVISION OF THE STANFORD-BINET SCALE (1937) FORM L..... 83

VII THE MAXIMUM PREDICTION OF THE CRITERIA..... 90

VIII SUMMARY AND CONCLUSIONS..... 110

1. The Investigation..... 110

2. The Test Reliabilities..... 111

3. The Test Validities..... 113

4. Reading Ability and Verbal Group Tests..... 117

5. The Stanford-Binet Test Form L..... 118

6. The Maximum Prediction of the Criteria..... 119

7. General Conclusions..... 120

Contents TABLES. Page

I	Essential Form A. Reliabilities by Two Methods....	38
II	The Reliability Coefficients and other Relevant Data.....	42
III	The Correlation Coefficients of the Tests and Assessments (Juniors).....	62
IV	The Correlation Coefficients of the Tests and Assessments (Juniors) with Age Constant.....	63
V	The Factor Loadings (Juniors) Unrotated.....	64
VI	The Correlation Coefficients of the Tests and Assessments (Seniors).....	65
VII	The Correlation Coefficients of the Tests and Assessments (Seniors) with Age Constant.....	66
VIII	The Factor Loadings (Seniors) Unrotated.....	67
IX	Description of the Items of the Essential "A" Test	77
X	The Block Item Analysis, Essential A Test.....	80
XI	Rank Order Correlations of the Orders of Difficulty of the Items of the Stanford-Binet Scale (Form L).....	85
XII	The "Easy" and "Hard" Items - The Stanford-Binet Scale Form L.....	89

Contents

HISTOGRAMS. Page.

1. Essential Verbal Group Test Form A..... 73
2. Burt's Reading Accuracy (Vocabulary) Test No.1..... 74
3. New Stanford-Revision of the Binet Scale (1937) Form L..... 75

BIBLIOGRAPHY..... 123

APPENDICES.

I A Representation of Card 1 Aycliffe I..... 127
II The Instructions issued to Housemasters to assist them in making their assessments of general intelligence..... 128
III Composition "A Holiday Adventure"..... 131

--oOo--

INTRODUCTION.

In this country, during the past ten years or so, classifying centres and classifying schools have been set up at various places at the instigation of the Home Office to provide a preliminary sorting out, prior to their transfer to appropriate institutions, of the adult persons and children who are committed by the courts to a period of detention in prisons, borstals or approved schools. The need for the classification of offenders appears now to be recognised as a sound approach to their more enlightened treatment in which emphasis tends to be transferred away from the mere aspect of preventive detention and punishment, to that of training with the ultimate goal of rehabilitation into society always clearly in mind. Classifying centres for juvenile offenders, providing diagnosis of individual problems and specific recommendations for training and treatment were discussed by Burt as early as 1925(1) as being likely developments of the future, and it can be said that the classifying schools now in existence to deal with the continuous flow of juvenile offenders committed to approved schools, function broadly along the lines he visualised. The classifying centres which deal with those persons who are committed to borstals and

---

(1) Burt, Sir. C. "The Young Delinquent". 4th. edition Appendix II pp. 617 - 627.

prisons have adopted procedures and techniques similar to those in use in the classifying schools for child delinquents, though the organisation of the former establishments is somewhat different, as much greater care has to be exercised in maintaining the older offenders in safe custody(1).

The Aycliffe Classifying School for Boys, which was opened in 1943 and which was the first of a series of similar schools to be opened in this country(2), functions as a collecting and dispersing centre for all boys committed to approved schools by the Juvenile Courts of Northumberland, Durham and North Yorkshire. The boys stay for a period up to eight weeks or so in the classifying school, living in small groups with housemasters. The housemasters are responsible for observing the boys' behaviour, for investigating home circumstances when necessary, and for preparing case histories. During this period each boy is tested by psychological and educational tests and is also thoroughly examined by a doctor. The senior boys in addition spend a week in the vocational selection workshop where they are allowed to try their hands at various kinds of practical tasks, and it is

---

(1) Mannheim, H. and Spencer, J. "Problems of Classification in the English Penal and Reformatory System", published by the Institute for the Scientific Treatment of Delinquency in 1950, gives a clear picture of the process of classification in the various types of institutions.

(2) The two other recently opened classifying schools for boys are Redbank, near Liverpool, which covers the North-Western Counties, and Kingswood, near Bristol, which covers the South-Western Counties and South Wales. There are also two classifying schools for girls.

in this workshop that their practical ability is assessed. A certain proportion of the boys, for whom a psychiatric examination appears necessary, are referred to the consultant psychiatrist who visits the school at intervals.

When sufficient information has been gathered about a batch of boys, a meeting is held of all those persons who have handled them, and after discussion each boy is disposed to the type of Approved School where it is considered he will receive the education and training which will best meet his case.

A classifying school thus provides a unique opportunity for the study of juvenile delinquency in its many aspects, and it was during the two years, 1948 and 1949, that the writer was privileged to reside in Aycliffe Classifying School and to carry out the investigation into the use of mental tests with approved school boys, which is the subject of this work.

—oOo—

I - THE AIM OF THE INVESTIGATION.

The process of classification in an establishment such as the Aycliffe Classifying School, consists of bringing together as much information as possible about each child's environmental background, main personality characteristics, attitudes, behaviour, emotional attachments, health and potentialities in schoolroom or workshop, so that from the total picture obtained, a reasonable diagnosis of the causes of the delinquent behaviour can be made. Once the diagnosis has been arrived at, recommendations for treatment automatically follow, and each child is despatched with a full report to the training school where the prescribed treatment can best be carried out. In diagnosing the causes of delinquent and anti-social behaviour in a child and in recommending a specific type of treatment, a great deal depends on the level of intelligence of the child concerned. The accurate assessment of the intelligence of all boys who pass through a classifying school on their way to training schools is therefore of fundamental importance. From the general point of view, a knowledge of the distribution of intelligence in the "population" of boys who are committed to approved schools is necessary in order that educational policy and practice may be shaped along progressive lines. From the individual point of view, an accurate assessment of intelligence enables each boy to

be directed to the kind of education for which he is best fitted; it enables mental deficiency to be detected and dealt with; it helps the educational psychologist to discover those cases whose educational backwardness is not due to poor intellectual ability, and finally, it provides a most important indication of the method of approach with regard to the individual treatment of delinquent behaviour.

Boys who are committed to approved schools for a period of training and education have normally appeared several times before a juvenile court prior to the occasion on which committal was considered necessary. The chief misdemeanours are stealing and breaking and entering and it would be a fair generalisation to say that the majority of these boys, by the time they reach an approved school are well on the way to becoming habitual criminals. Considered as a group they have several characteristics in common. For instance, reports from the ordinary day schools which they attended before committal, show that most of them are chronic truants. This may account to some extent for the fact that nearly all show serious educational backwardness. The homes from which they come tend largely to be "broken" or unsatisfactory homes and it is clear, from a perusal of the case histories, that cultural and socialising influences have been, to a great extent, lacking in the environment in the majority of cases. Inside the approved schools they prove somewhat troublesome and difficult to handle

and present many problems of discipline which do not arise with ordinary schoolboys.

The behaviour characteristics, general educational backwardness and limited cultural background of approved school boys are factors which might well be expected to modify their performances in intelligence tests so as to impair the accuracy of the assessments made. As a group, it would appear that their attention is over easily distracted; that they lack the ability to persevere and that they have a positive distaste for any task which is long and tedious or is similar in nature to school work. In Burt's[1] opinion, the low estimates for general ability among delinquents obtained so repeatedly are due to the influences of the factors outlined above, and he states that "unless special manoeuvres be tactfully tried to circumvent their suspicion, and secure their goodwill, their apparent prowess will fall much below their veritable powers".

With regard to the limitations imposed by educational backwardness and limited cultural background, it would be reasonable to expect that a number of the items appearing in intelligence tests, especially verbal tests, would be invalid. In other words, the boys would be unable to answer the items, not because of lack of intelligence but because of their restricted knowledge and experience. It follows therefore, that tests which have been

---

(1) Burt, Sir C. "Mental and Scholastic Tests" 2nd.Edition Oct.1947 pp.201 - 2.

standardised on a large representative sample of the total population may not provide satisfactory estimates when they are used with certain sections of the community. That tests, in their application to delinquents are not themselves beyond cavil has long been recognised by Burt[1]. For instance, in the New Revision of the Stanford-Binet Scale (1937), already heavily criticised by him[2] on the grounds that "the standardisation of each problem in terms of mental age assumes that the order of difficulty is constant for the two sexes, for different types of children and above all for different localities", the whole principle of age assignment to problems would appear to make the test unsound in its use with delinquent boys whose educational attainment is on the average two years below mental age and whose environmental experience and opportunity has been sadly limited. More recently, Blackburn[3] has stated that "the principal source of error in intelligence testing is the influence of different social environments upon test scores".

The evidence presented in the foregoing discussion would seem to lead clearly to the conclusion that there are difficulties in the way of obtaining accurate assessments, by means of

---

(1) Burt, Sir. C. "Mental and Scholastic Tests". 2nd. Edition Oct. 1947 p.198.

(2) Burt, Sir. C. "The Latest Revision of the Binet Intelligence Tests", The Eugenics Review XXX 1939 pp. 255-260.

(3) Blackburn, J. "The Influence of Social Environment on Intelligence Test Scores". 1948, British Social Hygiene Council.

intelligence tests, of the innate ability of delinquent boys such as pass through a classifying school. Firstly, subjective opinion suggests that the attitude of such boys to both test materials and the test situation would, in general, be one of mistrust and non-co-operation, and secondly, because of their peculiar limitations of environmental experience and education, intelligence tests, standardised on a representative sample of the total population are not entirely satisfactory as a means of measuring their innate mental ability.

After spending several months administering a variety of tests to numerous boys who were in transit through Aycliffe Classifying School, the writer decided that a planned investigation into the reliabilities and validities of a selection of these tests in relation to their use with approved school boys would provide, not only a mass of information of general psychological interest, but also information of practical value regarding the most suitable tests and techniques for use in classifying schools and other similar institutions(1). Accordingly, a battery of tests was prepared and put into operation in September 1948 with the assistance of the Classifying School teaching staff. The

---

(1) Approved Schools, of which there are some 170 in this country, present hitherto an almost untouched field for research into the problems of juvenile delinquency. The writer, for example, was probably the first psychologist to be appointed to carry out, in an approved school, an investigation of a psychological nature.

intake to the school at this time averaged ten boys per week and it was found that this number formed a convenient unit to deal with from the point of view of a weekly testing programme. The battery was continued in use until the end of May 1949 and during the period of eight and a half months, 327 consecutive entrants to the classifying school were tested by it. In addition, subjective assessments of general intelligence were made by the housemasters, and in the case of the older boys, subjective assessments of practical ability were made by the instructor in charge of the vocational selection workshop.

The 327 boys tested were considered to be fairly representative of the boys who pass through Aycliffe Classifying School. The range of chronological age was from 9 to 17 years and as this was considered to be rather wide, in the statistical treatment of the data, the sample was divided arbitrarily into "Juniors" and "Seniors", the line of demarcation being taken at 14 years 6 months.

The Battery of Tests and Assessments.

I Verbal

1. Essential Form A.
2. Essential Form B.
3. Simplex.
4. Stanford-Binet Form L.

II	Non-Verbal	5. Progressive Matrices.
		6. N.I.I.P. 70/23.
		7. V.S.10.
		8. T.S. 8.
		9. Aycliffe I.
III	Performance	10. Passalong.
		11. Kohs' Blocks.
		12. Cube Construction.
		13. Blocks Performance.
IV	Educational	14. Dictation.
		15. Composition.
		16. Arithmetic.
V	Assessments	17. General Intelligence.
		18. Practical Ability (Seniors only).

--oOo--

II - THE TESTS AND ASSESSMENTS USED IN THE INVESTIGATION.

VERBAL GROUP TESTS.

1. The Essential Intelligence Test.

Prepared and standardised by F.J. Schonell and R.H. Adams. This test, which was prepared and standardised some years ago is now referred to as Form A of the test, as the same authors published a parallel form known as Form B in 1948. It contains 100 items and is designed for use with children between the ages of 7 and 12 years. A time limit of 45 minutes is allowed to complete the test. A short practice test is given on the back of each booklet. Published by Oliver and Boyd.

2. The Essential Intelligence Test, Form B.

Mentioned in note 1 above.

3. The Simplex Junior Intelligence Scale.

Prepared and standardised by C.A. Richardson. Like the two Essential tests it contains 100 items. A time limit of 45 minutes is allowed. No practice test is provided. Published by G. Harrap and Co. Ltd.

4. The New Revised Stanford-Binet (1937).

By L.M. Terman and M.A. Merrill. Form L of this scale only was used. Published by G. Harrap and Co. Ltd.

NON-VERBAL GROUP TESTS.

5. Progressive Matrices (1938).

Prepared and standardised by J.C. Raven. This test contains 60 items, each printed on quarto size paper, the whole forming a booklet. Each item consists of a design or group of figures with a part missing. By observing the relationships which exist between the various parts of the design or figures forming the group, it is possible to select from a number of pieces at the bottom of the page, one which will correctly complete the "matrix". There are 5 sets of 12 items in the test labelled A,B,C,D and E respectively. Each set develops a different theme. The initial items in each set are easy enough to be self evident, the others follow on becoming increasingly difficult. The test is suitable for children and adults. Although no time limit is made, the time taken for each individual to complete the test must be noted. It is essentially a test of an individual's capacity to form comparisons and reason by analogy. Published by H.K. Lewis and Co. Ltd.

6. National Institute of Industrial Psychology Group Test 70/23.

This non-verbal group test was designed by Slater[1] in 1941. It is composed of 2 sub-tests as follows:-

---

(1) The construction of Group Test 70/23 is described by Slater in "Tests for Selecting Secondary and Technical School Children". Occupational Psychology (1941) XV (1), 10.

Part I : 25 items, time limit 5 minutes. Each item requires an analogy to be completed.

Part II : 28 items, time limit 8 minutes. Each item requires the arrangement of 5 components so that they shall be put in series order.

Practice examples are worked before both parts are attempted. Group Test 70/23 is obtainable from N.I.I.P.

7 & 8. Vocational Selection No.10 & Technical Selection No.8.

These two tests were designed by Peel(1) and are known briefly as V.S.10 and T.S.8. The same kinds of items are used in both tests and these are:-

(i) Pairs of patterns marked "A" and "B". "A" is the correct pattern, "B" is identical except for a small mistake. By comparing "B" with "A" the incorrect or incomplete portion of "B" must be discovered and a cross placed on the spot where the mistake occurs.

(ii) Pairs of shapes or designs which are mirror images. These are marked "A" and "B". A mistake has been made purposely in "B" and a cross must be placed on the exact spot where the mistake occurs.

(iii) Repetitive patterns. A mistake has been made purposely in the repetition of each pattern. A cross has to be

---

(1) A description of the items used in V.S.10 and T.S.8 together with illustrations is given by Professor E.A. Peel in "Evidence of a Practical Factor at the Age of Eleven", B.J.Ed.P. XIX Part I. Feb. 1949. p.6. In this article he refers to these items as belonging to an earlier test which he designed and which he calls T.G.T.

placed on the exact spot where the mistake in repetition occurs.

Each of the tests is divided into three sections and each section has a 5 minute time limit. The tests are set out as follows:-

V.S.10.	1.	9 items - pairs of patterns	- 5 minutes.
	2.	10 items - mirror images	- 5 minutes.
	3.	10 items - repetitive patterns	- 5 minutes.
T.S. 8.	1.	10 items - repetitive patterns	- 5 minutes.
	2.	12 items - pairs of patterns	- 5 minutes.
	3.	10 items - repetitive patterns	- 5 minutes.

A practice example is given before each section is attempted. Both of these tests are obtainable from Professor E.A. Peel, Birmingham University.

9. Aycliffe I. (Spatial judgment).

This test, designed by the writer, is composed of 30 items which are presented 6 at a time on large sheets of white cardboard. The items are of three types and 4 alternative solutions are offered in each case. Items 1 to 12 (Cards I and II) require the mirror images of given shapes to be found; items 13 to 18 (CardIII) deal with finding the shape which is the same as a given shape except that it has been rotated through an angle; items 19 to 30 (Cards IV and V) deal with finding the shapes which are not only the mirror images but which have been rotated through an angle as well.

At the top of each card is a practice item which is explained and demonstrated by means of a cardboard model before the items on the card are attempted. No time limit is made and the supervisor waits until everyone has finished a card before putting up the next one. The four responses offered for each item are lettered a, b, c and d, and the method of answering is to select the solution considered correct and write down its letter on the mark sheet provided.

The method of presentation was devised to ensure that interest would be sustained throughout the whole test and that the boys would have a clear presentation of the problems to be solved at regular intervals. It is desirable that the test should not be given to more than 8 boys at one time so that all may have a good view of the card set up before them.

The 30 items comprising the test were selected after a number of experiments had been carried out and an answer pattern made from the scores obtained by a sample of approved school boys. A representation of Card I is given in Appendix I.

PERFORMANCE TESTS.

10, 11 and 12. Alexander's Performance Scale.

This scale, which comprises three individual performance tests, was devised and standardised by W.P. Alexander. The tests making up the battery are Passalong, Kohs' Blocks and Cube Construction, and together they take about 40 minutes to administer

to one individual. Each of the three tests is made up of a number of items which are carefully graded in order of difficulty. The scoring is done by taking the time in seconds required to complete each item successfully, and then to transform this time to a score by means of a standardised table provided by Dr. Alexander. The total score for each of the three tests is obtained by adding together the marks obtained in the respective items.

The material for the performance scale is obtainable from Thomas Nelson and Sons Ltd.

13. Peel's Practical Abilities Test.

This practical abilities test designed by Peel[1] is composed of two sub-tests, P.T.1 and P.T.2, the former containing 14 items and the latter 15 items. The principle involved in both sub-tests is the same. In each item of the sub-tests four or five irregular shaped blocks are set out before the individual who is being tested. He must select two of the blocks and assemble them on the table so that they match, in the case of P.T.1, a half scale model made of concrete, and in the case of P.T.2, a perspective picture of the model required. Thus sub-test P.T.1 consists of 14 sets of 4 or 5 blocks of wood with a small concret model beside each set, and sub-test P.T.2 consists of 15 sets of 4 or 5 blocks of wood with a picture

---

(1) Peel, E.A. B.J.Ed.P. XIX Part I Feb.1949 pp. 5 and 6. A description of P.T.1. and P.T.2 is given together with photographs of one of the items from P.T.1. The version of P.T.1 used by the writer contained two fewer items than the version originally used by Professor Peel.

(about postcard size) beside each set. The items are laid out on tables in order of difficulty and the individual doing the test progresses without loss of time from one item to the next, leaving his completed assemblies on the table to be marked by the supervisor.

To ensure that the individual tested shall understand what is required of him, two practice demonstration items are given before each sub-test.

The time limit set by Professor Peel, for technical school students, was 9 minutes but this was found to be too short for approved school boys. Eventually, after some experiments, a time limit of 12 minutes was fixed for each sub-test. The score obtained in each sub-test is, of course, the number of correct solutions in the time limit.

The test material may be borrowed from Professor Peel for research purposes.

ATTAINMENT TESTS.

14. Burt's Test No.7. Dictation (Continuous Graded Test).(1)

The set piece devised by Burt is made up of 97 words which make a total of 500 letters. The words which increase steadily in difficulty, are strung together in phrases and sentences which carry some degree of meaning. Burt's method of marking

---

(1) Burt, Sir.C. "Mental and Scholastic Tests". 2nd. Edition Oct. 1947. Test Material Appendix I. p. 383.

the test is to give one mark for every letter correct. The writer, however, adopted the method of giving one mark for each word completely correct.(1)

15. Composition.

The conditions for writing the composition were standardised as far as possible by having a set piece printed at the head of each boy's paper. The piece provides the opening sentences of a story which has to be completed in 15 minutes.

In the case of the sample of boys tested at Aycliffe School, the 'marking' was done subjectively by the writer by reading through all the compositions and sorting them into grades according to the quality of their content. Fertility of ideas and fluency of expression were the main bases of the gradings, inaccurate spelling, lack or misuse of punctuation and ungrammatical construction were more or less ignored. After several readings, 13 grades were arrived at and each composition was then given a mark corresponding to its grading on a 13 point scale. (See Appendix III).

16. Mechanical Arithmetic.

This test was composed of 40 small sums (12 of them oral mental arithmetic) which covered fairly adequately the addition, subtraction, multiplication and division of numbers and money. One mark was given for each correct answer.

---

(1) In the sample of approved school boys tested, the writer's method of marking gave a symmetrical distribution of scores which was more acceptable for calculating correlations than the distribution obtained by Burt's method which gave a distribution heavily bunched towards the "top" end.

SUBJECTIVE ASSESSMENTS.

17. General Intelligence.

All boys who were tested by the battery of tests described above, were also subjectively assessed for general intelligence by their housemasters. The assessments were made on a 15 point scale as shown below:-

MD	Very dull		Below average						Above average		Superior		Very superior indeed.
-7	-6	-5	-4	-3	-2	-1	0	±	2	3	4	5	6	7
--------Below---------					------Average---					---------Above----------				

The four housemasters of the Aycliffe Classifying School took part in the experiment and made assessments for the boys who passed through their respective houses. Thus, each housemaster contributed approximately one quarter of the total assessments made.

It was hoped that these subjective assessments of general intelligence would provide some sort of criterion for estimating the validity of the tests composing the battery. That these assessments themselves would be doubtful quantities was fully realised. Nevertheless, the amount of agreement or otherwise which would be found to exist between the objectively measured test scores and the assessments was considered to be of sufficient interest to warrant them being carried out.

A copy of the written instructions and suggestions issued to the housemasters to help them in making the assessments is shown in Appendix II.

18.. Practical Ability.

The subjective assessments of practical ability were made for senior boys only and were graded on a 15 point scale, identical with that used for general intelligence, except that in the case of Practical Ability gradings, -5 and -6 were labelled "very poor" and -7 was labelled "no ability whatsoever".

The assessments were made by the instructor in charge of the vocational selection workshop.

The purpose of the assessments was to provide a set of gradings for comparison with the scores on those tests in the battery which purported to measure practical ability. It was considered that as the assessments were made after the boys had been observed closely for the period of a week while attempting various practical tasks in the workshop they would provide a reasonably satisfactory criterion.

In addition to the above tests, an estimate of each boy's reading age was obtained. Burt's Test No.1. Reading (Accuracy)[1] was used for this purpose. This test consists of 110 words which are arranged in order of increasing difficulty. To each age from 4 to 14, 10 words are assigned. The child tested has to read aloud the words in succession till he can read no more.

---

(1) Burt, Sir. C. "Mental and Scholastic Tests" 2nd. Edition October 1947. Test Material Appendix I. pp. 367 - 9.

His score, which is the total number of words pronounced correctly, can readily be converted to a reading age by the formula -

$$\text{Reading age} = \left(4 + \frac{\text{Words}}{10}\right) \text{ years.}$$

—oOo—

III - THE TEST RELIABILITIES.

1. The Reliability Coefficient.

The scores obtained in psychological tests are always affected by what are known as errors of measurement. No matter how carefully a test is constructed and standardised, or how carefully the conditions of administration are controlled, it is inevitable that the scores obtained will contain a certain amount of error. This means, in other words, that if a group of persons produce scores on a test, then it can never be certain that these scores represent either their true ability or what they would obtain if it were possible for them to do the test again under comparable conditions.

By applying a test on two occasions under similar conditions to the same group of individuals two sets of scores are obtained. If there is a high degree of correspondence between these two sets of scores the test is said to be reliable and to measure with an acceptable degree of accuracy the function or ability which it purports to measure. If, on the other hand, there is little correspondence between the sets of scores, the test is said to be unreliable and little faith can be put in the scores obtained on it. By calculating the correlation between the sets of scores obtained by two

applications of a test the amount of agreement between them can be expressed mathematically, the figure obtained being called the Reliability Coefficient.

In practice, repetition of the test is not always a very satisfactory method of estimating the reliability coefficient. If the time interval between the two applications is too short, the answers to the items are remembered and the correlation will tend to be higher than it should be. If, on the other hand, the time interval is too long, factors of growth and learning and other causes will operate and the correlation will tend to be reduced. Most of these difficulties can be overcome by carrying out the re-test with a parallel form of the original test. This method, however, is limited in its usefulness since few tests are published with parallel forms. When a parallel form is not available the "Split-half" method can be adopted. This has the advantage that the reliability coefficient can be calculated from a single application of the test to a group of individuals. The procedure in this method is to administer the test to a group in the normal manner and then to record as separate totals the number of marks obtained by each individual in the odd and even items. The correlation between the two sets of marks thus obtained is the reliability coefficient of a test half as long as the original test. Since reliability depends on the length of a test, a correction must be applied to obtain

the reliability coefficient of the test as a whole. This is done by the Spearman-Brown formula as follows:-

$$R = \frac{2r}{1 + r}$$

where r is the correlation obtained from the split-halves, and R is the correlation to be expected had it been possible to compare the whole of the test with another similar test.

The conditions under which the reliability coefficient of a test is estimated by the three methods outlined above, are different in each case and the resulting estimates of reliability are of course different. The split-half method probably gives the coefficient of reliability nearest the true value, though in actual fact it is very likely too high.

2. The Kuder Richardson Method of Estimating Test Reliability.

A rather different approach to the problem of estimating test reliability was made by Kuder and Richardson[1] in 1937. Their method, which makes use of data normally required in item analysis, provides another method of calculating the reliability coefficient from a single application of a test to a group of individuals.

A series of formulae are derived (the best known perhaps

---

(1) Kuder, G.F. and Richardson, M.W. "The Theory of the Estimation of Test Reliability". Psychometrika II (1937).

being Formula 20) from what Burt considers to be a rather formidable and highly speculative set of assumptions[1]. The reliability coefficient is defined as "the correlation between one experimental form of a test and an hypothetically equivalent form".

Two tests (or two forms of a test) are defined as being equivalent when corresponding items in either test

(i) have the same difficulty

(ii) have the same correlation with each other as they have with themselves

(iii) have the same correlation with all other corresponding items,

(iv) and are, in fact, generally interchangeable.

It is then assumed that for all practical purposes all the inter-item correlations may be taken as approximately equal to the average item self correlations, and finally, that their standard deviations are approximately equal.

Formula 20 as derived by Kuder and Richardson gives the reliability coefficient of a test of n items as

$$r_{tt'} = \frac{n}{(n-1)} \left\{ \frac{\sigma_t^2 - n\overline{p_i q_i}}{\sigma_t^2} \right\}$$

---

(1) Burt, Sir.C. "The Reliability of Teachers' Assessments of Their Pupils". B.J.Ed.P. Vol.15. 1945 pp. 80-92.

where σ_t^2 = the variance of the test as a whole and

$\overline{p_j q_j}$ = the mean variance of the n items.

The Kuder Richardson formulae, because of the rather artificial and restrictive nature of the basic assumptions have been severely criticised and Kelley(1) considers that in their final form they are utterly suspect and seems disposed to reject them altogether. Ferguson(2) has also derived the Kuder Richardson formula, adopting similar yet less restrictive assumptions and invoking as an alternative final postulate that the "average inter-item covariance" may be taken as being equal to what may be called the "average item self-covariance". Burt(3) however, considers that it would be as difficult to demonstrate for any given test that the component items obey Ferguson's alternative requirement as to prove that they conform to those laid down by Kuder and Richardson.

The theoretical "weaknesses" involved in the derivation of these formulae have, however, been no deterrent to their practical application and most test constructors have made use of them at one time or another in calculating test reliabilities.

---

(1) Kelley, T.L. "The Reliability Coefficient". 1942 Psychometrika VII p.81.

(2) Ferguson, G.A. "The Reliability of Mental Tests". 1940 p.31.

(3) Burt, Sir C. "The Reliability of Teachers Assessments of their Pupils". B.J.Ed.P. Vol.15 1945 pp. 80 - 92.

More recently Gulliksen[1] has derived the Kuder Richardson formulae by starting with two tests that are parallel item for item and by assuming that the average covariance among non-parallel items is equal to the average covariance among parallel items. This assumption which is the same as Ferguson's final postulate, stated mathematically means that

$\overline{r_{ij}\sigma_i\sigma_j}$ is taken to be equal to $\overline{r_{ii}\ \sigma_i^2}$

Lawley[2] has shown however that

$\overline{r_{ij}\sigma_i\ \sigma_j}$ is always $\leqslant \overline{r_{ii}\ \sigma_i^2}$

which means that estimates of reliability by the Kuder Richardson formulae are always likely to be under-estimates.

Gulliksen points out that Guttman,[3] who has presented a theory of reliability in terms of estimation of "lower bounds" for reliability, has derived a formula (L_3) which is identical with Formula 20 of Kuder and Richardson, and which in Guttman's opinion gives a lower bound or under-estimate of reliability.

---

(1) Gulliksen, H. "Theory of Mental Tests" (Chapman and Hall Ltd. 1950) pp. 221 - 227.

(2) An unpublished memorandum (19th.Aug.1945) at Moray House Room 70, Edinburgh, by D.N. Lawley.

(3) Guttman, L. "A Basis for Analysing Test-retest Reliability". Psychometrika X (1945) pp.255 - 282.

3. The Application of the Technique of Analysis of Variance to the Problem of Estimating Test Reliability.

A simple but fundamental principle involved in the study of test reliability is the concept that persons have "true" scores from which their actual scores deviate, due to errors of measurement. The variance of the empirical scores can be analysed, therefore, into two components, that due to the true marks and that due to errors of measurement.

If the variance of the actual scores be denoted by σ_t^2

then $\sigma_t^2 = \sigma_g^2 + \sigma_e^2$

where σ_g^2 = the variance of the true scores

and σ_e^2 = the variance due to errors of measurement.

The correlation between the true marks g and actual marks t is equal to the ratio of the standard deviation of the true marks and the actual marks, thus

$$r_{tg} = \text{the index of reliability} = \frac{\sigma_g}{\sigma_t} = \sqrt{\frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2}} \quad (1)(2)$$

If we imagine the test to be attempted by the same group of persons on two precisely equivalent occasions, the assumption being that the proportionate disturbance due to error will be the

---

(1) Guildford, J.P. "Psychometric Methods". Mc.Graw-Hill Publications 1936, p. 304 and p. 413.

(2) Gulliksen, H. "Theory of Mental Tests". Chapman and Hall Ltd. 1950 p. 23.

same on both presentations, then, by the product theorem,

$$r_{tt'} = \text{the reliability coefficient} = r_{tg} \times r_{t'g} = \frac{\sigma^2_g}{\sigma^2_g + \sigma^2_e} \quad (1)$$

This is another definition of reliability and as Burt has shown(2), it opens the way for the application of the technique of analysis of variance to the problem of estimating test reliability. Hoyt(3) and Jackson and Ferguson(4) have also discussed the use of analysis of variance in this field but as Burt deduces formulae which have already been derived from quite different premises, his theoretical discussion is given below:-

Let there be N persons assessed by n sub-tests or test items (referred to as tests).

Let X_{ij} denote the i^{th} person's raw score in the j^{th} test.

Let x_{ij} be the same reduced to deviation form.

The basic assumption is that X_{ij} (and therefore x_{ij}) may be analysed as the unweighted sum of three varying components due to

(i) the person tested

(ii) the test used

(iii) a random error

---

(1) Hartog, Rhodes and Burt. "The Marks of Examiners". 1936 Memorandum I p. 278.

(2) Burt, Sir C. "The Reliability of Teachers' Assessments of Their Pupils". B.J.Ed.P. Vol.15 1945 pp.80 - 92.

(3) Hoyt, C. "Test Reliability obtained by Analysis of Variance". 1941 Psychometrika, 6, pp. 153 - 160.

(4) Jackson, R.W.B. and Ferguson, G.A. "Studies on the Reliability of Tests". Bulletin No.12 Dept. of Ed.Research, Toronto University.

This is expressed as

$$x_{ij} = X_{ij} - A = p_i + t_j + e_{ij}$$

where A is the average of all the raw scores,

p_i the i^{th} person's average mark in the n tests,

t_j the average mark of the N persons in the j^{th} test, and

e_{ij} the error of measurement

Squaring both sides of the above equation, the total sum of squares can be split into three components each consisting of N x n squares

$$\sum\sum x_{ij}^2 = n\sum p_i^2 + N\sum t_j^2 + nN \sum\sum e_{ij}^2$$

or $S = P + T + E$

To find the variances, the square sums are divided by the corresponding degrees of freedom

$$\bar{P} = \frac{P}{(N-1)} \qquad \bar{T} = \frac{T}{(n-1)} \qquad \bar{E} = \frac{E}{(n-1)(N-1)}$$

If t_{gi} denotes the "true value" of i's total mark

then $\bar{P} = \sigma_g^2 + \sigma_e^2$ and $\bar{E} = \sigma_e^2$.

Thus $$r_{tt'} = \frac{\sigma_g^2}{\sigma_g^2 + \sigma_e^2} = \frac{\bar{P} - \bar{E}}{\bar{P}}$$

$$= \frac{nP - (S-T)}{(n-1)\,P}$$

$$= \frac{n}{n-1}\left\{1 - \frac{(S-T)}{nP}\right\} \qquad \text{.....1.}$$

Equation 1 above provides a speedy method of calculating the reliability coefficient of a test which is composed of sub-tests each of which has a range of marks. Burt points out that the fundamental assumption is that the standard deviations of the sub-tests do not differ significantly. This condition is not likely to be attained, however, Burt considers that even wide discrepancies in this respect entail no serious difficulty.

In tests composed of items instead of sub-tests, and where the items are given marks 1 or 0, T = 0 and Equation 1 reduces to

$$r_{tt'} = \frac{n}{n-1}\left\{1 - \frac{S}{nP}\right\}$$

$$= \frac{n}{n-1}\left\{1 - \frac{\sum \sigma_j^2}{\sigma_t^2}\right\} \qquad \text{.....2.}$$

where σ_j^2 = the variance of the j^{th} test item.

It can be shown that $\sigma_j^2 = p_j\, q_j$

where p_j = the proportion of testees answering the item correctly

and $q_j = 1 - p_j$

Hence equation 2 becomes

$$r_{tt'} = \frac{n}{n-1}\left\{1 - \frac{\sum p_j q_j}{\sigma_t^2}\right\}$$

$$= \frac{n}{n-1}\left\{\frac{\sigma_t^2 - n\overline{p_j q_j}}{\sigma_t^2}\right\} \qquad \ldots\ldots\ldots\ldots 3.$$

Equation 3 is identical with Formula 20 of Kuder and Richardson.

4. Factors Influencing Test Reliability.

Guildford[1] gives a list of 22 factors which affect test reliability. Most of the factors he enumerates can be regarded as extremely useful points to be borne in mind in constructing and administering tests. Three of them, however, are of more fundamental importance and are worthy of special mention. They are - the length of the test; the degree of heterogeneity of the group whose scores provide the reliability coefficient, and, function fluctuation. Vernon[2] also considers these three factors to be of primary importance.

(i) The Length of the Test.

The reliability of a test depends to a very large extent on its length. The longer the test, that is the more items it

---

(1) Guildford, J.P. "Psychometric Methods. (McGraw-Hill Publications 1936). pp. 417 - 418

(2) Vernon, P.E. "The Measurement of Abilities". (U.L.P.Ltd., 1940) pp. 145 - 149.

contains, the more reliable it will be. Theoretically it is possible to increase the reliability of a test almost indefinitely by adding more and more items, but of course, practical considerations prevent this being done.

If a test has a reliability coefficient r, then it is often desirable to find out how much longer the test must be to raise the reliability coefficient to a more acceptable level. This is done by applying the general form of the Spearman-Brown Formula.

$$R = \frac{nr}{1 + (n-1)\,r}$$

where n is the number of times the test must be lengthened.

(ii) Group Heterogeneity.

Scores obtained from a group with a wide range of chronological age and ability will give a higher reliability coefficient than would be obtained from the scores of a group with a more restricted range ability. For this reason it is desirable to relate a reliability coefficient to some standard group of individuals, and it is now accepted by most authorities that a single year group, that is all children whose ages range over one year (or a representative sample of such a group), is the most reasonable to use for this

purpose. This fact must be borne in mind when interpreting the reliability coefficients obtained in the present investigation, as the range of age and ability in both Senior and Junior groups was considerable.

(iii) Function Fluctuation.

When a group of individuals is retested with either the same test or a parallel form, part of the discrepancy between the two sets of test scores is due to the level of achievement of the individuals having changed between the two applications of the test. This variation of performance between test and retest is called function fluctuation.

Naturally, function fluctuation has a greater effect on reliability coefficients obtained by retesting or by applying a parallel form than on those obtained by the split half method or by the Kuder-Richardson formula. For any particular test, the difference between the reliability coefficients calculated by the former and latter methods serves to indicate the tendency of the group concerned to fluctuate with regard to the function or ability measured by the test. Thouless(1) has provided a formula for estimating function fluctuation from this difference.

---

(1) Thouless, R.H. "Test Unreliability and Function Fluctuation". B.J.P. 1936 XXV pp. 325 - 343.

5. The Methods Adopted for Calculating the Reliability Coefficients of the Tests in the Battery.

In the preceeding discussion it has been shown that the various methods available for estimating reliability coefficients are based on different assumptions and therefore give different results. As the aim of the investigation was to determine the extent to which intelligence tests are reliable when used with approved school boys, it was considered that, if possible, one method should be used for all the tests in the battery as this would permit valid comparisons to be made between the coefficients obtained for the different tests.

From this standpoint -

(i) the test-retest method could not be entertained seriously for general application to the battery, firstly, because of the amount of labour involved, and secondly, because of the varying and unpredictable influences of practice effect and function fluctuation.

(ii) The parallel form method could be used only with Essential A since parallel forms are not available for the other tests.

(iii) The split-half method could be applied only to Progressive Matrices; Aycliffe I and Kohs' Blocks, as the other tests do not lend themselves to being split into equivalent halves.

It was therefore decided to use the Kuder-Richardson method

since it could be applied to the majority of the tests without great difficulty.

The reliability coefficients of the following tests were calculated by means of Formula 20 -

Essential A.
Stanford-Binet Form L.
Progressive Matrices.
N.I.I.P. 70/23.
V.S.10.
T.S. 8.
Aycliffe I.
Blocks Performance.

For Passalong and Kohs' Blocks, however, the split-half method was used. In both these tests the items are really small sub-tests and in the case of Kohs' Blocks the total scores made by the individuals in the group on items 1, 3, 5, 7 and 9, were correlated with the total scores on items 2, 4, 6, 8 and 10. The Spearman-Brown formula was used to boost the correlation obtained to what would be expected for the full-length test. In the Passalong there are only nine sub-tests and in this case, the total scores on items 1, 3, 5, 7 and 9 were correlated with the total scores on items 2, 4, 6 and 8. The Spearman-Brown formula was again used to lift the correlation obtained to what would be expected for the full-length test. The use of the split-half method for

calculating the reliability coefficient of the Passalong test is open to criticism in view of the unbalanced nature of the two equivalent "halves". The values obtained therefore, are offered as nothing more than rough estimates.

Cube Construction is made up of three items or sub-tests and in this case it was considered that the most suitable method of calculating the reliability was by the method of analysis of variance recommended by Burt. In the case of the Junior group, a 3 x 159 analysis of variance table was constructed and the total sum of squares and the sums of squares due to persons and tests calculated. The sum of squares due to error was obtained by straightforward subtraction. From the variance $\overline{P}$ and $\overline{E}$, $r_{tt'}$ was calculated quite simply from the formula

$$r_{tt'} = \frac{\overline{P} - \overline{E}}{\overline{P}}$$

The reliability coefficient for the Senior group was calculated in exactly the same way though in this case, a 3 x 168 analysis of variance table was used.

Reliability coefficients were not calculated for Essential Form B and Simplex as it was considered they would be of the same order as that obtained for Essential Form A.

The use of Formula 20 to calculate the reliability of

the Stanford-Binet test is something of an innovation. Normally, the reliability of this test is estimated after retesting with Form M of the scale. As has been pointed out, the Kuder-Richardson method has the advantage of permitting the reliability coefficient to be calculated from one application of the test. The procedure was as follows. For the Junior group, all boys passed the items in Year VI and all failed the last three items in Superior Adult III. The test was therefore regarded as a test containing 71 items ranging from Year VII, 1 to S.A.III, 3. The reasonable assumption is of course, that each individual in the group passes all items below his basal year and fails all those in the years above the year in which he fails all items. By giving one mark for each item passed out of the total of 71 items and zero for each item failed, the proportion of individuals in the group who passed each item was readily calculated. The standard deviation of the total scores of the individuals in the group was also easily obtained. From these data, the reliability coefficient can be calculated by means of Formula 20. It should be noted that in the calculation of the reliability coefficient, the standard deviation used was obtained from scores on the test in which 1 or 0 were allocated to the items. It is therefore quite different from the standard

deviation of the test in terms of mental age (quoted in Table II) which was calculated separately. The reliability coefficient for Form L for the Senior group was calculated in exactly the same manner.

As already mentioned, the total sample of boys tested was divided into two groups for the purpose of calculating the reliability coefficients of the tests. It was considered necessary to do this since coefficients obtained from the total sample would have been spuriously boosted by the extreme range of ability. Apart from this consideration, however, it was felt that the two coefficients for each test would provide important information regarding the reliabilities at different age levels. It is admitted that the range of ability in these Junior and Senior groups is still large and perhaps greater than would exist in a single year group, so, to avoid criticism on this account, the "Standard Error of Measurement" has been quoted with each reliability coefficient.

The Standard Error of Measurement $= \sigma\sqrt{1 - r_{11'}}$

where $r_{11'}$ is the reliability coefficient of the test and σ is the standard deviation of the scores from which the reliability has been calculated. It has the advantage that it is more or less independent of the range of ability in the group. Its interpretation together with an illustration are given in the next section.

. The reliability coefficients and other relevant information are given in Table II (p.42).

6. Comments on the Reliability Coefficients Obtained.

It has been pointed out by Vernon(1) that a reliable test must have a reliability coefficient of at least 0.9 and that coefficients lower than this would indicate that the scores are too unstable to be trusted. A glance at Table II shows that only three tests in the battery pass this requirement, namely, Essential Form A; Stanford-Binet and Progressive Matrices. Furthermore, these three tests are acceptably reliable for both Junior and Senior groups. It is assumed that Essential Form B and Simplex would also have reliability coefficients greater than 0.9.

It was considered that to approved school boys, the three verbal group tests of intelligence, of all the tests in the battery, might well appear the most dull and uninteresting, requiring as they do, about 45 minutes each of close application to verbal material and that the boys' unfavourable reactions to these tests might have the effect of lowering the reliabilities. The fact that Essential Form A has a high reliability would tend to show, therefore, that approved school boys do apply themselves with consistent effort to such tests. This conclusion

---

(1) Vernon, P.E. "The Measurement of Abilities". 1940 p.145.

is further borne out by the high correlation between Essential Form A and the parallel Form B for both Junior and Senior Groups. The correlations can be regarded of course as reliability coefficients obtained by the parallel form method. Owing to the time interval between test and retest, function fluctuation has some effect and the coefficients are somewhat lower than those obtained by the Kuder-Richardson method.
The values are shown below in Table I.

TABLE I. - Essential Form A reliabilities by two methods.

Method of calculating the Reliability Coefficient.	Juniors $r_{11'}$	Seniors $r_{11'}$
Form A (Kuder-Richardson)	.932(N = 116)	.938(N = 117)
Form A - Form B	.886(N = 100)	.883(N = 135)

By using Fisher's "Z" transformation(1) it can be shown that the coefficients obtained by the two methods are different by amounts which are statistically significant.

To sum up, im the writer's opinion, the results obtained for Essential A, are little different from what would be expected from a hundred or so ordinary school boys of comparable age range. It

(1) Fisher, R.A. "Statistical Methods for Research Workers". 1946 edition. pp. 197 - 210.

must be remembered however, that verbal group intelligence tests are limited in their application to approved school boys since about one third of them are unable to produce valid scores because of reading difficulties. This problem is discussed in Chapter V on page 68 . The reliability coefficients of N.I.I.P.70/23, V.S.10, T.S.8, Aycliffe I, and Blocks Performance would appear to be limited by the fact that each of these tests contains too few items. On estimating what the reliabilities would have been(1) had each of the tests contained 100 items, it was found that most of them reached or passed the value of 0.9. This is of theoretical interest only as the nature of the test material would prevent the tests being lengthened by this amount, especially, for example, in the case of Blocks Performance.

In Alexander's Performance Scale, Kohs' Blocks would appear to be extremely promising as a performance test, and it is suggested that the extended test of 17 items as originally used by Kohs(2) would prove to have a highly satisfactory reliability coefficient.

Before discussing the Stanford-Binet Form L, it is necessary to explain how the "Standard Error of Measurement"

---

(1) The general form of the Spearman-Brown Formula was used for this calculation $R = \frac{nr}{1 + (n-1)\,r}$ (See page 30).

(2) Kohs, S.C. "Intelligence Measurement". (1923).

may be interpreted. An individual who obtains a score on a test also possesses what may be called a "true" score, that is a score which represents his true ability in the test. True scores, of course, can never be measured and it is unlikely that an actual score will coincide with the true score. If the test be given to the individual a large number of times (assuming this to be practically possible) the actual scores obtained would distribute themselves normally and the mean of the ddistribution would, in fact, be his true score. Two thirds of the actual scores obtained would be expected to fall within the range of one Standard Error of Measurement on either side of the true score. For example, if a boy in the Junior group having a true score of 30 in Progressive Matrices were to repeat the test a large number of times, then two thirds of these scores would fall within the range of 30 plus or minus 2.79 points. The Standard Error of Measurement is to a large extent independent of the range of ability (measured by the standard deviation) in the group of individuals from whose scores the reliability coefficient is calculated.

The Standard Error of Measurement of I.Q. for the Stanford-Binet Scale was found by Terman and Merrill[1] to vary, being greater at the higher I.Q. levels than at the lower. The

---

(1) Terman, L.M. and Merrill, M.A. "Measuring Intelligence". (1937) pp. 45 - 46.

average of their figures gives a mean value of just over 4 points of I.Q. This is extremely low in view of the fact that the figures were obtained by testing with Form L and then retesting with Form M. The Standard Errors of Measurement obtained for the Junior and Senior groups (5.90 and 6.69 respectively) given in terms of mental age, have roughly similar values in terms of I.Q. and although the figures are higher than those obtained by Terman and Merrill, it is concluded that the Scale is still highly reliable when applied to approved school boys.

—oOo—

TESTS	Number of items	JUNIORS					SENIORS				
		Number of boys in sample	Mean Score	σ	$r_{11'}$	$\sigma\sqrt{1-r_{11'}}$	Number of boys in sample	Mean Score	σ	$r_{11'}$	$\sigma\sqrt{1-r_{11'}}$
Essential Form A	100	116	68.73	13.27	.932	3.46	117	72.25	12.10	.938	3.01
Stanford-Binet Form L	-	156	134.30 *	26.71 *	.951	5.90 *	169	159.50 *	28.85 *	.946	6.69 *
Progressive Matrices	60	157	32.34	10.34	.927	2.79	172	37.59	8.35	.894	2.72
N.I.I.P. 70/23	53	158	20.39	6.81	.868	2.47	169	22.28	6.28	.842	2.49
V.S.10	29	157	15.54	5.07	.829	2.10	170	18.89	4.30	.781	2.01
T.S. 8	32	156	15.36	5.61	.857	2.12	170	18.50	5.08	.834	2.07
Aycliffe I	30	156	15.08	4.91	.766	2.38	168	18.43	5.51	.829	2.28
Alexander (1) Passalong	9	159	44.55	15.10	.679	8.56	169	50.81	13.40	.600	8.47
(2) Kohs Blocks	10	159	29.67	17.88	.869	6.47	169	43.57	20.28	.862	7.52
(3) Cube Construction	3	159	64.73	17.18	.722	9.05	168	76.82	12.46	.444	9.30
Blocks Performance	29	120	10.68	5.66	.837	2.29	140	14.20	5.49	.823	2.31

TABLE II. THE RELIABILITY COEFFICIENTS AND OTHER RELEVANT DATA.

* Means and standard deviations for the Stanford-Binet are given in months of mental age.

IV THE VALIDITIES OF THE TESTS.

1. The Meaning of Test Validity.

A test must show firstly that it is a reliable measuring device, and secondly, that it measures the function or ability which it is supposed to measure, in other words that it is a valid test. Test reliability and test validity are connected intimately with one another, the latter being very largely dependent on the former. Garret(1) shows that the upper limit of a test's validity is given by the formula -

$$r_{cx_\alpha} = \frac{r_{cx_1}}{\sqrt{r_{x_1x_1}}}$$

where r_{cx_α} is the correlation between the criterion (c) and the true scores x_α in the test x_1,

r_{cx_1} is the correlation between the criterion (c) and the test,

and $r_{x_1x_1}$ is the reliability coefficient of the test.

Thus it is important to bear in mind that an unreliable test, that is one in which the scores are affected by considerable errors of measurement, can never be highly valid. It is not

---

(1) Garrett, H.E. "Statistics in Psychology and Education". (Longmans Green and Co. 1945). p. 327.

easy to validate a mental test since the functions or abilities which it is supposed to measure are in general not measureable in their pure state. It is the practice, therefore, in the face of this difficulty to set up a criterion which appears to be closely related to these functions or abilities, and to compare the test scores with the criterion scores. The kinds of criteria often used are, success in school examinations; practical achievement in a workshops course, and teachers' estimates. The correlation between test scores and criterion scores provides a kind of validity coefficient for the test, but since criterion scores are themselves of questionable reliability and validity, the validation of tests by this method generally leads to unsatisfactory results. There is, however, another approach to the problem of test validation by factorial analysis and the method will be discussed later in this chapter.

2. The Importance of General Intelligence.

The majority of psychologists would agree that intelligence can be defined as "innate general cognitive efficiency"(1) and that it is distinct from knowledge or skill that is acquired. The existence of special abilities is also recognised but their exact definition and measurement is still perhaps a controversial

---

(1) Burt, Sir.C. "Mental and Scholastic Tests" 1947 edition. Appendix III. p. 129.

issue. There can be no doubt too, that it is the factor of general intelligence which very largely determines an individual's level of achievement in everything he does. Although he may possess certain specific abilities and through them be induced to take up this or that activity, nevertheless, it is the ubiquitous general factor which decides the extent of his success. Thus a great musician must have a high general intelligence as well as special talent and a "stupid" man with musical ability can never become a first-rate artist. The layman has long recognised that there is such a thing as general intelligence and uses many familiar synonyms when referring to it. The teacher is also aware that intelligence is general in nature and has no hesitation in allocating children to various types of education according to whether they are "bright" or "dull". It is only in a group of children of approximately the same level of general intelligence, say in the same class in a grammar school, that special abilities appear to stand out. special abilities in children must not be disregarded however, but must be nurtured with care and attention since it is the fundamental aim of education to provide the means whereby every child may develop his individual gifts to the fullest extent. Special abilities of real significance are, however, not so frequently met with among school children as is generally

supposed, and among the population of boys committed to approved schools, anyone with a decided talent in a certain direction is very definitely the exception rather than the rule. An estimation of a child's general intelligence is therefore of overwhelming importance in deciding his potentialities for education and training, especially in an approved school.

3. Intelligence as Measured by Tests.

Psychological research has established clearly that when a wide variety of tests, ranging from tests of simple sensory discrimination to those involving complex mental processes are given to the same group of individuals the correlations between the scores obtained are always positive. This fact in itself is sufficient to justify the assumption that a common factor enters into all the tests. It has also been found that the tests which correlate most highly with the common factor and with independent assessments of intelligence, are those which involve the more complex mental processes. Thus it can be said that the more complex the process tested, the higher will be its correlation with the factor of general intelligence. General intelligence therefore would appear to be identifiable with the "number, variety and compactness of the relations which an individual's mind can perceive and integrate into a coherent whole".(1)

---

(1) Burt, Sir. C. "Mental and Scholastic Tests". Appendix III p. 132.

When the common element due to general intelligence has been removed from a set of test correlations, small correlations will still remain between certain groups of tests. This indicates that over and above general intelligence, the tests measure group factors or special abilities. Finally, each test measures something which is specific to itself. To sum up, intelligence tests may be said to measure (i) a general ability which enters into all performances to a greater or lesser degree but highest of all into those which require complex relation eduction. This general factor is assumed to correspond to what the general public understands by "intelligence". (ii) Certain group factors which cover verbal, numerical, spacial, practical, musical and other special abilities. (iii) An ability which is absolutely specific to each particular test.

4. Comments on the Tests in the Battery.

The Binet Scale and its various revisions, including the 1937 revision by Terman and Merrill, were designed by their authors to measure intelligence, and although these tests are still in popular favour, the items or sub-tests of which they are composed have been severely criticised of recent years. It is considered that many of the items in these scales do not involve the high mental processes at all and therefore have low correlations with "g" the factor of general intelligence. The

effect of this, naturally, is to impair the correlation of the test as a whole with the general factor. Cattell[1] further considers that these scales are overloaded with life experience and scholastic skill.

The three verbal group tests can be considered to measure the factor of general intelligence and in addition, a special factor of verbal ability. Varying degrees of verbal ability in the individuals attempting such tests can, of course, affect their scores. This difficulty is got over to a large extent in tests like Progressive Matrices and N.I.I.P. 70/23 where no words occur in the test material. The items in these two tests require mainly the eduction of quite complex relationships in terms of shapes and figures and are clearly designed to measure general intelligence and nothing else apart from the specific factor peculiar to each of them. Recent research[2] however, suggests that over and above general intelligence, the N.I.I.P. 70/23 calls into play a special factor of spatial judgment to a small extent.

The remaining three non-verbal group tests V.S.10, T.S. 8 and Aycliffe I, deal largely with the recognition and imaginative manipulation of shapes and patterns. They were designed to measure some kind of spatial ability as well as general intelligence

---

(1) Cattell, R.B. "AGuide to Mental Testing". 2nd. edition 1948. p.XV (Introduction).

(2) Emmett, W.G. "Evidence of a Space Factor at 11+ and Earlier". B.J.P.(Stat. Section). Vol. II. Pt.I. March 1949.

and represent an effort on the part of their authors to meet the demand for tests which will select those children who at the age of 11+ would be more suitable for secondary education of a technical rather than an academic nature. The assumption is that those who score highly on "space" tests and poorly in verbal tests will naturally do better at tasks of a practical kind. Alexander's battery and Peel's Blocks Performance test are further efforts by means of manipulative tests to select "practical" rather than "academic" types of children. Alexander[1] believes firmly in a factor of practical ability but many psychologists are sceptical that such a factor exists and consider that practical ability is compounded of spatial, manual, physical and other special abilities. These tests, both spatial and performance, measure mainly the general factor of intelligence and the group factors assumed necessary for success in practical work to a very much smaller extent and it is doubtful, in view of this, whether such tests can pick out the "practical" types with a high degree of discrimination. The evidence in favour of their use for this purpose is still inconclusive as the recent articles on the subject by various authors will show[2]. Although Spearman's Two-Factor Theory[3],

---

(1) Alexander, W.P. "Intelligence: Concrete and Abstract". (1935) B.J.P. Monograph Supplement.

(2) Burt, Sir.C. Part IX, conclusion to the "Symposium on the Selection of Pupils for Different Types of Secondary Schools". B.J.Ed.P. Vol.XX. Pt.I. Feb.1950. pp. 1 - 10 with reference to Parts I - VIII of the "Symposium" contributed by various authors.

(3) Spearman, C. "The Abilities of Man". (Macmillan 1927).

with its complete denial of the existence of subsidiary group factors is now completely rejected it would seem, as Vernon points out[1], that "in point of fact Spearman has been proved much more nearly right than vocational and educational psychologists would wish him to be".

5. The Validation of the Tests in the Battery by Factor Analysis.

The first step in carrying out the process of factor analysis was to calculate all the inter-correlations of each test in turn with all the other tests in the battery. Product-moment correlation coefficients were calculated, the diagonal method being used[2]. In theory, the product-moment formula requires that the variables which are to be correlated shall have a normal distribution, however in practice, it is the custom to use this method if the score distributions are reasonably symmetrical or at least do not differ very significantly from the normal distribution. In the present investigation no estimates were made to find out to what extent the score distributions differed from the normal distribution. They were fairly symmetrical in shape in the majority of cases and it was considered that the product-moment formula would give a satisfactory estimate of the amount of agreement which existed

---

(1) Vernon, P.E. "The Structure of Human Abilities". (Methuen (1950). p.15.

(2) Chambers: Statistical Calculation (1945) p.50.

between the sets of test scores. Tables III and VI give the correlations of the distributions of the raw scores for Juniors and Seniors respectively. The scores obtained on Burt's Reading Vocabulary Test were heavily bunched to the top and (i.e. negatively skewed) for both Juniors and Seniors, and it was considered that the skewing was such that the product-moment formula could not be used. It was decided not to calculate the inter-correlations of this test with the other tests in the battery and therefore, it does not appear in the lists in the above mentioned tables. For the same reason Cube Construction does not appear in the list of tests given in Table VI. In the case of the Seniors the distribution of the scores in Progressive Matrices also showed a certain amount of negative skewness, but nevertheless, it was included. It is worthwhile pointing out that Burt's Vocabulary Test was designed to provide Reading Ages, and by the nature of its construction will always give a negatively skewed distribution when applied to a similar population as that used here. On the other hand the negative skewness of Cube Construction and Progressive Matrices when applied to the Senior group, would tend to show that these tests do not discriminate highly among the older boys, in other words, the older boys all tend to score highly. It will be noted that Practical Ability appears only in Table VI (Seniors). This is of course, due to the fact that only the Senior boys

received Practical Ability assessments.

Column 18 in each of Tables III and VI gives the correlation which the tests have with the distribution of chronological age in the Junior and Senior Groups respectively. It was necessary to calculate these correlations with age in order to apply a correction to the test inter-correlations. If two tests correlate with one another and each of them also correlates positively with age, then part of the agreement between the two sets of test scores is due to the fact that each correlates with age. The correction which is applied makes allowance for the correlation which each test has with age and enables the test inter-correlations to be recalculated on the assumption that age is constant in the sample which provided the test scores. Tables III and VI give the test inter-correlations of the raw scores of the distributions which are naturally influenced by age and Tables IV and VII give the values which would be obtained if age were constant in Junior and Seniors groups(1). The correlations shown in Tables IV and

---

(1) If r_{xy} is the correlation of the raw scores of test x and y, and r_{xa}, r_{ya} the correlation which each has with the distribution of chronological age in the sample tested, then the correlation of x and y with age constant is given by

$$r_{xy} \text{ (age constant)} = \frac{r_{xy} - r_{xa} \cdot r_{ya}}{\sqrt{1 - r_{xa}^2} \cdot \sqrt{1 - r_{ya}^2}}$$

VII are those upon which the factor analysis was carried out.

The correlation which each test has with chronological age (column 18 in Tables III and VI) is worthy of special comment. The first point which stands out is that with the Juniors the correlations are positive and quite clearly significantly different from zero, whereas with the Seniors, the majority are small and not significantly different from zero. The conclusion to be drawn from this is that in the Junior group, older boys tend to score higher than younger boys, and in the Senior group, chronological age has little effect on the scores, since the boys in this group have attained intellectual maturity, and the variation of chronological age within the group has little effect on the scores obtained. This result of course, was to be expected. In the case of the Senior boys the exceptions to this expected result are the three performance tests and the assessment of practical ability which still show appreciable positive correlations with chronological age. This shows that within the Senior group, the older boys do better at performance tests and in the workshop than do the younger ones, and it may be argued from this that the level of achievement in practical tasks is dependent on age and perhaps experience even after intellectual maturity, as measured by the other types of tests, has been attained. It will be noted that in both Junior and Senior groups the assessments of general intelligence by the Housemasters

correlate negatively with age. This is perhaps due to the fact for some reason or other that the Housemasters concerned tended to under assess the older boys and over assess the younger boys.

The Matrices of correlation coefficients given in Tables IV and VII were each factorised by Thurstone's Centroid method, using guessed communalities[1]. Three factors were extracted and the loadings obtained are shown in Tables V and VIII for Juniors and Seniors respectively. The figures in column I (Tables V and VIII) are the loadings of the tests in the common factor. If the battery used contains a wide variety of tests and the population tested is sufficiently heterogeneous with regard to ability (which was the case in the present investigation) the common factor is nearly the same as "g", the factor of general intelligence as defined by factor analysis. It is not, however, the same as "g" as Thurstone's centroid method normally requires the factors to be rotated to give psychological significance to the bi-polar factors. Rotating the factors has the effect of reducing the loading in the first factor, the extra variance being distributed among the other factors. In Vernon's opinion[2], the classification into the general and the bi-polar factors very often gives all the information about an

---

(1) Thomson, Sir.G.H. "The Factorial Analysis of Human Ability". pp. 161 - 170.

(2) Vernon, P.E. "The Structure of Human Abilities" (Methuen 1950) p. 24.

analysis which is required. The writer, after examining the rotational possibilities of the factors obtained in the analyses of the Junior and Senior groups came to the conclusion that they were best left in their unrotated condition. It was then assumed that the first factor in each case approximated to the general factor of intelligence and that the loadings could be regarded as the validity coefficients of the tests concerned.

Column II (Tables V and VIII) is the second factor and the loadings obtained show the extent to which the tests measure special abilities. It is a bi-polar factor with the verbal and educational tests (those with the positive loadings) at one end and the spatial and manual tests (those with the negative loadings) at the other end. This second factor serves to show the clear contrast between verbal and educational tests on the one hand and spatial and manual test on the other. It is interesting to note that the Housemasters' assessments of general intelligence are grouped with the verbal and educational tests in both Junior and Senior groups and that the assessment of Practical Ability (Senior groups only) is grouped with the spatial and manual tests. This would suggest firstly, that the Housemasters are perhaps influenced by verbal ability and educational attainment in their judgments, and secondly that there is a very clear connection between the spatial and

performance tests and practical ability as assessed in the workshop.

Column III (Tables V and VIII) gives the third factor obtained which is a further analysis of the correlations into special abilities. They represent a higher degree of differentiation than do the second factor loadings. It was doubtful whether it was strictly legitimate to extract a third factor since the correlations were obtained from quite small samples, 100 boys in the case of the Junior group and 135 in the case of the Senior group, and the correlation coefficients obtained had quite large standard errors. It may be, therefore, that the loadings in factor III are considerably affected by sampling errors in the original coefficients.

A scrutiny of the validity coefficients of the tests, that is their loadings in the general factor (column I, Tables V and VIII) shows immediately that the three verbal group tests and the Stanford-Binet Form L are superior to the remainder of the tests in the battery. It is most important to point out here that to the best of the writer's knowledge, this is the first occasion on which the Stanford-Binet test has been included for factorisation in a battery of miscellaneous tests. The results show that it appears to be very little different from verbal group intelligence tests. Its loading

in the general factor is about the same and in Factor II it appears to show less dependence on verbal and educational attainment than do the verbal group tests. The above remarks apply to both Junior and Senior groups, and from this it is concluded that the Stanford-Binet test is successfully vindicated as a means of measuring general intelligence. The four above mentioned tests are dependent to some extent on verbal facility and educational attainment and this is shown by the fact that in Factor II the positive loadings obtained are similar in nature to those of Dictation, Composition and Arithmetic. The three latter tests of educational attainment were included in the battery solely for the purpose of making this comparison. It is felt that this question of verbal facility and educational attainment is important when judging whether or not an intelligence test is suitable for use with approved school boys, and in view of the loadings obtained it is not unreasonable to state that the Stanford-Binet Form L does not appear overweighted in Factor II.

With regard to the measurement of general intelligence, there is little to choose between the non-verbal group tests and it is difficult to decide from the factorial data alone which of them bears the highest validity.

Of the performance tests, Kohs' Blocks stands out alone as being a very superior test.

At this stage it is necessary to say something about the column headed "h^2", in each of Tables V and VIII. The values given in this column are the proportions of the total variances of each of the tests measured by factors I, II and III combined. For example, in Essential A Table V, "h^2" is obtained by summing the squares of the three factor loadings of this test. The value 0.846 is the amount of variance out of a total variance of unity which is measured by factors I, II and III together. h^2 is readily converted to a percentage and perhaps the simplest way of regarding this figure is to state that the three factors extracted represent 84.6% of the total variance of the test. The remainder of the test variance, 15.4% is represented by error and something that the test measures which is specific to the test itself. If we consider that factors II and III, which measure special abilities over and above the common factor, are of some importance and should be included in our estimates of test validity, then the values given in the "h^2" column can also be regarded as validity coefficients. The contribution which the general factor makes to the total variance of a particular test is obtained by squaring the loading in the common factor. For example, in Essential A Table V, the loading in the common factor is 0.842, squared this becomes 0.709 or 70.9% of the total test variance. For this particular test, therefore, we

can say that 70.9% of the variance is taken out by the common factor, but if we include factors II and III then 84.6% of the test variance is taken out. The use of the values in the h^2 column as validity coefficients makes the assumption that the special abilities measured by factors II and III are to be combined with the general factor in the measurement of intelligence.

6. The Validities by the Criteria.

(i) The assessment of general intelligence.

The correlations of the tests with the assessments of general intelligence are given in column 17, Table IV, for the Juniors and in column 16, Table VII for the Seniors. The values given are of course those which were recalculated on the assumption that age was constant in the two samples. None of the correlations are high but it will be seen at a glance that the three verbal group tests and the Stanford-Binet Form L show the highest agreement. The smallness of the correlations in general is perhaps due to the somewhat unsatisfactory nature of the criterion itself. The assessments were made by four different housemasters who each assessed approximately one quarter of the boys in each of the samples. This in itself would reduce the validity of the criterion since it is impossible to have four independent persons with identical powers of judgment. It would

have been better if the housemasters could have provided assessments for all the boys, but this was not possible. An attempt was made to check the reliability of the housemasters' assessments by obtaining an independent assessment from the Warden of the Classifying School for 54 boys. The correlation between his assessments and those provided by the housemasters for the same 54,boys was 0.77.

(ii) The assessment of practical ability.

The correlations of the tests with the assessments of practical ability in the workshop (Seniors only) is given in column 17, Table VII. The correlations are low, Kohs' Blocks having the highest value of all the tests in the battery. The assessments of practical ability were made by the instructor are considered to be reasonably valid. Although perfect discrimination cannot be expected from a subjective assessment nevertheless, it was clear to the writer that after watching the boys working for a week in the workshop, the instructor was very well aware of the amount of aptitude, skill and ability possessed by each boy. The factor analysis (Table VIII) shows that the assessment of practical ability is factorisable into a comparatively large loading in the general factor and quite small loadings in factors II and III. It is to be expected that the general factor must play quite a large part in success or failure in practical tasks, and the small loadings in factors II

and III show that whatever are the special abilities which go to make up practical ability they are not apparently isolated by factorising the battery of tests along with the subjective assessments as far as this particular analysis is concerned.

—oOo—

TABLE V. THE UNROTATED FACTORS. (JUNIORS).

	I	II	III	h^2
1. Essential A.	842	348	127	846
2. Essential B.	858	336	104	860
3. Simplex.	809	350	088	785
4. Stanford-Binet (L).	841	188	055	746
5. Progressive Matrices.	681	-274	-233	593
6. N.I.I.P. 70/23.	720	-131	087	543
7. V.S. 10.	734	-210	-195	621
8. T.S. 8.	667	-308	-180	572
9. Aycliffe I.	644	-301	-133	523
10. Passalong	361	-252	211	238
11. Kohs' Blocks.	700	-463	185	739
12. Cube Construction.	521	-414	130	460
13. Blocks Performance.	609	-348	099	502
14. Dictation.	598	543	-225	703
15. Composition.	516	478	-194	532
16. Arithmetic.	681	213	306	603
17. General Intelligence (Ass).	620	201	-301	515
Column totals.	7.922	1.893	.566	10.381
Percentages of the total variance.	46.60	11.14	3.33	61.06

(Decimal points omitted from all loadings).

RESIDUALS.

	I	After 1st.Factor	After 2nd.Factor	After 3rd.Factor
Greater than 3 x std. error	113/136	8/136	0/136	0/136
Greater than 2 x std. error	125/136	22/136	0/136	0/136
Greater than 1 x std. error	131/136	83/136	10/136	8/136

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Essential A	1.	-	883	865	772	609	549	542	478	500	481	248	314	579	577	732	618	394	-058
Essential B	2.		-	876	777	684	609	571	565	512	543	277	366	543	583	743	595	437	019
Simplex	3.			-	755	626	566	527	480	540	487	214	352	571	632	766	592	407	-053
Stanford-Binet (L)	4.				-	539	535	532	535	481	564	275	408	520	575	651	552	449	099
Progressive Matrices	5.					-	488	640	573	552	625	466	446	301	339	549	420	471	106
N.I.I.P. 70/23	6.						-	539	548	561	622	356	535	278	407	501	474	342	-052
V.S.10	7.							-	661	597	684	425	501	319	357	500	403	410	021
T.S. 8	8.								-	532	658	412	473	225	269	451	389	395	-025
Aycliffe I	9.									-	624	347	520	223	319	393	362	399	011
Kohs' Blocks	10.										-	494	645	213	319	461	396	548	207
Passalong	11.											-	360	041	105	255	121	274	194
Blocks Performance	12.												-	247	151	296	293	421	205
Dictation	13.													-	622	535	441	163	031
Composition	14.														-	560	488	336	026
Arithmetic	15.															-	505	331	-096
Gen. Intelligence (Ass).	16.																-	363	-246
Practical Ability (Ass).	17.																	-	231
Chronological age	18.																		-

TABLE VI. (SENIORS). CORRELATION COEFFICIENTS OF THE TESTS AND ASSESSMENTS.

(Decimal points omitted from all correlations).

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Essential A	1.	-	886	865	783	620	548	545	478	501	505	267	334	585	580	731	624	419
Essential B	2.		-	879	779	686	611	571	565	512	551	281	370	546	583	748	619	445
Simplex	3.			-	765	636	564	529	479	541	510	231	371	576	634	765	598	432
Stanford-Binet (L)	4.				-	534	544	533	540	482	555	264	398	522	576	666	597	440
Progressive Matrices	5.					-	497	641	579	554	620	461	436	301	338	565	463	462
N.I.I.P. 70/23	6.						-	541	547	562	647	377	558	282	409	499	476	364
V.S. 10	7.							-	662	597	695	433	508	320	356	504	421	417
T.S. 8	8.								-	533	678	429	489	227	270	451	395	412
Aycliffe I	9.									-	636	355	529	224	318	396	376	407
Kohs' Blocks	10.										-	478	629	212	321	494	477	525
Passalong	11.											-	337	036	103	283	179	243
Blocks Performance	12.												-	247	149	324	362	392
Dictation	13.													-	625	543	465	161
Composition	14.														-	566	511	339
Arithmetic	15.															-	499	364
Gen. Intelligence (Ass).	16.																-	445
Practical Ability (Ass).	17.																	-

TABLE VII. (SENIORS). CORRELATION COEFFICIENTS OF THE TESTS AND ASSESSMENTS WITH AGE CONSTANT.

(Decimal points omitted from all correlations).

TABLE VIII.

THE UNROTATED FACTORS. (SENIORS).

		I	II	III	h^2
1.	Essential A.	848	340	151	859
2.	Essential B.	878	269	189	880
3.	Simplex.	856	330	094	851
4.	Stanford-Binet (L).	815	227	-023	717
5.	Progressive Matrices.	758	-164	276	678
6.	N.I.I.P. 70/23.	724	-165	-113	565
7.	V.S.10.	749	-283	089	649
8.	T.S. 8.	703	-316	092	602
9.	Aycliffe I.	681	-287	-047	549
10.	Kohs' Blocks.	770	-418	-091	776
11.	Passalong.	437	-372	186	364
12.	Blocks Performance.	590	-376	-267	561
13.	Dictation.	542	489	-129	545
14.	Composition.	611	423	-215	598
15.	Arithmetic.	765	267	153	660
16.	Gen. Intelligence (Ass).	679	178	-211	536
17.	Practical Ability (Ass).	567	-130	-120	353
	Column totals.	8.673	1.652	.434	10.740
	Percentages of the total variance.	51.02 %	9.72 %	2.56 %	63.18 %

(Decimal points omitted from all loadings).

RESIDUALS.

	I	After 1st. Factor	After 2nd. Factor	After 3rd. Factor
Greater than 3 x std. error	127 /136	6/136	-/136	-/136
Greater than 2 x std. error	134/136	24/136	1/136	-/136
Greater than 1 x std. error	135/136	87/136	15/136	2/136

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
Essential A	1.	-	886	852	812	591	567	575	514	464	348	470	322	465	705	616	731	469	416
Essential B	2.		-	855	807	570	604	603	547	472	328	526	369	430	702	656	727	482	429
Simplex	3.			-	745	562	606	556	480	451	276	440	331	404	732	587	696	446	357
Stanford-Binet L	4.				-	569	636	571	526	556	284	560	362	557	595	586	640	547	323
Progressive Matrices	5.					-	515	592	535	600	382	583	489	519	416	279	453	485	511
N.I.I.P. 70/23	6.						-	573	551	543	389	631	354	474	359	377	551	371	241
V.S.10	7.							-	685	518	344	622	410	521	472	421	463	333	266
T.S. 8	8.								-	579	358	612	380	540	311	356	386	303	296
Aycliffe I	9.									-	274	533	527	481	299	301	367	293	215
Passalong	10.										-	442	260	313	104	146	406	072	294
Kohs' Blocks	11.											-	587	647	199	222	452	282	214
Cube Construction	12.												-	539	112	053	284	163	104
Blocks Performance	13.													-	216	279	400	216	260
Dictation.	14.														-	695	569	409	314
Composition	15.															-	498	336	436
Arithmetic	16.																-	272	369
Gen. Intelligence (Ass).	17.																	-	-101
Chronological Age	18.																		-

TABLE III. (JUNIORS). CORRELATION COEFFICIENTS OF THE TESTS AND ASSESSMENTS.

(Decimal points omitted from all correlations).

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17
Essential A	1.	-	862	827	788	483	529	530	451	422	260	429	309	407	665	532	683	565
Essential B	2.		-	832	781	452	571	562	487	431	234	492	361	365	662	577	677	580
Simplex	3.			-	713	473	573	512	419	410	192	399	317	345	699	513	650	518
Stanford-Binet (L)	4.				-	496	607	532	476	527	209	531	349	518	546	522	592	616
Progressive Matrices	5.					-	470	550	468	583	282	564	510	465	314	072	331	628
N.I.I.P. 70/23	6.						-	544	518	518	343	611	341	439	307	311	512	409
V.S.10	7.							-	659	490	289	601	399	486	424	352	407	375
T.S. 8	8.								-	552	297	589	367	502	241	264	312	351
Aycliffe I	9.									-	226	511	520	451	249	236	317	324
Passalong	10.										-	406	241	257	013	021	335	107
Kohs' Blocks	11.											-	582	627	142	147	411	313
Cube Construction	12.												-	533	084	009	266	176
Blocks Performance	13.													-	146	191	339	252
Dictation	14.														-	653	514	467
Composition	15.															-	403	424
Arithmetic	16.																-	334
Gen. Intelligence (Ass).	17.																	-

TABLE IV. (JUNIORS). CORRELATION COEFFICIENTS OF THE TESTS AND ASSESSMENTS WITH AGE CONSTANT.

(Decimal points omitted from all correlations).

V - THE READING ABILITY NECESSARY TO PRODUCE A VALID SCORE ON A VERBAL GROUP INTELLIGENCE TEST.

1. The findings of Professor Schonell and M.A.Mellone.

The problem of deciding what level of reading ability a child shall have before he can be expected to produce a valid score on a verbal group intelligence test, is one which seems to have attracted few investigators in the field of educational psychology. Schonell(1) states that children with reading ages under 8½ years should not be given such tests, and mentions that Mellone(2) considers the minimum reading age to be 9½ years. Mellone arrived at this conclusion after testing four small samples of children, the ages of the children in the four groups being 8, 9, 10 and 11 years respectively. The tests used were the Sleight Non-verbal Intelligence Test, Moray House Intelligence Test No.26, and the Burt-Vernon Reading Test. The scores obtained by the children were converted to Intelligence Quotients by means of conversion tables. The Sleight test was used as a criterion and it was found that the mean verbal I.Q. of the 8 year group was depressed somewhat. It was concluded that this was due to verbal difficulties with the test material. Since the 9 year

---

(1) Schonell, F.J. "Development of Educational Research". B.J.Ed.P. Feb.1948. p. 14.

(2) Mellone, M.A. "Reading Ability and I.Q". B.J.Ed.P. June 1942. pp. 128 - 135.

group (and those above it) showed no such depression, it was further concluded that the verbal test only measured the true I.Q. in the 9 year group and upwards. In other words, 9½ years (the mean age of the 9 year group) is the minimum age at which a verbal group test should be used to estimate a child's I.Q.

The writer feels that Mellone's work is open to criticism in the method used to convert the verbal group test raw scores into intelligence quotients. The conversion table for the Moray House Test No.26 only allows intelligence quotients to be quoted for children with chronological ages between 10 and 12 years. The method of standardisation ensures that the graph of I.Q. against chronological age is a straight line and in order to obtain verbal intelligence quotients for the 8 and 9 year groups, Mellone extended the graph on the assumption that it would still be a straight line down to the 8 year level. This arrangement meant that a large proportion of the children in the younger age groups were given intelligence quotients which were based on raw scores of less than 10 marks. The writer considers that such low scores are so invalid that the work of calculating intelligence quotients from them would not appear to be worth carrying out.

Mellone points out clearly the weaknesses of this part of her statistical analysis and goes on to show that her general conclusions are not really invalidated by them.

2. The Present investigation.

The writer's approach to the problem was different from that of Mellone in that intelligence quotients were not used at all. An attempt was made to determine what level of reading age was necessary to produce a valid raw score on a verbal group intelligence test which is after all the problem in its simplest terms. The introduction of intelligence quotients was considered to be an unnecessary complication.

The test chosen for the experiment was the Essential Form A, chiefly because supplies of this test were to hand. The experiment was of course, a by-product of the major research into test reliabilities and validities.

It was realised that many of the boys would not be able to produce satisfactory scores on this verbal group test because of serious reading difficulties. Nevertheless, it was determined at the outset that consecutive entrants to the Classifying School would be tested by it, regardless of whether they could really attempt the test satisfactorily or not. This plan was carried out and altogether 327 boys were tested. The scripts obtained from this sample ranged from boys who could not make any score at all to those who obtained nearly full marks.

Before a boy attempted the group test, he received an individual practice test which covered the various types of items appearing in the Essential A Test. If after coaching, a boy still

had difficulty in understanding the printed instructions relating to the various types of items and the nature of the items themselves, it was decided that his reading ability was such that he would be unable to do himself justice in the test proper. In other words he would be unable to do many of the group items through failing to comprehend the printed instructions, and not necessarily because he had poor intelligence. By means of this practice test it was fairly easy to distinguish those who could and those who could not do the verbal group test. The decision was of course a subjective judgment to some extent on the part of the person carrying out the individual practice test. Although estimates were difficult in what might be called borderline cases, all the boys were recorded as being "Yes" or "No". Of the total of 327 boys tested, it was considered that 94 of them had not the reading ability necessary to produce a valid score on the verbal group test. This decision was made in all cases before the boys actually attempted the group test.

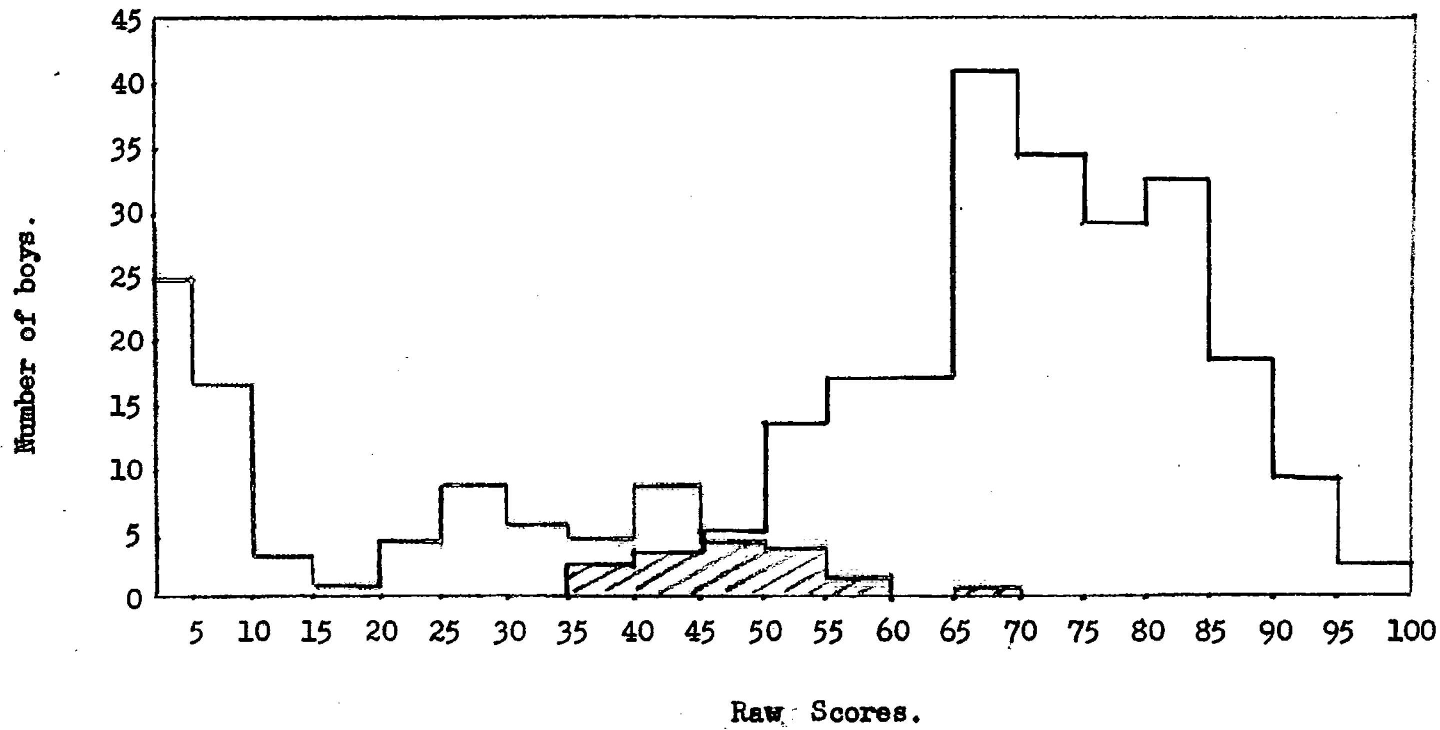
The 327 boys comprising the sample were also tested with Burt's Reading Accuracy (Vocabulary) Test No.1 and mental ages as obtained on the New Stanford Revision of the Binet Scale (1937) Form L, were also available.

On the completion of the testing and the marking of the scripts, histograms were prepared as shown on pages 72, 73, 74 and 75.

A glance at these histograms shows that in all three tests, the score distributions obtained by the boys who were considered to have insufficient reading ability to produce valid scores on the verbal group intelligence test, overlapped to some extent with the score distributions of those who were expected to produce valid scores. The overlapping is accounted for largely by the fact that the division of the boys into these two groups depended to some extent on the subjective opinion of the person carrying out the individual practise test. If the method by which the boys were divided into the two groups is accepted as being sound in principle then it seems reasonable to assume that the "true" line of demarkation is at the centre of the overlap. Thus in the Essential Form A, scores below about 47 would appear to have doubtful validity. This level of score seems to be related to a reading age of about 9½ years on Burt's test and a mental age of about 10½ years on the Stanford Binet.

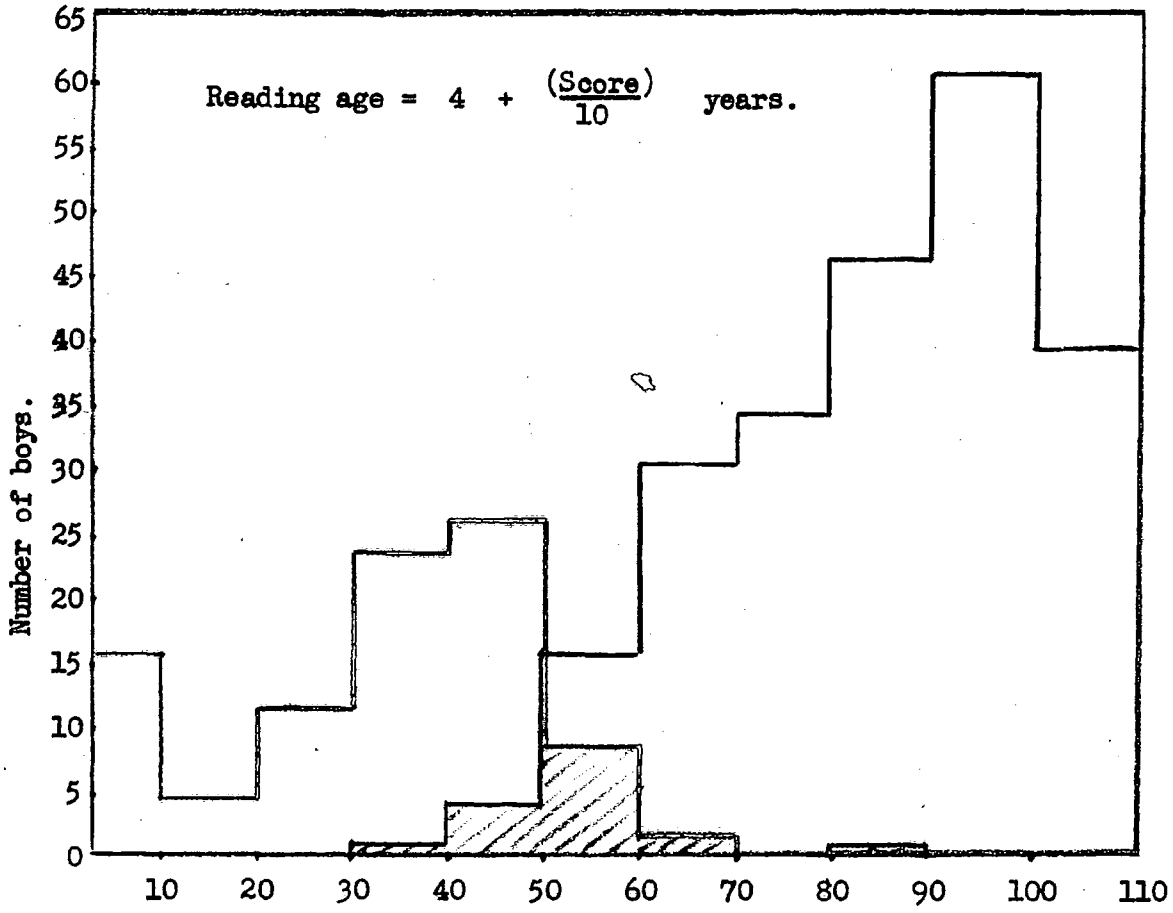
The empirical approach of the writer to the problem precluded the use of neat and orthodox statistical methods for the treatment of the data. Nevertheless, the conclusion that a reading age of 9½ years is necessary before a child can be expected to produce a valid score on a verbal group intelligence test, agrees remarkably with the findings of Mellone. The conclusion arrived at by the writer is, of course, based on

Essential Group Intelligence Test Form A.



25	17	3	1	5	9	7	6	9	5	4	2	-	1							= 94	"No"
							3	4	6	14	18	18	42	35	29	33	19	9	3	= 233	"Yes"
																				327	Total

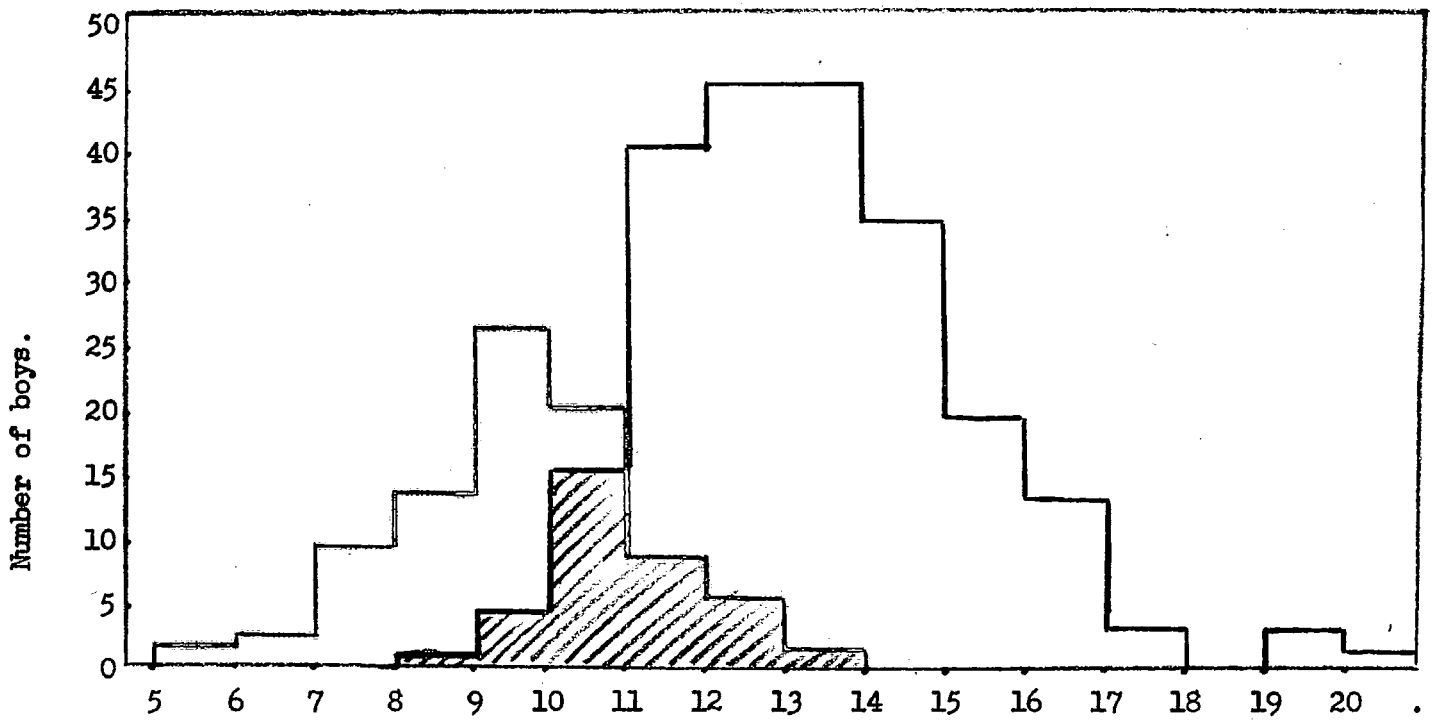
Burt's Reading Accuracy (Vocabulary) Test No.1.



Score - Number of words read correctly.

16	5	12	23	27	8	2	-	1	=	94	"No"
			1	4	16	31	34	46	61	40	= 233 "Yes"
											327 Total

New Stanford Revision of the Binet Scale (1937) Form L.



2	3		10	14	27	22		9	6	1								= 94	"No"
				1	4	16	40	46	46		35	21		14	4	-	4	2	= 233	"Yes"
																		327	Total

results obtained from a sample of approved school boys. It is considered however, that experiments on the same lines with the normal school population would produce similar results.

3. An enquiry into the structure of Essential Form A.

The conclusion that scores below about 47 on the Essential test are of doubtful validity is of such a serious nature that further enquiry was considered to be most necessary. The approach to this new investigation was also made empirically. The procedure adopted was quite straightforward. The scripts of all the boys who attempted the test were arranged in order of merit, then each boy's script was examined item by item. A description of the types of items comprising the Essential test is given in Table IX.

The examination of the scripts brought to light two important facts. Firstly, there was a general tendency for boys to leave unattempted whole blocks of items. This suggested that the written instructions relating to particular blocks of items were difficult for certain boys to comprehend, and that a barrier was erected which either prevented or deterred these boys from attempting the individual items comprising the blocks. Secondly, a large proportion of the boys, particularly those with the lower scores, attempted but failed to answer whole blocks of items correctly because they had clearly misunderstood the instructions.

TABLE IX.

Blocks of Items	Description.
1 - 9	Sentence Completion I. A missing word has to be written in the space provided.
10 - 15	Sentence Completion II. A choice of words to complete the sentence correctly is provided and the correct word must be underlined.
16 - 20	Alphabet Items. These items deal mainly with the position of certain letters of the alphabet. The alphabet is printed above the items to provide a standard situation for their solution.
21 - 26	Find the smallest. A series of objects or quantities is provided in jumbled order for each item. The smallest object or quantity has to be underlined in each case.
27 - 35	Find the Similar Word. Multiple choice items in which the word similar in meaning to the given key word has to be underlined.
36 - 39	Always has. Multiple choice items in which two words describing two properties always possessed by the given key word, have to be underlined.
40 - 49	Opposites. Multiple choice items in which the word opposite in meaning to the given key word has to be underlined.

TABLES IX. continued.

Blocks of Items	Description.
50 - 53	Three belonging together. In each item, 3 words which have a clear relation to one another have to be underlined. A multiple choice is provided in each case.
54 - 57	Two like the first three. In each item, three key words are given which are related in some way to one another. From a selection provided, two more words have to be underlined in each case, which are themselves related to the key words.
58 - 62	Miscellaneous Problems I.
63 - 71	Analogies. Multiple choice items based on word analogies.
72 - 75	Find the different word. In the group of words presented for each item all except one are related. The different word has to be underlined.
76 - 80	Miscellaneous Problems II
81 - 87	Series Completion. A series is given, and the two next consecutive items have to be added in each case.
88 - 100	Miscellaneous Problems III

When a boy answers an individual item incorrectly, after having adopted the proper procedure for dealing with the item,

for example, by underlining two words out of the given five in the "Always has" items, it is impossible to know whether his failure was due to lack of reading ability or inadequate intelligence. On the other hand, if a boy fails a whole block of items and he has adopted the same wrong procedure for answering these items throughout, then it seems reasonable to assume that he has not understood what was required of him.

Having arrived at this conclusion, it was decided to carry out an analysis of the blocks of items in the Essential test to record for each boy those which were "not attempted" and those which were "not understood". Of the 327 scripts obtained, only 212 were analysed in this way. The first 19 scripts were discarded as they had not been attempted at all, and the top 95 scripts (marks ranging from 75 to 97) showed that all items had been attempted and the instruction understood correctly. The analysis of the remaining 212 scripts is shown in Table X.

Table X overleaf.

TABLE X THE BLOCK ITEM ANALYSIS - Essential Test.

1 Blocks of items.	2 Description	3 Not Attempted	4 Not Understood
1 - 9	Sentence Completion I	-	-
10 - 15	Sentence Completion II	20	2
16 - 20	Alphabet Items	11	9
21 - 26	Find the smallest	20	1
27 - 35	Find the similar word	23	6
36 - 39	Always has	21	46
40 - 49	Opposites	18	7
50 - 53	Three belonging together	24	5
54 - 57	Two like the first three	21	55
58 - 62	Miscellaneous Problems I	33	?
63 - 71	Analogies	36	5
72 - 75	Find the different word	32	6
76 - 80	Miscellaneous Problems II	33	?
81 - 87	Series completion	61	25
88 - 100.	Miscellaneous Problems III	35	?

The figures shown in columns 3 and 4 of Table X have no absolute value of course and are relative to the 212 cases analysed. It is immediately apparent that there must be something

wrong with the instructions of items 36-39 "Always has", and 54-57 "Two like the first three". The instructions of "Series completion" would also appear to be unsatisfactory, though to a lesser degree. These three blocks of items show high values in column 4, and it was interesting to note that boys with scores as high as 75 out of the maximum of 100 showed that they had not understood these instructions. It is considered that this simple method of analysing items in blocks, is of great importance in test construction as it shows up weaknesses which would not be shown up by the normal methods of item analysis. It is the custom for the items in group tests, particularly verbal group tests, to be grouped together in blocks, and for tests constructed in this way, an analysis such as that described above is clearly a necessity.

In the analysis of the blocks of items, the responses "N.A." (not attempted) and "N.U." (not understood) were recorded for each boy in the reduced sample of 212. The responses were recorded from scripts which had been placed in order of merit commencing at the lowest score which was 1 and rising to a score of 75. A scrutiny of the responses of the lowest scorers showed that practically the whole of the test, apart from the first dozen or so items, were not attempted. As the scores increased so the proportion of blocks of items not attempted was reduced.

At the same time, the blocks of items recorded as not understood increased. As the score continued to rise the "N.A's" and "N.U's" thinned out and at a score level of between 45 and 50 it could be said that all blocks of items had been attempted and all had been understood with the exception of "Always has", "Two like the first three" and "Series completion". For these particular blocks of items, the respone "N.U" kept cropping up continually and it was considered that this was due to a defect in the manner of presentation of these items and not to imperfect understanding on the part of the boys concerned.

It is considered therefore that the above finding is further evidence to support the conclusion that scores below about 45 on the Essential test are of doubtful validity. This statement, which at first glance may appear extravagant, merely indicates that a score of about 45 represents the minimum level of ability at which an individual attempting the test can be expected to sample all the items in the test and show that he also understands the instructions.

---oOo---

VI - THE REVISION OF THE STANFORD-BINET SCALE (1937) FORM L.

The order of difficulty of the items.

Many psychologists have expressed the opinion that quite a number of the items in this test are not in their proper order of difficulty as far as children in this country are concerned. Burt pointed out in 1939[1] that between the ages of 4 and 14, out of the 66 tests, 32 would appear to be misplaced by at least one year, and further stated that with an externally graded scale such as this, everything turns on the relative difficulty of the test problems. The standardisation of each problem in terms of mental age assumes that the order of difficulty is constant for the two sexes, for different social classes, for different ages, for different types of child, and above all for different localities. In view of the fact that the test was standardised in the United States, it is not surprising to find that when used in this country, certain anomalies appear.

In testing a large number of approved school boys, the writer was able to form certain opinions regarding the order of difficulty of the items as far as their application to approved school boys was concerned. In other words certain items came to be regarded as difficult and others as easy. In general it could

---

(1) Burt, Sir.C. "The Latest Revision of the Binet Intelligence Tests". The Eugenics Review XXX 1939 pp. 225 - 260.

be said that vocabulary items, items requiring reading or a knowledge of the meaning of words were too difficult while "Frontier days", "Purse and field" and "The shadow", were too easy.

By about September 1949, 468 consecutive entrants to the Classifying School had been tested by Form L of the Stanford-Binet Scale[1] and it was decided to analyse some of the scripts to see if the subjective opinions of the writer were confirmed. It was decided to group the data in terms of mental age[2] as individuals of the same mental age have more trials or attempts in common than those of the same chronological age. The numbers in the groups taken for the analysis were as follows:-

Mental age	10.0-10.11	11.0-11.11	12.0-12.11	13.0-13.11
Number in group	59	61	67	58

In each of the groups, the proportion of boys answering each sub-test or item correctly was recorded. It was found, of course, that the ranges covered by the four groups were successively

---

(1) After May 1949, when the testing programme for research purposes was completed, certain of the tests used, among them the Stanford-Binet, were continued in use as part of the normal process of classification.

(2) Mc.Nemar, Q. "The Revision of the Stanford-Binet Scale" 1942. Mc.Nemar grouped his data in this way when investigating "spread" in the scale.

higher up the scale as they progressed from 1 to 4. By limiting the items to a range from Year IX 1 to Year XIV 6, it was possible to record a response for every boy in all the groups. It was assumed that every boy would pass all items below his basal year and would fail all items in the year above that in which he had failed all items. From the probability values obtained it was a simple matter to calculate the order of difficulty of the items within each group. The orders of difficulty were compared with one another, and the order published recently by Cole[1]. The results are shown below. The rank order method for calculating correlations was used.

TABLE XI - Rank Correlations of orders of difficulty of the Stanford-Binet Scale (Form L).

	10.0-10.11	11.0-11.11	12.0-12.11	13.0-13.11	Cole.
10.0 - 10.11	-	.87	.80	.70	.71
11.0 - 11.11		-	.90	.83	.73
12.0 - 12.11			-	.92	.72
13.0 - 13.11				-	.75
Cole					-

(1) Cole, R. "An item analysis of the Terman-Merrill Revision of the Binet Tests". B.J.P. (Statistical Section) November 1948 pp. 137 - 151.

From an inspection of Table XI it will be seen that (a) the orders of difficulty as obtained from the 4 groups of approved school boys tend to correlate higher amongst themselves than each does in turn with the order obtained by Cole, and (b) the correlations between adjacent groups are higher than between remote groups. These results suggest firstly, that approved school boys are a special population for whom the normal standardisation is not entirely satisfactory, and secondly, that the order of difficulty of the items varies at different levels of mental age, the amount of variation increasing the wider the difference in the level of mental age of the children tested.

In order to determine whether the subjective impressions of the writer, already referred to, had any foundation in fact, the orders of difficulty of the items in each of the four samples were examined and any items for which the age assignment, according to Terman and Merrill, appeared to be incorrect to a marked degree were noted. The criterion of displacement was decided upon quite arbitrarily and items which were displaced by an amount greater than one year were recorded as being either too easy or too difficult. The result of the analysis is shown in Table XII. It is interesting to note that the items which Cole found to be displaced are fewer in number and in most cases quite different from thosefound to be either too easy or too difficult according

to the order of difficulty obtained by testing approved school boys.

An inspection of the results shown in Table XII indicates that tentative conclusions only can be drawn regarding the pattern of difficult and easy items in this test, with the exception of the Minkus item which is clearly too difficult for these boys and the Reading and Report item which is also perhaps on the difficult side. The items which, on the evidence available, appear to be either too easy or too difficult are shown below.

Too Easy		Too Difficult.	
XI,3	Abstract Words I.	X,3	Reading and Report.
XII,5	Abstract Words II.	X,6	Memory (6 numbers).
XIII,1	Plan of Search.	XII,1	Vocab. (14 words).
XIV,3	Picture Absurdities III.	XII,4	5 Digits Reversed.
XIV,6	Abstract Words III.	XII,6	Minkus.

It is not surprising that the "Reading and Report" and "Minkus" items, which require a certain measurable standard of accuracy in English attainment, prove to be stumbling blocks for approved school boys, however, failure at items requiring the use of words is not general throughout the test as will be seen by

the fact that three items involving "Abstract Words" tend to be rather easy.

The easy items, in particular "Plan of Search" and "Picture Absurdities III", are unfortunately placed in the scale as they tend to spread out the testing unnecessarily. For instance, a child who passes only one item at the twelve year level and who, one might expect had reached his scoring limit, will often pass "Plan of Search" and Picture Absurdities III" which requires that he shall go on to attempt the Average Adult level.

The conclusion then is, as far as the range of items studied is concerned, that while non-readers and poor readers will obviously be penalised on a few of the items, the Stanford-Binet test, contrary to expectation, cannot be criticised severely on the grounds that the material of which it is composed is not validly applicable to approved school boys. The order of difficulty of the items, which is closely related to the age assignment, does however, appear to be somewhat variable, and in the case of approved school boys different from that shown in the recently published data on the test. It is considered, however, that gross errors in the calculation of the mental ages of approved school boys would be obviated, to some extent, by careful testing at the "top end" of the scale and by extending the range of items attempted to the year above that in which all items have been failed. This recommendation would apply particularly to a child who failed all items in Year XII.

TABLE XII. X - Items too difficult by an amount greater than 1 year.
0 - Items too easy by an amount greater than 1 year.

Items.	10.0-10.11	11.0-11.11	12.0-12.11	13.0-13.11	Stat. Journal.
Year IX 1. Paper Cutting I				X	
2. Verbal Absurdities II					
3. Memory for Designs.					
4. Rhymes.					
5. Giving Change.					
6. Memory (4 nos.reversed).					
Year X. 1. Vocabulary (11 words).					
2. Pictures Absurdities II.					
3. Reading and Report.	X		X	X	
4. Reasons					X
5. Word Naming (28 points).					
6. Memory (6 nos).			X	X	0
Year XI. 1. Memory for Designs.			X		
2. Verbal Absurdities III.				0	
3. Abstract Words I.			0	0	
4. Memory for Sentences IV.					
5. Word Naming (30 points).					
6. Similarities - 3 things.				X	X
Year XII. 1. Vocabulary (14 words).	X	X			
2. Verbal Absurdities II.					
3. Picture (Telegraph boy).					
4. Digits Reversed (five).		X		X	
5. Abstract Words II.			0	0	
6. Minkus.	X	X	X	X	
Year XIII 1. Plan of Search.	0		0		
2. Memory for Words.					
3. Paper Cutting I.				X	
4. Problems of Fact		0	0	X	
5. Dissected Sentences.	X				
6. Beads II.	0				
Year XIV 1. Vocabulary (16 words).				0	0
2. Induction.					
3. Picture Absurdities III.			0	0	
4. Ingenuity					
5. Orientation Direction I.					
6. Abstract Words II.			0	0	

VII - THE MAXIMUM PREDICTION OF THE CRITERIA.

One of the many subsidiary aims of this investigation was to determine which of the large number of tests in the battery would prove most suitable for use with approved school boys. In this respect the Stanford-Binet Scale, Progressive Matrices and Kohs' Blocks, because of their high reliability coefficients and high loadings in the general factor, stand out above the other tests. It is considered therefore, that these three tests would themselves form a battery adequate for use in a classifying school or similar institution. This conclusion, however, is discussed more fully in the final chapter of this work.

With reference only to the sample of senior boys tested, the Stanford-Binet Scale, Progressive Matrices and Kohs' Blocks have the additional property of possessing higher correlations with the assessments of general intelligence and practical ability than most of the other tests in the battery (see Table VII). If the assessments be regarded as criteria, then this fact can be taken as evidence of the comparatively high validity of these tests. The actual values of these correlation coefficients, which indicate the extent to which the tests are able, individually, to predict the criteria, are not themselves very high. In view of this, the writer considered that it would be worthwhile

investigating whether the general level of prediction could be raised advantageously by brigading the three tests together, each test being suitably weighted. It was decided, therefore, to calculate, firstly, the maximum prediction of each of the criteria separately, and secondly, the maximum prediction of the two criteria together as a compound criterion. In each of the former cases, where the criterion is a single assessment, maximum prediction is obtained by using the regression coefficients as test weights(1). In the latter case, where the criterion can be regarded as a battery of two assessments, weights must be given to the assessments as well as the tests in order to obtain maximum prediction. The method of finding sets of weights for both assessments and tests which yield, mathematically, the highest possible correlation between the batteries of assessments and tests, has been devised by Hotelling(2)(3). By this method, however, it is likely that the weights to be applied to the assessments, that is the components of the criterion, may not be acceptable on psychological grounds, and in actual practice, arbitrary weights

---

(1) Thomson, Sir.G.H. "The Factorial Analysis of Human Ability". Second edition. pp. 87 - 95.

(2) Hotelling, H. "The Most Predictable Criterion". J.E.P.XXVI 1935. pp. 139 - 142.

(3) Thomson, Sir.G.H. "The Maximum Correlation of Two Weighted Batteries". B.J.P. (Stat.Section) Vol.I. Pt.I, October 1947. This article gives examples of the application of Hotelling's method.

are generally assigned to the assessments. The problem, therefore, resolves itself into finding the weights to apply to the battery of tests to give maximum prediction of the criterion, the components of which are arbitrarily weighted.

For instance, if we have a battery of assessments referred to as the "a" variates, and a battery of tests referred to as the "b" variates, the matrix of correlations may be symbolised as

$$\begin{bmatrix} R_{aa} & R_{ab} \\ R_{ba} & R_{bb} \end{bmatrix}$$

Furthermore, if weights "u" are assigned to the assessments and weights "w" to the tests, the above Matrix can be rewritten as a pooling square, thus

	u'	w'
u	R_{aa}	R_{ab}
w	R_{ba}	R_{bb}

from which can be calculated the correlation between the two weighted batteries

$$r = \frac{u' R_{ab} w}{\sqrt{u' R_{aa} u \quad x \quad w' R_{bb} w}} \qquad (1)$$

If now the weights u assigned to the assessments are fixed according to some psychological consideration it can be shown that the weights w to be assigned to the tests to give maximum prediction can be obtained from the equation

$$w' = u'R_{ab} R_{bb} - 1 \quad . \qquad (2)$$

By substituting in equation (1) for w, it becomes

$$r = \sqrt{\frac{u'R_{ab} R_{bb} - 1 \; R_{ba} u}{u'R_{aa} u}} \quad (1) \qquad (3)$$

Equation (3), which gives the maximum prediction, is an expression which contains only the observed matrices of correlation coefficients R_{aa}, R_{ab} and R_{bb} and the arbitrarily assigned values of the assessment weights u.

When the criterion is composed of a single assessment the matrix of correlation coefficients

$$\begin{bmatrix} R_{aa} & R_{ab} \\ R_{ba} & R_{bb} \end{bmatrix} \text{ reduces to } \begin{bmatrix} 1 & r'_{oi} \\ r_{oi} & R \end{bmatrix} \quad (u = 1)$$

---

(1) Peel, E.A. "Prediction of a Complex Criterion and Battery Reliability". B.J.P. (Statistical Section) Vol. I. Part II July 1948.

and the equation for w becomes

$$w' = r'_{oi} R^{-1}$$

The maximum prediction, or multiple correlation

$$r = \sqrt{w' r_{oi}}$$

In estimating the maximum prediction of a criterion by a battery of tests, whether the criterion be single or compound, it is always advisable to calculate the reliability of the weighted battery of tests, for, as Thomson points out(1) the weights for maximum prediction are different from those which give the maximum reliability. Indeed, the best prediction weights may give poor reliability and the best reliability weights may give poor prediction. In view of this, in the calculations on maximum prediction which follow, for each set of weights, an estimate of battery reliability is given.

The Maximum Prediction of the Assessment of Practical Ability.

	Pr.Ab.	St. Binet	Matrices	Kohs
Pr. Ab.	1.000	.440	.462	.525
St. Binet	.440	1.000	.534	.555
Matrices	.462	.534	1.000	.620
Kohs	.525	.555	.620	1.000

Prac. Ability

(1) Thomson, Sir G.H. "Weighting for Battery Reliability and Prediction". B.J.P. Vol. XXX 1939-40. pp. 357 - 366.

$r'_{oi} R^{-1}$ is calculated by Aitken's method of pivotal condensation(1).

1.000	.534	.555	-1.000			1.089
.534	1.000	.620		-1.000		1.154
.555	.620	1.000			-1.000	1.175
.440	.462	.525				1.427
	.715	.324	.534	-1.000		.572
	1.000	.453	.747	-1.399		.800
	.324	.692	.555		-1.000	.571
	.227	.281	.440			.948
		.545	.313	.453	-1.000	.312
		1.000	.574	.831	-1.834	.572
		.178	.270	.318		.766
			.168	.170	.326	.664

$$w' = r'_{oi} R^{-1} = \begin{bmatrix} .168 & .170 & .326 \end{bmatrix} \quad (2)$$

$$r^2_{max} = w' r_{oi} = \begin{bmatrix} .168 & .170 & .326 \end{bmatrix} \begin{bmatrix} .440 \\ .462 \\ .525 \end{bmatrix} = .3236$$

$$r_{max} = .569$$

A useful check on the above calculation is made by applying the weights obtained to the original matrix and by means of the pooling square, calculating r_{max} by an

---

(1) Thomson, Sir G.H. "The Factorial Analysis of Human Ability". Second edition. pp. 92 - 94 and p. 361.

(2) As the tests are already weighted in the ratio of their respective standard deviations, the weights shown here and also in subsequent calculations are not absolute weights.

alternative method. The stages in this calculation are as follows:-

		.168	.170	.326
	1.000	.440	.462	.525
.168	.440	1.000	.534	.555
.170	.462	.534	1.000	.620
.326	.525	.555	.620	1.000

1.000	.0739	.0785	.1712
.0739	.0282	.0153	.0304
.0785	.0153	.0289	.0344
.1712	.0304	.0344	.1063

1.000	.3236
.3236	.3236

$$r_{max} = \frac{.3236}{\sqrt{.3236}} = .569$$

The Reliability of the Battery with the Weights which give Maximum Prediction of the Assessment of Practical Ability.

The method of calculating battery reliability, which is based on the principle of the pooling square, has been described by Thomson(1).

---

(1) Thomson, Sir G.H. "Weighting for Battery Reliability and Prediction". B.J.P. Vol. XXX 1939-40 pp. 357 - 366.

The weights giving maximum correlation are .168 .170 and .326 but to simplify the calculation they have been transformed to 1.000 1.012 and 1.940, which are in the same ratio as the original weights. The reliability coefficients of the Stanford-Binet Scale, Progressive Matrices and Kohs' Blocks are respectively .946 .894 and .862. The stages in the calculation are as follows:-

	1.000	1.012	1.940	1.000	1.012	1.940
1.000	1.000	.534	.555	.946	.534	.555
1.012	.534	1.000	.620	.534	.894	.620
1.940	.555	.620	1.000	.555	.620	.862
1.000	.946	.534	.555	1.000	.534	.555
1.012	.534	.894	.620	.534	1.000	.620
1.940	.555	.620	.862	.555	.620	1.000

1.000	.540	1.077	.946	.540	1.077
.540	1.024	1.218	.540	.915	1.218
1.077	1.218	3.764	1.077	1.218	3.245
.946	.540	1.077	1.000	.540	1.077
.540	.915	1.218	.540	1.024	1.218
1.077	1.218	3.245	1.077	1.218	3.764

11.458	10.766
10.766	11.458

Battery Reliability $= \frac{10.766}{11.458} = .940$

The Maximum Prediction of the Assessment of General Intelligence.

	Gen. Int.	St. Binet	Matrices	Kohs.
Gen. Int.	1.000	.597	.463	.477
St. Binet	.597	1.000	.534	.555
Matrices	.463	.534	1.000	.620
Kohs.	.477	.555	.620	1.000

By the same methods used for the assessment of Practical Ability, the following figures were arrived at:-

$$w' = \begin{bmatrix} .444 & .133 & .148 \end{bmatrix}$$

$$r_{max} = .630$$

Battery reliability = .961

The Maximum Prediction of the Complex Criterion composed of the Assessments of General Intelligence and Practical Ability.

	Gen. Int.	Pr. Ab.	St. Binet	Matrices	Kohs
Gen. Int.	1.000	.445	.597	.463	.477
Pr. Ab.	.445	1.000	.440	.462	.525
St.Binet	.597	.440	1.000	.534	.555
Matrices	.463	.462	.534	1.000	.620
Kohs	.477	.525	.555	.620	1.000

$R_{ab}\ R_{bb}^{-1}$ is calculated by Aitken's method of pivotal condensation.

1.000	.534	.555	-1.000			1.089
.534	1.000	.620		-1.000		1.154
.555	.620	1.000			-1.000	1.175
.597	.463	.477				1.537
.440	.462	.525				1.427
	.715	.324	.534	-1.000		.572
	1.000	.453	.747	-1.399		.800
	.324	.692	.555		-1.000	.571
	.144	.146	.597			.887
	.227	.281	.440			.948
		.545	.313	.453	-1.000	.312
		1.000	.574	.831	-1.834	.572
		.081	.490	.200		.772
		.178	.270	.318		.766
$R_{ab}R_{bb}^{-1}$			.444	.133	.148	.725
			.168	.170	.326	.664

$$w_1 = .444\ u_1 + .168\ u_2$$

$$w_2 = .133\ u_1 + .170\ u_2$$

$$w_3 = .148\ u_1 + .326\ u_2$$

It was decided to weight the assessments in the ratio 1 : 1, thus the weights become:-

$$w_1 = .612 \qquad w_2 = .303 \qquad w_3 = .474$$

$$w' = u' R_{ab} R_{bb}^{-1},$$

hence $u' R_{ab} R_{bb}^{-1} R_{ba} u = \begin{bmatrix} .612 & .303 & .474 \end{bmatrix} \begin{bmatrix} .597 & .440 \\ .463 & .462 \\ .477 & .525 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$

$$= \underline{1.390}$$

$$u' R_{aa} u = \begin{bmatrix} 1 & 1 \end{bmatrix} \begin{bmatrix} 1 & .445 \\ .445 & 1 \end{bmatrix} \begin{bmatrix} 1 \\ 1 \end{bmatrix}$$

$$= \underline{2.890}$$

$$r_{max} = \sqrt{\frac{u' R_{ab} R_{bb}^{-1} R_{ba} u}{u' R_{aa} u}} = \sqrt{\frac{1.390}{2.890}}$$

$$= \underline{.694}$$

The pooling square once more provides a useful check on this calculation -

	1.000	1.000	.612	.303	.474
1.000	1.000	.445	.597	.463	.477
1.000	.445	1.000	.440	.462	.525
.612	.597	.440	1.000	.534	.555
.303	.463	.462	.534	1.000	.620
.474	.477	.525	.555	.620	1.000
	1.000	.445	.365	.140	.226
	.445	1.000	.269	.140	.249
	.365	.269	.375	.099	.161
	.140	.140	.099	.092	.089
	.226	.249	.161	.089	.225

2.890	1.390
1.390	1.390

$$r_{max} = \sqrt{\frac{1.390}{2.890}} = .694$$

The reliability of the battery with the weights which give maximum prediction of the compound criterion was calculated by the method already described, and was found to be .956 .

The Weights which give Maximum Battery Reliability[1].

To calculate maximum battery reliability, the test correlations are set out in a six by six matrix with the test reliability coefficients replacing unity in the diagonals of the North-East and South-West sub-matrices

1.000	.534	.555	.946	.534	.555
.534	1.000	.620	.534	.894	.620
.555	.620	1.000	.555	.620	.862
.946	.534	.555	1.000	.534	.555
.534	.894	.620	.534	1.000	.620
.555	.620	.862	.555	.620	1.000

The matrix can be represented

A	C
C	A

To find the maximum reliability, the determinant $|CA^{-1}C - \lambda A| = 0$ must be solved for λ_1, its largestroot, since λ_1 is the square of the required coefficient.

The evaluation of $CA^{-1}C$ can be simplified by writing

$$D = A - C$$

Thus $CA^{-1}C$ becomes $A - 2D + DA^{-1}D$.

---

(1) Thomson, Sir G.H. "Weighting for Battery Reliability and Prediction". B.J.P. Vol. XXX 1939-40 pp. 357 - 366.

$DA^{-1}D$ can be obtained by Aitken's method of pivotal condensation.

The matrix is rewritten

A	D
-D	.

where $D = \begin{bmatrix} .054 & & \\ & .106 & \\ & & .138 \end{bmatrix}$

The pivotal condensation to obtain $DA^{-1}D$ is shown below:-

1.000	.534	.555	.054			2.143
.534	1.000	.620		.106		2.260
.555	.620	1.000			.138	2.313
-.054						-.054
	-.106					-.106
		-.138				-.138
	.715	.324	-.029	.106		1.116
	1.000	.453	-.041	.148		1.561
	.324	.692	-.030		.138	1.124
	.029	.030	.003			.062
	-.106					-.106
		-.138				-.138
		.545	-.017	-.048	.138	.618
		1.000	-.031	-.088	.253	1.134
		.017	.004	-.004		.017
		.048	-.004	.016		.059
		-.138				-.138
$DA^{-1}D$			.005	-.003	-.004	-.002
			-.003	.020	-.012	.005
			-.004	-.012	.035	.018

$$DA^{-1}D = \begin{bmatrix} .005 & -.003 & -.004 \\ -.003 & .020 & -.012 \\ -.004 & -.012 & .035 \end{bmatrix}$$

$$CA^{-1}C = A - 2D + DA^{-1}D$$

$$= \begin{bmatrix} 1.000 & .534 & .555 \\ .534 & 1.000 & .620 \\ .555 & .620 & 1.000 \end{bmatrix} - 2 \begin{bmatrix} .054 & & \\ & .106 & \\ & & .138 \end{bmatrix} + \begin{bmatrix} .005 & -.003 & -.004 \\ -.003 & .020 & -.012 \\ -.004 & .012 & .035 \end{bmatrix}$$

$$= \begin{bmatrix} .897 & .531 & .551 \\ .531 & .808 & .608 \\ .551 & .608 & .759 \end{bmatrix}$$

It is required to find the largest root of

$$\left| CA^{-1}C - \lambda A \right| = 0$$

that is of

$$\begin{vmatrix} .897 - \lambda & .531 - .534\lambda & .551 - .555\lambda \\ .531 - .534\lambda & .808 - \lambda & .608 - .620\lambda \\ .551 - .555\lambda & .608 - .620\lambda & .759 - \lambda \end{vmatrix} = 0$$

The largest root λ_1 must be between the square of the highest individual test reliability and unity.

That is, between $.946^2$ or .895 and 1.000

By substituting various values for λ and calculating the value of the determinant in each case, the value for λ_1 can be arrived at by interpolation:-

λ =	.9500	Δ =	-.001205
λ =	.9250	Δ =	-.000015
λ =	.9248	Δ =	-.000013
λ =	.9245	Δ =	-.000010
λ =	.924425	Δ =	-.000001 (By interpolation).
λ =	.9244	Δ =	+.000004
λ =	.9242	Δ =	+.000027

$$\lambda_1 = .924425$$

Maximum Battery Reliability $= \sqrt{.924425} = .961$

$$\therefore \quad \left| CA^{-1}C - \lambda_1 A \right| = \begin{vmatrix} -.0274 & .0374 & .0380 \\ .0374 & -.1164 & .0349 \\ .0380 & .0349 & -.1654 \end{vmatrix} = 0$$

The weights which give the maximum battery reliability of .961 are proportional to any row of the adjugate of:-

$$\left[CA^{-1}C - \lambda_1 A \right]$$

That is , of $\begin{bmatrix} .0180 & .0075 & .0057 \\ .0075 & .0031 & .0024 \\ .0057 & .0024 & .0018 \end{bmatrix}$

Thus the weights are proportional to

1.000 .417 .317

These weights were applied to the six by six matrix shown on page 102, and by means of the pooling square, the accuracy of the calculation which arrived at a maximum battery reliability of .961 was confirmed.

Thomson's method for calculating maximum battery reliability used above, is somewhat slower and more cumbersome than the method recently published by Peel(1), and Gulliksen(2), in discussing the two methods states that "the solution given by Thomson can be shown to be equivalent to that given by Peel", but advocates the use of Peel's method as it is much simpler.

---

(1) Peel, E.A. "Prediction of a Complex Criterion and Battery Reliability". B.J.P. (Stat. Section) 1948. Vol.I. Part II. pp. 87 - 89.

(2) Gulliksen, H. "Theory of Mental Tests". Chapman & Hall 1950 p. 346.

Summary of Results.

Criterion	Weights			Maximum Prediction	Battery Reliability
	1. St.Binet	2 Matrices	3 Kohs		
Practical Ability	1.000	1.012	1.940	.569	.940
General Intelligence	1.000	.300	.333	.630	.961
Practical Ability with General Intelligence	1.000	.495	.544	.694	.956
For maximum Battery Reliability	1.000	.417	.317	-	.961

In order that comparisons can be made easily the correlations of the three tests with the two assessments together with the test reliabilities are given below:-

	Prediction of		Test Reliabilities
	P. Ability	Gen. Int.	
1. St. Binet	.440	.597	.946
2. Matrices	.462	.463	.894
3. Kohs	.525	.477	.862

It will be seen that the highest prediction of the assessment of practical ability, by a single test is attained by Kohs' Blocks with a correlation with the assessment of .525 and a reliability of .862. Brigading the three tests and weighting them to given maximum prediction of the assessment of practical ability, raises the prediction to .569 which is not in fact substantially higher than that attained by the single test. The reliability of the battery is, however, much higher than that of the single test.

The Stanford-Binet Scale is the best single-test predictor of the assessment of general intelligence with a correlation with the assessment of .597 and a reliability of .946. The battery of three tests with each test suitably weighted gives a maximum prediction of the assessment of .630. The reliability of the battery is approximately the same as that of the single test.

In both the above cases, the weighted battery has the two-fold advantage over the single tests, of giving higher prediction of the assessments at an acceptable level of reliability. It could be argued from this that the weighted battery has therefore a higher validity in estimating "intelligence".

The maximum battery reliability of .961 would not appear

to be a very critical value and in rounding off decimal places in the above calculations it can be attained within a small range of variation of the weights applied to the battery for various purposes. This accounts for the fact that the weights which give maximum prediction of the assessment of general intelligence also give a battery reliability of .961.

The criterion compounded of the equally weighted assessments of practical ability and general intelligence has a maximum prediction by the battery of .694 which is a high value considering the very subjective nature of the two assessments. From the practical point of view, the older boys who pass through a classifying school are more or less beyond the stage where further education in the classroom can be provided and the main problem in their case is to discover potentialities for vocational training. The criterion compounded of practical ability and general intelligence, would therefore appear to be a useful concept, as far as the senior approved school boys are concerned.

In a practical application of the results of this chapter, weights used would, of course, be whole number as shown below:-

Prediction of	Weights		
	1 St. Binet	2 Matrices	3 Kohs
Practical Ability	1	1	2
General Intelligence	3	1	1
Practical Ability with General Intelligence	2	1	1

The technique of battery weighting having been fully explored, the writer is of the opinion that the work described above can be regarded as a pilot exploration only of the possibilities of using a battery of tests with various weights to estimate different aspects of the potentialities of approved school boys. The results in general are encouraging though it is considered that the implementation of the technique would require of itself a major research investigation.

The technique of battery weighting for maximum prediction (or maximum validity) has been used in this country principally in connection with the selection of children for grammar and technical schools(1).

—oOo—

(1) Peel, E.A. and Rutter D. "The Predictive Value of the Entrance Examination as Judged by the School Certificate Examination". B.J.Ed.P. Vol. XXI Part I. pp. 30 -35. February 1951.

VIII - SUMMARY AND CONCLUSIONS.

1. The Investigation.

A sample of 327 approved school boys who were consecutive entrants to the Aycliffe Classifying School, were tested by a miscellaneous battery of intelligence tests and attainment tests. In addition, subjective assessments of the boys' general intelligence were made by the housemasters, and in the case of the Senior boys (about half the total sample), subjective assessments of practical ability in the workshop were also made.

The purpose of the investigation was to discover whether the intelligence tests used gave reliable and valid estimates of the intelligence of approved school boys and also whether certain of the tests could indicate special ability in practical work.

No attempt was made to estimate the reliability and validity of the four tests of educational attainment. These were included in the battery to provide information regarding the influences which level of educational attainment might have in the scores obtained in certain of the tests.

The tests used were:-

Psychological:

1. Essential Verbal Group Test Form A.
2. Essential Verbal Group Test Form B.
3. Simplex Group Intelligence Test.
4. Stanford-Binet Form L.
5. Progressive Matrices.
6. N.I.I.P. 70/23 (Non-Verbal).
7. V.S.10.
8. T.S. 8.
9. Aycliffe I.
10. Passalong.
11. Kohs' Blocks.
12. Cube Construction.
13. Blocks Performance.

Attainment:

1. Burt's Reading Test No.1.
2. Burt's Dictation.
3. Composition (Story completion).
4. Arithmetic.

Assessments:

1. General Intelligence.
2. Practical Ability (senior boys only).

For statistical treatment the sample of 327 boys was divided into two parts:-

Juniors - 9 years to 14½ years.

Seniors - 14½ years to 17 years.

2. The Test Reliabilities.

It was decided that in estimating reliability coefficients, it would be advisable to select a method which could be applied

to as many of the tests as possible as this would permit valid comparisons to be made between the coefficients obtained for the different tests. The Kuder-Richardson method was therefore selected and it was applied to all the tests except Passalong, Kohs' Blocks and Cube Construction. The Split-half method was used for Passalong and Kohs' Blocks and an analysis of variance method for Cube Construction. Coefficients were not estimated for the verbal group tests Essential B and Simplex, as it was assumed they would be of the same order as that obtained for Essential A. The methods selected had the additional advantage that the reliability coefficients could be estimated by a single application of each of the tests.

The findings were that only three tests in the whole battery, Essential A, Stanford-Binet Form L and Progressive Matrices achieved a satisfactory level of reliability. The coefficients for these tests were all greater than 0.9 for both Junior and Senior groups. It was noted too, that the coefficients of these tests, which were estimated by the Kuder-Richardson method, were likely to be under-estimates.

The use of the Kuder-Richardson method to estimate the reliability of the Stanford-Binet test was considered to be something of an innovation.

Kohs' Blocks obtained a coefficient of 0.87 and it was felt that the original form of this test, as devised by Kohs with 17 cards, would have a higher reliability than the shortened form by Alexander which was used in this investigation.

The remainder of the tests, it was considered, failed to achieve satisfactory levels of reliability, mainly because they contained an insufficient number of items. Although some of them could be lengthened, certain others, by the nature of their material, would be difficult to increase in length.

Out of the total battery, it could be said that only Essential A (and of course, the other two verbal group tests), the Stanford-Binet, Progressive Matrices and the original form of Kohs' Blocks, proved to have acceptable levels of reliability.

3. The Test Validities.

(i) By Factor Analysis.

The inter-correlations between the various tests and assessments were calculated for both Junior and Senior groups and after the effect of chronological age had been removed, the matrices of correlation coefficients were factorised by Thurstone's centroid method, three factors being extracted in each case.

It was hoped that the factor patterns would permit rotation to give, in the case of the Senior boys, clear evidence of a connection in the third factor between the performance tests (and some of the non-verbal tests) and the assessment of practical ability. The standard errors of the correlation coefficients were high due to the small numbers in the two samples and it was doubtful if it was strictly legitimate to proceed with the extraction of the third factor. It was decided, therefore, to leave the factors in their unrotated condition. In the case of the Senior group, therefore, it was not possible to provide clear evidence that the "space" and performance tests did in fact measure practical ability. It was noted, however, that the assessment of practical ability appeared at the negative "end" of the second bi-polar factor along with the non-verbal and performance tests.

As rotations were not performed, this meant that factor one in each analysis was not, strictly speaking, identifiable with "g", the factor of general intelligence. It was considered, however, that in the unrotated condition, firstly, valid comparisons between the loadings of individual tests could still be made, and secondly, it would not be completely unjustifiable to regard factor one in each case as approximating to "g" and the loadings as validity coefficients of the individual tests.

The three verbal group test and the Stanford-Binet test

stood out above all others with loadings of about 0.85 in the first factor for both Juniors and Seniors. To the best of the writer's knowledge, the Stanford Binet test has never before been analysed in a battery of miscellaneous tests. A scrutiny of the loadings in the three factors obtained showed that this test was not really distinguishable from the three verbal group tests as far as the factor analysis was concerned.

The non-verbal group tests all obtained loadings in the first factor of the order of 0.70. Among the performance tests Kohs' Blocks obtained a loading of 0.75 which was considerably higher than those obtained by the other performance tests.

The second factor (a bi-polar factor) showed clearly the dichotomy between the verbal and educational tests on the one hand, and the non-verbal and performance tests on the other. The assessment of general intelligence appeared grouped with the verbal and educational tests, while the assessment of practical ability appeared with the non-verbal and performance tests.

The third factor (also a bi-polar factor) possessed small loadings and it was considered that they were too unreliable for definite conclusions to be drawn from them.

(ii) By the Criteria.

The correlations of the tests with the assessment of general intelligence were comparatively small, being highest with the tests which proved themselves to be the most reliable and to have the highest loadings in factor one. The smallness of the correlations was in general, disappointing and it was felt that the manner in which the assessment had been made was, unhappily, not entirely satisfactory. It was realised in asking the housemasters to assess general intelligence on the basis of the "common-sense" shown in the daily routine, that they had been given a difficult task. In such a population as this, the factors likely to mislead an observer are numerous, and it was concluded that estimates by persons made solely from observation of boys in the daily routine should be accepted with some reserve.

The assessments of practical ability which were made by one person after observing the boys' efforts at various practical tasks in the workshop, it was considered must provide a reasonably satisfactory criterion of practical ability. The correlations which the tests obtained with the assessments showed that the tests were not able to predict practical ability to any great extent. It was considered that in this case the fault lay chiefly with the tests rather than the assessments,

as it was noticed that the test which obtained the highest correlation with the criterion was Kohs' Blocks, which was also the most reliable of the "space" and performance tests. The writer's conclusion is, therefore, that before practical ability, as such, can be predicted satisfactorily by means of tests, test constructors must produce tests which not only have the elements of "space" and performance in them but must also be highly reliable measuring devices.

4. Reading Ability and Verbal Group Tests.

Using the Essential Verbal Group Test Form A, an attempt was made to determine the minimum reading age at which a boy can be expected to produee a valid score on the test. The scripts of some 200 boys were examined item by item and it was discovered that a reading age of about 9½ years on Burt's Reading Test No.1 was required before a boy possessed the necessary ability to attempt all the items in the test. The writer considers that before a child can produce a valid score on such a test he must at least, be able to attempt all the items in it.

This conclusion raised a further problem, in that a large number of boys clearly misunderstood the instructions relating towhole blocks of items and thus failed all the items in the block or section. The writer suggests that when such tests are

being constructed, as well as an item analysis, an investigation into the efficiency of the instructions should always be carried out.

A very practical result of this particular investigation was the conclusion that verbal group tests of intelligence are very limited in their application to approved school boys, as about one third of them are unable to do themselves justice on such tests because of backwardness in reading.

5. The Stanford-Binet Test, Form L.

The orders of difficulty of the items from Year IX 1 to Year XIV 6 were calculated for four groups of boys. Each group was composed of boys of mental ages as follows:-

1. 10.0 - 10.11
2. 11.0 - 11.11
3. 12.0 - 12.11
4. 13.0 - 13.11

The orders of difficulty obtained were compared with each other and with the order of difficulty published recently by Cole. Rank order correlations showed that the orders of difficulty changed with increasing mental age and also that the "approved school" orders of difficulty correlated more highly amongst themselves than they did with Cole's order of difficulty.

A further investigation of the items which, from the point of view of approved school boys, appeared to be displaced in the

order of difficulty showed that in general approved school boys have difficulty with items involving reading and especially with the Minkus item.

The above findings suggested that (a) the assumption (fundamental to any test in which the items are assigned a fixed mental age level) that the order of difficulty does not change at different age levels, is untrue, (b) the test depends to some extent on scholastic attainment especially English, and with approved school boys, who are in general very backward in reading, an under-estimate of general intelligence will perhaps be obtained in a large number of cases.

6. The Maximum Prediction of the Criteria.

The Stanford-Binet Test, Progressive Matrices and Kohs' Blocks were brigaded together to form a small battery, the purpose of which was to determine whether by weighting the tests, the prediction of the criteria, formed respectively of the assessments of general intelligence and practical ability, could be improved. These three tests were selected to form the small battery because of their apparent high reliability and validity.

The weights for maximum prediction were used and it was noted that not only was the level of prediction of the criteria

improved but that the reliability of the battery was also at an acceptably high level in each case.

The conclusion arrived at by the writer from the results of this pilot investigation was that in cases where a reasonably satisfactory criterion can be established, the use of a weighted battery has considerable advantages in prediction and possible reliability over a single test, and that further research into the practical applications of the weighted battery technique could be profitably undertaken in a wide variety of fields.

7. General Conclusions.

The investigation showed that contrary to expectation approved school boys do in general apply themselves satisfactorily to intelligence tests. This is borne out, firstly, by the high reliabilities obtained for a number of the tests, in particular the verbal group intelligence tests which were considered to be relatively unattractive to approved school boys, and secondly, by the subjective impressions of the writer and his colleagues. The general impressions gained was that the majority of boys co-operated and were anxious to do as well as they could in the testing.

This conclusion is somewhat different from what other persons who have tested delinquent children have found, and

it is suggested that the reliable results obtained in the present investigation are in a large measure due to the generally pleasant and friendly atmosphere in the Classifying School. Anti-social attitudes and behaviour and states of emotional anxiety or maladjustment in approved school boys are in the main due to extrinsic factors in the past environment. In the Classifying School, the spot-light of sympathetic individual attention is focussed on them from the moment they arrive in the school and the response which they make is immediate almost without exception. The testing in this investigation was carried out by persons who had worked and played with the boys thus the relationship between staff and boys conducive to good rapport was actually established before the testing situation was approached.

It is considered that psychologists and psychiatrists to whom delinquent children are brought for assessment in child guidance clinics and hospitals are at a great disadvantage and it is not perhaps surprising, since the children are faced with the ordeal of an interview with a strange person in a strange place, that test results are sometimes found to be unreliable.

It is considered, therefore, that the residential and semi-informal "set-up" in a classifying school provides almost deal conditions for the carrying out of psychological work with delinquent boys.

The tests in the battery which proved to be acceptably reliable and valid were, the three verbal group tests, the Stanford-Binet, Progressive Matrices and Kphs' Blocks. Unfortunately the verbal group tests have a limited use with approved school boys as anything up to one third of them are unlikely to be able to produce valid scores because of reading difficulties. Furthermore, few verbal group tests are standardised for chronological ages greater than about 12 years.

The Stanford-Binet test stands vindicated as having high reliability and validity and proves to be a useful psychometric tool. Nevertheless, it stands very much in need of revision and re-standardisation for use with children in this country.

Progressive Matrices and Kohs' Blocks also proved themselves to be acceptably reliable and valid for both Junior and Senior groups. These two tests together with the Stanford-Binet test form a small battery which the writer considers to be most adequate for measuring the general intelligence of approved school boys.

—oOo—

BIBLIOGRAPHY.

ALEXANDER, W.P. "Intelligence: Concrete and Abstract". B.J.P. Monograph Supplement No.19. 1935.

BLACKBURN, J. "The Influence of Social Environment on Intelligence Test Scores". 1948. British Social Hygiene Council.

BURT, Sir C. "The Young Delinquent". Fourth edition 1948.

BURT, Sir.C. "Mental and Scholastic Tests". Second edition, October 1947.

BURT, Sir.C. "The Latest Revision of the Binet Intelligence Tests". The Eugenics Review XXX, 1939. pp. 255 - 260.

BURT, Sir C. "The Reliability of Teachers' Assessments of their Pupils". B.J.Ed.P. Volume 15. pp. 80 - 92.

BURT, Sir C. "The Psychological Implications of the Norwood Report". B.J.Ed.P. No.13. 1943. pp. 126 - 140.

BURT, Sir C. Part IX conclusion to the "Symposium on the Selection of Pupils for Different Types of Secondary Schools". B.J.Ed.P. Volume XX Part I. February 1950 pp.1-10.

CATTELL, R.B. "A Guide to Mental Testing". Second edition 1948.

CHAMBERS. "Statistical Calculation". (1945).

COLE, R. "An Item Analysis of the Terman Merrill Revision of the Binet Tests". B.J.P. (Statistical Section) November 1948. pp. 137 - 151.

EL KOUSSEY, A.A.H. "The Visual Perception of Space". B.J.P. Monograph Supplement No.20. 1938.

EMMETT, W.G. "Evidence of a Space Factor at 11+ and Earlier". B.J.P. (Statistical Section) Volume II Part I. March 1949. pp. 3 - 16.

FERGUSON, G.A. "The Reliability of Mental Tests". 1940.

FISHER, R.A. "Statistical Methods for Research Workers". 1946 edition.

GARRETT, H.E. "Statistics in Psychology and Education". Longmans Green and Co. 1945.

GUILDFORD, J.P. "Psychometric Methods". Mc.Graw-Hill publications, 1936.

GULLIKSEN, H. "Theory of Mental Tests". Chapman and Hall Ltd., 1950.

GUTTMAN, L. "A Basis for Analysing Test-retest Reliability". Psychometrika X. (1945) pp. 255 - 282.

HARTOG, RHODES and BURT. "The Marks of Examinders". 1936. Memorandum I.

HOTELLING, H. "The Most Predictable Criterion". J.E.P. XXVI. 1935. pp. 139 - 142.

HOYT, C. "Test Reliability obtained by Analysis of Variance". 1941 Psychometrika 6. pp. 153 - 160.

JACKSON, R.W.B. and FERGUSON, G.A. "Studies on the Reliability of Tests". Bulletin No.12 Dept. of Educational Research, Toronto University.

KELLY, T.L. "Crossroads in the Mind of Man". California Standford University Press 1928.

KELLY, T.L. "The Reliability Co-efficient". 1942 Psychometrika VII pp. 75 - 83.

KOHS, S.C. "Intelligence Measurements". 1923.

KUDER, G.F. and RICHARDSON, M.W. "The Theory of the Estimation of Test Reliability". Psychometrika II (1937). pp.151 - 160.

MANNHEIM, H. and SPENCER, J. "Problems of Classification in the English Penal and Reformatory System". Published by the Institute for the Scientific Treatment of Delinquency in 1950.

MELLONE, M.A. "Reading Ability and I.Q.". B.J.Ed.P. June 1942. pp. 128 - 135.

PEEL, E.A. "Evidence of a Practical Factor at the Age of Eleven". B.J.Ed.P. XIX Part I. Feb. 1949. pp. 1 - 15.

PEEL, E.A. "Prediction of a Complex Criterion and Battery Reliability". B.J.P. (Statistical Section) Volume I. Part II. July 1948. pp. 84 - 94.

PEEL, E.A. and RUTTER, D. "The Predictive Value of the Entrance Examination as Judged by the School Certificate Examination". B.J.Ed.P. Volume XXI. Part I. (Feb.1951). pp. 30 - 35.

SCHONELL, F.G. "Development of Educational Research". B.J.Ed.P. Volume XVIII. Part I. Feb. 1948. pp. 1 - 15.

SLATER, Patrick. "Tests for Selecting Secondary and Technical School Children". Occupational Psychology (1941) XV (1), 10.

SLATER, Patrick. "Some Group Tests for Spatial Judgment or Practical Ability". Occupational Psychology January 1940.

SLATER, PATRICK. "The Development of Spatial Judgment and its Relation to some Educational Problems". Occupational Psychology, July 1943.

SPEARMAN, C. "The Abilities of Man". Macmillan 1927.

TERMAN, L.M. and MERRILL, MA. "Measuring and Intelligence". 1937.

THOMSON, Sir G.H. "The Factorial Analysis of Human Ability". Second edition.

THOMSON, Sir G.H. "The Maximum Correlation of Two Weighted Batteries". B.J.P. (Statistical Section) Volume I. Part I. Oct.1947. pp. 27 - 34.

THOMSON, Sir G.H. "Weighting for Battery Reliability and Prediction". B.J.P. Volume XXX 1939 - 40. pp. 357 - 366.

THOULESS, R.H. "Test Unreliability and Function Fluctuation". B.J.P. 1936 XXV. pp. 325 - 343.

THURSTONE, L.L. "Primary Mental Abilities". Psychometric Monograph No.1. University of Chicago Press 1937.

VERNON, P.E. "The Measurement of Abilities". U.L.P.Ltd., 1940.

VERNON, P.E. "The Structure of Human Abilities". Methuen 1950.

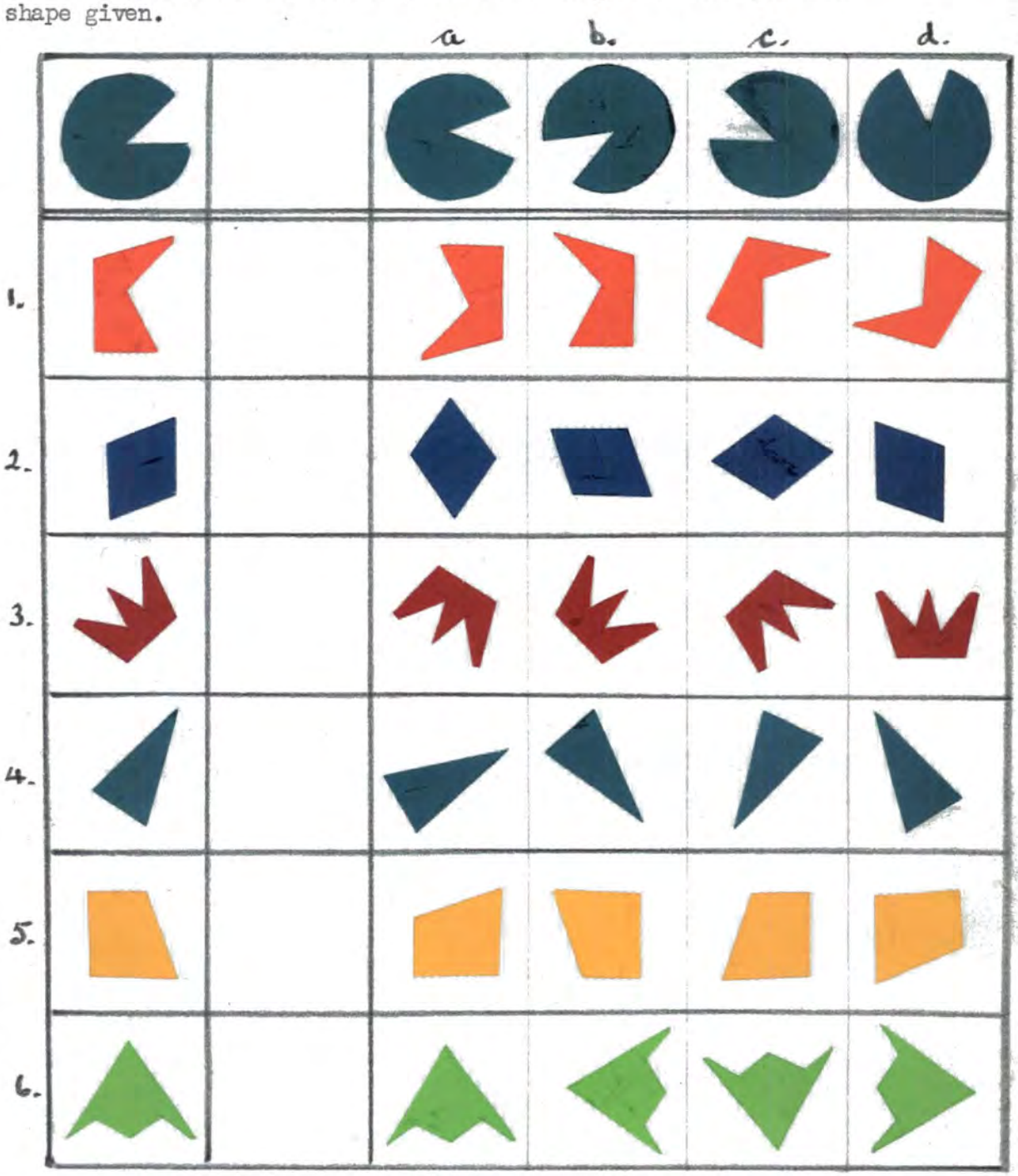
—oOo—

APPENDIX I.

SPECIMEN OF CARD 1 AYCLIFFE I.

It is required to find the mirror image of the shape shown on the left hand side of the card.

The top row is used for demonstration purposes. A piece of cardboard cut in the shape of the required mirror image is used to demonstrate and this is turned over like the page of a book and placed in the blank space on the right of the specimen shape given.



APPENDIX II.

BELOW IS A SPECIMEN OF THE "INSTRUCTIONS" GIVEN TO HOUSEMASTERS TO GUIDE THEM IN MAKING THE ASSESSMENTS OF GENERAL INTELLIGENCE.

"The Housemasters of the Classifying School are asked to contribute to the investigation by making subjective estimates of the intelligence of all the boys who pass through their respective houses. The estimates will be given on a 15 point scale as shown below:-

M.D.	Very Dull	Below Average	Average	Above Average	Superior	Very Superior Indeed.
-7	-6 -5	-4 -3	-2 -1 0 1 2	3 4	5 6	7
		BELOW	AVERAGE		ABOVE.	

A few suggestions about the correct use of the scale may be helpful. The estimates must be subjective estimates based solely on observations of the boys during their residence in the house. By watching the way they cope with the daily routine, by talking with them, working and playing with them, it should be possible to form some idea of the amount of "gumption", "brains" or "common sense" that each possesses.

A word or two of warning is perhaps necessary. Care must be taken not to over-score a boy who has an attractive appearance, pr who is quiet, biddable or co-operative, or who is a facile talker.

On the other hand, of course, the under-scoring of those who are reticent or non-co-operative must be guarded against. Personal likes and dislikes too, must not be allowed to influence judgment.

The age of the boy for whom an estimate is being made, is an important factor to be considered. The younger ones cannot be expected to be as knowledgeable about things in general as the older ones. Therefore, some allowance must be made for age, otherwise the younger boys would be under-scored and the older ones perhaps over-scored. Finally, it must be stressed that on no account must scholastic attainment or the results of previous or current mental tests be allowed to influence the assessments. Indeed, it would be better if housemasters allowed themselves to remain in ignorance of any test results until after they had made their own assessments.

In actually using the scale, the soundest way is, first of all to decide whether a boy is AVERAGE, ABOVE or BELOW, and having decided this, to consider his placing within the selected range. It is very advisable to look at the printed scale while making the placement.

The groupings on the scale refer of course, to the distribution of intelligence in the whole population and not to the Aycliffe standards.

It must not be thought that since the majority of boys

passing through the Classifying School appear to be on the dull side, that boys at the top end of the scale do not also crop up from time to time. At the other end of the scale, category '-7' (Mentally deficient) must be given if the housemaster considers this to be the true category, whatever might be the official opinion of this particular boy's mental state. The estimates on the scale will therefore cover the whole 15 points but will tend to bunch near the average mark".

—oOo—

APPENDIX III

Name.....

Date.....

A HOLIDAY ADVENTURE.

One fine day during the holidays, I set out to explore a big wood near my home. As I was walking through the wood I came across a broken-down cottage. Part of the roof had fallen in, the door was missing, and most of the window panes were broken. It looked as though no one had lived there for a long time. When I looked through the doorway, however, I was very surprised to see a pot cooking on a fire in the middle of the floor. Just as I was about to peep in at the window, I heard someone coming through the wood behind me, so I hid behind a bush to see who it might be.