THE RELIABILITY OF MENTAL TESTS

BY

GEORGE A. FERGUSON

Thesis submitted for the Degree of Doctor of Philosophy in Education at the University of Edinburgh.

May 15, 1940.

CONTENTS

PART I

PREFACE

The thesis here presented is divided into two parts. Part I is largely a theoretical discussion of problems concerning the reliability of mental tests. Suggestions are made for increasing the reliability and general efficiency of tests as instruments for the selection of individuals for specified purposes. Part II is experimental in type, and is devoted to a consideration of the reliability of Moray House Tests of Intelligence, Arithmetic, and English. Comparisons are made between the reliability of Moray House Group Tests of Intelligence, and the reliability of the Stanford Binet scale (new revision). Data are presented regarding the constancy of the Intelligence Quotient as measured by Group Tests of Intelligence.

Some discussion and calculation appears in Chapter 5 (pp. 67-90) which is a repetition of material appearing in the previous Chapter. Chapter 5, "A Bi-factor Analysis of Reliability Coefficients", has been submitted as it stands to the British Journal of Psychology for publication. The necessary clerical work involved in rewriting this section to eliminate slight overlap with previous sections did not seem justified.

The notation and terminology of Chapter 7, "Theories of Test Structure, and Methods for Improving the Efficiency of Tests", is not satisfactory, but is the best I could attain at the time of writing.

I wish to extend my sincere thanks to Professor Godfrey H. Thomson and Mr. W.G.Emmett for encouragement, assistance, and valuable criticism throughout the course of the work, and also for the use of statistics and other data in the Moray House records. Thanks are also due to Mr. D.N.Lawley for assistance in the solution of certain mathematical problems. I am also deeply indebted to the Doncaster Education Authority for permission to use statistics in their records.

George A. Ferguson, B.A., B.Ed.

Moray House,
Edinburgh,
May 2nd.,1940.

PART I.

Part I is largely concerned with the theoretical aspects of the reliability of mental tests. Some suggestions are made for increasing the reliability and the general efficiency of tests.

THE GENERAL CONCEPT OF RELIABILITY:

The estimation of quantitative values is in all science characterized by inaccuracies of observation. The concept of an inaccurate observation antithetically implies the existence of a true value to which a given series of observations may approximate in greater or less degree. The existence of a true value is in the last analysis a philosophical abstraction and cannot be known. None the less the scientist must accept the belief that true values of the quantities which he presumes to measure do exist, perhaps only in the mind of an omnipotent deity, otherwise the logical presumptions of his science become invalid, and his scientific observations become meaningless randomizations. The true value of any given quantity may be defined in the statistical sense as the mean of an infinite number of fallible observations of that quantity. Since an infinite number of observations can never be made, the true value is never exactly determinable. Given this concept we may define an error of measurement as the difference between this hypothetical true value, and any single fallible estimate of that value.

Now scientific measurements are of many different types. Certain quantities may be measured directly, while others can be measured only through a knowledge of certain functional relationships. The quantitative nature of certain phenomena can only be inferred indirectly by a knowledge of their effect

THE GENERAL CONCEPT OF RELIABILITY.

The estimation of quantitative values is in all science characterised by inaccuracies of observation. The concept of an inaccurate observation antithetically implies the existence of a true value to which a given series of observations may approximate in greater or less degree. The existence of a true value is in the last analysis a philosophical abstraction and cannot be known. None the less the scientist must accept the belief that true values of the quantities which he presumes to measure do exist, perhaps only in the mind of an omnipotent deity, otherwise
1 the logical presumptions of his science become invalid, and his scientific observations become meaningless randomisations. The true value of any given quantity may be defined in the statistical sense as the mean of an infinite number of fallible observations of that quantity. Since an infinite number of observations can never be made, the true value is never exactly determinable. Given this concept we may define an error of measurement as the difference between this hypothetical true value, and any single fallible estimate of that value.

Now scientific measurements are of many different types. Certain quantities may be measured directly, while others can
2 be measured only through a knowledge of certain functional relationships. The quantitative nature of certain phenomena can only be inferred indirectly by a knowledge of their effect

on certain other phenomena. In other cases quantitative
description is attained by measuring responses relative to a
specified set of circumstances. The measurement of mental
abilities in the field of psychological science is of this
latter type; that is, we describe the traits of individuals
in terms of their responses to a specified set of circumstances,
namely the test situation.

The presumption of mental measurement is that mental
traits exist in some amount, and that they can be quantitatively
3 described by the measurement of ability, an ability being
defined by what an individual can do. The inference is that
what an individual can do bears some correspondence to
certain characteristics of mind, which characteristics are
known as traits. Now, since any one individual can perform
a multiplicity of operations, we can never exactly determine
the extent of a person's ability by a test situation. The
only remaining course is to measure under certain specified
conditions a limited number of things a person can do,
regarding the performance of a person on a limited number of
4 tasks as representative of his hypothetical potential
performance. Thus a mental test samples a persons ability.
The more representative the abilities as measured by the test
are of all the abilities possessed by the individual the more
valid the test. Thus, low test validity may be described as
errors due to the sampling of ability. This concept of test
validity requires further consideration. We usually attempt

to measure the validity of tests by describing them with reference to external criteria, teachers' estimates, success in secondary school or in an occupation, but these criteria are themselves merely samples of the total population of things that persons can do. We presume, however, that these criteria, while they themselves are invalid due to errors in the sampling of ability, are in all likelihood based on larger and more representative samples than the sample of ability measured by a test or a battery of tests; consequently we regard them as more valid indices of a persons hypothetical potential performance.

As well as errors resulting from the unrepresentative sampling of ability, another fundamental type of error results from the inaccuracy with which a test measures the sample of ability which it measures. Errors of this type are embraced in the concept of test reliability. Due to a multiplicity of causes, certain tests are more accurate instruments of measurement than others. According to the magnitude of the errors made in measuring the sample of ability which a test measures, we describe it a s being more or less reliable. Reliability is not directly concerned with whether the sample of ability as measured by the test is a representative sample of all the abilities of any one person, but with the errors of observation made in defining that sample.

To revert to the concept of true values in scientific
7 measurement discussed above, the psychologist must assume
that true values of the quantities which he measures, exists,
although these true values are only defined relative to the
test. Thus we must presume that a true score exists on any
given test for any given person, certain specified conditions
being kept constant, from which a given observation may err
in greater or less degree. If errors of measurement are due
to a multiplicity of random causes they are believed to obey
certain well defined laws; that is, we find in practice
that errors of measurement approximate to the normal law
8 of errors. Errors of measurement in the measurement of
mental abilities are also assumed to obey this normal law of
(1) errors, and this assumption has been verified empirically.
By the computation of the appropriate parameters the
distribution of errors of observation made by any mental
test, may be determined. The parameters defining this
distribution of differences between the observed and true
values are used in determining the accuracy with which a test
measures the sample of ability which it measures. From a
knowledge of these parameters we can estimate the probability
that a given observation deviates by some given amount from
9 the hypothetical true value.

The normal law of error holds when there are a large
number of independent sources of error, each of which is
(2) normally distributed. The error variances of different

sources of error are directly additive when the errors are uncorrelated. Thus if $\mathcal{E}^2$ represents the total error variance, and $s_1^2, s_2^2, s_3^2, \cdots \cdots s_K^2$ are the variances of k independent sources of error, we may write

$$\mathcal{E}^2 = s_1^2 + s_2^2 + s_3^2 \cdots s_K^2 = \sum_{i=1}^{K} s_i^2$$

If, however, the errors are not independent but are correlated, the above equation becomes

$$\mathcal{E}^2 = \sum_{\substack{i=1 \\ i \neq j}}^{K} \sum_{j=1}^{K} r_{ij} s_i s_j + \sum_{i=1}^{K} s_i^2$$

The above functions enable us to measure what part of the total error variance is due to some particular source, when
10 that particular source of error can be isolated and controlled under experimental conditions. If, however, the distribution of errors were not found to obey the normal law, we should presume that one or more of the component variances were due to the operation of certain systematic factors, which in themselves were not normally distributed. We might, therefore, proceed to control such systematic factors and describe their distributions.

In estimating the magnitude of the errors involved in any measurement we can (a) make a large number of observations of a single quantity under constant conditions, and from the
11 distribution of the differences between each observation, and the mean of the observations estimate the error variances,

or we can (b) make two observations of a series of variable quantities, and from the distribution of differences between the two observations of each quantity estimate by an appropriate technique, on the assumption that the errors are random and uncorrelated, the variance of the errors involved. The variance of the distribution of differences between two series of fallible observations of a variable quantity is found to be twice the variance of the differences between a single series of observations and the true values. This
17 observation is directly apparent on reference to the additive nature of the variances of independent sources of error. With two series of observations, each assumed equally fallible, the variance of the difference between the two series is made up of two components, the variance of the differences between one series of observations and the true values, and the variance of the differences between the other series of observations and the true values.

The determination of reliability by a large number of observations of a single quantity is not applicable in the field of mental testing due to the influence of certain
13 psychological factors. Consequently reliability must be determined by making two series of observations of a single variable quantity. Thus the psychologist makes two series of observations of what is presumed to be the same mental abilities, and finds the correlation between the two series. This correlation between two series of fallible observations

is in general use, and is termed the reliability coefficient. It is, of course, possible to find the variance of the distribution of differences between the two series of observations, and find the error variance of a single observation by dividing this variance by two, but this
14 technique is not generally employed. The correlation between two series of observations as an indication of test reliability is influenced by certain psychological factors, which tend in some degree to invalidate its use as a parameter purely descriptive of test efficiency. The nature and extent of these psychological considerations will be discussed shortly.

Three methods of estimating the reliability of tests are in general use;

(1) Repetition of the same test.

(2) Application of parallel forms of the test.

(3) Split-half method.

A fourth method of estimating the reliability of tests from answer pattern data exists. This method, which has recently been derived, will be considered in detail elsewhere.

Each of the three general methods of estimating the reliability of tests is characterised by certain disadvantages,
15 psychological in type. If the same test is repeated after a short time interval many of the persons tested will recall on the second application of the test, some of their previous responses, and as a consequence their scores will be increased.

If this increase in score is uncorrelated, with ability, the reliability coefficient will be uninfluenced. Since , however, there is some reason to believe that bright persons tend to increase their score more on the second application of the test than dull persons, the reliability coefficient will be spuriously increased. If a sufficiently lengthy time interval is permitted to elapse between the successive
16 applications the influence of memory and practice on the reliability coefficient will be partly eliminated. If, however, the function tested exhibits a certain variability with time, the reliability coefficient cannot be regarded as a parameter purely descriptive of the efficiency of the test, but must be regarded as partly descriptive of the reliability of the abilities tested. The repetition method is not in general use in estimating the reliability of group tests. Reliability coefficients for individual intelligence tests and performance tests are frequently determined by this method.

The estimation of reliability coefficients by the
7 administration of two parallel forms is applicable when two forms of a test exist which may be regarded as exhibiting a high degree of equivalence. When the two forms are not equivalent the correlation coefficient will be reduced by the presence of specific factors, and cannot be regarded as a reliability coefficient. A tetrad criterion can readily be devised to determine whether the two forms may be regarded

as parallel.

Many of the disadvantages that apply to the estimation of the reliability of tests by the administration of the same form apply also to the method of estimating reliability by the administration of equivalent forms. Practice may
18 spuriously increase the reliability coefficient between equivalent forms when the time interval between the two testings is short. When a lengthy time interval is permitted to elapse the reliability coefficient becomes an index not only of the accuracy with which the test measures the function which it presumes to measure, but also of the constancy of that function.

Reliability coefficients are also frequently estimated by dividing a test into two halves, which are assumed equivalent, usually by summing the scores of the persons tested on the odd and even items, and then on certain
19 assumptions estimating from the correlation between the halves of the test what the correlation would be had each half been twice as long. It is now generally held that the split-half method yields estimates of test reliability that are too high, due to the correlation of errors. This method of estimating test reliability will be considered in detail later, and the concept of error correlation qualified.

Much of the confusion that exists among the literature on test reliability arises from failure to observe the distinction between the reliability of tests and the

reliability of persons. The adoption of the concept
20 'reliability of persons' indicates that we are of the opinion that mental abilities are not entirely constant, but are characterised by a quotidian variability. The existence of a quotidian variability of ability, indicated by common sense, has been definitely established. If now a reliability coefficient is estimated by the application of the same or parallel forms of the test on different days it cannot be regarded as a parameter purely descriptive of the accuracy with which the test measures the abilities which it measures, but must be regarded as in part an indication of the constancy of the abilities tested. It is true that for certain purposes
21 we wish to use the reliability coefficient not only as an indication of test efficiency, but also as an indication of the constancy of the abilities tested as well, but under other circumstances we may wish a parameter purely descriptive of the test. Consequently it becomes necessary for us to redefine the term 'reliability of tests.' The term 'reliability of tests' may be defined as the accuracy (not constancy) with which a test measures the abilities which it measures at the time when it measures them. The 'reliability of persons' may be described (not defined) as
22 the accuracy with which a persons ability at any point in time approximates to his 'true ability.'

On the assumption that errors due to the unreliability of tests are uncorrelated with errors due to the unreliability of persons, we may write;

$$\varepsilon^2 = S_t^2 + S_p^2$$

Where ε^2 = total error variance.

S_t^2 = error variance of the test.

S_p^2 = error variance of the persons.

If r_{11} is the correlation between two parallel forms given on the same day, and $r_{11'}$ the correlation between the same two forms given on different days, and on the assumption that the component sources of error that constitute S_t^2 are uncorrelated with each other, and similarly for S_p^2, we may write

$$S_t^2 = 1 - r_{11}$$

$$S_p^2 = r_{11} - r_{11'}$$

Thus, certain conditions being satisfied, we can estimate the
23 error variance of tests, and the error variance of person. 255

2307

BASIC RELIABILITY FORMULAE.

THE SPEARMAN-BROWN FORMULA.

The Spearman-Brown formula is in general use for estimating the reliability of a whole test from a knowledge of the correlation between the test halves, and also for demonstrating the relationship between the length of a test and its reliability. The Spearman-Brown formula is capable of ready proof from the formula for the correlation of sums. We shall firstly consider the case where the test is doubled in length, and secondly the case where the length of the test is increased n times.

The assumption underlying the Spearman-Brown formula for double length is that if the test were given a second time the variance of each test half would be the same, and all the intercorrelations between the four test halves would be the same. On this assumption it only remains to determine the correlation between the sum of two equally intercorrelated variables with the sum of the same two equally intercorrelated variables. A formula for such a correlation may be readily derived from a pooling square in which all the values of r are equal as follows:-

$$\begin{array}{c|cc|cc|}
 & Z_1 & Z_2 & Z_1' & Z_2' \\
\hline
Z_1 & 1 & r & r & r \\
Z_2 & r & 1 & r & r \\
\hline
Z_1 & r & r & 1 & r \\
Z_2 & r & r & r & 1 \\
\hline
\end{array}$$

where Z_1 and Z_2 refer to the odd and even items on the test, and Z_1' and Z_2' to the odd and even items on a hypothetical second application of the test. The correlation is then given by dividing the sum of the elements in the north-east quadrant of the pooling square by the square root of the product of the sum of the elements in the north-west quadrant and the sum of the elements in the south-east quadrant. Writing $r_{(1+2)(1'+2')} = r_{11}$, then

$$r_{11} = \frac{2r}{1 + r}$$

where r_{11} = reliability coefficient.

r = correlation between the odd and even items on a test.

This formula is the Spearman-Brown formula for estimating what the reliability of a test would be if it were doubled in length, and represents a special case of the more general formula for estimating the influence on reliability of lengthening a test n times.

In deriving the general formula it is also necessary to assume that all the n parts of our hypothetical lengthened test are equally intercorrelated. Thus we again write the intercorrelations between the parts of our test in the form of a pooling square.

$$
\begin{array}{c|cccc|cccc|}
 & z_1 & z_2 & \cdots & z_n & z_1' & z_2' & \cdots & z_n' \\
\hline
z_1 & 1 & r & \cdots & r & r & r & \cdots & r \\
z_2 & r & 1 & \cdots & r & r & r & \cdots & r \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
z_n & r & r & \cdots & r & r & r & \cdots & r \\
\hline
z_1' & r & r & \cdots & r & 1 & r & \cdots & r \\
z_2' & r & r & \cdots & r & r & 1 & \cdots & r \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
z_n' & r & r & \cdots & r & r & r & \cdots & 1 \\
\hline
\end{array}
$$

Z_1, Z_2, ... Z_n refer to the n parts of the test, and Z_1', Z_2', ... Z_n' to the n parts of the test on its hypothetical second application.

Writing $r_{(1+2+\cdots n)(1'+2'\cdots n')} = r_{nn}$ we immediately derive

$$r_{nn} = \frac{n\,r}{1+(n-1)r}$$

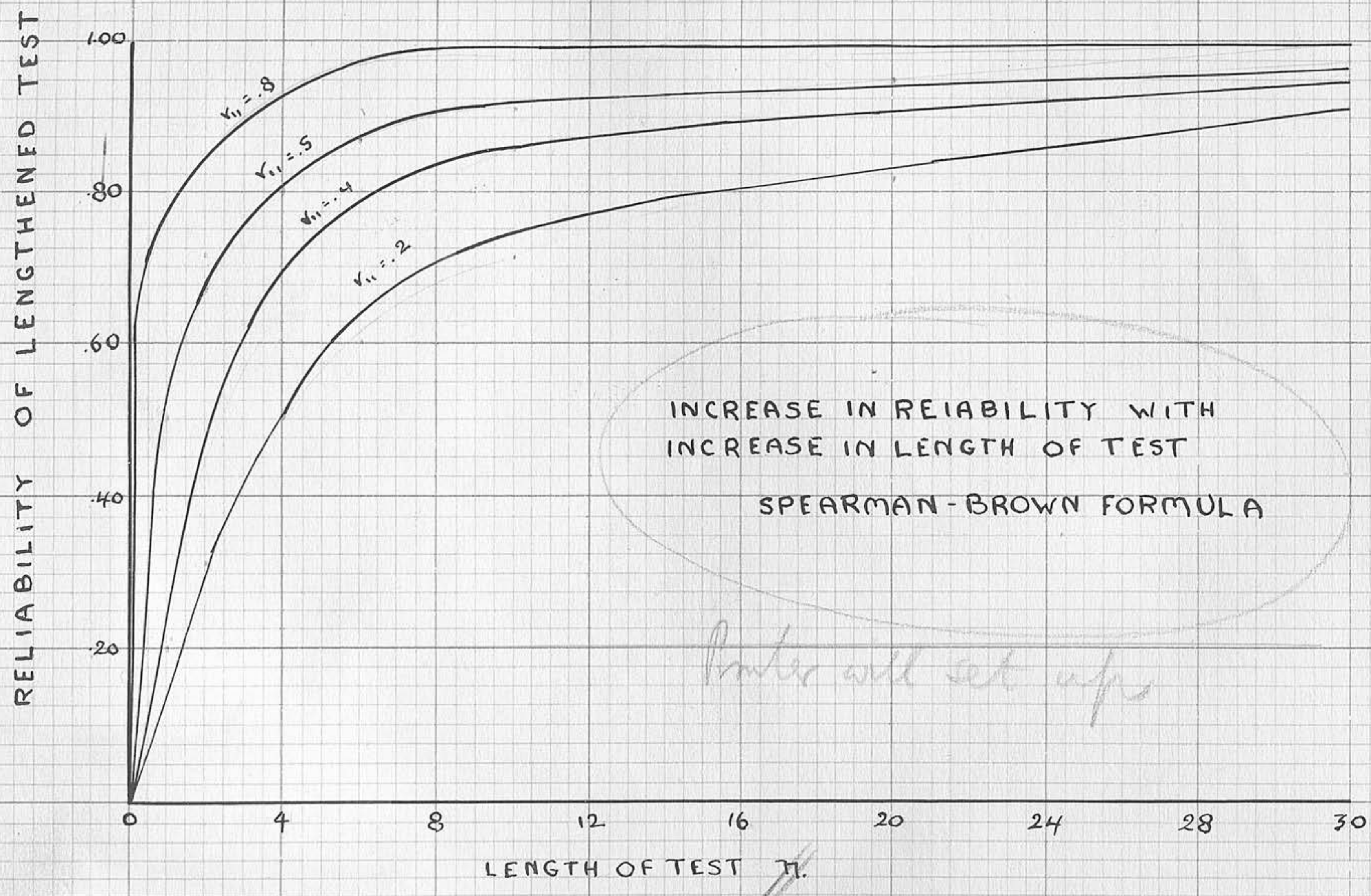
This formula is the usual Spearman-Brown formula for estimating what the reliability of a test would be if it were lengthened n times.

Examination of these formulae for estimating the influence of length of test on reliability indicates that as r $\longrightarrow$ 0 the test must be lengthened many times before a substantial increase in reliability can be attained.

Conversely as $r \longrightarrow 1$ increasing the length of the test results in no great increase in the reliability coefficient.

These observations will be rendered apparent on reference to Figure 1 where reliability is plotted against length of test for different values of r. All the members of this family of curves pass through the origin and become asymptotic as the length of the test is increased towards infinity.

FIG. I.



INCREASE IN REIABILITY WITH
INCREASE IN LENGTH OF TEST
SPEARMAN-BROWN FORMULA
RELIABILITY OF LENGTHENED TEST
1.00
.80
.60
.40
.20
$r_{11} = .8$
$r_{11} = .5$
$r_{11} = .4$
$r_{11} = .2$
0
4
8
12
16
20
24
28
30
LENGTH OF TEST n.
Printer will set up

THE INDEX OF RELIABILITY.

The index of reliability is at times used instead of the coefficient of reliability as a parameter descriptive of test efficiency. The coefficient of reliability on the one hand is the correlation between two series of fallible observations of a series of true values, while the index of reliability is the correlation between a single series of observations and a series of true values. The distinction between these two concepts will be clarified on reference to Figure 2. The test vectors Z_1 and Z_1' in two dimensional space represent two series of fallible observations of a single series of true values, represented by the vector Z_t.

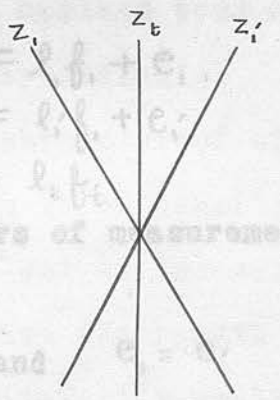


Fig. 2

The cosine of the angle between the two vectors Z_1 and Z_1' is the reliability coefficient. The cosine of the angle between the vector of true values Z_t, and either Z_1 or Z_1' is the index of reliability. The vector Z_t is not in the same two dimensional space as Z_1 and Z_1' but is in a third dimension.

The correlation between a single series of fallible observations and a series of true values may be shown to be equal to the square root of the correlation between two series of fallible observations of the same true values, when the errors of observation are random and equal in variance; that is, the index of reliability is equal to the square root of the reliability coefficient.

The proof is simple in type. If Z_1 and Z_1' represent two fallible series of observations, and Z_t represents the true values, then

$$Z_1 = l_1 f_1 + e_1$$
$$Z_1' = l_1' f_1 + e_1'$$
$$Z_t = l_t f_t$$

But if the errors of measurement are purely random and equal in variance

$$l_1 = l_1' \quad \text{and} \quad e_1 = e_1'$$
$$\text{but} \quad r_{11} = l_1 l_1' = l_1^2$$
$$\text{and} \quad r_{11} = 1 - e_1^2$$

Further, the correlation between Z_1 or Z_1' and Z_t may be written

$$r_{1t} = l_1 l_t$$

but $\quad l_t = 1$

therefore $\quad r_{1t} = l_1 l_t = l_1$

but $\quad r_{11} = l_1^2$

hence $\quad r_{1t} = \sqrt{r_{11}}$

(Formula for the index of Reliability)

It is apparent that no matter what other variable the series of observations z_1 were correlated with the factor loadings of that variable common to Z_1 could never be unity. Consequently the index of reliability of a test represents the maximum correlation that a test is capable of yielding with any other test or battery of tests in the whole universe of tests. The reliability index represents the correlation between a test which is an imperfect instrument of measurement, and another test measuring the same abilities which is perfectly reliable.

A less algebraic proof of the index of reliability can be attained which is of considerable interest. As we increase the length of a test we increase its reliability, so that if we were to increase the length of a test an infinite number of times, its reliability would become unity; that is, the test would be a perfect measure of the abilities which it measured, and each of the test vectors Z_1 and Z_1' would lie directly along the vector Z_t. The problem then becomes one of determining the correlation between a single fallible test, and the same test lengthened an infinite number of times.

Let Z_1 be a test and Z_1', Z_1'', ...Z_∞ an infinite number of parallel forms. Let the intercorrelations be written in the form of a pooling square, as follows;

$$
\begin{array}{c|c|ccccc|}
 & Z_1 & Z_1' & Z_1'' & \cdot & \cdot & Z_{1\infty} \\
\hline
Z_1 & 1 & r & r & \cdot & \cdot & r_\infty \\
\hline
Z_1' & r & 1 & r & \cdot & \cdot & r \\
Z_1'' & r & r & 1 & \cdot & \cdot & r \\
\cdot & & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & & \cdot & \cdot & \cdot & \cdot & \cdot \\
Z_{1\infty} & r_\infty & r & r & \cdot & \cdot & 1 \\
\hline
\end{array}
$$

The average value of the elements in the north-east quadrant, when the test is lengthened an infinite number of times is of course r_{11}. It is also apparent that as $n \longrightarrow \infty$ the average value in the south-east quadrant approximates to r_{11}. We may, therefore, write the correlation between a test, and an infinite number of parallel forms of the test in the form

$$r_{1t} = \frac{r_{11}}{\sqrt{r_{11}}} = \sqrt{r_{11}}$$

(Formula for index of reliability)

THE CORRECTION FOR ATTENUATION.

The general effect of random errors of observation is to reduce correlation; that is, the presence of random errors tends to attenuate the correlation between observed values away from the correlation between the true values of the quantities observed. The greater the magnitude of the errors of observation the greater the attenuation effect. As the length of a test is increased an infinite number of times $r_{11} \longrightarrow 1$; that is when $n = \infty$ the test becomes a true measure. The problem, therefore, of determining the correlation between two series of true values involves the determination of what the correlation between two tests would be had each test been lengthened an infinite number of times.

Let us assume that Z_1 and Z_2 are two tests lengthened an infinite number of times, and that all the intercorrelations are written in the form of a pooling square as follows;

$$
\begin{array}{c|cccc|cccc|}
 & Z_1 & Z_1' & \cdots & Z_{1\infty} & Z_2 & Z_2' & \cdots & Z_{2\infty} \\
\hline
Z_1 & 1 & r_{11} & \cdots & r_{11} & r_{12} & r_{12} & \cdots & r_{12} \\
Z_1' & r_{11} & 1 & \cdots & r_{11} & r_{12} & r_{12} & \cdots & r_{12} \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
Z_{1\infty} & r_{11} & r_{11} & \cdots & 1 & r_{12} & r_{12} & \cdots & r_{12} \\
\hline
Z_2 & r_{12} & r_{12} & \cdots & r_{12} & 1 & r_{22} & \cdots & r_{22} \\
Z_2' & r_{12} & r_{12} & \cdots & r_{12} & r_{22} & 1 & \cdots & r_{22} \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
Z_{2\infty} & r_{12} & r_{12} & \cdots & r_{12} & r_{22} & r_{22} & \cdots & 1 \\
\hline
\end{array}
$$

The average value in the north-east quadrant of the pooling square is equal to r_{12}. It is furthermore apparent that as $n \rightarrow \infty$ the average value in the north-west quadrant approximates to r_{11}, so that when $n = \infty$, the average value of the elements in that quadrant is r_{11}. Similarly when $n = \infty$ the average value of the elements in the south-east quadrant is r_{22}. We may, therefore, write

$$r_{12\infty\infty} = \frac{r_{12}}{\sqrt{r_{11}\, r_{22}}}$$

(Formula for correcting a correlation coefficient for attenuation)

Another proof of the formula for correcting a correlation coefficient for attenuation, more algebraic in type, exists, which exhibits some interesting properties.

Let z_1 and z_2 be two tests expressed in terms of r linearly independent common factors, such that

$$Z_1 = l_1 x + l_1' y + \quad \cdots \quad b_1 s_1 + e_1$$
$$Z_2 = l_2 x + l_2' y + \quad \cdots \quad b_2 s_2 + e_2$$

Then

$$r_{12} = \sum_{r=1}^{r} (l_i l_j)$$

If z_1 and z_2 were perfect measures $e_1 = 0$, $e_2 = 0$

Hence

$$l_{1\infty} = \frac{l_1}{\sqrt{1 - e_1^2}} \quad ; \; l_{2\infty} = \frac{l_2}{\sqrt{1 - e_2^2}} \quad ; \text{ etc.}$$

(3) where and are values of and uninfluenced by random

errors of measurement.

Therefore

$$r_{12\infty\infty} = \frac{\sum_{r=1}^{r}(l_i l_j)}{\sqrt{(1-e_1^2)(1-e_2^2)}} = \frac{r_{12}}{\sqrt{r_{11}\, r_{22}}}$$

(Formula for correcting a correlation coefficient for attenuation)

The correction for attenuation is used to determine the degree of intrinsic relationship between two variables; that is, to determine a correlation coefficient that is not a function of the errors of measurement involved.

Investigators have on occasion found that correlation coefficients corrected for attenuation exceeded unity, and on these grounds the formula has at times suffered condemnation. Spearman has shown that a sampling error of a coefficient corrected for attenuation is considerably greater than the attenuated coefficient, and that we should expect under certain circumstances coefficients to exceed unity within the limits of their sampling error. Corrected coefficients greater than unity may at times be obtained when the reliability coefficients and correlation coefficients used in the attenuation formula have not been sonsistently determined. Thus, certain sources of error may be exerting an influence on the coefficients in the denominator of the attenuation formula, which sources of error are not influencing the coefficients in the numerator, and vice versa. Under such circumstances we should expect to obtain over:estimates and underestimates, respectively, of the true relationship between the correlated variables. Such inconsistencies have been adequately treated by Thouless. (Robert H. Thouless, The effect of errors of measurement on correlation coefficients, B.J.P. XX1X, 1938.)

When the corrected coefficient determined by the use of consistent correlations is in the neighbourhood of unity, we may state taht the departure of the obtained coefficient from unity is due to the presence of random errors, and not specific factors. Spearman has demonstrated that when the tetrad criterion holds for coefficients uncorrected for attenuation it will also hold for corrected coefficients. By generalizing this theorem we may state that the rank of any correlation matrix remains unchanged when its elements are corrected for attenuation. In order to transform the factor loadings obtained from uncorrected coefficients into the loadings that would have obtained from corrected coefficients, we merely pre-multiply the factorial matrix by a diagonal matrix with elements $\frac{1}{\sqrt{r_{ii}}}$, where r_{ii} is the reliability coefficient of test i. This amounts to dividing the factor loadings of each test by the square root of the reliability coefficient of that test. This technique indicates whether specific factors are real specifics or purely error variance.

THE STANDARD ERROR OF A TEST SCORE.

The error variance of a test score is the variance of the difference between an infinite number of observations of that score and the mean of the observations. On the assumption that persons and trials are uncorrelated we may use the variance of the difference between a series of observed scores, and the series of corresponding true scores as an estimate of the error variance. Now, as discussed previously, if we make two series of observations the variance of the difference between these two series is made up of two components, the variance of the difference between one series of observations and the true scores, and the variance of the difference between the other series of observations and the true scores. Hence on the assumption that each series of observations is equally fallible, we may write

$$\sigma^2_{(1-1')} = 2\mathcal{E}^2$$

where $\sigma^2_{(1-1')}$ = the variance of the difference between two series of observations.

$\mathcal{E}^2$ = the variance of the difference between one series of observations and the true values.

but $\sigma^2_{(1-1')} = \sigma_1^2 + \sigma_{1'}^2 - 2r_{11}\sigma_1\sigma_{1'}$

since

$$\sigma_1^2 = \sigma_{1'}^2$$

therefore $\sigma^2_{(1-1')} = 2\sigma^2(1 - r_{11})$

$$\varepsilon^2 = \sigma^2(1 - r_{11})$$

(Formula for the error variance of a test score)

and $$\varepsilon = \sigma\sqrt{1 - r_{11}}$$

(Formula for the standard error of a test score)

If the two series of observations are reduced to standard measure $\sigma = 1$. Therefore the standard error of a standard score is given by

$$\varepsilon = \sqrt{1 - r_{11}}$$

(Formula for the standard error of a standard score)

If the errors of measurement are purely random the error variance of a test score should be uninfluenced by the degree of selection of the sample. This observation is capable of simple demonstration on reference to the Otis-Kelly formula for correcting a reliability coefficient for selection. This formula is given by

$$\frac{\sigma_1^2}{\Sigma_1^2} = \frac{1 - R_{11}}{1 - r_{11}}$$

where σ_1^2, Σ_1^2 and r_{11}, R_{11} represent the variance and reliability coefficient obtained from the sample and the population respectively. If ε^2 represents the error variance of a test score estimated from the sample, and

E^2 the error variance estimated from the population

$$\varepsilon^2 = \sigma_1^2(1 - r_{11})$$
$$E^2 = \Sigma_1^2(1 - R_{11})$$

but $$\sigma_1^2(1 - r_{11}) = \Sigma_1^2(1 - R_{11})$$

therefore $$\varepsilon^2 = E^2$$

Since the error variance of a test score is independent of the degree of selection of the group it furnishes under certain circumstances a more useful index of test efficiency than the reliability coefficient. It is of particular value in comparing the results of different investigators who have employed samples of different degrees of selection.

The standard error of a test score, and indeed, the standard errors of all types of parameters, is frequently interpreted as implying that the probability is 68/100 that the true value lies within the range defined by once the standard error on either side of the observed score, or 95/100 that the true value lies within the range defined by twice the standard error. This method of interpretation is not quite correct. A given observation x may take any value between $\pm 2\sigma$ of a distribution centred on a hypothetical true value x_∞, where twice the standard error is taken as the criterion of acceptability. The implication is that with any given observation x we may state with reasonable certainty that the true value lies within $x \pm 2\sigma$. If, however, I were to make a large number of observations x, $x_1, x_2, x_3, \ldots\ldots\ldots\ldots x_n$, it does not follow that 95 out of 100 of such observations lies within the limits $x \pm 2\sigma$. Indeed if the given observation x were at the extreme right of the $\pm 2\sigma$ range sampling distribution centered on the mean of a large number of observations the probability is only 50/100 that any other single observation will lie within the limits $x \pm 2\sigma$. This type of problem involves the distinction between inverse and fiducial probability.

RELATIONSHIP BETWEEN STANDARD ERROR AND LENGTH OF TEST.

As we increase the length of a test to increase its reliability we also increase the variance of raw scores of the test. The variance of raw scores on the lengthened test is readily derived from the appropriate pooling square, and is given by the formula

$$\sigma_{sn}^2 = n\sigma^2\left[1 + (n-1)r\right]$$

where σ_{sn}^2 = the variance of raw or deviation scores on a test lengthened n times.

σ^2 = the variance of scores on a test of unit length.

r = the reliability coefficient of a test of unit length.

n = number of times the test is lengthened.

But the reliability of a test lengthened n times as given by the Spearman-Brown formula is

$$r_{nn} = \frac{nr}{\left[1 + (n-1)r\right]}$$

Combining these two equations we may write

$$\frac{\sigma_{sn}^2}{\sigma^2} = \frac{n^2 r}{r_{nn}}$$

This equation shows the relationship between the variance of a test lengthened n times and a test of unit length in terms of the reliability of a test lengthened n times and the reliability of a test of unit length.

It now remains to derive the relationship between the error variance of a test score on a test of unit length, and the error variance of the same test lengthened n times. If $\mathcal{E}^2$ and $\mathcal{E}_n^2$ are the error variances of a test of unit length, and the same test lengthened n times, then

$$\mathcal{E}^2 = \sigma^2(1-r)$$

$$\mathcal{E}_n^2 = \sigma_{sn}^2(1-r_{nn})$$

Hence

$$\frac{\mathcal{E}^2}{\mathcal{E}_n^2} = \frac{\sigma^2(1-r)}{\sigma_{sn}^2(1-r_{nn})}$$

but

$$\frac{\sigma_{sn}^2}{\sigma^2} = \frac{n^2 r}{r_{nn}}$$

therefore

$$\frac{\mathcal{E}^2}{\mathcal{E}_n^2} = \frac{r_{nn}(1-r)}{n^2 r(1-r_{nn})}$$

Substituting the Spearman-Brown formula for r_{nn} in this equation we find that

$$\mathcal{E}_n^2 = n\mathcal{E}^2$$

Thus we may say that the error variance of a test score on a test lengthened n times is equal to n times the error variance of a test of unit length.

THE STANDARD ERROR OF THE DIFFERENCE BETWEEN TWO TEST SCORES.

The error variance of the difference between the test scores of two persons is the variance of the difference between the scores obtained by the two persons on an infinite number of trials. If the trials are uncorrelated we may write

$$\mathcal{E}^2_{(1-1')} = \mathcal{E}^2_1 + \mathcal{E}^2_{1'}$$

If we are testing the significance of the difference between the scores obtained by two persons on the same test, then

$$\mathcal{E}^2_{(1-1')} = 2\mathcal{E}^2_1 = 2\sigma^2(1-r_{11})$$

If we adopt the 95 per cent probability sampling distribution as the criterion of acceptability, we may state that the difference between the scores of two persons on the same test must be 2.828 times the standard error of a single score, before the abilities of the two persons tested may be regarded as differing significantly. This indicates that mental tests must yield very high reliability coefficients before they may be regarded as discriminating with much accuracy between the persons tested.

If now we wish to determine the significance of the difference between scores of the same person, or different persons, on two different tests, on the assumption that the correlation between trials is zero, we may write

$$\mathcal{E}^2_{(1-2)} = \mathcal{E}^2_1 + \mathcal{E}^2_2$$

where $\mathcal{E}^2_{(1-2)}$ = the error variance of the difference between a score on z_1 and a score on z_2.

$\mathcal{E}^2_1$ = the error variance of z_1.

$\mathcal{E}^2_2$ = the error variance of z_2.

Hence $$\mathcal{E}^2_{(1-2)} = \sigma_1^2 + \sigma_2^2 - r_{11}\sigma_1^2 - r_{22}\sigma_2^2$$

The above relationship may be adapted to standard measure. The standard error of the difference between the standard scores of two persons on the same test is given by

$$\mathcal{E}_{(1-1')} = \sqrt{2 - 2r_{11}}$$

while the standard error of the difference between the standard scores of the same or different persons on different tests is given by

$$\mathcal{E}_{(1-2)} = \sqrt{2 - r_{11} - r_{22}}$$

THE TRUE VARIANCE OF A TEST.

Errors of measurement tend to increase the variance of obtained scores. On the assumption that such errors are purely random, by the additive nature of the variances of uncorrelated variables we may write

$$\sigma^2 = \sigma^2_\infty + \mathcal{E}^2$$

where σ^2 = obtained variance.

σ^2_∞ = true variance (variance uninfluenced by random errors)

$\mathcal{E}^2$ = error variance

but $\varepsilon^2 = \sigma^2(1 - r_{11})$

hence $\sigma_\infty^2 = \sigma^2 r_{11}$

INCREASED VALIDITY WITH INCREASED RELIABILITY.

Since random errors of measurement tend to attenuate the correlation of a test with a criterion the validity of a test may be increased by increasing its reliability. A formula is readily derived showing the influence on the correlation of a test with a criterion of lengthening the test any number of times. If r_{01} is the correlation of a test with a criterion, and r_{11} the reliability of the test, we may write the intercorrelations between the criterion and n tests of unit length in the form of a pooling square.

$$
\begin{array}{c|ccccc}
z_0 & z_1 & z_2 & \cdots & & z_n \\
1 & r_{01} & r_{01} & \cdots & \cdots & r_{01} \\
\hline
r_{01} & 1 & r_{11} & \cdots & \cdots & r_{11} \\
r_{01} & r_{11} & 1 & \cdots & \cdots & r_{11} \\
\vdots & \vdots & & \ddots & & \vdots \\
r_{01} & r_{11} & r_{11} & \cdots & \cdots & 1
\end{array}
$$

From this square we immediately derive the formula

$$r_{01(n)} = \frac{n\, r_{01}}{\sqrt{n + n(n-1)\, r_{11}}}$$

where $r_{01(n)}$ is the correlation with the criterion of the test lengthened n times.

By writing this equation explicitly for n we may estimate the number of times that a test must be lengthened in order to attain a specified validity, when the specified validity lies between r_{01} and $r_{01\infty\infty}$.

$$n = \frac{1 - r_{11}}{\frac{r_{01}^2}{r_{01(n)}^2} - r_{11}} \rightarrow 0 \quad n \rightarrow \infty$$

We may on occasion wish to estimate the maximum possible correlation between a test and a criterion; that is the correlation that would have obtained had the test been perfectly reliable, or had the test been lengthened an infinite number of times. Examination of the pooling square given above will show that as $n \longrightarrow \infty$ the average value of the elements in the south-east block approximates to ~~r_{01}~~ r_{11},

3 hence

$$r_{01.\infty} = \frac{r_{01}}{\sqrt{r_{11}}}$$

This formula yields the correlation between criterion and true scores. If, however, the criterion is itself not a perfectly reliable measure, and if its reliability coefficient is known, we may estimate the correlation between the true criterion scores and true test scores by the usual attenuation formula.

306

THE ESTIMATION OF RELIABILITY FROM ANSWER PATTERN DATA.

The interpretation of a test not as a unit in itself, but as a large composite battery of small item tests, each having its own variance and intercorrelations with all the other items on the test, and contributing by virtue of its variance and correlation with other items to the action of the test as a whole, not only indicates certain concepts which are fundamental in the theory of reliability, but also suggests new methods for the estimation of reliability from the usual parameters computed for the selection of test items from answer pattern data.

The correlation of a test z_1 of n elements with another test z_2 of n' elements may be interpreted as the correlation of the sum of the n elements of z_1 with the n' elements of z_2. Thus the correlation r_{12} is a simplification of the complex interaction of all the n elements of z_1 with each other, the variance of z_1, the interaction of all the n' elements of z_2 with each other, the variance of z_2, and the interaction of all the n elements of z_1 with the n' elements of z_2, the covariance. The correlation between any two tests may, therefore, be described as a simplification of a complexity of interactions between test elements.

In terms of the above theory the correlation between the tests z_1 and z_2 may be written from formulae:-

$$r_{12} = \frac{\sum_{i=1}^{n}\sum_{j'=1}^{n'} r_{ij'}\sigma_i\sigma_{j'}}{\sqrt{\left[\sum_{i=1}^{n}\sum_{j=1}^{(n-1)} r_{ij}\sigma_i\sigma_j + \sum_{i=1}^{n}\sigma_i^2\right]\left[\sum_{i'=1}^{n'}\sum_{j'=1}^{(n'-1)} r_{i'j'}\sigma_{i'}\sigma_{j'} + \sum_{i'=1}^{n'}\sigma_{i'}^2\right]}} \qquad (1)$$

where σ_i^2 = the variance of item one on the test z_1 of n elements.

$\sigma_{i'}^2$ = the variance of item i' on the test z_2 of n' elements.

r_{ij} = the correlation between the items i and j on z_1.

$r_{i'j'}$ = the correlation between the items i' and j' on z_2.

$r_{ij'}$ = the correlation between the item i on z_1 and the item j' on z_2.

The term in the numerator of equation (1) is equal to $r_{12}\sigma_1\sigma_2$ while the terms in the denominator are respectively σ_1^2 and σ_2^2. Equation (1) indicates that the correlation between two tests is a complex function of the item variances and inter-item covariances.

Let us now consider the case of a test of n elements given twice to the same sample of persons. From answer patterns constructed for each application of the test it is possible with great arithmetical labour to calculate the variance of each item on each application of the test, the reliability of each item, and all the $4n^2 - 3n$ other inter-item correlations of the 2n test elements. From these values by formulae for the correlation of sums, the correlation between the scores of the persons tested on each application of the test could be found. A correlation coefficient thus calculated should agree exactly with the coefficient obtained by correlating raw scores, when the item variances are estimated by the formula $p_i q_i$, and the inter-item correlations by the formula:-

$$r_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i q_i \, p_j q_j}} \qquad (2)$$

where p_{ij} = the proportion of persons passing both item i and j.

p_i = the proportion of persons passing item i.

p_j = the proportion of persons passing item j.

q_i = the proportion of persons failing item i.

q_j = the proportion of persons failing item j.

Although the process of estimating reliability outlined above does not lend itself to ordinary computational purposes the general theory of this process suggests methods whereby reliability coefficients may be estimated from certain parameters commonly computed for purposes of item selection. These methods have been devised by G.F.Kuder and M.W.Richardson, (Psychometrika vol.2, no.3, Sept. 1937 p 151-160) and are considered in detail below. The formulae given here are substantially similar to those given by Kudor and Richardson, although the methods of derivation differ slightly.

The intercorrelations between all the n items on a test, and the n items on a hypothetical equivalent form of the test, may be written in the form of a pooling square as follows:-

$$
\begin{array}{c|ccccc|ccccc|}
 & \sigma_1 & \sigma_2 & \cdots & & \sigma_n & \sigma_1 & \sigma_2 & \cdots & & \sigma_n \\
\hline
\sigma_1 & 1 & r_{12} & \cdots & & r_{1n} & r_{11} & r_{12} & \cdots & & r_{1n} \\
\sigma_2 & r_{12} & 1 & \cdots & & r_{2n} & r_{12} & r_{22} & \cdots & & r_{2n} \\
\cdot & \cdot & \cdot & \cdots & & \cdot & \cdot & \cdot & \cdots & & \cdot \\
\cdot & \cdot & \cdot & \cdots & & \cdot & \cdot & \cdot & \cdots & & \cdot \\
\cdot & \cdot & \cdot & \cdots & & r_{n(n-1)} & \cdot & \cdot & \cdots & & r_{n(n-1)} \\
\sigma_n & r_{1n} & r_{2n} & \cdot & r_{n(n-1)} & 1 & r_{1n} & r_{2n} & \cdot & r_{n(n-1)} & r_{nn} \\
\hline
\sigma_1 & r_{11} & r_{12} & \cdots & & r_{1n} & 1 & r_{12} & \cdots & & r_{1n} \\
\sigma_2 & r_{12} & r_{22} & \cdots & & r_{2n} & r_{12} & 1 & \cdots & & r_{2n} \\
\cdot & \cdot & \cdot & \cdots & & \cdot & \cdot & \cdot & \cdots & & \cdot \\
\cdot & \cdot & \cdot & \cdots & & \cdot & \cdot & \cdot & \cdots & & \cdot \\
\cdot & \cdot & \cdot & \cdots & & r_{n(n-1)} & \cdot & \cdot & \cdots & & r_{n(n-1)} \\
\sigma_n & r_{1n} & r_{2n} & & r_{n(n-1)} & r_{nn} & r_{1n} & r_{2n} & & r_{n(n-1)} & 1 \\
\hline
\end{array}
$$

The sum of the weighted elements in the north-east quadrant divided by the square root of the product of the sum of the weighted elements in the north-west quadrant and the sum of the weighted elements in the south-east quadrant is the correlation between the two forms of the test. Since the two forms of the test are assumed parallel then the wieghted elements in the north-west quadrant may be regarded as the same as the weighted elements in the south-east quadrant. Also the the weighted elements in the north-east and south-west quadrants may be regarded as the same as the elements in the other two quadrants, with the exception of the elements down the diagonals. It is known that the sum of the weighted elements in the north-west quadrant is equal to the variance of the test. The correlation between the test and its hypothetical parallel form may then be written as follows:-

$$r_{tt} = \frac{\sigma_t^2 - \sum_{i=1}^{n} \sigma_i^2 + \sum_{i=1}^{n} r_{ii}\sigma_i^2}{\sigma_t^2} \qquad (3)$$

where r_{tt} = the reliability coefficient of the whole test.

σ_t^2 = the variance of the test.

σ_i^2 = the variance of the item i.

r_{ii} = the reliability coefficient of the item i.

All the terms in equation (3) may be determined from a single application of the test except the item reliabilities r_{ii}, which cannot be known without giving the test a second time to the same sample of persons. Since, however, the term $\sum_{i=1}^{n} r_{ii}\sigma_i^2$ is small in comparison with the term $\sum_{i=1}^{n} \sigma_i^2$ small discrepancies in reasonably guessed values of r_{ii} will have no great influence on the value of r_{tt}. With Moray House Tests the mean value of the item reliability, $\overline{r_{ii}}$, is about .40 or .50.

By making certain assumptions a number of other formulae better adapted to calculation may be derived. If we are willing to assume that the average inter-item covariance, $\overline{r_{ij}\sigma_i\sigma_j}$, is equal to the average value of the product of the item reliability and the item variance, $\overline{r_{ii}\sigma_i^2}$, formula (3) may be written in the form

$$r_{tt} = \frac{n^2\,\overline{r_{ij}\sigma_i\sigma_j}}{\sigma_t^2} \tag{4}$$

where $\overline{r_{ij}\sigma_i\sigma_j}$ = the average inter-item covariance.

But

$$\sigma_t^2 = \sum_{\substack{i=1\\ i\neq j}}^{n}\sum_{j=1}^{n} r_{ij}\sigma_i\sigma_j + \sum_{i=1}^{n}\sigma_i^2$$

$$= n(n-1)\,\overline{r_{ij}\sigma_i\sigma_j} + \sum_{i=1}^{n}\sigma_i^2 \tag{5}$$

Therefore

$$\overline{r_{ij}\sigma_i\sigma_j} = \frac{\sigma_t^2 - \sum_{i=1}^{n}\sigma_i^2}{n(n-1)} \tag{6}$$

Formula (6) is not an approximation but an exact measure of the average inter-item covariance, and is in itself an illuminating index of test efficiency. The greater the average inter-item covariance the greater the variance of raw scores. Furthermore the tendency exists for the reliability of a test to increase as some direct function or other of the sum of the inter-item covariances. The quantity $\overline{r_{ij}\sigma_i\sigma_j}$ varies from 0.0 to .25. For Moray House Tests $\overline{r_{ij}\sigma_i\sigma_j}$ has a value of about .04.

Substituting equation (6) in equation (4) we have

$$r_{tt} = \frac{n}{n-1} \quad \frac{\sigma_t^2 - \sum_{i=1}^{n} \sigma_i^2}{\sigma_t^2} \qquad (7)$$

This formula is similar to Kuder and Richardson's formula (20), although their process of derivation is much more elaborate than the simple derivation given here. Furthermore, the derivation given by these authors requires three very broad assumptions, (1) that the matrix of inter-item correlations has a rank of one, (2) that all the intercorrelations are equal, (3) that the item variances are equal. If the validity of this formula were dependent on the accuracy with which a test approximated to these three conditions its value as a measure of reliability would be seriously impaired, since few tests approximate either to unit rank or equality of either inter-item correlation or item variance. As we have attempted to show, the valid use of this formula for the estimation of test reliability need not necessarily depend on any of the assumptions made by Kuder and Richardson, but rather upon the more conservative assumption that the average inter-item covariance, $\overline{r_{ij}\sigma_i\sigma_j}$, is equal to $\overline{r_{ij}\sigma_i^2}$. Although $\overline{r_{ij}\sigma_i\sigma_j}$ may in actual practice be a discrepant estimate of $\overline{r_{ii}\sigma_i^2}$, the order of discrepancy that is likely to arise will have no great influence on the estimated reliability coefficient.

Certain suggestions may be made here to facilitate the computation of the term $\sum_{i=1}^{n}\sigma_i^2$ in formula (7). $\sum_{i=1}^{n}\sigma_i^2$ may be calculated directly by finding values of p_iq_i and summing over n items. If, however, a calculating machine is available capable of multiplying and adding in a single operation, since $\sum_{i=1}^{n}\sigma_i^2 = \sum_{i=1}^{n}p_iq_i = \sum_{i=1}^{n}p_i - \sum_{i=1}^{n}p_i^2$ the shortest method is to sum values of p and subtract from this sum the sum of the squared values of p.

An interesting variation of equation (7) is obtained if we assume that all the items in the test have equal values of p_i. When $p_i = p_j$ the quantity $\overline{p_iq_i} = \overline{p_i}\,\overline{q_i}$, that is the average variance is equal to the product of the average of p_i and the average of q_i. On this assumption formula (7) may be written in the form

$$r_{tt} = \frac{n}{n-1} \cdot \frac{\sigma_t^2 - npq}{\sigma_t^2}$$

but
$$p = \frac{\sum_{i=1}^{n} X_t}{nN} = \frac{M_t}{n} \qquad (8)$$

(9)

where N = number of persons.

n = number of items.

$\sum_{i=1}^{n} X_t$ = the sum of the scores of N persons.

M_t = the mean score of all the persons on the test.

therefore
$$r_{tt} = \frac{n}{n-1} \cdot \frac{n\sigma_t^2 - M_t(n - M_t)}{n\sigma_t^2} \qquad (10)$$

When $\overline{p_i q_i} = \bar{p}_i \bar{q}_i$ formula (10) will yield an underestimate of the reliability coefficient.

In order to test the comparative merits of some of the formulae given above, reliability coefficients were calculated for a number of Moray House Tests by formulae (3), (7), and (10).

The tests used were M.H.T. 23, 26, 27, and 30, M.H.A. 11, and M.H.E. 12. Reliability coefficients were calculated for M.H.A.11 for parts 1 and parts 2, separately and combined. In estimating reliability coefficients by formula (3) guessed values of $\overline{r_{ii}}$ were used. These guessed values were .20, .30, .40 and .50. The reliability coefficients estimated by these three formulae are given in Table 1. The boosted split-half reliabilities of M.H.T.23 and 26 are also given. Table 2 shows the standard deviation of the raw scores in each test, the mean of raw scores, the number of items on each test, and the number of cases upon which each coefficient is based.

Examination of Table indicates the following:- (1) Formula (7) yields values of the reliability coefficient slightly smaller than the boosted split-half reliabilities. This may possibly be attributed to the fact that $\overline{r_{ij}\sigma_i\sigma_j}$ is an underestimate of $\overline{r_{ii}\sigma_i^2}$. The boosted split-half reliability cannot, however, be regarded as a criterion. The actual process of selecting the odd and even items will tend with certain types of tests to make the scores on the odd items more nearly similar to the scores on the even items than is compatible with a valid estimate of test reliability. (2) Formula (10) yields estimates of the reliability coefficient that are too small. This is directly due to the fact that with Moray House Tests

$\overline{p_i q_i} \neq \bar{p}_i \bar{q}_i$. This tends to reduce the estimate of test reliability as given by formula (10). (3) Formula (4) gives estimates for various values of r_{tt} differing at most by .03. An estimate of $\overline{r_{ii}}$ equal to .40 or .50 will give values of reliability coefficients in close correspondence to the coefficients that would have obtained by the split-half method. If a value of $\overline{r_{ii}} = .20$ is used formula (4) will yield values in close correspondence with those obtained by formula (7). (4) Reliability coefficients estimated by any one method are consistent with each other and directly comparable. That is, the largest coefficient calculated by formula (7) is also the largest coefficient calculated by formulae (3) and (10). In the examples given in Table I there is one exception to this which can readily be explained. We can conclude, therefore, that all these methods are useful for comparing the relative reliabilities of different tests.

174

Table 1.

Test	r_{11}=.2	Formula (3) r_{11}=.3	r_{11}=.4	r_{11}=.5	Formula (7)	Formula (10)	Split-half Reliability
M.H.T.23	.9614	.9662	.9710	.9759	.9613	.9427	.9775
" 26	.9637	.9683	.9728	.9773	.9643	.9476	.9721
" 27	.9585	.9637	.9689	.9741	.9577	.9316	--
" 30	.9668	.9709	.9751	.9792	.9682	.9615	--
M.H.A.11							
Part 1	.9272	.9363	.9454	.9545	.9312	.9273	--
Part 2	.9538	.9596	.9654	.9711	.9582	.9334	--
(1+2)	.9688	.9727	.9766	.9805	.9705	.9593	--
M.H.E.12	.9649	.9693	.9737	.9781	.9642	.9511	--

Table 2.

Test	S.D.	Mean	n	N
M.H.T. 23	19.36	48.93	100	171
M.H.T. 26	20.07	47.53	100	162
M.H.T. 27	19.12	49.15	100	221
M.H.T. 30	22.25	39.00	100	271
M.H.A. 11				
Part 1	10.38	24.43	42	222
Part 2	13.12	22.77	60	222
(1+2)	22.50	47.27	102	222
M.H.E. 12	23.34	39.95	120	200

REFERENCES

(1) Kuder, G.F., and Richardson, M.W., (1937), "The Theory and Estimation of Test Reliability", Psychometrika, vol. 2, pp. 151-160.

(2) Kuder, G.F., and Richardson, M.W., (1939), "The Calculation of Test Reliability based on the Method of Rational Equivalence", J. Educ. Psychol. xxx, 681-687.

FACTORS INFLUENCING RELIABILITY

FACTORS INFLUENCING RELIABILITY.

The present discussion is concerned with an examination of the factors influencing reliability coefficients. As specified previously much of the prevailing confusion that characterises the reliability concept is clarified by arbitrarily distinguishing between the 'reliability of tests' and the 'reliability of persons'. The validity of the concept 'reliability of persons' depends on the existence of a quotidian variability of mental function resulting from the action of a multiplicity of causes upon the persons tested. If such quotidian variability exists it will tend to make reliability coefficients calculated by the split-half method, and boosted by the Spearman-Brown formula, greater than reliability coefficients calculated by correlating parallel forms of a test with a time interval between successive testings.

The present enquiry was initiated to determine whether or not mental functions were characterised by a quotidian variability, and if mental functions exhibit such variability to estimate the influence of its presence upon reliability coefficients calculated by different methods. A preliminary discussion is presented, dealing with the variability of cognition, and methods of measuring such variability.

THE VARIABILITY OF COGNITIVE FUNCTION. C.r.s.c.

An examination of available relevent data indicates that possible variations in cognitive function may be classified into two categories. The first category includes those variations that may be described as quotidian. These variations are the resultant of the action of a multiplicity of random environmental influences upon the mental structure. One theory suggests that variations of this type may be of central physiological origin, and may be characterised by periodic fluctuation or oscillation. The second category includes variations in cognitive function over longer time intervals. These alleged long term variations are regarded as causually determined by environmental factors.

Spearman,[x] while accepting variations of the former type, repudiated the latter. With reference to these alleged variations over long time intervals he writes that "these "variations really derive from the operation of measurement, "not from the g itself which is measured."

[x] Spearman, C., (1932), "Abilities of Man", p.366.

Numerous enquiries have been conducted to determine the constancy or lack of constancy of the Stanford-Binet I.Q. These experiments indicate that there is a marked increase in variation with increase in the time interval between successive applications of the test. Robert L.Thorndike[x], by pooling the results of numerous investigators in this field, found that the correlation between test and retest varied from .889 for time intervals less than one month, to .698 for a time interval of 60 months.

Retests with certain Moray House Intelligence Tests at varying time intervals show a slight decrease in correlation with increase in time interval, but this decrease is of such a small order as to be insignificant. The following table contains in summary the available data on the constancy of the I.Q. as measured by Moray House Tests.

t	r	N
1 week	.931	629
1 week	.940	629
1 week	.935	629
7 weeks	.937	1030
15 months	.935	394
26 months	.929	363
38 months	.895	195

x Thorndike, Robert L.,(1933), "The Effect of the Interval between Test and Retest on the Constancy of I.Q." J.Educ.Psychol. vol. xxlv, pp. 543-549.

The last three coefficients in the above table are corrected for selection. These results indicate that the abilities measured by Moray House Tests exhibit no appreciable variation capable of detection by correlational technique with increase in time interval, and lend considerable weight to Spearman's hypothesis regarding the constancy of g over lengthy time intervals.

The above data throw no light on quotidian variations in cognitive function which may exist quite independent of long term variations. We shall firstly consider the various methods for isolating and measuring such variations.

METHODS of MEASURING FUNCTIONAL VARIABILITY. C.r.s.e.

Numerous methods have been devised for measuring functional variability. Some of these methods are considered briefly here.

ital The Double Test-retest of Function Fluctuation.

A method of measuring functional variability has been indicated by Thouless.[x] This method involves the administration of two intercorrelated tests at the same time, and correlating the arrays of scores thus found with arrays of scores found by administrating the same two tests, or parallel forms, again together at some other time. If z_1 and z_2 are the measurements obtained at the first administration of the tests, and z_1' and z_2' are the measurements obtained at the second administration, then $r_{12'}$ and $r_{1'2}$ will be less than r_{12} and $r_{1'2'}$ if functional varaibility is present. If the unreliability of the tests used is the only cause of variation, and errors of measurement do not correlate then r_{12}, $r_{1'2'}$, $r_{12'}$ and $r_{1'2}$ will tend to be equal. If functional variability is found to be present then r_{12} and $r_{1'2'}$ will have in common a factor of temporal contiguity increasing their inter: :correlations which factor is not common to $r_{12'}$ and $r_{1'2}$.

x Thouless, Robert H., (1936), "Test Unreliability and Function Fluctuation", B.J.P., xxvl, pp325- .

Thouless points out that the correlation between the differences between test and retest is demonstrative of functional variability. If there is no variation in the function tested then $r_{(1-1')(2-2')}$ will be positive if the correlation between the two tests is positive. This technique was first used by Brown and Thomson in detecting the presence of correlation between errors of measurement. Values of $r_{(1-1')(2-2')}$ can be conveniently calculated from a pooling square of intercorrelations between tests given on the same day and tests given on different days. Each test must be weighed according to its standard deviation, and appropriate negative signs introduced.

As an index for measuring the amount of fluctuation of function, Thouless proposes a method which takes into consideration the size of the intercorrelation between the tests. This is necessitated by the fact that $r_{(1-1')(2-2')}$ is not independent of the size of r_{12}. If r_{12} is small, then $r_{(1-1')(2-2')}$ will be small. He proposes to take as his index the correlation between the differences between test and retest divided by the mean of the same time correlations between z_1 and z_2. The resulting index is given by the formula

$$\frac{r_{(1-1')(2-2')}}{\frac{1}{2}(r_{12}+r_{1'2'})}$$

If this quantity is significantly different from zero then function fluctuation is present.

ital

The Coefficient of Trait Variability.

Another quantitative criterion for measuring functional variability has been proposed by G.B.Paulsen.* He advances the view that variability in the trait tested is responsible for the discrepancy between reliability coefficients calculated by thee split-half method, and coefficients calculated by correlating the scores on the same or parallel forms after a time interval. He proposes to correct the test retest coefficients for attenuation, using the boosted split-half reliability coefficients in the denominator of the attenuation formula. This corrected test re-test coefficient is called the coefficient of trait variability. Thus

$$C.T.V. = \frac{r_{11'}}{\sqrt{r_{11}\, r_{1'1'}}}$$

where $r_{11'}$ is the correlation obtained by test re-test by the same or equivalent forms, r_{11} the boosted split-half reliability of one form, and $r_{1'1'}$ the boosted split-half reliability of the other. If no trait variability is present, this coefficient will have a value of unity. It will be less than unity when trait variability is present. Thouless points out that this method is a special case of his test re-test criterion, the pairs in Paulsen's method being not different tests but pairs of the same test.

* Paulsen, G.B., (1931) "A Coefficient of Trait Variability" Psychol. Bulletin, xxvii, p.218.

Analysing the Error Variance of a Test.

It is possible to analyse the error variance of a test into two components, one component being the variance of the fluctuation in the ability tested, the other component being the error variance due to the incapacity of the test as an instrument of measurement. Thus we can write

$$s_e^2 = s_t^2 + s_f^2$$

where s_e^2 = total error variance of the test.

s_t^2 = the variance due to the incapacity of the test as an instrument of measurement.

s_f^2 = variance due to fluctuation in the ability tested.

If r_{11} is the correlation between two parallel forms given on the same day, and $r_{1'1'}$ is the correlation between the same two forms given on different days, then

$$s_e^2 = 1 - r_{11}$$

and $$s_f^2 = r_{11} - r_{1'1'}$$

Factors of Temporal Contiguity.

The use of some of the measures outlined above are invalidated as pure measures of functional variability due to the possible correlation of errors. In Paulsen's coefficient of trait variability it is unlikely that the boosted split-half reliability is equal to the reliability that would have obtained if the time interval between the tests were zero, and functional variability were absent.

Errors probably correlate to some small extent, and thereby spuriously increase the obtained reliability coefficients. Furthermore errors on two different tests given on the same day may also correlate. Since no method is apparent at the moment for adequately discriminating between the correlation of errors, and the absence of functional variability, we propose to use the term 'factors of 'temporal contiguity', a term first proposed by Thouless, factors of temporal contiguity being defined as those factors which tend to increase the correlation between tests given on the same day, and to reduce the correlation between tests given on different days. The existence of a factor of temporal contiguity may be due to the fluctuation of the abilities measured from day to day, or to the correlation of errors between tests given on the same day, or to some other cause as yet unpostulated. A technique is here developed for the measurement of such factors.

The Measurement of Factors of Temporal Contiguity.

The measurement of factors of temporal contiguity is a relatively simple procedure. It involves subtracting the matrix of intercorrelations between tests given on different days from the matrix of intercorrelations of the same tests given on the same day. The matrix of residuals is then examined. If these residuals can be considered as significantly greater than zero, then factors of temporal contiguity are known to exist common to the tests given on the same day. If the residual correlations are not significantly greater than zero, then we must assume that such factors are not present. If we conclude that our residuals are significant, we can then proceed to estimate the loadings of our factors of temporal contiguity by averaging all possible combinations of

$$r_{ib}^2 = \frac{r_{ij}\, r_{ik}}{r_{jk}}$$

in our residual matrix, where r_{ib} is the loading of our factor of temporal contiguity in test i. The assumption is made that our table of residual correlations, found by subtracting the matrix of intercorrelations between tests given on different days from the matrix of intercorrelations of the same tests given on the same day has a rank of 1.

To illustrate the procedure outlined above, a fictitious tabke of intercorrelations was drawn up between three tests given on the same day, and given on different days. Let z_1, z_2, and z_3 be three tests given on the same day, and z_1', z_2' and z_3' be the same tests or their parallel forms given on some other day. Let their matrix of intercorrelations be as shown in Table 3.

Table 3

	z_1	z_2	z_3	z_1'	z_2'	z_3'
z_1	--	.387	.407	.881	.293	.345
z_2	.387	--	.328	.297	.666	.202
z_3	.407	.328	--	.341	.200	.901
z_1'	.881	.297	.341	--	.386	.410
z_2'	.293	.666	.200	.386	--	.326
z_3'	.345	.202	.901	.410	.326	--

It will be observed from this fictitious matrix of intercorrelations that the intercorrelations between tests given on the same day are greater than the intercorrelations between tests given on different days. We, therefore, postulate the existence of factors of temporal contiguity. With only three different tests in a battery, we must assume that there is only one general factor, and no group factors. Although for purposes of simplicity only three tests are used in this illustration, the method outlined is entirely general and may be used with any number of tests, and any number of factors. Examination of the matrix of intercorrelations given in Table 3 leads us to expect the factor pattern given in Table 4.

Table 4

	a	b	c	s_1	s_2	s_3	error specefics
1	x	x		x			x
2	x	x			x		x
3	x	x				x	x
1'	x		x	x			x
2'	x		x		x		x
3'	x		x			x	x

The assumption is made that the data used is fallible, and that r_{12}', r_{13}', and r_{23}' are not exactly equal to r_{12}', r_{13}'', and r_{23}', respectively, but are very nearly so. Similarly r_{12}, r_{13} and r_{23} are not exactly equal to r_{12}'', r_{13}'', and r_{23}'', respectively. The reliability coefficients r_{11}', r_{22}', and r_{33}' are placed down the diagonals of the south-west and north-east quadrants. We have therefore, two matrices of intercorrelations between tests given on the same day, and two matrices of intercorrelations given on different days, and four possible matrices of residuals due to the existence of factors of temporal contiguity. Since z_1, z_2, and z_3 are the same tests or parallel forms of z_1', z_2' and z_3', we assume that the first factor loading of z_1 is the same as the first factor loading of z_1'; similarly with z_2, z_2' and z_3, z_3'. We, therefore, calculate the first factor loadings of z_1 and z_1' by averaging the two values of

$\dfrac{r'_{12}\,r'_{13}}{r_{23}'}$ and $\dfrac{r'_{12}\,r'_{13}}{r'_{23}}$, and taking the square root of this average; similarly with z_2, z_2' and z_3, z_3'.

Given the first factor loadings we can then calculate the matrix of intercorrelations accounted for by the first factor. Subtracting this matrix from the matrices of intercorrelations given on different days, we obtain a table of residuals, which are nearly zero. Subtracting the same matrix from the table of intercorrelations between tests given on the same day, we obtain a table of somewhat larger residuals. These matrices of residuals are given in Table 5.

Table 5

	z_1	z_2	z_3	z_1'	z_2'	z_3'
z_1	--	.0920	.0641	.3776	-.0020	.0021
z_2	.0920	--	.1270	.0020	.4931	.0010
z_3	.0641	.1270	--	-.0019	-.0010	.6674
z_1'	.3776	.0020	-.0019	--	.0910	.0671
z_2'	-.0020	.4931	-.0010	.0910	--	.1270
z_3'	.0021	.0010	.6674	.0671	.1270	--

The large residuals in the north-west and south-east quadrants must be accounted for by factors of temporal contiguity. Analysing the residuals in the north-west quadrant of Table 5, we obtain the factors of temporal contiguity common to z_1, z_2 and z_3, while the residuals in the south-east quadrant of Table yield similar factors common to z_1', z_2' and z_3'. The factor loadings thus calculated are given in the b and c columns of Table 6. The specefics and error specefics have also been calculated, and their loadings appear also in Table 6.

Table 6

Factor Pattern

variable	factor I a	factor II b	factor III c	specifics s_1	s_2	s_3	error specific
1	.7095	.2155		.5755			.3450
2	.4158	.4270			.5575		.5773
3	.4833	.2976				.7609	.3146
1'	.7095		.2193	.5740			.3450
2'	.4158		.4150		.5620		.5773
3'	.4833		.3060			.7575	.3146

The above fictitious example illustrates how the absence of functional variability may be measured as a factor of temporal contiguity. The method may be used with any number of tests, and any method of obtaining factors may be employed. If Thurstone's method is used, the centroid solution, calculated from the intercorrelations between tests given on different days, may be rotated into any psychologically significant configuration independent of the temporal contiguity factors, which must be regarded as already psychologically significant.

EXPERIMENTAL.

In order to determine the influence of 'factors of 'temporal contiguity' upon reliability coefficients the scores of 212 persons on the odd and even items of three Moray House Tests, M.H.T. 21, 23, and 26, were found. These three tests were administered with a time interval of one week between their successive administrations. Moray House Tests are known to exhibit a high degree of equivalence, and for the purpose of this investigation the three tests used are regarded as parallel forms. The theory underlying the experiment was that if factors of temporal contiguity existed, the correlation between parts of the same test would be higher than the correlation between parts of different tests.

The intercorrelations between the six test halves were calculated. These intercorrelations together with their standard errors are given in Table 7. Examination of this table indicates that the correlation between halves of the same test are markedly higher than the correlation between halves of different tests. Each coefficient was boosted by the Spearman-Brown formula for double length. These coefficients together with their standard errors calculated by the Shen formula are given in columns 3 and 4 of Table . The correlations between the whole tests are given in Table column 5.

Table 7

Tests	$r_{\frac{11}{22}}$	S.E.$_{\frac{11}{22}}$	r_{11}	S.E.$_{r11}$	
21 odd-23 Odd	.8806	.0154	.9365	.0087	
21 even-23 odd	.8993	.0132	.9470	.0073	r_{21-23} = .9078
21 Odd-23 even	.8527	.0187	.9207	.0109	
21 even-23 even	.8764	.0172	.9341	.0090	
21 odd-26 odd	.8643	.0174	.9272	.0100	
21 even-26 odd	.8932	.0139	.9436	.0077	r_{21-26} = .9076
21 odd-26 even	.8609	.0178	.9253	.0103	
21 even-26 even	.8969	.0134	.9456	.0075	
23 odd-26 Odd	.9086	.0122	.9521	.0066	
23 even-26 odd	.8937	.0138	.9439	.0077	r_{23-26} = .9284
23 odd-26 even	.9081	.0121	.9518	.0066	
23 even-26 even	.8953	.0137	.9448	.0076	
21 odd-21 even	.9278	.0095	.9625	.0051	
23 odd-23 even	.9393	.0081	.9687	.0043	
26 odd-26 even	.9457	.0072	.9721	.0038	

The standard deviation of raw scores for the whole tests and for each test half are as follows:-

Test	S.D.
M.H.T. 21	19.955
M.H.T. 23	17.953
M.H.T. 26	17.349
M.H.T. 21 even	10.1550
M.H.T. 21 odd	10.171
M.H.T. 23 even	8.792
M.H.T. 23 odd	9.477
M.H.T. 26 even	8.922
M.H.T. 26 odd	8.668

It will be observed from Table 7 that the boosted split-half reliability coefficients are in all cases greater, than the coefficients obtained by correlating parallel forms of the whole tests. The reasons for this are obviously that the correlation between the odd and even items of a test is higher than the correlation between corresponding parts of tests given on different days. Thus if we consider each parallel form of a test as composed of two separate variables, one variable representing the odd, the other the even items, then the correlation between the two whole tests may be written in the form

$$r_{(1+2)(1'+2')} = \frac{r_{11'} + r_{12'} + r_{12'} + r_{22'}}{2 - 2r_{12}}$$

where each variable is equally weighted. The Spearman-Brown formula makes the assumption that the coefficients in the numerator of the above equation are equal to each other, and equal to the coefficient in the denominator. When, however, a time interval separates the two testings the elements in the numerator may be substantially less than the elements in the denominator. Consequently the value $r_{(1+2')(1+2')}$ will tend to be less than reliability coefficients estimated by the Spearman-Brown formula.

286

A BI-FACTOR ANALYSIS OF RELIABILITY COEFFICIENTS.

George A. Ferguson

From the Education Department, Moray House,

University of Edinburgh.

A BI-FACTOR ANALYSIS OF RELIABILITY COEFFICIENTS.

1. Introduction

In the estimation of test reliability investigators have usually found that reliability coefficients obtained by correlating test halves, and boosting the obtained correlations by the Spearman-Brown formula, were higher than those obtained by correlating parallel forms administered on different days. Presumably, factors operate which determine an increase in the correlation between test halves given on the same day over the correlation between test halves given on different days. Whether these factors result from a quotidian variability of mental function or from the correlation of errors, provided quotidian variability and error correlation can, in themselves be considered as distinct concepts, or from a cause as yet unpostulated, is not clear. Whatever the cause the present enquiry was initiated to isolate and measure such factors, and to determine their influence on reliability coefficients. In the measurement of the factors in question, Holzinger's extension of the Spearman technique was used. Differences of opinion exist as to the legitimacy of the term 'bi-factor', since investigators apparently used a similar procedure to factorise matrices of correlations of rank greater than 1 before Holzinger advanced his systematic treatment of the method. Whatever the historical issues involved the term 'bi-factor' is used for convenience throughout this paper. A brief summary of the bi-factor method is given here to clarify later discussion.

11. Holzinger's Bi-factor Method.

Holzinger's method of bi-factor analysis attempts to describe a matrix of correlations in terms of one general factor, a number of group factors common to two or more variables, and as many specific factors as there are variables. This reduces the matrix to a minimum factorial description of one general factor, n specific factors, where n is the number of tests, and q group factors, q being smaller than n. The procedure is to examine the matrix of correlations to be factorised in order to isolate any groups of tests that correlate more highly among themselves than they do with the remaining tests in the battery, and grouping those tests together whose intercorrelations constitute elements in vanishing tetrads.

In allocating tests to group Holzinger uses what is termed a B-coefficient. A B-coefficient is defined as, "the average of all intercorrelations of tests 1,2,.......K, "divided by the average of all correlations of tests "1,2,........K, with the remaining tests not in the group."

Having allocated the tests to groups, the next procedure is to remove the general factor. This is accomplished in a manner similar to that employed by Spearman in estimating g loadings by averaging all possible combinations of

$$r^2_{ig} = \frac{r_{ij}r_{ik}}{r_{jk}} \qquad (1)$$

In the bi-factor method only those values of r are used that are elements in tetrads approximating zero.

Let the following represent a hypothetical bi-factor pattern with six variables.

	a	b	c	d
1	a_1	b_1		
2	a_2	b_2		
3	a_3		c_3	
4	a_4		c_4	
5	a_5			d_5
6	a_6			d_6

Examination of this factor pattern will show that certain tetrads such as $r_{13}r_{24} - r_{14}r_{23}$ will be zero, while certain others such as $r_{12}r_{34} - r_{14}r_{23}$ will be greater than zero. In the above factor pattern there will be four values of r_{ia}, which with fallible data must be averaged. Thus the formula for the general factor loading of the first variable becomes:-

$$r_{1a}^2 = \frac{r_{13}r_{15} + r_{14}r_{16} + r_{13}r_{16} + r_{14}r_{15}}{r_{35} + r_{46} + r_{36} + r_{45}}$$

Having removed the general factor a teble of residual correlations is calculated, and the group factors removed successively.

<u>111. A Bi-factor Analysis of the Intercorrelations between the Halves of Three Equivalent Test Forms.</u>

The data used in the present enquiry resulted from the administration of three Moray House Tests of Intelligence, M.H.T.21, M.H.T.23 and M.H.T.26 to some 1800 children in West Yorkshire. The administration of these three tests constituted part of an experiment conducted by the West Yorkshire National Union of Teachers into the relative effectiveness of different types of examinations for selecting children for secondary school education. These data were made available, and lent themselves adequately for the purposes of the enquiry described in this paper. The time interval separating the successive administrations of the three tests was one week.

Since the procedure of the present experiment involved the laborious task of calculating the scores of each child on the odd and even items of each test, a random sample of 212 children was selected from the number available.

The standard deviations of raw scores in the sample and in the population for the three tests were as follows:-

Tests	Sample	Population	N
M.H.T. 21	19.96	22.16	212
M.H.T. 23	17.95	20.29	212
M.H.T. 26	17.35	19.74	212

Each test contained 100 items, and required 45 minutes to administer. The three tests were similar in structure, and are regarded as parallel forms. The scores of each child on the odd and even items of each test were found. The standard deviations of scores on the six test halves were as follows:-

Test	odd	even
M.H.T. 21	10.15	10.17
M.H.T. 23	8.79	9.48
M.H.T. 26	8.92	8.67

The fifteen different intercorrelations between the six halves of the three tests were calculated. Three of these intercorrelations are between halves of tests given on the same day. The remaining twelve intercorrelations are between halves of tests given on different days. Since the three tests are regarded as parallel forms each correlation may be regarded as a reliability coefficient of a half test. None of the coefficients have been boosted by the Spearman-Brown formula. Evidence will be advanced later in this paper to show that the three forms used exhibited a high degree of equivalence.

Examination of the matrix of intercorrelations (Table 8) between the halves of three parallel forms of the same test shows immediately that the correlations between the halves of the same tests are higher than the correlation between the halves of different tests; that is, between the halves of tests given on different days.

Table 8

	1	2	3	4	5	6
1	-	.9457	.9086	.8937	.8643	.8932
2	.9457	-	.9081	.8953	.8609	.8969
3	.9086	.9081	-	.9393	.8806	.8993
4	.8937	.8953	.9393	-	.8527	.8764
5	.8643	.8609	.8806	.8527	-	.9278
6	.8932	.8969	.8993	.8764	.9278	-

NOTE

Variables 1and 2 refer to the odd and even items, respectively, of M.H.T.26, variables 3 and 4 to the odd and even items of M.H.T.23, and variables 5 and 6 to the odd and even items of M.H.T.21.

The correlations between halves of the same test have been marked off in Table 8 by diagonal blocks, and they form non vanishing tetrads with the other coefficients in the matrix. The correlations between halves of the tests given on different days form tetrad differences whose values do not differ significantly from zero. It is evident, therefore, that it is possible to describe the present matrix of correlations in terms of one general factor, and three group factors. Since the coefficients in Table represent the correlations between parallel forms of the same test no specific factor variance other than error factor variance is to be expected. If the test used had not approximated to a high degree of equivalence, specific factors would have required consideration. The close correspondence of the intercorrelations of the halves of the tests is suggestive that adequate parallelism was secured.

In the present analysis the first factor loadings were estimated by formula (2), and are recorded in the first column of the factor pattern, Table 10. The residuals $\underline{r_{ij}} = r_{ij} - a_i a_j$ were then calculated. The table of residuals after removal of the general factor is given in Table 9.

First Residual Correlations

Table 9

	1	2	3	4	5	6
1	-	.0426	-.0026	.0024	-.0010	.0012
2	.0426	-	.0050	.0036	-.0047	.0045
3	-.0026	.0050	-	.0396	.0071	-.0012
4	.0024	.0036	.0396	-	-.0016	.0044
5	-.0010	-.0047	.0071	-.0016	-	.0727
6	.0012	.0045	-.0012	.0044	.0727	-

Examination of the first residual matrix (Table 9) indicates that the general factor loadings have described with a high degree of accuracy the majority of the inter:
:correlations. The residuals $\underline{r_{12}}$, $\underline{r_{34}}$, and $\underline{r_{56}}$ are, however, considerably larger than the remaining residuals, and indicate the expected tendency for further overlap between the variables 1 and 2, 3 and 4, 5 and 6. The largest residual among the non diagonal elements where zero tetrad differences were presumed, $\underline{r_{35}}$, is only .97 times the standard error of the initial correlation. All the residuals, excluding those in the diagonal blocks, are insignificant, if a comparison with the standard errors of the initial correlations can be regarded as a criterion. The residuals in the diagonal blocks, $\underline{r_{12}}$, $\underline{r_{34}}$, and $\underline{r_{56}}$ are all significant

when judged by the same criterion, the smallest diagonal residual $\underline{r_{34}}$ being 9.2 times as large as the standard error of the initial correlation.

The next step in the calculation was to find the error variance of each variable by the formula $e_i^2 = 1 - r_{ii}$, where e_i^2 is the error variance of variable i, and r_{ii} the reliability coefficient of variable i. The loadings of the error factors were thus found, and these are recorded in the staggered e_i column of Table 10. In estimating these loadings the odd-even item correlation of each test was taken as r_{ii}, and the assumption made that the odd items of each test had an error variance equal to that of the even items. This is, indeed, a justifiable assumption, and the only one that can be made in the present analysis.

The remaining group factor loadings were then readily calculated by the following simple formula:-

$$r_{ib}^2 = 1 - r_{ia}^2 - e_i^2$$

where r_{ib}^2 is the variance of factor b in test i, r_{ia}^2 the variance of the general factor, and e_i^2 the error factor variance of test i.

Factor Pattern
Table 10

variable	Factor I a	Factor II b	Factor III c	Factor IV d	Error Factor Loadings e_i^2						h_i^2
1	.9501	.2074			.2330						.9457
2	.9505	.2045				.2330					.9457
3	.9591		.1393				.2464				.9393
4	.9381		.2435					.2464			.9393
5	.9107			.3137					.2687		.9278
6	.9389			.2152						.2687	.9278

The factor pattern of Table 10 describes with considerable accuracy the original correlation matrix. Some estimation of how closely the final factor pattern accounts for the original correlations is given by examination of the final residuals in Table 11.

Final Residual Correlations.
Table 11

	1	2	3	4	5	6
1	-	.0000	-.0026	.0024	-.0010	.0012
2	.0000	-	.0050	.0036	-.0047	.0045
3	-.0026	.0050	-	.0057	.0071	-.0012
4	.0024	.0036	.0057	-	-.0016	.0044
5	-.0010	-.0047	.0071	-.0016	-	.0052
6	.0012	.0045	-.0012	.0044	.0052	-

IV. Interpretation of Factors.

The factors isolated by the above analysis require interpretation. Close correspondence of the general factor loadings, and also of the group factor loadings, is a good criterion of test form equivalence. Variable 5 (the odd items of M.H.T.21) manifests the highest degree of inequivalence, but this inequivalence is not sufficiently prominent to introduce a specific factor loading approximating anywhere near significance. The close correspondence of factor loadings as calculated above is a better index of test equivalence than the correspondence of the intercorrelations between the halves of tests. If the halves of the various tests used are equivalent then the intercorrelations of the halves of the tests given of different days should be equal within the limits of sampling error. The converse, however, does not hold. The fact that the correlations between the halves of tests given on different days are equal is no indication of test equivalence. If A were a test of intelligence and B a test of ability to do arithmetic, and a_1, a_2 are the odd and even items respectively of test A, while b_1, b_2 are the odd and even items respectively of test B, then r_{a1b1}, $r_{a_1b_2}$ $r_{a_2b_1}$, $r_{a_2b_2}$ could all readily be equal, and yet it is obvious that A is a test of different structure from B.

What is indicated, however, is that the odd and even items of test A are equivalent, and the odd and even items of test B are equivalent, but the halves of A are not necessarily equivalent to the halves of B. Close correspondence of the factor loadings of two forms of a test, when used in the same battery of tests, both parallel forms being applied to the same group of children, is a reliable index of the equivalence of the two forms. In the above analysis the absence of anything approximating to a significant specific is demonstrative that good equivalence has been obtained.

The group factors isolated by the above analysis may be termed factors of temporal contiguity, a term first used by Thouless. If we could conclude that the function measured were a non-fluctuating one, then these group factors could be interpreted as largely the result of error correlation. If we could conclude that in correlating the halves of the same test the errors are uncorrelated , then the group factors could be described as manifestations of the absence of functional variability between those tests having group factors in common. Since, however, it is not unlikely that both the correlation of errors, and functional variability are exerting a positive influence on the size of the group factors, and since no method of determining the relative importance of these two influences is at the moment apparent, it is only possible to describe these factors as factors of temporal contiguity, and to regard them merely as the resultant of those influences that tend to reduce the correspondence between test scores on parallel forms of the same test with increase in the time interval between their applications.

V. A Comparison of Multiple Orthagonal Factors with Bi-factors.

To obtain a comparison between the factors obtained by the above bi-factor analysis, and those obtained by multiple factor analysis the intercorrelations given in Table 8 between the halves of three parallel forms of the same test were analysed by Thurstone's method. The largest correlation in each row was used as the diagonal element, and was maintained unchanged throughout the analysis in that it represented a very close approximation to the true comunality. It was found that this matrix of correlations could be adequately described in terms of three multiple orthagonal factors instead of four bi-factors. This is in complete correspondence with the findings of Holzinger that four bi-factors can be described in terms of three multiple orthagonal factors. The centroid solution of the Thurstone analysis is given in Table 12.

Table 12

Tests	Loadings of the factors Centroid Solution			Communality
	1	11	111	h_i^2
1	.9560	-.1044	.1401	.9445
2	.9563	-.1049	.1498	.9480
3	.9602	-.0690	-.1217	.9416
4	.9465	-.1297	-.1687	.9411
5	.9320	.2472	-.0166	.9299
6	.9508	.1604	.0111	.9299

The communalities of the centroid solution are in close agreement with the communalities of the bi-factor solution, and both patterns describe the correlations of the original matrix with a close degree of accuracy.

The factor pattern of Table 12 was now rotated to remove negative loadings, and to obtain as many zero loadings as possible, while still maintaining a factor space of three dimensions. This was done by rotating two factors at a time graphically, factors 1 and 2 being rotated first, and then factors 1 and 3. Each pair of columns of loadings was post-multiplied by a 2x2 orthogonal matrix representing a rotation of rectangular axes in two dimensions through a given angle θ. The elements of this orthogonal matrix

$$\begin{bmatrix} \cos\theta & \sin\theta \\ -\sin\theta & \cos\theta \end{bmatrix}$$

were found by regarding the loadings of the test through which the axes were rotated as co-ordinates of a point in a plane, and by these co-ordinates calculating the sine and cosine of the angle of rotation. The rotated factor loadings are given in Table 13.

Table 13

Tests	Rotated Factor Loadings			Communality
	1	11	111	h_1^2
1	.9229	.0004	.3052	.9449
2	.9216	-.0001	.3149	.9485
3	.9687	.0360	.0475	.9419
4	.9700	-.0258	.0000	.9416
5	.8881	.3474	.1451	.9304
6	.9119	.2631	.1723	.9305

The factor pattern of Table 13 is one of a large number that could be obtained by using different angles of rotation. Four of the loadings of Factor 11, and two of the loadings of Factor 111 are regarded as zero. These loadings are under: :lined in Table 13. All other loadings are positive. No system of rotation can produce more than six zeros in this pattern in this three dimensional factor space. The bi-factor solution describes the observed correlations in terms of four factors and twelve factor loadings. The rotated multiple factor pattern describes the same correlations in terms of three factors and twelve factor loadings.

By the method described by Holzinger in "Student Manual of "Factor Analysis" the relationship between the two factor patterns can be found, the relationship being expressed in terms of a set of three linear equations. This involves the reduction of the original tests to as many new variables as there are group factors in the bi-factor solution. In this case the six original tests are expressed in terms of three composite tests z_a, z_b and z_c. The first factor loading of the composite test z_a for both bi-factor and multiple orthagonal patterns is found by adding the first factor loadings of variables 1 and 2, and dividing this sum by the combined standard deviation of these tests. The formula for the combined standard deviation of n variables when each variable is given unit weight is as follows:-

$$\sigma_{1+2+3\cdots n} = \sqrt{n + 2(r_{12} + r_{13} + r_{14} \cdots r_{n-1,n})}$$

The values in the present case are $\sigma_{1+2} = 1.9729$, $\sigma_{3+4} = 1.9694$ and $\sigma_{5+6} = 1.9635$.

The reduced factor pattern calculated from the bi-factor solution is found to be as follows:-

			h_i^2
$z_a = .9635 + .2093b$			.9721
$z_b = .9633a$	$+ .1944c$		.9657
$z_c = .9420a$		$+ .2694d$	.9599

Taking the same composite tests for the multiple factor solution we obtain the following set of equations:-

h_i^2

$z_a = .9350Z_1 + .0001Z_2 + .3143Z_3$ .9730

$z_b = .9844Z_1 + .0052Z_2 + .0241Z_3$ .9697

$z_c = .9167Z_1 + .3109Z_2 + .1671Z_3$ .9631

The communities in both sets of equations are in close correspondence. The intercorrelations of the reduced tests are given in Table 14.

Table 14

Bi-factor			Multiple		
	z_a	z_b		z_a	z_b
z_a			z_a		
z_b	.9281		z_b	.9280	
z_c	.9076	.9074	z_c	.9080	.9079

The two sets of correlations given in Table 14 are in close agreement and indicate that both patterns are equally good fits of the observed correlations.

By equating these two sets of equations, and solving for Z_1, Z_2, and Z_3 we can obtain a set of equations which shows the relationship between the two sets of factors by describing the multiple factors in terms of bi-factors. These three equations are found to be

$Z_1 = .9471a + .1059b + .2144c - .0043d$

$Z_2 = .0712a - .3291b - .3057c + .8880d$

$Z_3 = .1654a + .7226b - .6258c - .0153d$

The standard deviations of Z_1, Z_2 and Z_3 in the above equations approximate to unity, and the intercorrelations of the Z's approximate to zero.

The relative importance to be attached to each bi-factor in describing the Z's may be found by squaring all the values in the above equations obtaining the following:-

$$\sigma_{Z_1}^2 = .9489\,\sigma_a^2 + .0112\,\sigma_b^2 + .0460\,\sigma_c^2 + .0000\,\sigma_d^2$$
$$\sigma_{Z_2}^2 = .0051\,\sigma_a^2 + .1083\,\sigma_b^2 + .0935\,\sigma_c^2 + .7885\,\sigma_d^2$$
$$\sigma_{Z_3}^2 = .0274\,\sigma_a^2 + .5222\,\sigma_b^2 + .3950\,\sigma_c^2 + .0000\,\sigma_d^2$$

From these equations it is apparent that nearly all the variance of Z_1 is attributable to the bi-factor a. Z_2 is made up largely of the bi-factor #, while Z_3 is composed largely of the bi-factors b and c.

Vl. Some Oberservations Regarding the Above Comparison.

The above enquiry commenced with the initial hypothesis that factors of temporal contiguity existed, tending to make the intercorrelations between tests given on the same day greater than the intercorrelation between tests given on different days. The necessary intercorrelations were calculated, and the postulated factors of temporal contiguity isolated and measured by a bi-factor analysis. It was found that the bi-factor solution furnished a factorial configuration in complete agreement with the postulated psychological hypothesis. The compatibility between the factorial configuration and the psychological hypothesis was sufficient to regard the initial hypothesis as proved.

When we now come to analyse our table of inter:
:correlations by multiple factor methods we find that an

equally accurate mathematical description can be obtained in terms of a pattern of three factors, but no matter what method of rotation is adopted these three factors can never be transformed into a psychologically meaningful configuration within a factor space of three dimensions, a factor space of foxr dimensions being required before our factor pattern can become compatible with our initial hypothesis. It is of course clear that an orthagonal transformation can in theory be obtained capable of rotating the three multiple factors into a psychologically meaningful four factor space. This would involve post-multiplying the factorial matrix of order 6 x 3 with known elements by an orthagonal matrix of order 3 x 4 of unknown elements. The estimation of the elements of the orthagonal matrix capable of brining the multiple factor pattern into agreement with the bi-factor pattern is a matter of considerable mathematical difficulty, and of great mathematical labour.

In the present example the simplicity of our factor pattern renders the inadequacy of a three dimensional factor space, and the necessity of an additional space readily observable. Furthermore, the difficulty of attaining a meaningful interpretation of our three rotated multiple factors is also apparent. With more complicated factor patterns, however, this difficulty is not readily observed, and the psychologist has no clue to guide him to the conclusion that his factor pattern must be rotated into additional dimensions to obtain meaningful factors. The assumption is usually made that a minimum number of factors with as many zero loadings as possible is likely to be the most meaningful configuration attainable. In our present example such a configuration has little, if any, meaning, and it does not seem likely that in more complicated patterns the reduction of the number of factors to a minimum would necessarily lead to the most meaningful solution. Our conclusion is, therefore, that under certain circumstances by reducing the number of factors to a minimum we will arrive at an invalid interpretation of the mental factors involved in the performance of certain tests, and that under these circumstances bi-factor solutions will tend to more meaningful results than orthagonal solutions.

The fundamental difference between the Thurstone method of obtaining factors and the bi-factor method seems to be this. The former attempts to fit a psychological interpretation to a mathematical hypothesis. The latter attempts to fit a mathematical interpretation to a psychological hypothesis. Since we are primarily interested in proving or disproving psychological hypothesis the bi-factor method would seem, from the point of view of psychology, to be the more valid scientific method, and more likely to produce useful results.

VI. SUMMARY.

1. The intercorrelations between the split-halves of three equivalent group tests of intelligence given on different days are analysed by Holzinger's bi-factor method, and factors of 'temporal contiguity' isolated and measured.
2. The existence of factors of 'temporal contiguity' may be due to the absence of the influence of functional varaibility on the correlations between tests given on the same day, or to the correlation of errors, or both.
3. The existence of factors of 'temporal contiguity' explain why reliability coefficients calculated by the split-half method, and 'boosted' by the Spearman-Brown formula are unusually higher than reliability coefficients obtained by correlating parallel forms.
4. A comparison is made between bi-factor and multiple factor techniques.
5. Reasons and calculations are advanced to show that the reduction of the number of factors to a minimum may under certain circumstances lead to meaningless factors, quite incompatible with a previously established psychological hypothesis.
6. The argument is presented that from the point of view of psychology the fitting of a mathematical interpretation to a psychological hypothesis, rather than the converse, is the more valid scientific method, and likely to lead to more meaningful results.

REFERENCES.

(1). Holzinger, Karl J. (1937) "Student Manual of Factor "Analysis." Chicago.

(2) Holzinger, Karl J., and Swineford, Frances (1937). "The Bi-factor Method," Psychometrika, 2,1, pp. 41-54.

(3) Thouless, Robert H. "Test Unreliability and Function "Fluctuation." Brit. J. Psychol., XXVl, pp. 325-343.

(4) Holzinger, Karl J., (1938) "Relationship between "Three Multiple Orthagonal Factors and Four Bi-factors." J. of Educ. Psychol., Vol. XX1X, pp. 513-519.

(5) Thurstone, L. L. (1935) "The Vectors of Mind." (Chicago).

37

THE INFLUENCE OF THE USE OF MULTIPLE/CHOICE ITEMS ON TEST RELIABILITY

THE INFLUENCE OF THE USE OF MULTIPLE-CHOICE ITEMS ON TEST RELIABILITY.

One source of test unreliability derives from the use of test items of the multiple-choice type. In the specific case of a test constructed of true-false items a person's score will vary from trial to trial due to the influence of chance alone, quite apart from other contributing sources of error. Thus, with a test constructed entirely of true-false items, the probability is one half that completely ignorant or very unintelligent persons will attain a score of $N/2$ by pure guess work, where N is the number of items on the test, and provided all items are attempted. The mean score made by such a hypothetical population of persons on such a test will be $N/2$ and the variance of scores $N/4$. Thus with a test of 100 items of the true-false type all of which are attempted the distribution of scores made by our

NOTE These values are calculated from formulae for the mean and variance of the point binomial. The mean of the point binomial is Np, and its variance Npq. In the present argument N is the number of items on the test, p is the probability of getting an item correct by chance, and q is the probability of getting it wrong by chance. When the items are of the true-false type $p=q=\frac{1}{2}$.

completely ignorant population will have a mean of 50 and a variance of 25. If the test were given again to the same population we should expect the same mean and variance, and a correlation between test and retest of zero, since all scores on both applications of the test are made by chance alone. With a test constructed of 100 multiple-choice items where the number of alternatives offered is five this hypothetical population will have scores normally distributed about a mean of 20 with a variance of 16. If such a test were given a second time to the same population, we should again expect a correlation between test and retest, of zero.

With a test constructed of true-false or multiple-choice items we may make the assumption that all individuals, with the exception of those who make perfect scores, secure some of their scores by chance and some as a result of their knowledge or ability. Thus , disregarding for the moment other sources of variable error, we may assume that every person's score on a multiple-choice test is capable of division into two parts;

$$z = x\ y \qquad (1)$$

where z = obtained score

x = score resulting from ability

y = score resulting from chance.

If the test is given a number of times to the same individual z will vary because of chance variations in y. It is apparent, therefore, that apart from other sources of variable error, chance is a factor contributing to unreliability in tests constructed of items of the multiple-choice type.

The usual formula for correcting a test score for chance is

$$x = z - w/(n-1) \qquad (2)$$

where x and z are as above, w is the total number of incorrect responses, and n is the number of alternative responses for each item, the number of alternative responses for every item on the test being the same. It may be mentioned here that this formula is usually written in different notation. This formula is based on the assumption that if an individual scores x points without the aid of chance the probability is $\frac{1}{2}$ that he will increase his score $\frac{z-x+w}{n}$ points by chance alone. If the procedure of administering the test is such that we may regard all items not passed as attempted, the relationship is simplified, and we may state that the probability is $\frac{1}{2}$ that an individual who scores x points by ability alone will score $\frac{N-x}{n}$ additional points by chance, where N is the number of items on the test. Thus the odds are even that an individual who scores 50 points by ability on a test constructed of 100 items with 5 alternatives for each item will increase his score 10 points by chance, thus making a total score of 60.

Since chance is a source of unreliability in tests of the multiple-choice type, it is possible to estimate the maximum reliability attainable by such tests if chance were the only source of unreliability. It is also possible to estimate the importance of chance as a factor in test unreliability relative to other sources of variable error.

Let $z = x+y$

where z, x, and y are as above.

For any given value of x the variance of y is equal to

$$(N-x)pq \qquad (3)$$

where N = the number of items on the test.

p = the probability of success on an item.

q = the probability of failure on an item.

Averaging this component over normally distributed values of x we obtain

$$s_y^2 = (N-M_x)pq \qquad (4)$$

where s_y^2 = variance of y for normally distributed values of x.

M_x = mean of x.

It may also be shown that

$$M_x = \frac{M_z - Np}{q} \qquad (5)$$

where M_z = the mean of z

so that

$$s_y^2 = (N-M_z)p \qquad (6)$$

Now the usual formula for the error variance of a test score is

$$E_z^2 = s_z^2(1 - r_{zz'}) \qquad (7)$$

where E_z^2 = error variance of a score in test z

s_z^2 = variance of z.

$r_{zz'}$ = the reliability coefficient of test z.

Therefore

$$r_{zz'} = 1 - E_z^2/s_z^2 \tag{8}$$

If chance is the only souree of unreliability

$$E_z^2 = s_y^2 \tag{9}$$

The maximum reliability that can be attained by a test constructed of multiple-choice items will be given by substituting equation (6) in equation (8) obtaining the following formula:-

$$r_{zz'\,(\text{max.})} = 1 - \frac{(N-Mz)p}{s_z^2} \tag{10}$$

where $r_{zz'\,(\text{max.})}$ = the maximum reliability that can be attained with a test constructed of multiple-choice items.

If n is the number of alternative responses $p = 1/n$, and we can write the above formula in the form

$$r_{zz'\,(\text{max.})} = 1 - \frac{(N - M_z)}{ns_y^2} \tag{11}$$

If chance is not the only source of unreliability, and other sources of variable error are present, on the assumption that such errors are uncorrelated, the variances are additive, and we have the relation

$$E_z^2 = e_z^2 + s_y^2 \qquad (12)$$

where e_z^2 = the variance of other sources of error.

Hence

$$r_{xx'} = \frac{r_{zz'}}{1 - \frac{N - M_z}{ns_z^2}} \qquad (13)$$

Where $r_{xx'}$ is the reliability that would have obtained if the probability of scoring a certain number of points by chance were zero.

We are, therefore, in a position to analyse the total error variance of a test into two components, (a) that due to some unknown source of error, (b) that due to the use of multiple-choice items.

By way of illustrating formula (5) Table 15 was constructed showing the maximum reliability that can be attained with a test of 100 items for varying numbers of alternative responses, and different standard deviations. The mean score in this Table is taken as 50.

The formulae developed in this paper are largely of theoretical interest in that they disclose the influence of certain chance factors on test reliability. For practical purposes a variety of complications may tend to invalidate their use, if they are used without due regard for the assumptions upon which they are based. Firstly, it is assumed that all items on a test are attempted by all individuals in the sample tested. This will only be the case when unlimited time is given for the completion of the test. When, however, a time limit is set so that speed of performance is regarded as an index of ability, the influence of the use of multiple-choice items on test unreliability will be somewhat reduced, because the less capable persons will not attempt the items near the end of the test.

Furthermore, by increasing the number of alternatives, although we increase the reliability of the test, we also increase the difficulty values of the items. Apart from the influence of chance altogether, we cannot regard an item containing 4 alternatives as directly comparable with the same item with another alternative added. An individual who is quite capable of selecting the proper response from 4 alternatives, might experience difficulty in selecting the proper response from 5 alternatives. The nature of the alternative added may tend to increase the difficulty value of the item.

The situation is further complicated by the fact that guessing is seldom an entirely chance process. Degrees of certainty exist, and all alternatives may not seem equally plausible to the testee. It would seem, therefore, that an individual who should fail an item of n alternatives has a probability greater than 1/n of responding correctly. One counteracting influence is that the ability to guess the correct answer may be correlated with the ability measured by the test.

Table 15

A Table of maximum reliabilities attainable for a test of 100 items for different numbers of alternative responses, and different values of the standard deviation.
The mean is taken as 50.

Alternatives	Standard Deviation				
	5	10	15	20	25
2	.0000	.7500	.8889	.9357	.9600
3	.3333	.8333	.9259	.9583	.9733
4	.5000	.8750	.9444	.9687	.9800
5	.6000	.9000	.9556	.9750	.9840
6	.6667	.9166	.9630	.9792	.9866
7	.7143	.9286	.9683	.9821	.9886
8	.7500	.9375	.9722	.9844	.9900
9	.7778	.9444	.9753	.9861	.9911
10	.8000	.9500	.9778	.9875	.9920

THEORIES OF TEST STRUCTURE

AND

METHODS OF IMPROVING THE EFFICIENCY OF TESTS

INTRODUCTION.

The interpretation of a test as a large composite battery of small unit tests, each unit contributing by virtue of its interaction with the other units of the test to the functioning of the test as a whole, indicates methods whereby the basic factors within the test structure influencing the efficacy of the whole test may be analysed. Such a concept suggests methods and guiding principles in the construction of mental tests whereby reliability and discriminate power, may be increased, and the worth of the test as an instrument for educational selection improved in some degree. The present discussion is developed to investigate the properties of the fundamental interactions within the test structure which determine the functioning of the whole test. Such a discussion involves a detailed analysis of the properties of the answer pattern of tests.

BASIS FORMULAE.

A study of answer pattern structure involves the use of certain formulae in general use for purposes of item selection. The most fundamental of these are the formulae for the variance of a dichotomously scored variable, and the inter: :correlation of such dichotomously scored variables.

The variance of a single dichotomously scored test item is given by the formula, pq, where p is the proportion of persons passing the test item, and q the proportion of persons failing the item.

The correlation between any two dichotomously scored items is given by the formula

$$r_{ij} = \frac{p_{ij} - p_i p_j}{\sqrt{p_i q_i p_j q_j}}$$

where r_{ij} = correlation between items i and j.

p_{ij} = proportion of persons passing both items.

p_i = proportion of persons passing item i.

p_j = proportion of persons passing item j.

q_i = proportion of persons failing item i.

q_j = proportion of persons failing item j.

Given the item variances and the inter-item correlations determined by the above formulae, the variance of the whole test is obtained by writing the inter-item correlations in the form of a pooling square with 1's down the diagonal, weighting each item according to its standard deviation, and summing the weighted elements. The sum of the weighted elements is the variance of scores on the whole test; thus the variance of test scores is written as a function of n independent item variances, and n (n-1) inter-item covariances, as follows;

x Thomson, Godfrey H.,"The Factorial Analysis of Human Ability". University of London Press, pp. 83-101.

where σ_t^2 = variance of raw scores on whole test.

n = number of test items.

$$\sigma_t^2 = \sum_{i=1}^{n} \sigma_i^2 + \sum_{\substack{i=1 \\ i \neq j}}^{n} \sum_{j=1}^{n} r_{ij}\,\sigma_i \sigma_j$$

This equation indicates that to increase the variance of a test, without increasing the value of n, thereby increasing the tests capacity for discriminating between the persons tested, we must increase the item variances and the inter-item covariances. Since the item variances represent only 1/n per cent of the elements in the initial pooling square, we conclude that when n is large the inter-item covariances are the basic determiners of test variance.

NOTE

A note may be appended here regarding the answer pattern matrix. The answer pattern of a test is written in the form of a matrix in which each row represents an item, each column represents a person, and each element a_{ij} has a value of either zero or unity when the items are scored dichotomously, as follows;

$$A = \begin{bmatrix} a_{11} & a_{12} & a_{13} & \cdot & \cdot & \cdot & a_{1N} \\ a_{21} & a_{22} & a_{23} & \cdot & \cdot & \cdot & a_{2N} \\ a_{31} & a_{32} & a_{33} & \cdot & \cdot & \cdot & a_{3N} \\ \cdots & \cdots & \cdots & \cdot & \cdot & \cdot & \cdot \\ \cdots & \cdots & \cdots & \cdot & \cdot & \cdot & \cdot \\ a_{n1} & a_{n2} & a_{n3} & \cdot & \cdot & \cdot & a_{nN} \end{bmatrix}$$

Denoting this matrix by A we may write

$$AA' = P$$

where P is the matrix of the number of persons passing both items i and j. The matrix of the proportion of persons passing both items i and j is denoted by

$$\lambda AA' = \lambda P$$

where $\lambda = \frac{1}{N}$, N being the number of persons.

The matrix of the inter-item covariances is then denoted by

$$\lambda AA' - \lambda^2 QQ' = C$$

where Q is the column vector of the number of persons passing each item, and C the matrix of inter-item covariances.

Just as it is possible to estimate the correlation between the rows of the answer pattern matrix, so the correlation between columns, i.e. the correlation between persons, may be estimated. By arguments similar to those used above the matrix of inter-person covariances may be found and denoted by

$$kA'A - k^2LL' = D$$

where $k = 1/n$, L a column vector of raw scores, and D the matrix of inter-person covariances.

No simple reciprocal relationship is apparent between the correlation of the rows of the answer pattern matrix, and the correlation between the columns.

A UNIQUE ANSWER PATTERN MATRIX.

David A. Walker[x] has investigated some of the properties of answer pattern matrices, and the relationship between such properties and the distribution of raw scores. He points out that any person's score x on a test may be made up in a large number of different ways. Theoretically at least ${}^{x}C_{n}$ possible ways existof making a score x on a test of n items. Firstly the score x may be made by responding correctly to the x easiest items on the test. When the score x of every person tested is composed of correct responses to the x easiest items on the test, where the x easiest items are described by the responses of all persons tested, and when the y persons passing a given item are the y most capable

x Walker, D.A., (1931), "Answer Pattern and Score Scatter in Tests and Examinations", B.J.P. xxll, pp. 73-86.
(1936), "Answer Pattern and Score Scatter in Tests and Examinations", B.J.P. xxll, pp. 301-308.
(1940)"Answer Pattern and Scores Scatter in Tests and Examinations", B.J.P. xxx, pp.248-260.

persons in the sample, where the y more capable persons are described by the performance of all persons tested on the whole test, the answer pattern matrix may be described as unique. In practice, however, such a unique answer pattern matrix is never attained, since an element of 'higgledypiggledyness' enters into the composition of all but zero and perfect scores. The answer pattern of every test approximates in greater or less degree to such a unique theoretical configuration, and we shall demonstrate below that the closer this approximation the more efficacious the test.

Walker points out that when the answer pattern matrix is unique the distribution of raw scores is completely determined by the difficulty values of the items, the distribution of raw scores being equal to the first differences of the distribution of the number of persons passing each item correctly, the items being arranged in order of difficulty. Thus, if P_0, P_1, P_2, P_3, ...P_k represent the number of persons passing each item, the items being arranged in order of difficulty, then the frequencies of the distribution of raw scores may be found by taking the first differences of this distribution, as follows;

item	no. persons passing item	frequencies of raw scores
0	P_0	$P_0 - P_1 = f_1$
1	P_1	$P_1 - P_2 = f_2$
2	P_2	$P_2 - P_3 = f_3$
3	P_3	$P_3 - P_4 = f_4$
4	P_4	$P_4 -$
.	.	
.	.	
k	P_k	$P_{(k-1)} - P_k = f_k$

where f_1, f_2, f_3..........f_k are the frequencies of the distribution of raw scores on the test. It is thus apparent that when the answer pattern matrix is unique the distribution of the number of persons passing each item, the items being arranged in order of difficulty, is the same as the cumulative frequency distribution of raw scores. The distribution of raw scores are therefore completely determined by the difficulty values of the test items.

Walker has devised an index to measure the amount of divergence of the answer pattern of any test from the unique answer pattern that would have obtained had the score x of every child been made by answering correctly the x easiest items on the test. Walker termed this index the 'coefficient of hig'. In a later article, however, he expressed some scepticism of its utility, and at present no convenient quantitative measure is available for estimating the divergence of an obtained answer pattern matrix from a theoretically unique matrix.

PROPERTIES OF ANSWER PATTERN MATRICES.

The above discussion has been presented preparatory to the development of certain associated theorems fundamental in the theory of test construction. These theorems permit more of demonstration than of rigorous proof.

THEOREM 1 Lack of uniqueness in the answer pattern matrix tends to reduce the variance of raw scores.

Consider the hypothetical answer pattern matrix of a test of 4 items given to a sample of 16 persons. Let $C_1, C_2, \ldots C_{16}$ refer to persons, and Q_1, Q_2, Q_3, Q_4, refer to items

Table 16

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	no. persons passing items
Q_1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	15
Q_2						1	1	1	1	1	1	1	1	1	1	1	11
Q_3												1	1	1	1	1	5
Q_4																1	1
raw score	0	1	1	1	1	2	2	2	2	2	2	3	3	3	3	4	32

Table 16

Each row in the above answer pattern matrix shows the number of persons passing each item. Each column shows the number of items passed by each person. Thus the sum of the elements in the column vector C_1 is the raw score of the ith. person. It will be observed that the distribution of raw scores is binomial, and that the frequencies of this distribution are equal to the first differences obtained from the distribution of the number of persons passing each item.

By interchanging any number of rows or any number of columns in the above answer pattern the uniqueness of the answer pattern remains unchanged. Interchanging columns amounts merely to rearranging individuals; interchanging rows amounts to rearranging the items in a different order of difficulty. Any rearrangement of the elements in the above answer pattern matrix which does not correspond to an interchanging of complete rows or columns will reduce the inter-item correlation. Thus if the element $a_{3.12}$ is moved to a position $a_{3.5}$, the inter-item correlation r_{23} will be reduced, and the variance of raw scores reduced from 1 to .875. By changing the position of the elements in any given row such that the answer pattern matrix ceases to be unique certain inter-item correlations, and covariances are reduced. A reduction in the sum of the inter-item covariances is, as previously established, accompanied by a reduction in the variance of the whole test. We must, therefore, conclude that lack of uniqueness in the answer pattern matrix tends to reduce the variance of raw scores.

<u>THEOREM 2</u>. Lack of uniqueness in the answer pattern matrix tends to reduce the reliability of the test. Conversely by increasing the degree to which the answer pattern approximates to a unique solution we tend to increase the reliability of the test.

The reliability of a test is a function, not only of the independent item reliabilities, but also of the inter-item covariances except in the theoretical case when the test is perfectly reliable. This statement is capable of adequate demonstration on reference to a pooling square containing the intercorrelations between all the n items on a test, and the n items on a hypothetical equivalent form of the test, as follows;

$$
\begin{array}{c|cccc|cccc|}
 & \sigma_1 & \sigma_2 & \cdots & \sigma_n & \sigma_1 & \sigma_2 & \cdots & \sigma_n \\
\hline
\sigma_1 & 1 & r_{12} & \cdots & r_{1n} & r_{11} & r_{12} & \cdots & r_{1n} \\
\sigma_2 & r_{12} & 1 & \cdots & r_{2n} & r_{12} & r_{22} & \cdots & r_{2n} \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
 & & & & r_{n(n-1)} & & & & r_{n(n-1)} \\
\sigma_n & & & r_{n(n-1)} & 1 & & & r_{n(n-1)} & r_{nn} \\
\hline
\sigma_1 & r_{11} & r_{12} & \cdots & r_{1n} & 1 & r_{12} & \cdots & r_{1n} \\
\sigma_2 & r_{12} & r_{22} & \cdots & r_{2n} & r_{12} & 1 & & \\
\vdots & \vdots & \vdots & & \vdots & \vdots & \vdots & & \vdots \\
 & & & & r_{n(n-1)} & & & & r_{n(n-1)} \\
\sigma_n & & & r_{n(n-1)} & r_{nn} & & & r_{n(n-1)} & 1 \\
\hline
\end{array}
$$

From this pooling square it is apparent that

$$r_{11} = \frac{\sum_{i=1}^{n} r_{ii} + \sum_{\substack{i=1\\ i\neq j}}^{n} \sum_{j=1}^{n} r_{ij}\sigma_i\sigma_j}{\sum_{i=1}^{n} \sigma_i^2 + \sum_{\substack{i=1\\ i\neq j}}^{n} \sum_{j=1}^{n} r_{ij}\sigma_i\sigma_j}$$

Examination of this equation indicates that when n is large the sum of the n(n-1) inter-item covariances greatly outweighs the other terms in the equation as determiners of r_{11}. Increasing the quantity $\sum_{\substack{i=1\\ i\neq j}}^{n} \sum_{j=1}^{n} r_{ij}\sigma_i\sigma_j$ independent of the other terms in the equation, without increasing the value of n, will increase the test reliability except in the special case where the test is perfectly reliable. The greater the value of n the more the sum of the inter-item covariances tends to outweigh the other elements. This explains analytically why the reliability of a test is increased by increasing its length. We have already demonstrated that the further the answer pattern of a test digresses from a unique solution the smaller the value of the summed inter-item covariances, the number of items being kept constant. The conclusion is, therefore, that the greater the lack of uniqueness in the answer pattern matrix the lower the reliability of the test. Conversely by increasing the degree to which the answer pattern approximates to a unique solution we increase the reliability of the test.

THEOREM 3. Lack of uniqueness in the answer pattern matrix tends to reduce the correlation of a test item with the whole test.

This proposition is capable of ready demonstration on reference to the formula for bi-serial r, or the corresponding formula for the Pearson product-moment r for the correlation between a dichotomously scored variable and a polytomously scored variable. The usual formula for bi-serial r is written as follows:

$$r_{bis.} = \frac{M_p - M_q}{S.D.} \cdot \frac{pq}{z}$$

where M_p = mean score on the whole test of persons solving the item correctly.

M_q = mean score on the whole test of persons failing the item.

S.D. = standard deviation of raw scores.

p = proportion of whole group passing the item.

q = proportion of whole group failing the item.

z = ordinate of the normal curve cutting off p proportion of cases.

The corresponding product-moment formula for the correlation between a dichotomous and a polytomous variable is written in the form

$$r = \frac{M_p - M_q}{S.D.} \sqrt{pq}$$

This formula is capable of ready derivation from the formula for the calculation of a correlation coefficient from raw scores on the assumption that one of the variables is dichotomously distributed.

Reference to any answer pattern will show that the quantity M_p is a maximum for any item of given difficulty when the x persons passing that item are the persons scoring the x highest marks on the test, or when the item vector of the answer pattern matrix is unique. Thus lack of uniqueness in the answer pattern matrix can decrease, but never increase the value of M_p. The converse holds for M_q. It follows, therefore, that $M_p - M_q$ is a maximum for an item of any given difficulty when the answer pattern matrix is unique. Hence we conclude that lack of uniqueness may decrease, but never increase, the correlation of an item with the whole test.

A Note on the Matrix of Inter-item Correlations Obtained from a Unique Answer Pattern Matrix.

The matrix of inter-item correlations obtained from a unique answer pattern matrix has certain interesting and unusual properties which are considered briefly here.

Consider a hypothetical test of n items arranged in ascending order of difficulty, and let the difficulty values (the proportion of persons passing each item) of the items be $p_1, p_2, p_3 \ldots\ldots\ldots p_n$. Since the answer pattern matrix is unique $p_1 > p_2 > p_3 \ldots\ldots\ldots p_n$, and $p_{12} = p_2$, $p_{13} = p_3 \ldots\ldots p_{(n-1)n} = p_n$. Therefore the inter-item covariances $p_{ij} - p_i p_j = p_j q_i$, where $p_i > p_j$. The matrix of inter-item covariances is then as follows:-

$$
\begin{array}{cccccccccc}
 & 1 & 2 & 3 & 4 & 5 & & & & n \\
1 & \underline{p_1 q_1} & p_2 q_1 & p_3 q_1 & p_4 q_1 & p_5 q_1 & \cdot & \cdot & \cdot & p_n q_1 \\
2 & p_2 q_1 & \underline{p_2 q_2} & p_3 q_2 & p_4 q_2 & p_5 q_2 & \cdot & \cdot & \cdot & p_n q_2 \\
3 & p_3 q_1 & p_3 q_2 & \underline{p_3 q_3} & p_4 q_3 & p_5 q_3 & \cdot & \cdot & \cdot & p_n q_3 \\
4 & p_4 q_1 & p_4 q_2 & p_4 q_3 & \underline{p_4 q_4} & p_5 q_4 & \cdot & \cdot & \cdot & p_n q_4 \\
5 & p_5 q_1 & p_5 q_2 & p_5 q_3 & p_5 q_4 & \underline{p_5 q_5} & \cdot & \cdot & \cdot & p_n q_5 \\
 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
n & p_n q_1 & p_n q_2 & p_n q_3 & p_n q_4 & p_n q_5 & \cdot & \cdot & \cdot & p_n q_n
\end{array}
$$

The item variances have been inserted in the diagonal.

Examination of this matrix of inter-item covariances indicates immediately that all the tetrad differences formed from elements all of which lie on one side of the diagonal are zero, while all tetrads formed from elements which lie on both sides of the diagonal are not zero.

By inserting the item variances in the principal diagonal all tetrads which include one diagonal element are zero. if they do not cross the diagonal Those which include two diagonal elements are of course not zero. (4)

The matrix of inter-item correlations is obtained by dividing each element in the covariance matrix by the standard deviation of the two items involved. The matrix of inter-item correlations obviously exhibits the same properties as the matrix of inter-item covariances.

Consider for clarity of illustration a numerical example. Let the following represent a unique answer pattern matrix of a test of 8 items administered to a sample of 20 persons.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}	C_{18}	C_{19}	C_{20}	p_i
Q_1		1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.95
Q_2				1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1	.85
Q_3							1	1	1	1	1	1	1	1	1	1	1	1	1	1	.70
Q_4									1	1	1	1	1	1	1	1	1	1	1	1	.60
Q_5													1	1	1	1	1	1	1	1	.40
Q_6																1	1	1	1	1	.25
Q_7																		1	1	1	.15
Q_8																			1	1	.10
		1	1	2	2	2	3	3	4	4	4	4	5	5	5	6	6	7	8	8	

The matrix of inter-item covariances obtained from the above answer pattern is as follows. The item variances have been inserted in the diagonal.

	1	2	3	4	5	6	7	8
1	.0475	.0425	.0350	.0300	.0200	.0125	.0075	.0050
2	.0425	.1275	.1050	.0900	.0600	.0375	.0225	.0150
3	.0350	.1050	.2100	.1800	.1200	.0750	.0450	.0300
4	.0300	.0900	.1800	.2400	.1600	.1000	.0600	.0400
5	.0200	.0600	.1200	.1600	.2400	.1500	.0900	.0600
6	.0125	.0375	.0750	.1000	.1500	.1875	.1125	.0750
7	.0075	.0225	.0450	.0600	.0900	.1125	.1275	.0850
8	.0050	.0150	.0300	.0400	.0600	.0750	.0850	.0900

The matrix of inter-item correlations is as follows:-

	1	2	3	4	5	6	7	8
1	1.0000	.5461	.3504	.2810	.1873	.1326	.0964	.0765
2	.5461	1.000	.6417	.5145	.3430	.2426	.1765	.1400
3	.3504	.6417	1.000	.8018	.5345	.3780	.2750	.2182
4	.2810	.5145	.8018	1.00	.6667	.4714	.3430	.2722
5	.1873	.3430	.5345	.6667	1.0000	.7071	.5145	.4082
6	.1325	.2425	.3780	.4714	.7071	1.0000	.7276	.5774
7	.0964	.1765	.2750	.3430	.5145	.7276	1.0000	.7935
8	.0765	.1400	.2182	.2722	.4082	.5774	.7935	1.0000

These two numerical matrices, the matrix of inter-item covariances and the matrix of inter-item correlations, reveal the unusual properties previously mentioned.

The properties which apply to the matrices of inter-item covariances and correlations formed from a unique answer pattern apply also to the matrices of inter-person covariances and correlations formed from such an answer pattern.

Whether these matrices of inter-item covariances and correlations can be described profitably in terms of factors, and what particular factorial configuration can best describe matrices of this type is not at the moment of writing immediately apparent.

One tentative factor pattern for 8 variables where $p_1 > p_2 > p_3 \ldots\ldots p_8$ is as follows:-

Tests	I	II	III	IV	V	VI	VII	VIII
1	x							
2	x	x						
3	x	x	x					
4	x	x	x	x				
5	x	x	x	x	x			
6	x	x	x	x	x	x		
7	x	x	x	x	x	x	x	
8	x	x	x	x	x	x	x	x

Factors or bonds (column heads above)

When the loadings in the above pattern are obtained from correlations resulting from a unique answer pattern, they maintain a constant ratio throughout the columns;

that is all possible tetrads that can be formed from the loadings in the above pattern are zero.

If such a factor pattern were psychological, meaningful it would imply that as the items increased in difficulty (the difficulty of an item being defined by the number of persons passing it) new mental factors are involved in the attainment of a correct response.

The whole question is closely linked with Professor Godfrey Thomson's sampling theory of ability. (see "The "Factorial Analysis of Human Ability" pp.267-284). With reference to our numerical example let us presume conditionally that the minds of the 20 members of our hypothetical sample of persons are comprised of innumerable bonds, and that the successful response to a particular item requires the formation of a certain number of such bonds. To answer item 1 correctly the formation of only one bond is required; to answer item two correctly requires the formation of the bond required to solve item one plus an additional bond and so on. Thus we may say that to solve item 1 is a relatively simple procedure requiring the formation of only a single bond, while to solve item 8 is a complex operation requiring the formation of 8 different bonds. It may be noted here that the term'bond' is used with all the limiting conditions imposed in Professor Thomson's discussion of the subject. The bond for instance, required to solve item 1 may be a complex of smaller bonds.

In the illustration given here we have made the assumption that our answer pattern matrix is unique, and heve consequently imposed a certain definite structure upon the minds of our 20 hypothetical persons. Furthermore we have imposed a certain definite structure upon our 8 hypothetical test items. In actual practice our answer pattern would not be unique but would only approximate to uniqueness in greater or less degree.

The answer pattern might be as follows:-

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	C_{17}	C_{18}	C_{19}	C_{20}	p_i
Q_1	1	1	1	1	1	1	1	1	.	1	1	1	1	1	1	1	1	1	1	1	.95
Q_2	1	1	1	1	1	.	.	.	1	1	1	1	1	1	1	1	1	1	1	1	.85
Q_3					1	1	1	1	.	1	1	1	1	1	.	1	1	1	1	1	.70
Q_4			1	1	.	.	.	.	1	1	1	1	1	1	.	.	.	1	1	1	.60
Q_5											1	1	1	1	1	.	.	1	1	1	.40
Q_6														1	1	.	.	1	1	1	.25
Q_7																	1	1	1	.	.15
Q_8																			1	1	.10

The configuration of bonds or factors derived from such a pattern would be very nearly as follows. The zeros would not be exactly zeros for certain mathematical reasons but they would be nearly zeros.

Items	Bonds or Factors							
	1	11	111	1V	V	V1	V11	V111
1	x							
2	.	x						
3	x	.	x					
4	.	x	.	x				
5	x	x	.	.	x			
6	x	x	.	.	x	x		
7	x	x	x	.	.	.	x	
8	x	x	x	x	x	x	.	x

NOTE. (The above pattern is not exact. Time has not permitted the working of an exact numerical example).

The argument, therefore, seems to indicate that lack of uniqueness in the answer pattern structure results in part at least from the way in which test items sample the bonds of the mind. Another source of lack of uniqueness results from the fact that different persons may employ different bonds in answering the same items correctly.

The fact that the elements in an answer pattern are not all inserted at random, but approximate in some degree or other to a unique configuration seems to indicate that the mind has a certain structure. As the answer pattern departs from uniqueness towards randomness the whole matrix of inter-item correlations is reduced in

rank. If the elements in the answer pattern were inserted purely at random all the inter-item correlations would tend to be zero, and would indicate that there was no linkage
3 between the innumerable elements or bonds of the minds of the persons tested.

If this structure which the mind seems to possess is in part imposed by education, and other environmental influences we would expect that the answer patterns of tests given to young children would depart more substantially from uniqueness than the answer pattern of tests given to older children. If this argument is correct we would expect the reliability of tests to increase with increase in age. Such a hypothesis is readily capable of experimental treatment.

The above discussion, written hurriedly under the pressure of much other work, must be regarded as purely tentative. The matter is at present undergoing further consideration.

OTHER FACTORS CONTRIBUTING TO ANSWER PATTERN UNIQUENESS.

Certain other factors contribute in some degree to lack of uniqueness in the answer pattern matrix, and thereby detract from the efficiency of the test as a selective instrument.

Firstly, if the xeasiest items on the test are not the first x items on the test, that is, if the items are not arranged in order of difficulty, it is less likely that a person who makes a score x will procure that score by answering correctly the x easiest items on the test. Thus the testee may waste time attempting items too difficult for him, and, if the test has a time limit, fail to reach items that he could readily do correctly. It is desirable, therefore, that the items on a test be arranged in order of difficulty, if the test is to attain a high degree of effectiveness.

Secondly, the use of items of the multiple-choice type will also tend towards lack of uniqueness in the answer pattern matrix. With items of this type there exists a probability that the testee will respond correctly by chance alone. The probability that an individual will respond correctly by chance alone is independent of the difficulty of the items, when the number of alternatives is constant. Thus an individual may make a score by chance on items that are beyond his level of ability. Such responses will be arranged in random manner in the answer pattern, and will tend to reduce the inter-item correlations. Hence by reducing the probability of making a certain score by chance we reduce the discrepancy between the obtained answer pattern matrix and the desired unique matrix.

In short, all purely random influences resulting from the interaction of test and testee which contribute to the unreliability of tests will increase the lack of uniqueness in the answer pattern matrix.

THE THEORY OF TEST DISCRIMINATION.

Every test item on which the persons tested may either pass or fail performs in itself a dichotomous function, namely that it divides the sample of persons tested into two groups; persons capable of passing the item, and persons incapable of passing the item. The level of ability at which the item is able to dichotomize the group, depends on the difficulty of the item. An item that divides the sample of persons into two equal categories may be described as discriminating about the mean. With two items of different difficulty the sample of persons tested would be divided into three ability categories. This statement is only true in the sense that if each item is scored one mark for a pass, and no marks for a failure, the total scores on the two items of the persons tested would be either 0, 1, or 2. If, however, we denote the two items as i and j, where item j is more difficult than item i, a person may fail both items, pass item i and fail item j, pass item j and fail item i, or pass both items. A pass on the more difficult item j does not necessarily imply a pass on the easier item i. For reasons previously discussed certain persons may find item j easier than item i, although item j may be more difficult than item i, where the term 'more difficult' is defined by the responses of the majority. Let us assume none the less for benefit of clarity at this point in our discussion that all persons passing item j also pass item i. Thus conditionally we may state that a test of two items of different degrees of difficulty will divide the persons tested into three ability categories, while a test of three items of different degrees of difficulty will divide the persons tested into four ability categories. The more items of different difficulty we add to our test the greater the number of categories into which the test is able to subdivide the group. Thus a test constructed of a large number of items of different degrees of difficulty, each item performing its own particular dichotomous function and discriminating at a particular level of ability, performs a polytomous function; that is, it divides the persons tested into a large number of categories, each category representing a different level of ability. Finally, having obtained items of varying difficulty, we reach a position where the items are maximally different from one anotherwith respect to difficulty. This position yields a rectangular distribution of raw scores, and will be discussed at greater length below.

The above discussion relates for clarity of illustration to the ideal situation where the answer pattern matrix is unique. In practice the discriminative power of an item is seriously blurred by lack of answer pattern uniqueness; that is, by the presence of group factors, and the action of numerous random influences. It is apparent, therefore, that when the answer pattern matrix is unique the test discriminates perfectly between the persons tested, and the more closely the answer pattern of a test can be made to approximate to this desired position the more efficient its discriminative power, and the greater its sensitivity in arranging the persons tested according to their measured capacity.

DISCRIMINATION AND THE CORRELATION BETWEEN PERSONS.

As mentioned previously we may calculate the correlations between the columns of the answer pattern matrix as well as the correlations between the rows. Thus, instead of correlating items we may correlate rows. As previously established the sum of all the inter-item covariances plus the item variances equals the variance of raw scores. Similarly the sum of the inter-person covariances plus the variances of the persons is equal to the variance of the distribution of the number of persons passing the items. As the variance of raw scores is increased the variance of the number of persons passing the items is decreased. Thus when a theoretical maximum variance is attained; that is, when half the persons tested make zero scores and the other half perfect scores; the number of persons passing each item is the same. Such a fictitious test discriminates perfectly about the mean, but does not discriminate perfectly between persons in the two broad categories. In this imaginary case all the correlations between items are perfect, while the correlations between persons are indeterminate.

As we reduce the variance of raw scores we increase the variance of the distribution of the number of persons passing each item. When in the theoretical case the raw scores form a rectangular distribution with standard deviation σ, the distribution of the number of persons passing the test items is also rectangular with standard deviation $n/N\sigma$, where n is the number of items, and N the number of persons. As we continue to decrease the variance of raw scores we increase the variance of the number of persons passing each item until an ultimate position is reached wher the variance of raw scores is zero, all persons making a score of n/2, and the variance of the distribution of persons passing the items is a maximum.

The conclusion resulting from the above argument is that by the selection of items which correlate highly among themselves we increase the variance of raw scores, and at the same time reduce the variance of the distribution of the number of persons passing each item; that is, we reduce the correlation between the persons tested, and make the persons tested appear more unlike one another. Thus, high test discriminative power involves high inter-item correlation, and low inter-person correlation. This observation furnishes an interesting addition to prevailing theories of test discrimination.

AN INDEX OF ITEM DISCRIMINATION.

Many existing techniques of item selection assist directly or indirectly in the elimination of lack of uniqueness in the answer pattern matrix. Among these are those techniques which require a division of the group tested into thirds or sixths. Methods of item selection which employ as a criterion the correlation of an item with the whole test are of no great value in the construction of tests of high discriminative power since the indices used are not independent of the difficulty values of the items.

As an index of the discriminative power of an item we propose to use the correlation of that item with a hypothetical item of corresponding difficulty, which is answered correctly by the x persons making the x highest scores on the whole test. Such an index furnishes an estimate of the accuracy with which a test item discriminates at the level of ability where it presumes to discriminate, and as such may be regarded as an indication of ~~the reliability~~ of the reliability of discrimination of a test item.

The correlation between two test items is gi[illegible] the formula

$$r_{ij} = \frac{p_{ij} - p_{i}p_{j}}{\sqrt{p_i q_i p_j q_j}}$$

Denoting our test item by i, and our hypothetical item of corresponding difficulty by a, and since $p_i = p_a$ we may write

$$r_{ia} = \frac{p_{ia} - p_i^2}{p_i q_i}$$

Since p_{ia} is equal to or less than p_i we may write $p_{ia} = p_i - w_i$. where w_i is the proportion of individuals failing item i who would have passed had the item discriminated perfectly, or the number of individuals passing item i who would have failed had the item discriminated perfectly. We may, therefore write our coefficient of item discrimination in the form

$$r_{ia} = 1 - \frac{w_i}{p_i q_i}$$

$1 - \frac{w_i}{p_i q_i}$

The coefficient r_{ia} varies as a correlation coefficient from -1 to 1. As an explanatory example consider the answer pattern of the following item i. Let $C_1, C_2 \ldots\ldots C_{10}$ refer to persons.

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}
item i			1	1	0	0	1	1	1	1
item a					1	1	1	1	1	1

In this example $p_i = .6$, and $p_i q_i = .24$, $w_i = .20$. Therefore, r_{ia}, the coefficient of item discrimination is .1668. We may say that such an item as this does not discriminate with sufficient accuracy at the level of ability where it presumes to discriminate.

In order to estimate values of r_{ia} exactly it is necessary to arrange the answer pattern in such a way that w_i is exactly determinable. This involves the construction of an answer pattern in which a column is assigned to each person, and a row to each item. With a test constructed of a large number of items, and given to a fairly large sample the construction of such an answer pattern is laborious. For the ordinary routine of item selection it is suffieient to divide the persons tested into six categories according to their scores on the whole test. From an answer patterm thus grouped w_i may be estimated by a process of interpolation. Values of r_{ia} calculated by this ready method should be sufficiently close approximations to serve as guiding parameters in the selection of test items of high discriminative power.

MEASURING LACK OF UNIQUENESS IN THE ANSWER PATTERN MATRIX.

The following embodies an attempt to measure the influence of lack of uniqueness in the answer pattern matrix upon the functioning of the whole test.

From the first differences of the distribution of persons passing each item we can obtain the actual scores that the persons tested would have made had the x persons passing each item been the persons making the x highest scores on the whole test. These scores we shall call for convenience D-scores. D-scores exhibit a number of interesting properties. The mean of the D-scores of the persons tested is the same as the mean of raw scores. The D-score of a person below the mean is always less than his raw score; the D-score of a person above the mean is always greater than his raw score. The discrepancy between D-score and raw score is due to the influence of lack of uniqueness in the answer pattern matrix. The further the answer pattern of a test digresses from a unique position the greater these discrepancies. We see, therefore, that lack of uniqueness tends to make the raw scores of the (5) persons tested regress towards the average, while approximating to a unique position pulls the scores apart, and increases the discriminative power of the test. This agrees with the previously established theorem that lack of uniqueness in the answer pattern matrix reduces the variance of raw scores.

The variance of D-scores is consequently always substantially greater than the variance of raw scores. With Moray House Tests the standard deviation of raw scores is about 20, while the standard deviation of the corresponding D-scores is about 30. It should be pointed out here that if the answer pattern matrix had, in the first instance, been unique, the variance of raw scores would not be 30, but it would be somewhere between 20 and 30, possibly about 25. It had been our original intention to use the correlation between raw scores and D-scores as a measure of lack of uniqueness, but in actual experiment the regression lines of the correlation table were found to exhibit a certain non linearity. The correlation between D-scores and raw scores of a random sample of 162 persons on M.H.T.26, disregarding the non-linearity of regression, was found to be .9789

A better indication of the amount of divergence of the obtained answer pattern matrix from the hypothetical unique matrix is given by the ratio of the variance of raw scores to the variance of D-scores. With M.H.T.26 this index was found to be .406. The less the divergence of the obtained matrix from the unique position the more closely does this ratio approximate to unity.

Figure 1 gives the distribution of raw scores of 162 persons on M.H.T.26, and the corresponding distribution of D-scores. The standard deviation of raw scores was found

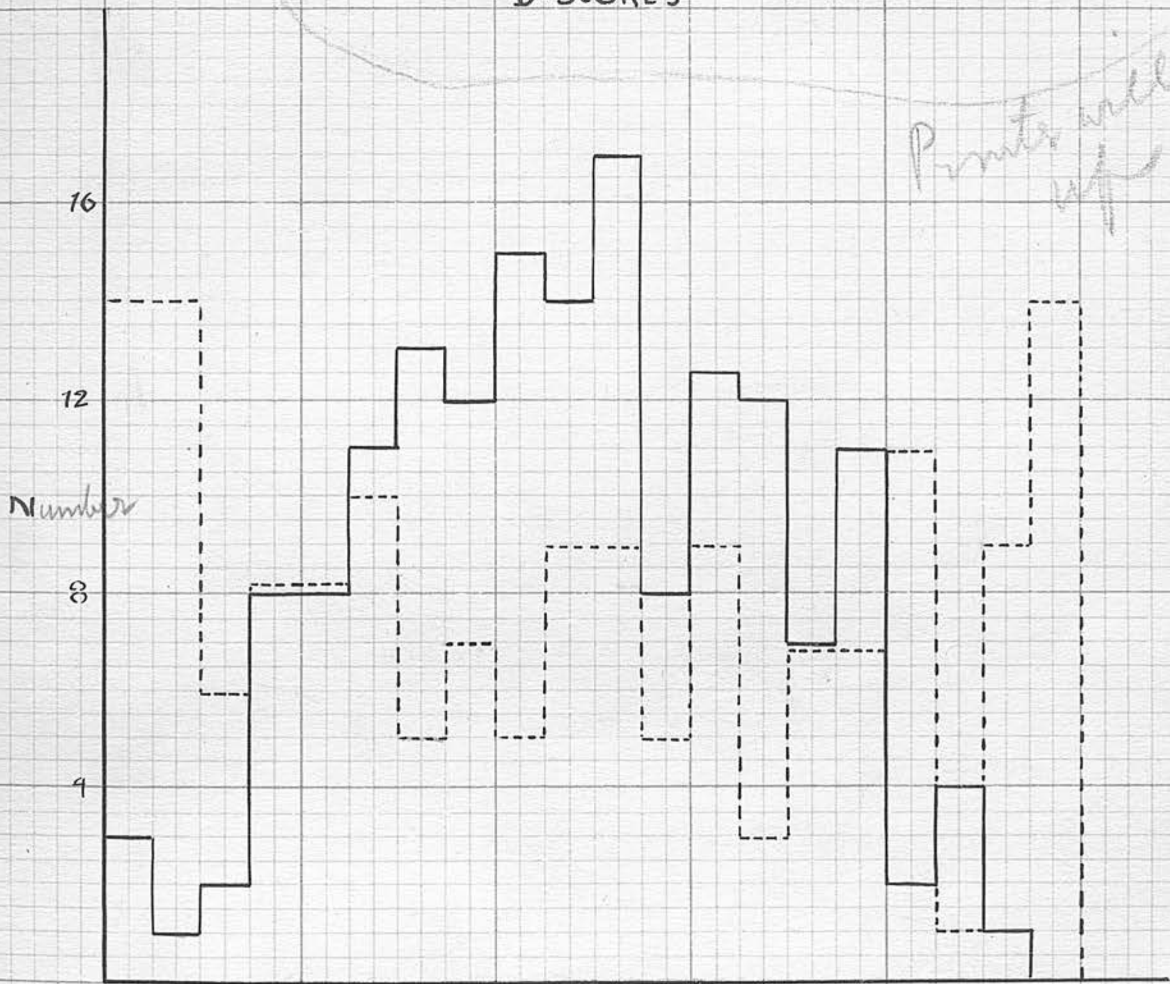
Fig 3

A COMPARISON BETWEEN A DISTRIBUTION OF RAW SCORES, AND CORRESPONDING D-SCORES.

RAW SCORES ————

D-SCORES - - - - - - -

Printer will set up



	0	5	10	15	20	25	30	40	45	50	55	60	60	65	70	75	80	85	90	95
RAW SCORE	3	1	2	8	8	11	13	12	15	14	17	8	13	12	7	11	2	4	1	0
D-SCORES	14	14	6	8	8	10	5	7	5	9	9	5	9	3	7	7	11	2	9	14

to be 20.06, and the standard deviation of D-scores 31.50. Examination of this figure indicates clearly the influence of lack of uniqueness in the answer pattern on the test structure, showing how such lack of uniqueness makes the scores of the individuals tested regress towards the average.

PLATYKURTIC DISTRIBUTION OF RAW SCORES.

In the argument developed above we have attempted to demonstrate that variance of raw scores, reliability, and discriminative power are functions of the item variances and covariances. These item variances and covariances are in themselves limited in magnititude by the type of distribution of raw scores which the test constructor predetermines, since by the appropriate selection of items many different types of distributions may be obtained. The belief has generally dominated educational measurement that some intrinsic desirability characterised normally distributed raw scores, and that various types of skewed, leptokurtic, and platykurtic distributions were to some degree at least less satisfactory than normal distributions. Adherence to distributions of the normal type has resulted, firstly, from the belief that ability is normally distributed in the population, and, secondly, because many statistical parameters are computed with greater facility, and are more intelligible, when the distributions of scores used in their computation are approximately normal. A belief, sometimes held and obviously false, is that correlation coefficients calculated by the product-moment method are invalid unless the correlated variables are normally distributed. This is not a necessary condition for the valid use of the product-moment formula, but linearity of regression, and variables distributed in a variety of ways other than normal may, when correlated yield regression lines which exhibit such linearity.

The purpose of the present discussion is to demonstrate that, since the item variances and covariances may be increased by the selection of items yielding types of distributions other than normal, the efficacy of tests as reliable, discriminative instruments for the selection of individuals for occupational and scholastic purposes may be substantially improved by the adoption of platykurtic and rectangular distributions.

The reasoned argument supporting this statement is as follows. By increasing the platykurtosis of a distribution we increase the variance of raw scores without increasing the number of items. This increased variance is accompanied, either causually or effectually, by increased inter-item covariance. This increased inter-item covariance, as previously established increases the reliability of the test. Furthermore, by increasing the platykurtosis and thereby increasing the variance of raw scores we reduce the correlation between the persons tested, making them appear more different from one another, thereby increasing the discriminative powers of the test.

The above discussion may be clarified with reference to the following fictitious example. Consider a test constructed of four test items of such a type as to yield a binomial distribution of raw scores when administered to a population of 16 persons. Let the answer pattern be as shown in Table 16 page 6, where $C_1, C_2, C_3, \ldots\ldots\ldots\ldots C_{16}$ refer to persons, and Q_1, Q_2, Q_3, Q_4 refer to items. We assume for the sake of simplicity that the answer pattern matrix is unique. The argument, however, is quite general.

The variances, covariances, and intercorrelations of the four items are as follows:-

covariances

	1	2	3	4
1.	.0586			
2.	.0430	.2148		
3.	.0195	.0977	.2148	
4.	.0039	.0195	.0430	.0586

inter-correlations

	1	2	3	4
1	---			
2	.3830	---		
3	.1741	.4547	---	
4	.0666	.1741	.3830	---

The item variances are written in the diagonal of the matrix of covariances. The variance of raw scores on this fictitious test is 1, while the variance of the distribution of the number of persons passing each item is 29.00.

Let us now consider the answer pattern of the type shown in Table 17, derived from a test constructed of four items administered to a sample of 16 persons. The distribution of raw scores is not binomial but platykurtic.

Table 17

	C_1	C_2	C_3	C_4	C_5	C_6	C_7	C_8	C_9	C_{10}	C_{11}	C_{12}	C_{13}	C_{14}	C_{15}	C_{16}	No. persons passing item
Q_1				1	1	1	1	1	1	1	1	1	1	1	1	1	13
Q_2							1	1	1	1	1	1	1	1	1	1	10
Q_3											1	1	1	1	1	1	6
Q_4														1	1	1	3
	0	0	0	1	1	1	2	2	2	2	3	3	3	4	4	4	(raw scores)

The variances, covariances, and intercorrelations of the four items of this fictitious test are as follows:-

covariances	1	2	3	4		1	2	3	4
1	.1523				1	--			
2	.1172	.2344			2	.6204	--		
3	.0703	.1406	.2344		3	.3722	.5998	--	
4	.0352	.0703	.1172	.1523	4	.2309	.3722	.6204	--

The variance of raw scores on this fictitious test is 1.8750. It will be observed that by increasing the platykurtosis of our distribution of raw scores we have increased the variance from 1 to 1.8750. The inter-item covariances and also the intercorrelations have been increased substantially. Furthermore, the variance of the distribution of the number of persons passing each item has been decreased from 29 to 14.5. This represents a very marked decrease in the magnititude of the inter-person covariances, and indicates that the test yielding the platykurtic distribution of scores is discriminating more effectively between persons than the test yielding the binomial distribution.

The split-half reliabilities of these two small hypothetical tests is also readily calculated. The 'boosted' split-half reliability of the test yielding the binomial distribution of raw scores was found to be .5625. The corresponding figure for the test yielding the platykurtic distribution was found to be .6750.

This simple hypothetical example demonstrates, therefore, that increasing the platykurtosis of the distribution of scores (a) increases the inter-item covariances, (b) increases the inter-item covariances, (c) increases the variance of raw scores, (d) increases the reliability of the test, (e) reduces the correlation between persons, (f) increases the discriminative power of the test, and from all points of view improves the efficacy of the test as an instrument of measurement.

TYPES OF DISTRIBUTIONS OF RAW SCORES.

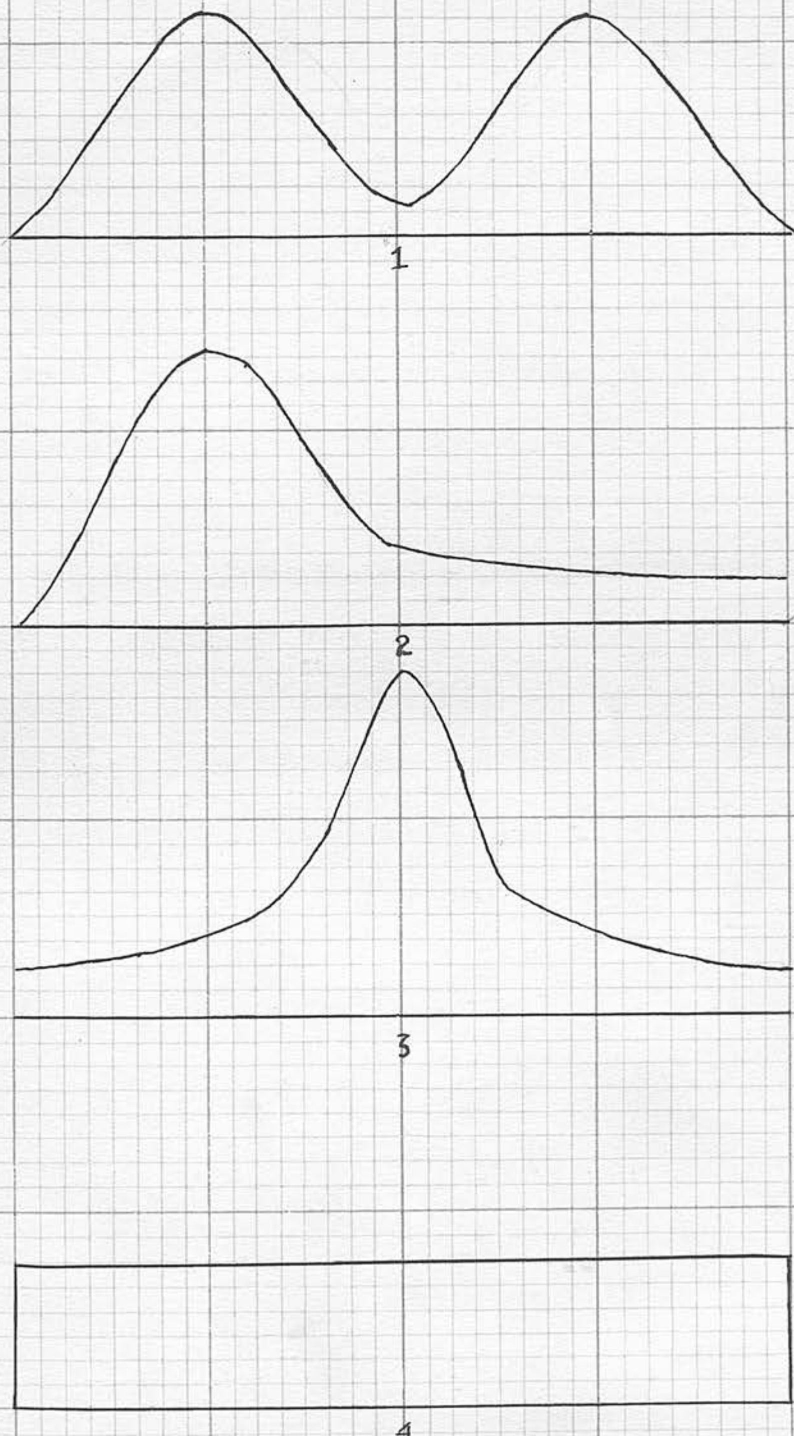
As indicated above by the selection of appropriate items the distribution of raw scores may be predetermined by the test constructor. We may, therefore, consider what type of distribution of the many possible types will produce the most efficient results in the field of mental testing. The answer to this problem is that the type of distribution which is selected must depend on the ultimate function which the test is intended to accomplish. Thus if we are selecting candidates for secondary schools, and wish the test to discriminate with a high degree of accuracy between the lower two thirds and the upper one third of the persons tested, items should be selected yielding a distribution of raw scores which is different in type from a distribution which would discriminate well between the lower one third and the upper two thirds. Distributions may be determined which will accomplish their respective functions more efficaciously than had the test been designed to yield a distribution of raw scores approximating to normality in a representative population. Similarly if a test is desired for the general purpose of discriminating at all levels of ability a particular distribution, namely rectangular, may be obtained which will accomplish this function with maximal efficiency.

The theory developed here depends on two generalizations; (a) the shorter the ordinate of the curve of the distribution at the point of selection the greater the discriminatory power of the test at that point, (b) the discriminatory power of a test may be increased at one level of ability at the expense of discriminatory power at other levels of ability.

It is theoretically possible to construct a test such that half the persons tested make zero scores and the other half make perfect scores. Such a test would have maximum inter-item correlation, and every item would have maximum variance of .25. The variance of raw scores would also be a maximum. A test of this theoretical type would discriminate perfectly about the mean, but would have no capacity for discriminating between the persons in each category. If we were to attempt to construct a test of this type we should find that due to lack of uniqueness in the answer pattern matrix the scores of the persons tested could not be made to fall into two main categories, but would be approximately symmetrical and bi-modal with the minimal ordinate between the two modes at the mean. Similarly if we wished to discriminate well at some other level of ability a test could be constructed yielding an asymmetrical bi-modal distribution with the minimal ordinate between the two modes at the point of selection.

A situation may arise, and does arise in the selection

FIG. 4



of candidates for certain types of secondary education, where we wish to select a certain proportion of individuals from a given population, and to discriminate between the relative abilities of the individuals selected. Let us presume that we wish to select the upper third of the candidates, and to discriminate between them. The test to accomplish this function should be constructed of items of such a nature that theoretically two thirds of the persons tested fail all items while the remaining third are distributed equally throughout the whole range of items. Thus with a test of 100 items administered to a group of 3000 candidates from which we wish to select a 1000 the ideal test would be one upon which 2000 persons scores zero marks, and the remaining 1000 persons scored marks ranging from 1 to 100 with ten persons in each of the 100 categories. In practice this ideal situation can never be attained but may be roughly approximated to be a positively skewed distribution of the form shown in Figure ~~11~~ 4 diagram 2. By constructing the test such that the scores pile up at the lower and average ranges of ability, and are spread out at the upper ranges of ability we increase the power of the test to discriminate bright candidates while decreasing its power to discriminate between average and dull candidates. Thus poor discriminative power at certain levels of ability is compensated for by increased discriminative power at other levels of ability. Similarly if a test is desired to discriminate efficiently between the relative abilities of a certain proportion of dull children items may be selected which will yield a negatively skewed distribution of raw scores.

A situation may arise where a test is required which will select a given proportion of bright persons and a given proportion of dull persons, and will discriminate between the relative abilities of persons arbitrarily described as dull and between the relative abilities of persons arbitrarily described as bright. Let us presume that we wish to select the upper third and lower third of persons in a given population and that we wish to discriminate with maximal efficiency between persons in the upper third, and also between persons in the lower third. We are not concerned with discriminating between persons in the middle third. It follows that we can increase discrimination in the upper third and in the lower third at the expense of discrimination in the middle third. To accomplish the purpose desired items must be selected which will yield a distribution of scores which is unimodal, symmetrical, and markedly leptokurtic, tailing of on both sides in the manner suggested in Figure ~~11~~ 4, diagram 3.

RECTANGULAR DISTRIBUTIONS.

If now a test is desired for general experimental purposes, that is if our interest in the persons at one level of ability is no greater than our interest in persons at other levels of ability, we requier a test which will discriminate with equal efficiency at all levels of ability. The discriminative power of a test attains this unpreferential uniformity when the distribution of raw scores is rectangular, or when the height of the ordinates of the distribution are the same at all levels of ability. All types of distributions other than rectangular sacrifice discriminative power at one level of ability for increased discriminative power at other levels of ability.

With a rectangular distribution every observation has an equal probability of being anywhere in the range from zero to n, where n is the number of items on the test. The standard deviation of scores on a test of this type is given in the theoretical case by the formula $n/\sqrt{12}$, there being $n+1$ possible categories into which the scores may fall. With a test of 100 items the standard deviation of scores is 28.86, while the standard deviation of scores of a corresponding test yielding a normal distribution of scores is usually about 17.

The values of B_1 and B_2 for a rectangular distribution, calculated from the first four moments, are respectively 0 and 1.8. Values of $B_1 = 0$ indicate that the distribution is symmetrical. Values of $B_1 < 1.8$ indicate that the distribution is tending to become bi-modal, while values of $B_2 > 1.8$ indicate that there is a tendency for the scores to be concentrated near the centre of the scale.

665

THE RELIABILITY OF MENTAL TESTS.

PART II.

Part II is largely experimental, and involves a detailed study of the reliability of Moray House Tests of Intelligence, English, and Arithmetic.

THE RELIABILITY OF MENTAL TESTS.

(1)

Object of Investigation.

The investigation presented below was undertaken to determine (a) the reliability of certain group tests of intelligence, (b) whether group tests of intelligence were more consistent instruments of measurement than individual tests. The tests considered in the present enquiry are given to some 150,000 children annually in the schools of Britain for the purpose of selecting candidates for certain types of secondary school education; consequently the question of their reliability is a matter of no little importance.

Data Used.

The data used in the present investigation were acquired in an experiment designed to determine the relative effectiveness of two types of examinations in selecting children for secondary school education. This experiment was conducted in West Yorkshire under the auspices of the National Union of Teachers, while the statistical work involved was carried out by Professor G.H. Thomson, and W.G. Emmett at Moray House Teachers' Training College. The West Yorkshire Experiment included the administration of three Moray House Intelligence Tests, M.H.T.21, M.H.T.23, and

and M.H.T.26 to the same group of roughly 1800 children. The statistical data resulting from the application of three group tests of intelligence to the same sample furnished comprehensive material for an investigation into the reliability of such tests.

The Group Tested.

All the children in 39 schools in West Yorkshire be:tween the ages 10:0 and 10:11 on March 1st 1937 were given the tests. One school did not complete the experiment, while about 200 children in the other schools did not do all three intelligence tests, thus reducing the number of cases included in the final statistical analysis to 1535.

Administration of the Tests.

To eliminate as far as possible the effect of practice on the standardisation the schools were divided into two groups, designated Group A and Group B. The number of children in Group A was approximately 1,020, and Group B approximately 720. The tests were administered in the following order:-

Group A. Schools

March 2nd. 1937--

Intelligence Test M.H.T.21

March 9th. 1937--

Intelligence Test M.H.T.23.

March 16th. 1937--

Intelligence Test M.H.T.26.

Group B. Schools.

March 2nd. 1937--

Intelligence Test M.H.T.23.

March 9th. 1937--

Intelligence Test M.H.T.21.

March 16th. 1937--

Intelligence Test M.H.T.26.

Each test consisted of 100 items, and the time of administration was 45minutes. The procedure of administering two tests to Group A, and administering the same two tests in reverse order to Group B, while tending to eliminate any mean increase in I.Q. due to practice when both groups are considered together, exerts an influence on the intercorrelations between the tests. This problem is discussed at greater length in the section on practice effect.

Standardisation.

The standardisations of the three tests were effected in the usual manner by finding the scores at the

the 5th., 16th., 50th., 84th., and 95th. percentile levels for each month of birth separately, plotting these scores against the ages, and fitting a least square line to the twelve points thus found. A standardised score of 100 is given to the child whose score is equal to the average score of all the children in his age group. The standard deviation of standardised scores is taken as 15 in all Moray House Tests. The slope of each least square line determines the increment of raw score for increase in age at each percentile level.

Standardised scores correspond very closely to I.Q's and in this enquiry are regarded as such.

The standardisation was based only on those children taking all three tests, 1586 in number. A table of norms was prepared for each test, and three Intelligence Quotients found for each child, these quotients being calculated to the nearest half point.

The distributions of raw scores (with frequencies expressed as percentages) mean scores, and standard deviations are given in Table 1.

TABLE 1.

Distribution of raw scores, Mean Score, and Standard Deviation for M.H.T. 21, 23, and 26.

Score Interval	M.H.T. 21	M.H.T. 23	M.H.T. 26
90-99	0.8	0.1	0.3
80-89-	3.6	2.6	3.8
70-79	8.7	9.2	9.3
60-69	12.2	14.3	14.9
50-59	16.2	17.0	17.7
40-49	14.5	17.6	20.1
30-39	14.7	14.7	14.3
20-29	12.0	12.1	9.9
10-19	9.5	7.3	6.1
0-9	7.8	5.1	3.6
Mean Score	43.15	44.76	47.06
Standard Deviation	22.16	20.29	19.74

<u>Note</u>--- The frequencies are expressed as percentages.

Analysis of Data.

Three group Intelligence Quotients for some 1800 children of a single age range, calculated by the application of three group tests of similar type with a constant time interval of one week, furnished data of a sufficiently comprehensive nature to warrant a detailed enquiry into the reliability of the tests used, and the associated topic, the constancy of the Intelligence Quotient.

In analysing the data in the present investigation the general technique was to calculate the variation in I.Q. between the three sets of Intelligence Quotients for each child separately. Thus three distributions of variations in I.Q. were obtained. These variations were then sub-classified according to brightness. Groups A and B were considered separately and combined. The standard deviations of variation in I.Q. were calculated for groups A and B, for sub-groups of Groups A and B, and also for the two Groups combined. From these standard deviations reliability coefficients and standard errors of I.Q. were obtained. The method by which these parameters are obtained from the standard deviations of variations in I.Q. will be discussed later.

Parallel Forms.

Any enquiry into test reliability by the correlation of parallel forms necessitates some assurance as to the strict equivalence of the forms used. Otherwise the presence of a specific factor will tend to reduce the size of the correlation between the forms, and such correlations cannot be regarded as valid reliability coefficients.

In this enquiry M.H.T. 21, 23, and 26 are regarded as parallel forms of the same test, and no reason exists to doubt the validity of this assumption. The items on each test are similar in type, namely analogies, number series etc. The number of items on each test (100), and the duration of each test (45minutes) are the same. The standardisations are based on exactly the same sample of the population. The high reliability coefficients found,also lend weight to the assumption that the three tests approximate very closely to equivalence. The equivalence of the test forms used is considered at greater length in this thesis.

Practice Effect.

When two parallel forms of a test are given to the same group of children, the scores on the second form will usually tend to be higher than the scores on the first form due to practice effect, and familiarity with the test situation. If the effect of parctice is uniform at all levels of ability, that is if the dull child tends to increase his score through practice as much as the bright child, the practice effect will have no influence on the reliability coefficient. If, however, the bright child gains more through practice than the dull child, or if the dull child gains more through practice than the bright child, the correlation between the two sets of scores will be spuriously increased by some small amount.

If the children in the upper ranges of ability gain more through practice than the children in the lower ranges of ability, the standard deviation of the second set of scores will tend to be higher than the standard deviation of the first. If the children in the lower ranges gain more through practice than the children in the upper ranges the standard deviation of the second set of scores will be less than the standard deviation of the first.

Allan G. Rodger[x] in a study based on only 76 cases reports that the increase in I.Q. due to practice effect varies directly according to brightness, the increase from test to retest being about one half point of I.Q. for children of I.Q. 80, one point of I.Q. for children of I.Q. 100 and one and a half points of I.Q. for children of I.Q. 120. W.G. Emmett in an unpublished enquiry, by converting the raw scores obtained on the three Moray House Tests used in the West Yorkshire Experiment into I.Q's, using norms based on the performance of children in another area, found that there existed no apparent systematic relationship between practice effect and level of ability. This finding is in direct disagreement with the finding of Rodger. Until more decisive evidence is forthcoming we must regard the problem of practice effect relative to ability as undetermined.

As previously explained, an effort was made in the West Yorkshire Experiment to eliminate the possible influence of practice on the test standardisation by dividing the schools tested into two Groups, Group A and Group B, administering M.H.T.21 to Group A schools and M.H.T.23 to Group B schools on the first day of testing, and reversing the procedure on the second day of testing.

x Rodgers, Allan G., (1936) "The Application of six Group Intelligence Tests to the Same Children, and the Effects of Practice", B.J.E.P. vol.vi, 291-305.

M.H.T.26 was administered to the two groups on the third day of testing, the pupils in both groups having the same amount of practice.

The plan of the experiment eliminates, therefore, any mean change in I.Q. from one test to another when Groups A and B are considered together. Unfortunately the procedure outlined above tends to introduce certain possible sources of error;

(a) The standard deviations of raw score for the two groups taken separately will be slightly less than for the combined groups.

(b) The correlation between the tests will be slightly greater for the two groups taken separetely than for the combined groups.

(c) The standard deviation of variation in I.Q. will tend to be slightly smaller when the two groups are taken separately than when the two groups are combined.

These conditions imply that the reliability coefficients found by correlating I.Q.'s on M.H.T.21 and M.H.T.23 for Groups A and B separately will be slightly higher than when both groups are combined; similarly but to a less degree with M.H.T.21 and M.H.T.26, and with M.H.T.23 and M.H.T.26. This was indeed found to be the case as an examination of the reliability coeffecients of the three tests for Groups A and B taken separately, and for the combined groups indicates. (see Table 11).

Furthermore, if our two tests are strictly equivalent we should expect the correlation between M.H.T.21 and M.H.T 26, and also the correlation between M.H.T.23 and M.H.T.26, for the whole group, to be slightly higher than the correlation between M.H.T.21 and M.H.T.23. This was indeed found to be the case.

Since the technique of the experiment was such as to introduce the difficulties discussed above, the standard deviations of variations in I.Q. and reliability coefficients were calculated for Groups A and B separately at different levels of ability. This procedure was justified since no systematic relationship was found between practice effect and level of ability was found in this data. The standard deviations and variations in I.Q. were also calculated for the two groups combined. The standard deviations of variation in I.Q. for the combined group will be overestimates, the reliability coefficients underestimates.

With reference to the parameters computed from the combined groups, it may be observed that those computed on the variation in I.Q. between any two tests will be consistent with one another and strictly comparable. The parameters computed on the variations between M.H.T.21 and M.H.T.26, and between M.H.T.23 and M.H.T.26, for the

the combined groups, are strictly comparable with one another, but not with those computed on variations between M.H.T.21 and M.H.T.23, the standard deviations of variation in I.Q. in the latter case being greater overestimates than the standard deviations in the former.

TABLE 2.

Table of Correlations

Tests correlated	correlation	Standard error
Group A.		
M.H.T.21/23	.933	.0043
M.H.T.21/26	.922	.0050
M.H.T.23/26	.940	.0039
Group B		
M.H.T.21/23	.931	.0053
M.H.T.21/26	.940	.0049
M.H.T.23/26	.935	.0050
Combined Groups		
M.H.T.21/23	.917	.0026.
M.H.T.21/23	.924	.0024
M.H.T.23/26	.937	.0022

Correlation of I.Q.

The correlation coefficients given in Table 11, found by correlating I.Q.'s may be regarded as reliability coefficients. Correlation coefficients found by correlating I.Q.'s may be regarded as more valid indices of reliability than coefficients calculated by correlating raw scores. The correlation of raw scores will yield a coefficient that is too high due to the influence of age, and such a coefficient cannot be regarded as a valid index of reliability until age has been partialled out. If a test has been effectively standardised the correlation of raw score with age partialled out will be the same as the correlation of I.Q.

With a single year group the correlation of raw score with age is very small. The correlation of raw scores between two parallel forms of a test will be approximately .002 higher than the correlation of the corresponding I.Q.'s A close estimate of the correlation of raw scores on M.H.T.21, 23, and 26 can be reached by the addition of .002 to the correlation of I.Q.'s given in Table 11.

The Normality of Distributions of Variations in I.Q.

Examination of the distributions given in Tables 3, 4, and 5 suggests that variations in I.Q. from test to retest are normally distributed.

Pearsons formulae for β_1 and β_2 were used to test the normality of some of these distributions.

These formulae with Sheppard's corrections are as

THE NORMALITY OF DISTRIBUTIONS OF VARIATIONS IN I.Q.

$$\mu_1 = 0$$

$$\mu_2 = \nu_2 - \nu_1^2 - \frac{1}{12}$$

$$\mu_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3$$

$$\mu_4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4 - \frac{1}{2}(\nu_2 - \nu_1^2) + \frac{7}{12}$$

where μ_1, μ_2, μ_3, and μ_4 are the first, second, third and fourth moments about the true mean, and ν_1, ν_2, ν_3, and ν_4 are the corresponding moments about an arbitrary point.

From the above formulae β_1 and β_2 may be computed as follows:-

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3}, \quad \beta_2 = \frac{\mu_4}{\mu_2^2}$$

When β_1 is equal to zero the distribution is symmetrical; when β_2 is less than 3 the distribution is platykurtic; when β_2 is greater than 3 the distribution is leptokurtic.

The Normality of Distributions of Variations in I.Q.

Examination of the distributions given in Tables 3, 4, and 5 suggests that variations in I.Q. from test to retest are normally distributed.

Pearsons formulae for β_1 and β_2 were used to test the normality of some of these distributions.

These formulae with Sheppard's corrections are as follows:-

$$\mu_1 = 0$$
$$\mu_2 = \nu_2 - \nu_1^2 - \frac{1}{12}$$
$$\mu_3 = \nu_3 - 3\nu_2\nu_1 + 2\nu_1^3$$
$$\mu_4 = \nu_4 - 4\nu_3\nu_1 + 6\nu_2\nu_1^2 - 3\nu_1^4 - \tfrac{1}{2}(\nu_2 - \nu_1^2) + \frac{7}{12}$$

where μ_1, μ_2, μ_3, and μ_4 are the first, second, third and fourth moments about the true mean, and ν_1, ν_2, ν_3, and ν_4 are the corresponding moments about an arbitrary point.

From the above formulae β_1 and β_2 may be computed as follows:-

$$\beta_1 = \frac{\mu_3^2}{\mu_2} \quad , \quad \beta_2 = \frac{\mu_4}{\mu_2^2}$$

When β_1 is equal to zero the distribution is symmetrical; when β_2 is less than 3 the distribution is platykurtic; when β_2 is greater than 3 the distribution is leptokurtic.

In the present enquiry values of β_1 and β_2 were computed for distributions of variations in I.Q. for Groups A and B combined. These values of β_1 and β_2 are as follows:-

	β_1	β_2	N
M.H.T.21/23	.00100	2.9502	1535
M.H.T.21/26	.00360	3.0303	1535
M.H.T.23/26	.00003	3.1316	1535

$\sqrt{\beta_1}$ has a standard error of $\sqrt{\frac{6}{N}}$ for samples of N in a normally distributed population. $\beta_2 - 3$ has a standard of $\sqrt{\frac{24}{N}}$. The following Table gives values of $\sqrt{\beta_1}$, $\beta_2 - 3$, $\sigma_{\sqrt{\beta_1}}$, and $\sigma_{(\beta_2 - 3)}$.

	$\sqrt{\beta_1}$	$\beta_2 - 3$	$\sigma_{\sqrt{\beta_1}}$	$\sigma_{(\beta_2 - 3)}$
M.H.T.21/23	.0316	.0498	.0624	.1249
M.H.T.21/26	.0600	.0303	.0624	.1249
M.H.T.23/26	.0055	.1316	.0624	.1249

In no case does the distributions of I.Q. variations exhibit any significant skewness, or either leptokurtic or platykurtic tendencies. We, therefore, conclude that the normal probability curve describes with a high degree of accurracy variations in I.Q. between successive applications of Moray House Group Tests of Intelligence, and that no systematic factor is operating in causing these variations.

Variations due to any inadequacy of the tests as instruments of mental measurement, and variations due to fluctuation in the capacities tested both seem to be normally distributed in a normal population. The above computations indicate also that errors made in the measurement of cognitive abilities in the field of psychometrics obey the normal curve of errors as used in the physical sciences.

TABLE 3 WEST YORKSHIRE

DISTRIBUTIONS OF DIFFERENCES IN I.Q.

GROUP A

I.Q. diff.	M.H.T.21/23	M.H.T.21/26	M.H.T.23/26
19.5			1
18.0			0
16.5	2	2	2
15.0	1	1	2
13.5	2	3	6
12.0	7	7	4
10.5	5	8	10
9.0	10	14	16
7.5	25	31	42
6.0	27	43	45
4.5	39	44	76
3.0	67	58	117
1.5	85	83	104
0	105	100	93
-1.5	100	89	92
-3.0	102	98	91
-4.5	77	77	71
-6.0	71	67	45
-7.5	63	51	35
-9.0	42	38	22
-10.5	33	35	16
-12.0	21	24	5
-13.5	12	17	7
-15.0	5	9	4
-16.5	2	3	
-18.0	1	2	
-19.5	2	2	
	906	906	906

TABLE 4 WEST YORKSHIRE

DISTRIBUTION OF DIFFERENCES IN I.Q.

GROUP B

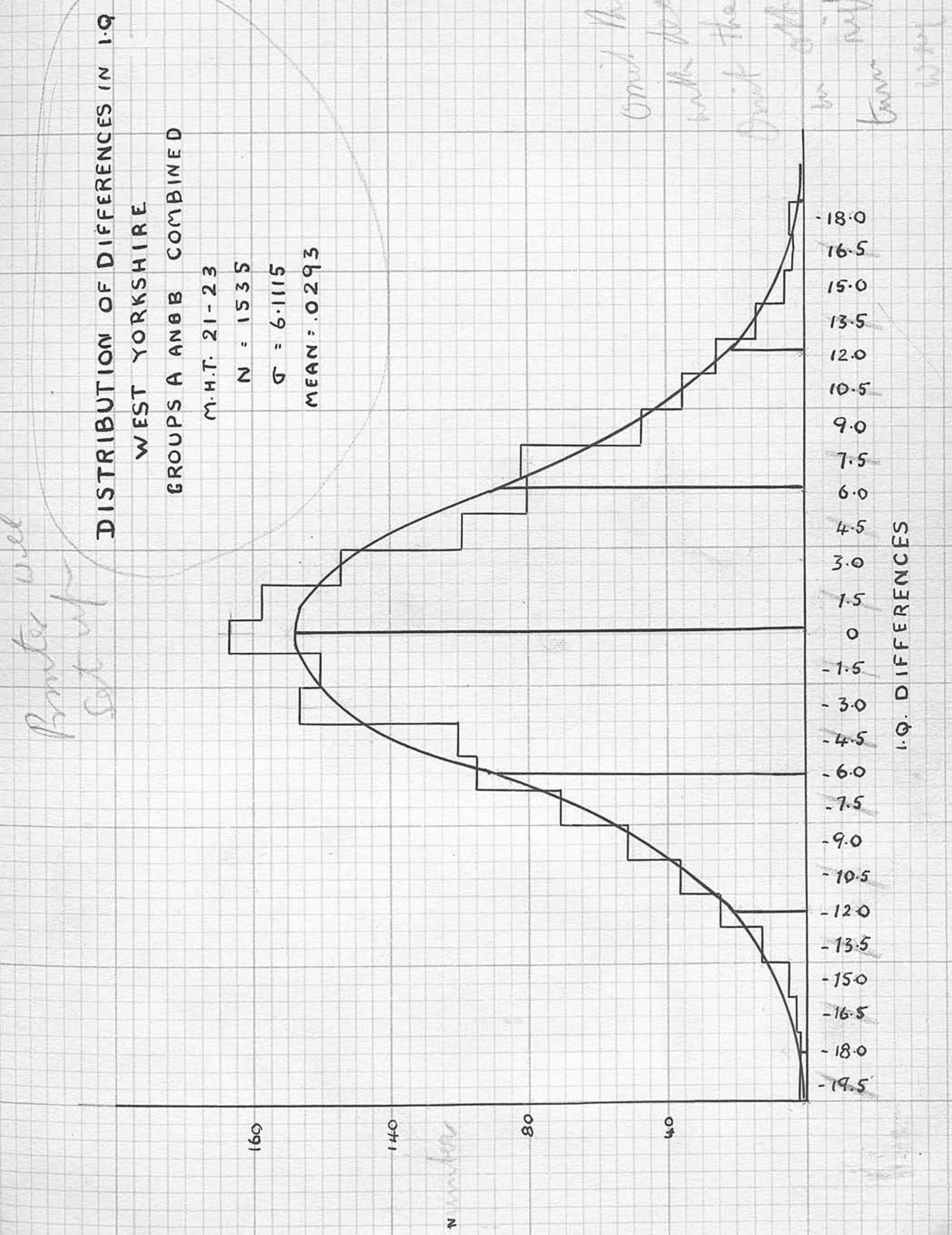
I.Q. diff.	M.H.T.21/23	M.H.T.21/26	M.H.T.23/26
19.5			
18.0	4	2	
16.5	1	1	
15.0	4	2	
13.5	11	6	6
12.0	18	12	3
10.5	30	16	7
9.0	36	30	8
7.5	57	35	11
6.0	53	49	30
4.5	59	56	46
3.0	67	66	58
1.5	71	76	70
0	62	67	51
-1.5	40	56	68
-3.0	44	54	68
-4.5	23	39	55
-6.0	24	26	51
-7.5	8	17	42
-9.0	9	8	18
-10.5	3	7	18
-12.0	3	1	10
-13.5	0	1	4
-15.0	0	1	3
-16.5	1	0	1
-18.0	1	1	0
-19.5			1
	629	629	629

TABLE 5 WEST YORKSHIRE

DISTRIBUTION OF DIFFERENCES IN I.Q.

GROUPS A AND B COMBINED.

I.Q. diff.	M.H.T.21/23	M.H.T.21/26	M.H.T.23/26
19.5			1
18.0	4	2	0
16.5	3	3	2
15.0	5	3	2
13.5	13	9	12
12.0	25	19	7
10.5	35	24	17
9.0	46	44	24
7.5	82	66	53
6.0	80	92	75
4.5	98	100	122
3.0	134	124	175
1.5	156	159	174
.0	167	167	144
-1.5	140	145	160
-3.0	146	152	159
-4.5	100	116	126
-6.0	95	93	96
-7.5	71	68	77
-9.0	51	46	40
-10.5	36	42	34
-12.0	24	25	15
-13.5	12	18	11
-15.0	5	10	7
-16.5	3	3	1
-18.0	2	3	0
-19.5	2	2	1
	1535	1535	1535



DISTRIBUTION OF DIFFERENCES IN I.Q
WEST YORKSHIRE
GROUPS A AND B COMBINED
M.H.T. 21-23
N : 1535
σ : 6.1115
MEAN : .0293
160
140
80
40
N
-18.0
16.5
15.0
13.5
12.0
10.5
9.0
7.5
6.0
4.5
3.0
1.5
0
-1.5
-3.0
-4.5
-6.0
-7.5
-9.0
-10.5
-12.0
-13.5
-15.0
-16.5
-18.0
-19.5
I.Q. DIFFERENCES

FIG. 5

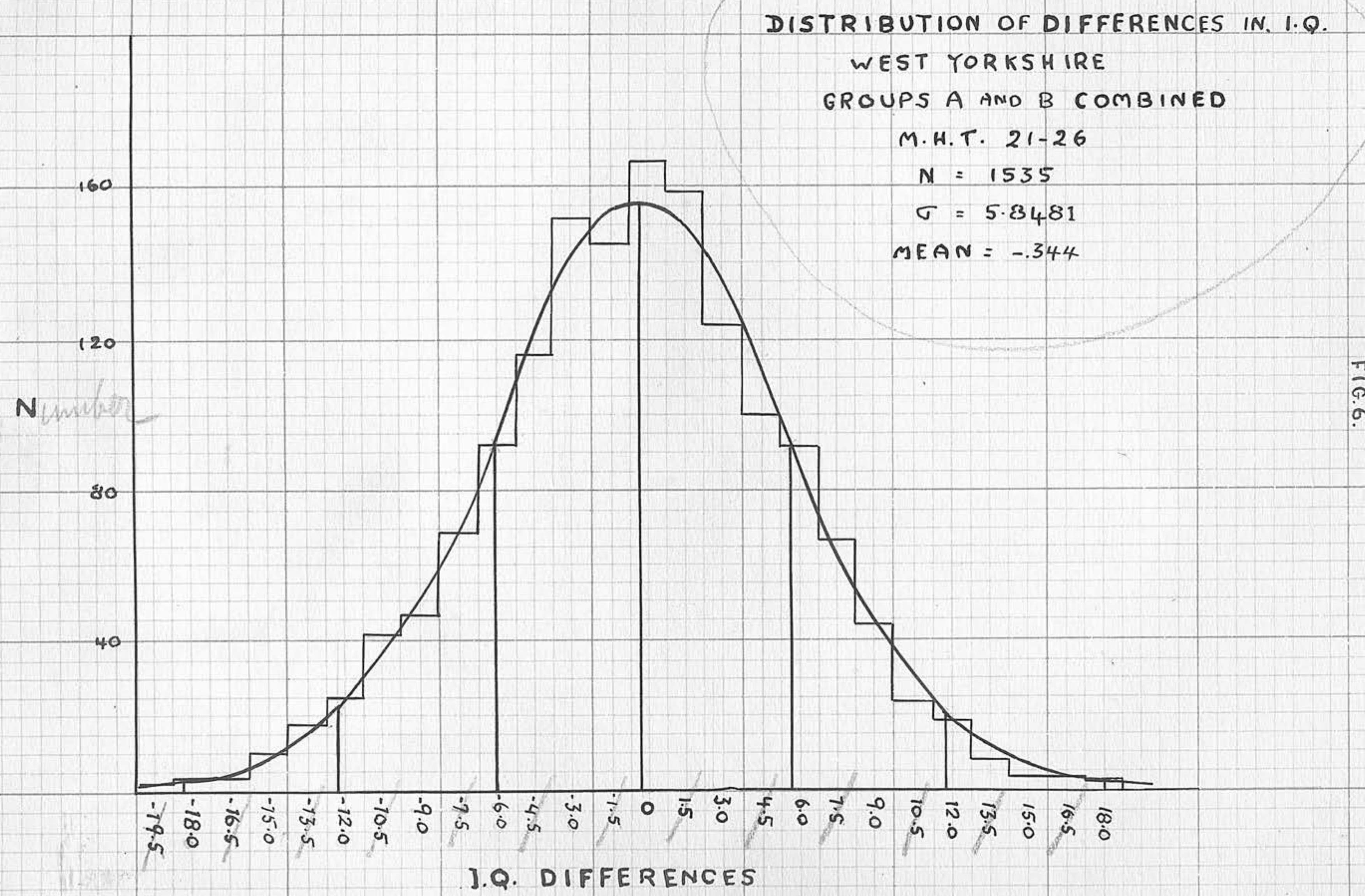

DISTRIBUTION OF DIFFERENCES IN I.Q.
WEST YORKSHIRE
GROUPS A AND B COMBINED
M.H.T. 21-26
N = 1535
σ = 5.8481
MEAN = -.344
160
120
80
40
N
-19.5
-18.0
-16.5
-15.0
-13.5
-12.0
-10.5
-9.0
-7.5
-6.0
-4.5
-3.0
-1.5
0
1.5
3.0
4.5
6.0
7.5
9.0
10.5
12.0
13.5
15.0
16.5
18.0
I.Q. DIFFERENCES

FIG. 6.

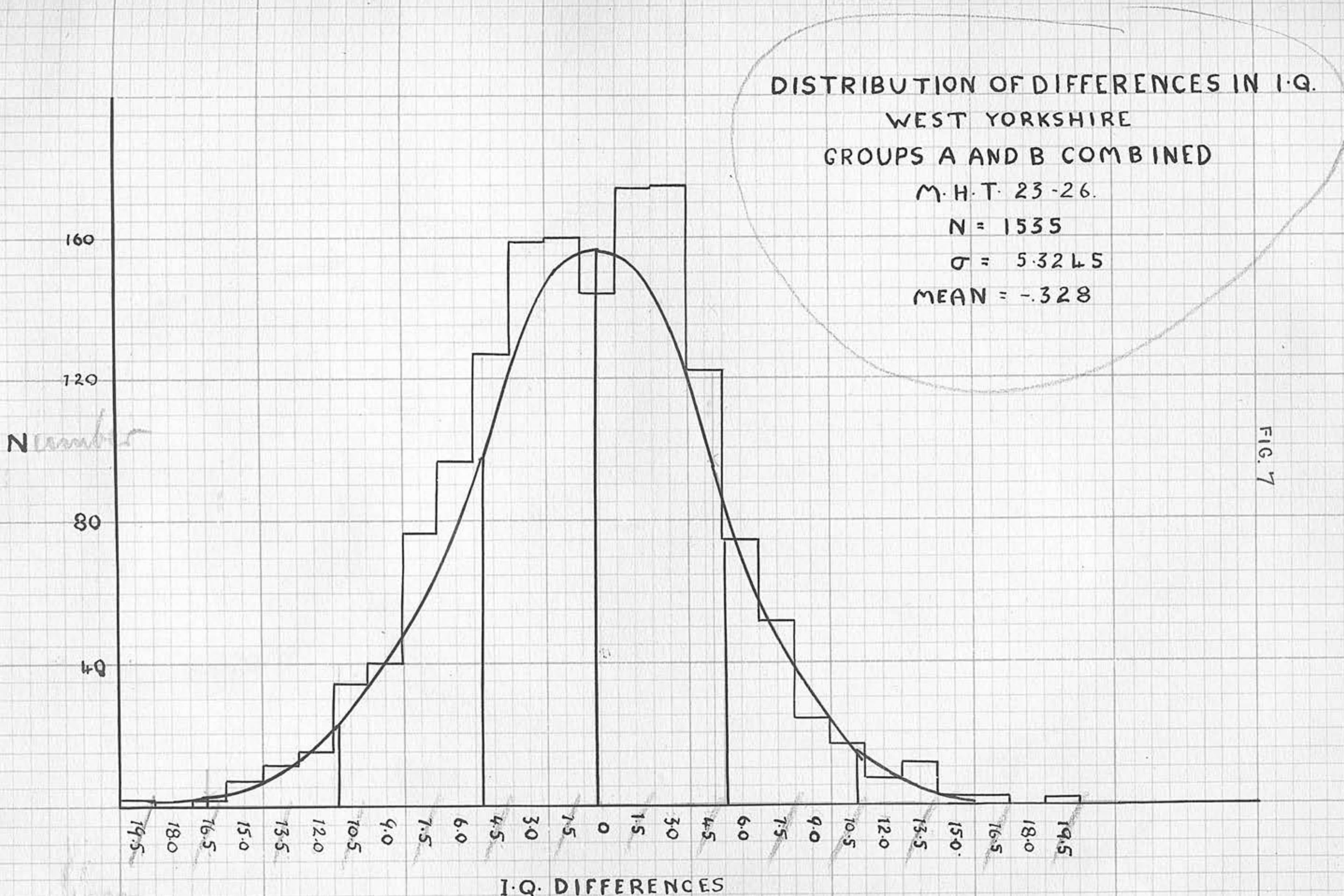

DISTRIBUTION OF DIFFERENCES IN I·Q.
WEST YORKSHIRE
GROUPS A AND B COMBINED
M·H·T· 23-26.
N = 1535
σ = 5.3245
MEAN = -.328
N
160
120
80
40
19.5
18.0
16.5
15.0
13.5
12.0
10.5
9.0
7.5
6.0
4.5
3.0
1.5
0
1.5
3.0
4.5
6.0
7.5
9.0
10.5
12.0
13.5
15.0
16.5
18.0
19.5
I·Q· DIFFERENCES

FIG. 7

VARIATIONS IN INTELLIGENCE QUOTIENT RELATIVE TO LEVEL OF ABILITY

VARIATION IN INTELLIGENCE QUOTIENT RELATIVE TO LEVEL OF ABILITY.

To determine whether I.Q. differences varied in relation to level of ability, all children were classified into 5 point I.Q. categories according to their average I.Q. as measured by the three tests, M.H.T. 21, 23 and 26. A child's I.Q. on any one of the three tests could have been taken as the basis for classification, but the average I.Q. on the three parallel forms furnished a more reliable estimate of each child's ability.

Since the number of cases above 130 and below 70 I.Q. was very small, and since the tests were not designed to discriminate accurately beyond these levels, the enquiry was confined to a consideration of the 12 catogaries between these limits, all cases above 130 and below 70 being deleted.

The standard deviations of differences in I.Q. between each of the three tests for Groups A and B separately, and for Groups A and B combined were calculated at each 5 point average I.Q. level of ability. Each standard deviation was corrected for grouping by Sheppard's correction. The assumption is made that

that intelligence is a continuous variate. The differences in I.Q. from test to retest were grouped with a class interval of 1.5 points of I.Q. Correcting for grouping reduced the standard deviation of differences by about .015.

Tables 6 to 14 give the distributions of differences in I.Q. for each 5 point I.Q. category between M.H.T.21 and M.H.T.23, M.H.T.21 and M.H.T.26, M.H.T.23 and M.H.T.23 and 26, for Groups A and B separately, and for Groups A and B combined.

Tables 15 to 23 give the uncorrected standard deviations of differences in I.Q. and the corresponding deviations corrected for grouping for I.Q. differences between M.H.T.21 and 23, M.H.T.21 and 26, and M.H.T.23 and 26, for Groups A and B separately and for Groups A and B combined.

From the distributions and tables given in this section many of the parameters given in later departments of this enquiry are computed.

Examination of these tables suggests that the I.Q. of dull children tends to be less variable than the I.Q. of bright children. The significance of this suggestion will be considered later when reliability relative to level of ability is discussed.

TABLE 6

DISTRIBUTIONS OF VARIATIONS IN I.Q. AT DIFFERENT LEVELS OF ABILITY.

Group A - M.H.T. 21/23

Interval	70.0-74.5	75.0-79.5	80.0-84.5	85.0-89.5	90.0-94.5	95.0-99.5	100-104.5	105-109.5	110-114.5	115-119.5	120-124.5	125-129.5
15.75-	-	-	1	-	1	-	-	-	-	-	-	-
14.25-	-	-	-	-	-	-	-	-	-	1	-	-
12.75-	-	-	-	-	1	1	-	-	-	-	-	-
11.25-	1	-	1	1	-	2	-	-	1	1	-	-
9.75-	-	1	2	-	-	-	1	-	1	-	-	-
8.25-	-	1	-	1	2	-	3	3	-	-	-	-
6.75-	-	2	-	4	3	3	3	5	4	-	1	-
5.25-	5	1	1	3	4	3	4	1	2	-	1	2
3.75-	1	1	2	4	6	4	6	5	4	3	3	-
2.25-	3	5	6	8	10	11	10	6	3	1	3	1
0.75-	8	5	9	5	14	11	10	5	9	6	2	1
-0.75-	7	7	9	12	18	19	10	8	7	1	5	2
-2.25-	7	8	11	13	11	8	12	9	10	4	5	2
-3.75-	4	1	12	14	10	14	11	14	14	4	3	1
-5.25-	1	4	5	9	14	17	7	8	8	2	-	2
-6.75-	1	3	7	7	10	12	7	13	7	2	2	-
-8.25-	-	3	4	5	8	8	6	10	12	3	3	1
-9.75-	-	3	2	4	7	9	9	5	2	1	-	-
-11.25-	-	2	4	4	4	4	5	2	2	4	2	-
-12.75-	-	1	0	3	2	3	5	3	3	-	1	-
-14.25-	-	-	1	2	2	5	-	1	1	-	-	-
-15.75-	-	-	1	-	1	-	2	-	-	-	1	-
-17.25-	-	-	-	1	-	1	-	-	-	-	-	-
-18.75-	-	-	-	-	-	-	-	-	-	1	-	-
-20.25-	-	-	-	-	-	-	-	2	-	-	-	-
Totals	38	48	78	100	128	135	111	100	90	34	32	12

TABLE 7

DISTRIBUTION OF VARIATIONS IN I.Q. AT DIFFERENT LEVELS OF ABILITY.

Group A - M.H.T. 21/26.

Inter val.	70.0-74.5	75.0-79.5	80.0-84.5	85.0-89.5	90.0-94.5	95.0-99.5	100-104.5	105-109.5	110-114.5	115-119.5	120-124.5	125-129.5
15.75-	-	-	-	1	-	-	-	-	1	-	-	-
14.25-	-	-	-	-	-	-	-	-	1	-	-	-
12.75-	-	-	-	1	1	-	1	-	-	-	-	-
11.25-	1	-	-	3	-	-	1	2	-	-	-	-
9.75-	-	1	1	1	2	2	-	-	1	-	-	-
8.25-	2	-	1	-	1	5	1	1	-	-	2	1
6.75-	2	1	3	1	4	4	5	2	2	3	3	1
5.25-	1	1	5	5	6	6	8	4	4	1	1	1
3.75-	1	3	5	1	6	3	8	3	5	1	7	1
2.25-	5	7	4	2	9	8	8	4	7	2	2	-
0.75-	6	2	4	14	11	9	9	13	8	4	1	2
-0.75-	10	7	9	11	11	16	10	8	12	2	4	-
-2.25-	4	8	8	7	14	19	7	9	8	1	2	2
-3.75-	2	5	4	7	16	13	10	16	16	6	2	1
-5.25-	1	5	13	9	12	9	6	9	8	4	1	0
-6.75-	-	-	9	11	7	9	8	13	5	2	2	1
-8.25-	3	1	4	6	8	8	7	7	3	3	1	-
-9.75-	-	2	4	2	7	6	7	4	3	1	2	-
-11.25	-	4	1	9	3	6	5	2	3	1	-	1
-12.75-	-	-	1	6	5	7	1	1	1	1	1	-
-14.25-	-	-	1	1	2	3	6	-	1	2	-	1
-15.75-	-	1	-	-	1	-	3	2	1	-	1	-
-17.25-	-	-	1	-	1	1	-	-	-	-	-	-
-18.75-	-	-	-	-	1	1	-	-	-	-	-	-
-20.25-	-	-	-	2	-	-	-	-	-	-	-	-
Totals.	38	48	78	100	128	135	111	100	90	34	32	12

TABLE 8

DISTRIBUTIONS OF VARIATIONS IN I.Q. AT DIFFERENT LEVELS OF ABILITY.

Group A - M.H.T. 23/26.

Inter val.	70.0-74.5	75.0-79.5	80.0-84.5	85.0-89.5	90.0-94.5	95.o-99.5	100-104.5	105-109.5	110-114.5	115-119.5	120-124.5	125 129.5
18.75-	-	-	-	-	-	-	-	-	1	-	-	-
17.25-	-	-	-	-	-	-	-	-	-	-	-	-
15.75-	-	-	-	-	-	-	-	1	1	-	-	-
14.25-	-	-	-	-	-	1	-	-	-	1	-	-
12.75-	-	-	-	1	-	-	1	2	-	2	-	-
11.25-	-	-	-	-	1	1	1	-	1	-	-	-
9.75-	-	-	2	-	1	2	2	1	2	-	-	-
8.25-	-	-	1	2	4	5	1	1	1	-	-	1
6.75-	-	3	4	2	5	7	6	6	6	1	2	-
5.25-	1	2	4	5	4	5	8	5	6	2	-	3
3.75-	1	2	8	7	15	13	9	11	5	4	1	-
2.25-	5	9	8	16	17	14	12	15	11	7	3	-
0.75-	8	3	9	12	18	17	15	6	10	1	4	1
-0.75-	11	6	12	12	13	11	10	5	7	2	3	1
-2.25-	5	9	6	6	8	12	13	12	13	3	2	3
-3.75-	1	5	8	8	11	17	11	12	12	-	5	1
-5.25-	5	3	4	12	11	14	4	10	4	2	2	-
-6.75-	-	2	2	6	7	7	6	3	3	2	6	1
-8.25-	1	3	3	2	3	5	8	4	4	1	1	-
-9.75-	-	1	5	3	3	3	2	-	1	2	1	1
-11.25-	-	-	1	3	4	-	-	5	-	1	2	-
-12.75-	-	-	-	1	1	-	1	-	1	1	-	-
-14.25-	-	-	-	2	1	1	1	1	1	-	-	-
-15.75-	-	-	1	-	1	-	-	-	-	2	-	-
Totals.	38	48	78	100	128	135	111	100	90	34	32	12

TABLE 9

M.H.T. 21/23

DISTRIBUTIONS OF DIFFERENCES IN I.Q. AT VARIOUS LEVELS OF ABILITY.

GROUP B

I.Q. diff.	70-74.5	75-79.5	80-84.5	85-89.5	90-94.5	95-99.5	100-104.5	105-109.5	110-114.5	115-119.5	120-124.5	125-129.5
18.0						1					2	1
16.5						0					1	0
15.0					2	0		2			0	0
13.5			1		2	1	2	2			2	1
12.0			0	4	4	1	3	3	1		0	2
10.5		1	1	1	7	3	4	6	2	3	1	1
9.0	1	1	2	2	5	1	5	5	3	9	2	0
7.5	2	1	3	4	8	5	11	7	5	4	2	5
6.0	1	3	3	1	6	9	8	6	7	2	2	5
4.5	0	0	4	7	7	10	14	5	4	5	2	1
3.0	2	1	1	9	6	5	7	13	9	7	5	2
1.5	2	2	3	5	14	13	5	6	8	6	6	1
.0	3	2	5	1	7	6	6	11	8	5	6	2
-1.5	1	1	2	6	4	4	7	2	5	5	3	0
-3.0	1	1	0	0	4	7	5	8	8	4	6	0
-4.5	0	0		0	3	3	0	5	2	5	1	4
-6.0				1	2	2	3	5	4	2	2	3
-7.5				2	0	0	0	2	1	2	0	1
-9.0				0	1	1	0	0	1	2	4	0
-10.5					1		0	0		0	1	1
-12.0							0	0		1	2	0
-13.5							0					0
-15.0							0					0
-16.5							0					1
-18.0							1					
	13	13	25	43	83	72	81	88	68	62	50	31

TABLE 10

M.H.T. 21/26

DISTRIBUTIONS OF DIFFERENCES IN I.Q. AT VARIOUS LEVELS OF ABILITY.

GROUP B

I.Q. diff.	70-74.5	75-89.5	80-84.5	85-89.5	90-94.5	95-99.5	100-104.5	105-109.5	110-114.5	115-119.5	120-124.5	125-130
18.0												2
16.5											1	0
15.0						1		1			0	0
13.5			1		1	0		0		1	2	1
12.0			3		1	1	4	1	2	0	0	0
10.5		1	0	2	2	0	5	1	1	1	1	2
9.0		0	1	0	4	3	9	3	3	3	1	3
7.5		1	5	2	2	5	6	7	3	2	2	0
6.0	2	1	2	6	7	4	3	9	4	6	2	3
4.5	1	0	3	6	7	7	6	5	8	6	5	2
3.0	3	0	2	7	8	16	6	6	4	5	6	3
1.5	3	3	3	2	10	12	12	13	7	6	5	0
.0	2	3	1	7	12	8	9	5	15	2	2	1
-1.5	2	2	2	2	9	5	3	6	7	10	7	1
-3.0	0	2	0	2	9	3	7	11	6	4	8	2
-4.5		0	1	2	6	4	3	8	2	6	4	3
-6.0			1	1	2	1	3	7	3	5	2	1
-7.5				1	2	1	3	3	2	3	0	2
-9.0				2	1	0	1	0	0	0	0	4
-10.5				1		0	1	1	1	1	1	1
-12.0						1		0		0	0	
-13.5								1		0	0	
-15.0										1	0	
-16.5											0	
-18.0											1	
-19.5												
	13	13	25	43	83	72	81	88	68	62	50	31

TABLE 11

M.H.T. 23/26.

DISTRIBUTIONS OF DIFFERENCES IN I.Q. AT

VARIOUS LEVELS OF ABILITY.

GROUP B

I.Q. diff.	70-74.5	75-79.5	80-84.5	85-89.5	90-94.5	95-99.5	100-104.5	105-109.5	110-114.5	115-119.5	120-124.5	125-130
16.5												
15.0												
13.5							2			2	2	
12.0		1					0		1	1	0	
10.5		0			1		0	1	1	1	2	1
9.0		0			0	1	3	0	1	0	2	1
7.5		0	1		1	2	1	0	3	0	2	1
6.0	1	0	3	1	3	2	2	4	4	7	2	1
4.5	0	1	2	3	6	9	8	7	3	4	1	2
3.0	3	0	4	4	5	7	6	11	10	3	1	4
1.5	1	1	2	7	5	9	13	12	5	7	5	3
.0	2	1	4	5	7	5	6	6	5	3	4	3
-1.5	2	3	3	5	5	7	9	8	10	4	7	5
-3.0	0	1	0	3	8	12	9	13	4	8	7	3
-4.5	2	1	2	4	12	5	6	4	8	6	5	0
-6.0	2	1	4	6	10	4	4	6	3	2	6	3
-7.5		1		2	9	5	5	5	6	6	2	1
-9.0		1		1	4	2	1	4	1	3	1	0
-10.5		1		0	6	2	1	1	3	3	0	1
-12.0				2	0		3	2		1	1	1
-13.5					1	0	1	2		0		0
-15.0							1	1		1		0
-16.5								0				1
-18.0								0				
-19.5								1				
	13	13	25	43	83	72	81	88	68	62	50	31

TABLE 12

M.H.T. 21/23.

DISTRIBUTIONS OF DIFFERENCES IN I.Q. AT VARIOUS LEVELS OF ABILITY.

GROUPS A AND B.

I.Q. diff.	70-74.5	75-79.5	80-84.5	85-89.5	90-94.5	95-99.5	100-104.5	105-109.5	110-114.5	115-119.5	120-124.5	125-130
19.5												
18.0						1					2	1
16.5			1		1	0					1	0
15.0			0		2	0		2		1	0	0
13.5			1		3	2	2	2		0	2	1
12.0	1		1	5	4	3	3	3	2	1	0	2
10.5	0	2	3	1	7	3	5	6	3	3	1	1
9.0	1	2	2	3	7	1	8	8	3	9	2	0
7.5	2	3	3	8	11	8	14	12	9	4	3	5
6.0	6	4	4	4	10	12	12	7	9	2	3	7
4.5	1	1	6	11	13	14	20	10	8	8	5	1
3.0	5	6	7	17	16	16	17	19	12	8	8	3
1.5	10	7	12	10	28	24	15	11	17	12	8	2
.0	10	9	14	13	25	25	15	19	15	6	11	4
-1.5	8	9	13	19	15	12	19	11	15	9	8	2
-3.0	5	2	12	14	14	21	16	22	22	8	9	1
-4.5	1	4	5	9	17	20	7	13	10	7	1	6
-6.0	1	3	7	8	12	14	11	18	11	4	4	3
-7.5		3	4	7	8	8	6	12	13	5	3	2
-9.0		3	2	4	8	10	9	5	3	3	4	0
-10.5		2	4	4	5	4	5	2	2	4	3	1
-12.0		1	0	3	2	3	5	3	3	1	3	0
-13.5			1	2	2	5	0	1	1	0	0	0
-15.0			1	0	1	0	2	0		0	1	0
-16.5				1		1	0	0		0		1
-18.0							1	0		1		
-19.5								2				
	51	61	103	143	211	207	192	188	158	96	82	43

TABLE 13

M.H.T. 21/26.

DISTRIBUTIONS OF DIFFERENCES IN I.Q. AT VARIOUS LEVELS OF ABILITY.

GROUPS A AND B.

I.Q. diff.	70-74.5	75-79.5	80-84.5	85-89.5	90-94.5	95-99.5	100-104.5	105-109.5	110-114.5	115-119.5	120-124.5	125-129.5
19.5												
18.0												2
16.5				1					1		1	0
15.0				0		1		1	1		0	0
13.5			1	1	2	0	1	0	0	1	2	1
12.0	1		3	3	1	1	5	3	2	0	0	0
10.5	0	2	1	3	4	2	5	1	2	1	1	2
9.0	2	0	2	0	5	8	10	4	3	3	3	4
7.5	2	2	8	3	6	9	11	9	5	5	5	1
6.0	3	2	7	11	13	10	11	13	8	7	3	4
4.5	2	3	8	7	13	10	14	8	13	7	12	3
3.0	8	7	6	9	17	24	14	10	11	7	8	3
1.5	9	5	7	16	21	21	21	26	15	10	6	2
.0	12	10	10	18	23	24	19	13	27	4	6	1
-1.5	6	10	10	9	23	24	10	15	15	11	9	3
-3.0	2	7	4	9	25	16	17	27	22	10	10	3
-4.5	1	5	14	11	18	13	9	17	10	10	5	3
-6.0	0	0	10	12	9	10	11	20	8	7	4	2
-7.5	3	1	4	7	10	9	10	10	5	6	1	2
-9.0	0	2	4	4	8	6	8	4	3	1	2	4
-10.5		4	1	10	3	6	6	3	4	2	1	2
-12.0		0	1	6	5	8	1	1	1	1	1	0
-13.5		0	1	1	2	3	6	1	1	2	0	1
-15.0		1	0	0	1	0	3	2	1	1	1	
-16.5			1	0	1	1					0	
-18.0				0	1	1					1	
-19.5				2								
	51	61	103	143	211	207	192	188	158	96	82	43

TABLE 14

M.H.T. 23/26.

DISTRIBUTIONS OF I.Q. DIFFERENCES AT VARIOUS LEVELS OF ABILITY.

GROUPS A AND B.

I.Q. diff.	70-74.5	75-79.5	80-84.5	85-89.5	90-94.5	95-99.5	100-104.5	105-109.5	110-114.5	115-119.5	120-124.5	125-129.5
19.5									1			
18.0									0			
16.5								1	1			
15.0						1		0	0	1		
13.5				1		0	3	2	0	4	2	
12.0		1		0	1	1	1	0	2	1	0	
10.5		0	2	0	2	2	2	2	3	1	2	1
9.0		0	1	2	4	6	4	1	2	0	2	2
7.5		3	5	2	6	9	7	6	9	1	4	1
6.0	2	2	7	6	7	7	10	9	10	9	2	4
4.5	1	3	10	10	21	22	17	18	8	8	2	2
3.0	8	9	12	20	22	21	18	26	21	10	4	4
1.5	9	4	11	19	23	26	28	18	15	8	9	4
.0	13	7	16	17	20	16	16	11	12	5	7	4
- 1.5	7	12	9	11	13	19	22	20	23	7	9	8
-3.0	1	6	8	11	19	29	20	25	16	8	12	4
-4.5	7	4	6	16	23	19	10	14	12	8	7	0
-6.0	2	3	6	12	17	11	10	9	6	4	12	4
-7.5	1	4	3	4	12	10	13	9	10	7	3	1
-9.0		2	5	4	7	5	3	4	2	5	2	1
-10.5		1	1	3	10	2	1	6	3	4	2	1
-12.0			0	3	1	0	4	2	1	2	1	1
-13.5			0	2	2	1	2	3	1	0		0
-15.0			1		1		1	1		3		0
-16.5								0				1
-18.0								0				
-19.5								1				
	51	61	103	143	211	207	192	188	158	96	82	43

TABLE 15

Table of Standard Deviations of Variations in I.Q. between M.H.T. 21 and M.H.T. 23 for various I.Q. levels with values of N for Group A.

I.Q. level.	S.D. uncorrected	S.D. corrected	Values of N.
125-130	3.9843	3.9608	12
120-124	5.4213	5.4041	32
115-119	6.5205	6.5061	34
110-114	5.1969	5.1788	90
105-109	5.7261	5.7098	100
100-104	5.7746	5.7584	111
95-99	5.5428	5.5259	135
90-94	5.4915	5.4743	128
85-89	5.4228	5.4054	100
80-84	5.3820	5.3645	78
75-79	5.2394	5.2217	48
70-74	3.5016	3.4748	38

TABLE 16

Table of Standard Deviations of Variations in I.Q. between M.H.T 21 and M.H.T. 26 for various I. Q. levels with values of N for Group A.

I.Q. level.	S.D. uncorrected	S.D. corrected	Values of N.
125-130	6.6896	6.6755	12
120-124	6.1973	6.1821	32
115-119	5.7260	5.7096	34
110-114	5.4843	5.4674	90
105-109	5.1363	5.1180	100
100-104	6.5087	6.4943	111
95-99	6.0240	6.0084	135
90-94	5.9166	5.9007	128
85-89	6.7116	6.6977	100
80-84	5.4596	5.4423	78
75-79	5.1555	5.1371	48
70-74	4.3106	4.2888	38

TABLE 17

Table of Standard Deviations of Variations in I.Q. between M.H.T. 23 and M.H.T. 26 for various I.Q. levels with values of N for Group A.

I.Q. level.	S.D. uncorrected	S.D. corrected	Values of N.
125-130	5.1722	5.1539	12
120-124	4.6650	4.6449	32
115-119	7.4865	7.4741	34
110-114	5.5505	5.5337	90
105-109	5.5791	5.5623	100
100-104	5.1361	5.1177	111
95-99	5.0183	4.9995	135
90-94	5.2265	5.2085	128
85-89	5.1486	5.1306	100
80-84	5.1540	5.1362	78
75-79	4.1193	4.0964	48
70-74	2.7945	2.7608	38

TABLE 18

Table of Standard Deviations of Variations in I.Q. between M.H.T. 21 and M.H.T. 23 for various I.Q. levels with values of N for Group B.

I.Q. level.	S.D. uncorrected	S.D. corrected	Values of N.
125-130	7.7304	7.7183	31
120-124	7.2765	7.2633	50
115-119	5.6589	5.6423	62
110-114	4.7427	4.7229	68
105-109	5.6646	5.6477	88
100-104	5.3184	5.3006	81
95-99	4.8729	4.8537	72
90-94	5.5332	5.5163	83
85-89	4.8041	4.7850	43
80-84	3.9573	3.9335	25
75-79	4.0566	4.0334	13
70-74	3.6342	3.6083	13

TABLE 19

Table of Standard Deviations of Variations in I.Q. between M.H.T. 21 and M.H.T. 26 for various I.Q. levels with values of N for Group B.

I.Q. level.	S.D. uncorrected	S.D. corrected	Values of N.
125-130	8.1671	8.1554	31
120-124	5.3657	5.3481	50
115-119	5.5371	5.5202	62
110-114	4.6986	4.6787	68
105-109	5.3988	5.3814	88
100-104	5.6285	5.6118	81
95-99	4.3785	4.3571	72
90-94	4.6867	4.6666	83
85-89	4.9958	4.9772	43
80-84	5.0322	5.0135	25
75-79	3.9468	3.9230	13
70-74	2.3927	2.3532	13

TABLE 20

Table of Standard Deviations of Variations in I.Q. between M.H.T. 23 and 26 for various I.Q. levels with values of N for Group B.

I.Q. level.	S.D. uncorrected	S.D. corrected	Values of N.
125-130	5.8962	5.8802	31
120-124	6.3909	6.3764	50
115-119	6.3107	6.2958	62
110-114	5.2400	5.2220	68
105-109	5.4828	5.4657	88
100-104	5.3028	5.2851	81
95-99	4.5890	4.5686	72
90-94	5.0522	5.0334	83
85-89	4.3913	4.3697	43
80-84	4.1508	4.1282	25
75-79	5.7510	5.7344	13
70-74	3.6995	3.6741	13

TABLE 21

GROUPS A AND B

Table of Standard Deviations of Variations in I.Q. between M.H.T. 21 and M.H.T. 23 for various I.Q. levels with Standard Errors and values of N.

	I.Q. Level	S.D. Uncorrected	S.D. corrected	Standard Error	Values of N.
(1)	125-130	7.0133	6.9993	.7545	43
(2)	120-124	6.7791	6.7650	.5283	82
(3)	115-119	6.2801	6.2660	.4524	96
(4)	110-114	5.3865	5.3693	.3023	158
(5)	105-109	6.3948	6.3807	.3292	188
(6)	100-104	6.2988	6.2847	.3205	192
(7)	95-99	5.8989	5.8832	.2889	207
(8)	90-94	6.1766	6.1611	.3000	211
(9)	85-89	5.8514	5.8350	.3448	143
(10)	80-84	5.6853	5.6691	.3951	103
(11)	75-79	5.4054	5.3886	.4877	61
(12)	70-74	3.6178	3.5907	.3537	51

TABLE 22

GROUPS A AND B.

Table of Standard Deviations of Variations in I.Q. between M.H.T.21 and MH.T. 26 for various I.Q. levels with Standard Errors and values of N.

	I.Q. Level	S.D. Uncorrected	S.D. Corrected	Standard Error	Values of N.
(1)	125-130	7.6821	7.6701	.8268	43
(2)	120-124	5.7177	5.7020	.4453	82
(3)	115-119	5.7687	5.7521	.4153	96
(4)	110-114	5.3061	5.2884	.2977	158
(5)	105-109	5.4530	5.4360	.2805	188
(6)	100-104	6.5634	6.5493	.3340	192
(7)	95-99	5.9028	5.8869	.3381	207
(8)	90-94	5.7183	5.7020	.2777	211
(9)	85-89	6.5369	6.5229	.3855	143
(10)	80-84	6.0534	6.0380	.4208	103
(11)	75-79	5.0705	5.0520	.4572	61
(12)	70-74	3.9348	3.9129	.3854	51

TABLE 23

GROUPS A AND B.

Table of Standard Deviations of Variations in I.Q. between M.H.T. 23 and M.H.T. 26 for various I.Q. levels with Standard Errors and values of N.

	I.Q. Level	S.D. Uncorrected	S.D. Corrected	Standard Errors.	Values of N.
(1)	125-130	5.7330	5.7168	.6163	43
(2)	120-124	5.9825	5.9669	.4660	82
(3)	115-119	6.7707	6.7569	.4878	96
(4)	110-114	5.4611	5.4440	.3065	158
(5)	105-109	5.6391	5.6228	.2901	188
(6)	100-104	5.2242	5.2062	.2655	192
(7)	95-99	4.9058	4.8861	.2399	207
(8)	90-94	5.3435	5.3249	.2593	211
(9)	85-89	4.9620	4.9431	.2921	143
(10)	80-84	4.9332	4.9140	.3425	103
(11)	75-79	4.5653	4.5456	.4114	61
(12)	70-74	3.0585	3.0278	.3013	51

583

The Estimation of Reliability.

The variance of variation in I.Q. as measured by two parallel forms of a test is given in the formula

$$\sigma^2_{(1-1')} = \sigma_1^2 + \sigma_{1'}^2 - 2 r_{11'} \sigma_1 \sigma_{1'}$$

where $\sigma^2_{(1-1')}$ the variance of the differences between the tests 1 and 1'.

THE ESTIMATION OF RELIABILITY

σ_1^2 = the variance of test 1

$\sigma_{1'}^2$ = the variance of test 1'

$r_{11'}$ = the correlation between tests 1 and 1'.

In the present enquiry $\sigma^2_{(1-1')}$ is the variance of the differences between two successive sets of I.Q. as found by the application of two parallel forms of the same test to the same group of individuals. σ_1 and $\sigma_{1'}$ are the standard deviations of I.Q. as measured by forms 1 and 1', respectively. Both σ_1 and $\sigma_{1'}$ are equal to 15, since Moray House Tests are standardised on this basis. Since the two forms of the test used were parallel, $r_{11'}$ is regarded as a reliability coefficient.

Since $\sigma_1 = \sigma_{1'} = 15$ formula (1) reduces to

$$\sigma^2_{(1-1')} = 2\sigma^2(1 - r_{11'})$$

But the formula for the standard error of a test score is known to be $\quad \varepsilon_1 = \sigma\sqrt{1 - r_{11'}}$

where ε_1 = the standard error of a test score.

The Estimation of Reliability.

The variance of variation in I.Q. as measured by two parallel forms of a test is given in the formula

$$\sigma^2_{(1-1')} = \sigma_1^2 + \sigma_{1'}^2 - 2r_{11'}\sigma_1\sigma_{1'}$$

where $\sigma^2_{(1-1')}$ the variance of the differences between the tests 1 and $1'$.

σ_1^2 = the variance of test 1.

$\sigma_{1'}^2$ = the variance of test $1'$.

$r_{11'}$ = the correlation between tests 1 and $1'$.

In the present enquiry $\sigma^2_{(1-1')}$ is the variance of the differences between two successive sets of I.Q. as found by the application of two parallel forms of the same test to the same group of individuals. σ_1 and $\sigma_{1'}$ are the standard deviations of I.Q. as measured by forms 1 and $1'$, respectively. Both σ_1 and $\sigma_{1'}$ are equal to 15, since Moray House Tests are standardised on this basis. Since the two forms of the test used were parallel, $r_{11'}$ is regarded as a reliability coefficient.

Since $\sigma_1 = \sigma_{1'} = 15$ formula (1) reduces to

$$\sigma^2_{(1-1')} = 2\sigma^2(1 - r_{11'})$$

But the formula for the standard error of a test score is known to be $\varepsilon_1 = \sigma\sqrt{1 - r_{11'}}$

where ε_1 = the standard error of a test score.

In the present enquiry $\mathcal{E}_1$ is the standard error of an I.Q. It follows therefore, that

$$\mathcal{E}_1 = \frac{\sigma_{(1-1')}}{\sqrt{2}}$$

The standard error of an I.Q. is, therefore, equal to the standard deviation of variation in I.Q. between two series of I.Q's, obtained by retest or by the application of two parallel forms to the same sample of the population, divided by $\sqrt{2}$. Thus from values of $\sigma_{(1-1')}$ calculated at different levels of ability, it is possible to calculate values of the standard error of I.Q. at each level of ability under consideration by merely multiplying the values of $\sigma_{(1-1')}$ by ·7071.

The quantities $\mathcal{E}_1$ and $\sigma_{(1-1')}$ must be interpreted correctly. The quantity $\mathcal{E}_1$ determines how closely an individualsI.Q. as measured by a fallible test approximates to his true I.Q. An individuals true I.Q. as measured by a given test may be defined as the mean of an infinate number of estimates of the individuals I.Q. as measured by the test inquestion.

Note:- In the present enquiry all statistical parameters are corrected for grouping. In the formula $\sigma^2_{(1-1')} = \sigma_1^2 + \sigma_{1'}^2 - 2r_{11'}\sigma_1\sigma_{1'}$ if the variances σ_1^2 and $\sigma_{1'}^2$ are uncorrected for grouping the variances of the differences, $\sigma^2_{(1-1')}$, must be corrected twice by Sheppard's correction. The same result can be obtained by correcting σ_1^2 and $\sigma_{1'}^2$, leaving the term $2r_{11'}\sigma_1\sigma_{1'}$ uncorrected. The product-moment $r_{11'}\sigma_1\sigma_{1'}$ is independent of grouping, and is the same for values σ_1, $\sigma_{1'}$, and $r_{11'}$, either corrected or uncorrected. Grouping increases the standard deviation of the variates, and reduces the correlation between them in such a manner that the product-moment $r_{11'}\sigma_1\sigma_{1'}$ is constant.

$\mathcal{E}_1$ is the standard deviation of variation in I.Q. as measured by two tests, one having a reliability coefficient less than unity, the other having a reliability coefficient equal to unity, and, therefore yielding true measures of I.Q. The quantity $\sigma_{(1-1')}$ determines how closely an individuals score as measured by a fallible test approximates to his score on a parallel form of equal fallibility to the first.

It may be shown that the standard deviation of the difference between two variables, where the two variables are two parallel forms of the same test and the correlation between them is regarded as a reliability coefficient, is equal to the standard error of the difference between two scores, or I.Q's, on the two forms. The standard error of the differencebetween two scores is expressed by the general formula

$$\mathcal{E}_{(1-1')} = \sqrt{\mathcal{E}_1^2 + \mathcal{E}_{1'}^2 - 2 r_{e_1 e_{1'}} \mathcal{E}_1 \mathcal{E}_{1'}}$$

where $\mathcal{E}_{(1-1')}$ is the standard error of the difference between two scores of I.Q.'s.

$\mathcal{E}_1$ is the standard error of a score or I.Q. as measured by form 1.

$\mathcal{E}_{1'}$ is the standard error of a score or I.Q. as measured by form $1'$.

$r_{e e_1}$ = error correlation

Note:- See L.L.Thurstone, 'The Reliability and Validity of Tests' P. 22.

Since the errors in the two forms are assumed to be un: :correlated, the correlational term in $V_{ee'}$ vanishes, and the formula reduces to

$$\mathcal{E}_{(1-1')} = \sqrt{\mathcal{E}_1^2 + \mathcal{E}_{1'}^2}$$

but $\mathcal{E}_1$ is equal to $\mathcal{E}_{1'}$ (which it must according to the method of calculating it) so that

$$\mathcal{E}_{(1-1')} = \mathcal{E}_1\sqrt{2}$$

but we know from formula (2) that

$$\mathcal{E}_1 = \frac{\sigma_{(1-1')}}{\sqrt{2}}$$

thus

$$\mathcal{E}_{(1-1')} = \sigma_{(1-1')}$$

<u>The Calculation of Mean absolute Deviations.</u>

If the variations in I.Q. from test to retest are normally distributed then $\sigma_{(1-1')}$ bears a relationship to the mean absolute deviation (sometimes called the average difference, mean variation, average deviation, variation taken regardless of sign) such that

$$M.A.D. = .7979\,\sigma_{(1-1')}$$

where M.A.D. = the mean absolute deviation.

thus

$$\mathcal{E}_1 = \frac{M.A.D.}{.7979\sqrt{2}}$$

where $\mathcal{E}_1$ = the standard error of a test score.

The Calculation of Reliability Coefficients.

Given values of $\sigma_{(1-1')}$ and σ we can calculate reliability coefficients at different levels of ability.

since
$$\sigma^2_{(1-1')} = 2\sigma^2(1 - r_{11'})$$

therefore
$$r_{11'} = 1 - \frac{\sigma^2_{(1-1')}}{2\sigma^2}$$

Given values of $\mathcal{E}_1$ and σ we can calculate reliability coefficients by the formula
$$r_{11'} = 1 - \frac{\mathcal{E}_1^2}{\sigma^2}$$

Similarly given values of the mean absolute deviation we can calculate reliability coefficients by the formula
$$r_{11'} = 1 - \frac{M.A.D^2}{1.2733\sigma^2}$$

Since $r_{11'}$ is a function of both the standard deviation of the test and the standard error of a test score two tests with the same reliability coefficients may have different standard errors, because each test may yield a different standard deviation of I.Q. It follows, therefore, that standard errors of I.Q.'s as measured by Moray House Tests, which are standardised on the bases that the standard deviation of I.Q. is 15, are not directly comparable with standard errors of I.Q.'s as measured by the New Revision of the Binet Scale, which yields a standard deviation of I.Q. equal to 16.4. It follows also that test~~s~~ scores on a test of low reliability may have a small standard error because of a small standard deviation.

The standard error of an I.Q. expressed in standard measure or of a standard score is a more useful index for comparing the efficiency of two tests than the standard error of a raw or deviation score, if the samples to which the tests have been given are representative. The formula for the standard error of a standard score is

$$\mathcal{E}_s = \sqrt{1 - r_{11'}}$$

where $\mathcal{E}_s$ = the standard error of a standard score.

Reliability in Relation to Ability.

In the present investigation reliability coefficients were calculated at different levels of ability using the formula

$$r_{11'} = 1 - \frac{\sigma^2_{(1-1')}}{2\sigma^2}$$

This method is directly comparable with the method used by Terman in calculating reliability coefficients for the New Revision of the Stanford Binet at different levels of ability. Terman calculated the mean absolute deviations in I.Q. at different levels of ability and used the appropriate formula

$$r_{11'} = 1 - \frac{M.A.D}{1.2733\sigma^2}$$

where σ is equal to 16.4

The reliability coefficients calculated from the standard deviations of the differences in I.Q. between the three tests M.H.T. 21, 23, and 26 for Group A at various levels of ability are given in Table 24. Corresponding data are given in Table 25. For Group B reliability coefficients were calculated for Groups A and B combined. These coefficients and their standard errors for the three tests are given in columns 2 and 3 of Tables 26, 27, and 28 respectively, for M.H.T. 21/23, 21/26, and 23/26.

Examination of these Tables indicates that no unique reliability coefficient exists for any one test, the general tendency being for tests to be more reliable at the lower than at the upper ranges of ability. For example in Table 26 the reliability coefficients vary from .891 for children of I.Q. between 125 and 130 to .971 for children with I.Q.'s between 70 and 74.

To test whether the suggested decrease in reliability with increase in level of ability was significantly different from zero the reliability coefficients calculated for Groups A and B combined were converted into z scores, and least square lines fitted to each series of z scores thus obtained. Fitting a least square line to the values of z is preferable to fitting the line to the values of r, because, since the values of r are very high, their sampling distributions will be badly skewed. The sampling distribution of z is approximately normal, and its standard error is independent of the values of the true correlation in the population. The equation for converting r's into z's is

$$z = \tfrac{1}{2}(\log(1+r) - \log(1-r)$$

Each point was weighted by (N-3), the reciprocal of its variance. The slopes of the least square lines were calculated by the formula

$$b = \frac{S(N-3)Sxy - SxSy}{S(N-3)Sx^2 - (Sx)^2}$$

where b = slope of the best fitting least square line.

N-3 = reciprocal of variance of z.

x = deviation from guessed mean.

y = z scores

The standard error of b is given by the formula

$$\frac{\sigma^2}{S(x-\bar{x})^2}$$

where σ^2 is the variance of z, and is equal to 1.

The slopes of the lines thus obtained for the three tests for Groups A and B combined with their standard errors and values of of t are as follows:-

Test	slope	$S.D._b$	t
M.H.T. 21/23	-.0247	.0097	2.546
M.H.T. 21/26	-.0075	.0097	0.773
M.H.T. 23/26	-.0373	.0097	3.845

In the case of tests 21 and 23, and 23 and 26 the slopes may be regarded as differing significantly from zero. This implies that in these two cases there exists a significant decrease in reliability with increase in ability. The slope of the values of z for tests 21 and 26 does not differ from zero

Smoothed values of z were obtained, and these smoothed values of z converted into smoothed values of r. The smoothed values of z and r are given in columns 6 and 4 respectively, of Tables 26, 27, and 28.

Figures 8, 9, and 10 give values of z plotted against varying levels of ability with the best fitting least square line. Some doubt exists as to whether the relationship is linear. An examination of the above figures would seem to indicate that a polynomial of the third degree would be a better fit than a least square line. The data, however, are not sufficiently comprehen:-:sive to warrant the arithmetical labour involved in fitting such a curve.

The reliability coefficients given in the above enquiry for Moray House Tests obtained by the application of parallel of the same tests after a time interval of one week must be regarded as highly satisfactory. The boosted split half reliabilities of these tests are considerably higher than the coefficients obtained by correlating parallel forms. The split half reliabilities of M.H.T. 26, 23, and 21, based on a sample of 212 cases, are respectively .9721, .9687 and .9625. The reliability coefficients calculated by the application of parallel forms after an interval of one week are reduced by variations in the function tested. The reliability coefficients obtained by by the split half method are increased possibly by the correlation of errors. The reliability coefficients that would have obtained for the tests used in the present investigation had the function tested exhibited no variability, and had errors of measurement been uncorrelated would be about .95.

It may be observed here that small differences in large reliability coefficients may correspond to fairly substantial differences in the standard errors of I.Q. A difference of one point in the second decimal place in coefficients above .90 may represent a considerable divergence in the degree of concomitant variation between the variates correlated, while a difference of one point in the second decimal place of coefficients of about .70 represents a very small change in the degree of such concomitant variation. (see Garrett, Statistics in Psychology and Education, p283, for further elaboration on this point). Thus a small change in a high reliability coefficient will correspond to a large difference in the standard error in I.Q., while a small change in low reliability coefficients will correspond to a small change in the standard error of I.Q.

It may be remarked here that a single test yielding a reliability coefficient less than .90 cannot be regarded as an efficient instrument of cognitive measurement,

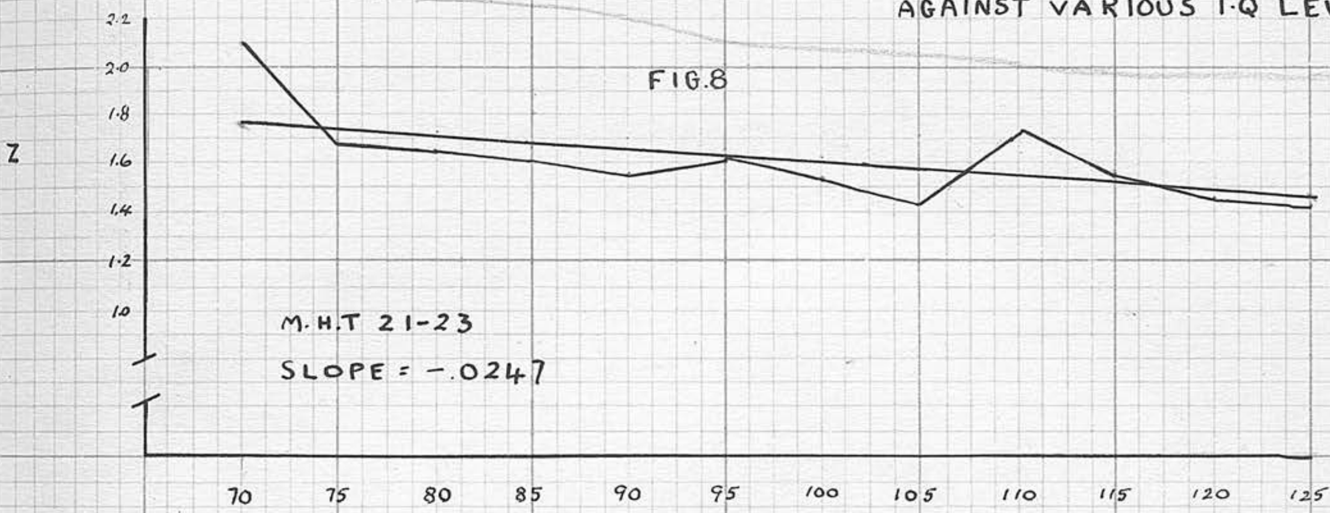
measurement, and should not be used in reaching any serious conslusions regarding a child's future educational career.

160

FIGURES SHOWING DECREASE IN RELIABILITY WITH INCREASE IN ABILITY

VALUES OF Z PLOTTED AGAINST VARIOUS I.Q LEVELS

FIG.8



I·Q.   M. H. T. 21-23

FIG.9

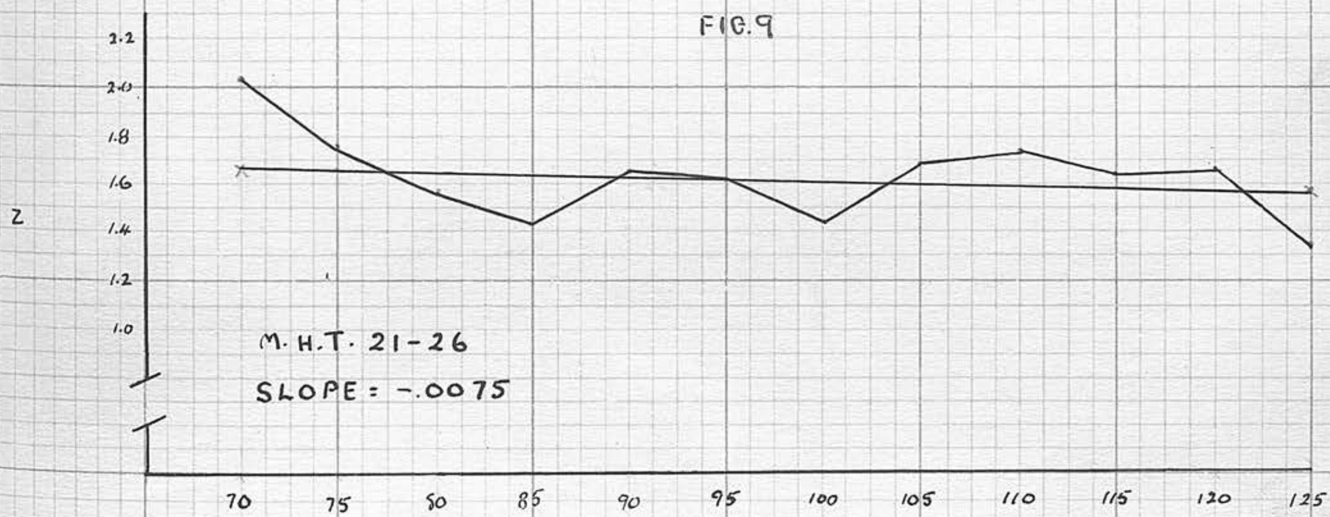
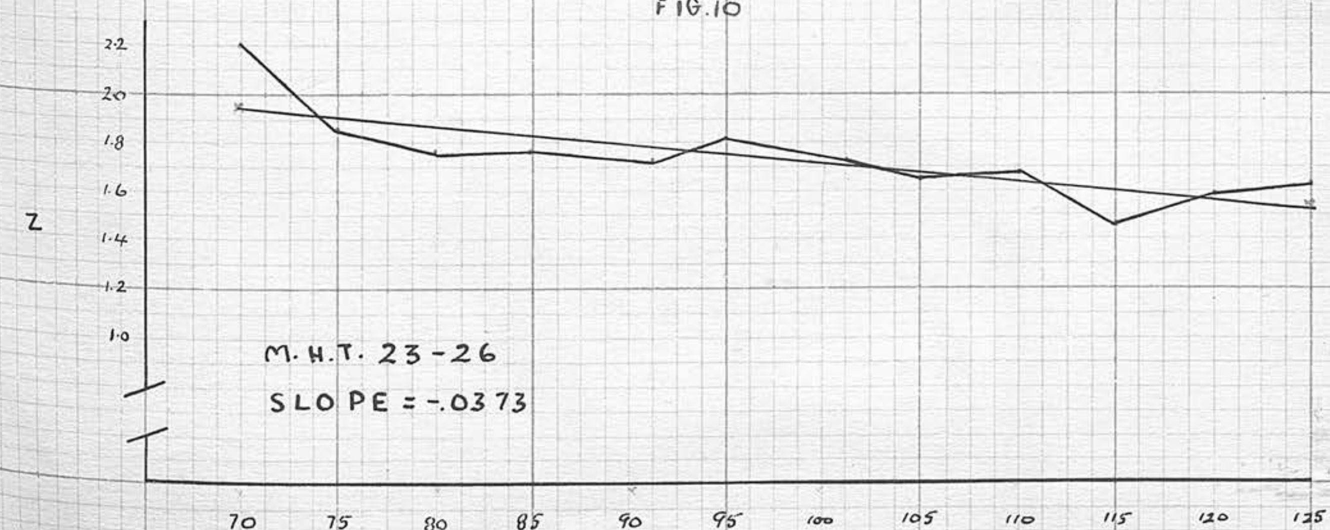


FIG.10



I·Q.

TABLE 24

RELIABILITY COEFFICIENTS CALCULATED AT

VARIOUS LEVELS OF ABILITY.

GROUP A.

I.Q. Level	rll M.H.T. 21/23	rll M.H.T. 21/26	rll M.H.T. 23/26	N.
70-74.5	.973	.959	.983	38
75-79.5	.939	.941	.963	48
80-84.5	.936	.934	.941	78
85-89.5	.935	.900	.942	100
90-94.5	.933	.923	.940	128
95-99.5	.932	.920	.945	135
100-104.5	.926	.906	.942	111
105-109.5	.928	.942	.931	100
110-114.5	.940	.934	.932	90
115-119.5	.906	.928	.876	34
120-124.5	.935	.915	.952	32
125-129.5	.965	.901	.941	12

by formula $1 - \frac{\sigma_D^2}{2\Sigma^2}$

TABLE 25

RELIABILITY COEFFICIENTS CALCULATED AT VARIOUS LEVELS OF ABILITY.

GROUP B.

I.Q. Level	rll M.H.T. 21/23	rll M.H.T. 21/26	rll M.H.T. 23/26	N
70-74.5	.971	.988	.970	13
75-79.5	.964	.966	.927	13
80-84.5	.966	.944	.962	25
85-89.5	.949	.945	.958	43
90-94.5	.932	.952	.944	83
95-99.5	.948	.958	.954	72
100-104.5	.938	.930	.938	81
105-109.5	.929	.936	.934	88
110-114.5	.950	.951	.939	68
115-119.5	.929	.932	.912	62
120-124.5	.883	.936	.910	50
125-129.5	.868	.852	.923	31

TABLE 26

TABLE SHOWING DECREASE IN RELIABILITY WITH INCREASE IN ABILITY

Groups A and B combined for M.H.T. 21/23.

I.Q. level	r	r	Smoothed values of r	Values of z	Smoothed values of z	N.
125-130	.891	.0314	.902	1.425	1.485	43
120-124	.898	.0214	.906	1.462	1.509	82
115-119	.913	.0170	.911	1.545	1.534	96
110-114	.936	.0099	.915	1.705	1.559	158
105-109	.910	.0125	.919	1.528	1.584	188
100-104	.912	.0121	.923	1.540	1.609	192
95-99	.923	.0103	.927	1.609	1.633	207
90-94	.916	.0111	.930	1.564	1.658	211
85-89	.924	.0122	.933	1.617	1.683	143
80-84	.929	.0135	.936	1.650	1.708	103
75-79	.935	.0161	.939	1.694	1.733	61
70-74	.971	.0080	.942	2.110	1.757	51

TABLE 27

TABLE SHOWING DECREASE IN RELIABILITY WITH INCREASE IN ABILITY.

Groups A and B combined for M.H.T. 21/26.

I.Q. Level	r	r	Smoothed Values of r	Values of z	Smoothed values of z	N
125-130	.869	.0373	.919	1.329 ~~1.581~~	1.581 ~~1.429~~	43
120-124	.928	.0153	.920	1.644 ~~1.588~~	1.588 ~~1.464~~	82
115-119	.926	.0145	.921	1.630 ~~1.596~~	1.596 ~~1.499~~	96
110-114	.938	.0096	.922	1.720 ~~1.603~~	1.603 ~~1.534~~	158
105-109	.934	.0093	.923	1.689 ~~1.610~~	1.610 ~~1.569~~	188
100-104	.905	.0131	.924	1.410 ~~1.618~~	1.618 ~~1.604~~	192
95-99	.923	.0103	.925	1.609 ~~1.626~~	1.626 ~~1.639~~	207
90-94	.928	.0096	.926	1.644 ~~1.633~~	1.633 ~~1.674~~	211
85-89	.905	.0151	.928	1.410 ~~1.640~~	1.640 ~~1.709~~	143
80-84	.919	.0153	.929	1.582 ~~1.648~~	1.648 ~~1.743~~	103
75-79	.943	.0142	.930	1.765 ~~1.656~~	1.656 ~~1.779~~	61
70-74	.966	.0094	.931	2.030 ~~1.663~~	1.663 ~~1.814~~	51

TABLE 28

TABLE SHOWING DECREASE IN RELIABILITY WITH INCREASE IN ABILITY.

Group (A and B) M.H.T. 23/26

I.Q. Level	r	r	Smoothed Values of r	Values of z	Smoothed Values of z	N
125-130	.927	.0215	.912	1.635	1.541	43
120-124	.921	.0168	.918	1.596	1.578	82
115-119	.899	.0196	.924	1.465	1.615	96
110-114	.934	.0102	.929	1.689	1.653	158
105-109	.930	.0099	.934	1.658	1.690	188
100-104	.940	.0084	.938	1.738	1.727	192
95-99	.947	.0072	.943	1.802	1.765	207
90-94	.937	.0084	.947	1.713	1.802	211
85-89	.946	.0088	.951	1.791	1.839	143
80-84	.946	.0104	.954	1.791	1.877	103
75-79	.954	.0115	.957	1.875	1.914	61
70-74	.980	.0055	.960	2.200	1.951	51

TABLE 29

TABLE OF STANDARD ERRORS OF I.Q. AT DIFFERENT LEVELS OF ABILITY. - GROUP A.

I.Q. LEVEL	M.H.T. 21/23	M.H.T. 21/26	M.H.T. 23/26
125-130	2.0887	4.7202	3.6443
120-124	3.3212	4.3714	3.2844
115-119	4.5005	4.0373	5.2849
110-114	3.6619	3.8660	3.9129
105-109	4.0374	3.6189	3.9331
100-104	4.0718	4.6023	3.6187
95-99	3.9074	4.2485	3.5351
90-94	3.3709	4.1724	3.6829
85-89	3.8222	4.7359	3.6278
80-84	3.7932	3.8483	3.6318
75-79	3.6923	3.6324	2.8966
70-74	2.4570	3.0326	1.9522

TABLE 30

TABLE OF STANDARD ERRORS OF I.Q. AT DIFFERENT LEVELS OF ABILITY. - GROUP A.

I.Q. LEVEL	M.H.T. 21/23	M.H.T. 21/26	M.H.T. 23/26
125-130	5.4576	5.7667	4.1579
120-124	5.1359	3.7816	4.5088
115-119	3.9897	3.9033	4.4518
110-114	3.3396	3.3083	3.6925
105-109	3.9935	3.8052	3.8648
100-104	3.7481	3.9681	3.7371
95-99	3.4321	3.0809	3.2305
90-94	3.9006	3.2998	3.5591
85-89	3.3835	3.5194	3.0898
80-84	2.7814	3.5450	2.9191
75-79	2.8520	2.7740	4.0548
70-74	2.5514	1.6639	2.5980

TABLE 31

TABLE OF STANDARD ERROR OF I.Q. AT DIFFERENT LEVELS OF ABILITY.

GROUPS A AND B.

INTERVAL.	M.H.T. 21/23	M.H.T. 21/26	M.H.T. 23/26
125-130	4.9492	5.4235	4.0423
120-124	4.7835	4.0319	4.2192
115-119	4.4307	4.0673	4.7778
110-114	3.7966	3.7394	3.8495
105-109	4.5118	3.8438	3.9759
100-104	4.4439	4.6310	3.6813
95-99	4.1600	4.1626	3.4550
90-94	4.3565	4.0319	3.7652
85-89	4.1259	4.6123	3.4953
80-84	4.0086	4.2695	3.4747
75-79	3.8103	3.5723	3.2142
70-74	2.5390	2.7668	2.1410

A NOTE ON RELIABILITY AND SELECTION

A NOTE ON RELIABILITY AND SELECTION.

Throughout the investigations described in the present thesis reliability coefficients have been estimated by the formula

$$R_{11'} = 1 - \frac{\sigma_D^2}{2\Sigma^2} \qquad (1)$$

$R_{11'}$ = reliability coefficient in unselected population.

σ_D^2 = the variance of the difference in I.Q., A.Q. or E.Q. between test and retest.

Σ^2 = the variance of I.Q., A.Q. and E.Q. in the unselected population (with all Moray House Tests $\Sigma = 15$).

If the variance of the differences between test and retest is calculated by the diagonal adding method it must be corrected twice by Sheppard's correction in order to furnish a best estimate of $R_{11'}$. If the variance, σ_D^2, is calculated by subtracting the actual quotients, and grouping in a convenient number of categories the usual form of Sheppard's correction is applied.

It may be demonstrated that the Otis-Kelley formula is a derivative of formula (1). The Otis-Kelley formula is usually written

$$\frac{\sigma_1^2}{\Sigma_1^2} = \frac{1 - R_{11'}}{1 - r_{11'}} \qquad (2)$$

where $\sigma_{1'}^2 =$ the variance of the test, whose reliability is being estimated, in the selected population.

$\Sigma_1^2 =$ the variance of the same test in the unselected population

$R_{11'} =$ the reliability coefficient found for the unselected population.

$r_{11'} =$ the reliability coefficient for the selected population.

Transposing formula (2) we have

$$R_{11'} = 1 - \frac{\sigma_1^2}{\Sigma_1}(1 - r_{11'})$$

but

$$\sigma_D^2 = \sigma_1^2 + \sigma_{1'}^2 - 2r_{11'}\sigma_1\sigma_{1'}$$

where $\sigma_D^2 =$ the variance of the differences between the two tests in the selected population.

Since the Otis-Kelley formula assumes that $\sigma_1^2 = \sigma_{1'}^2$, then

$$\sigma_D^2 = 2\sigma^2(1 - r_{11'})$$

therefore

$$R_{11'} = 1 - \frac{\sigma_D^2}{2\Sigma^2}$$

The above relationship should be fairly obvious given the knowledge that the standard error of a test score, formula $\varepsilon_1 = \sigma_1\sqrt{1 - r_{11'}}$ is independent of selection. Formula (1) may also be derived from the formula for the standard error of a test score.

It may be demonstrated also that the formula

$$R_{11'} = 1 - \frac{\sigma_D^2}{2\Sigma_1^2}$$

is independent of selection when $\sigma_1^2 \neq \sigma_{1'}^2$, but $\Sigma_1^2 = \Sigma_{1'}^2$.

Since

$$\Sigma_D^2 = \Sigma_1^2 + \Sigma_{1'}^2 - 2R_{11'}\Sigma_1\Sigma_{1'}$$

$$= 2\Sigma^2(1 - R_{11'}) \qquad (3)$$

and

$$\sigma_D^2 = \sigma_1^2 + \sigma_{1'}^2 - 2r_{11'}\sigma_1\sigma_{1'} \qquad (4)$$

and since σ_D^2 and Σ_D^2 are due to chance errors of measurement, and unrelated to the degree of selection we may write $\sigma_D^2 = \Sigma_D^2$. Thus σ_D^2 estimated from a selected population may be used as the best available estimate of Σ_D^2 in the unselected population. Equating (3) and (4) we have

$$R_{11'} = 1 - \frac{1}{2\Sigma^2}\left(\sigma_1^2 + \sigma_{1'}^2 - 2r_{11'}\sigma_1\sigma_{1'}\right)$$

$$= 1 - \frac{\sigma_D^2}{2\Sigma^2}$$

Thus the conclusion is reached that if $\sigma_1^2 \neq \sigma_{1'}^2$ formula (1) is still valid. In the majority of reliability coefficients given in this thesis it is unlikely that $\sigma_1^2 = \sigma_{1'}^2$, although we are justified in the assumption that $\Sigma_1^2 = \Sigma_{1'}^2 = 225$, since the tests used were standardised on that basis.

In summary we may state that formula (1) is useful in the estimation of test reliability because (a) it automatically corrects for selection when $\sigma_1^2 = \sigma_{1'}^2$, and when $\sigma_1^2 \neq \sigma_{1'}^2$, (b) it short circuits the computation of a large number of unnecessary statistical parameters, and eliminates much arithmetical labour.

A COMPARISON OF THE RELIABILITY OF MORAY HOUSE TESTS USED IN THE PRESENT ENQUIRY WITH THE RELIABILITY OF THE NEW TERMAN REVISION OF THE STANFORD-BINET SCALE

<u>A comparison of the reliability of the Moray House Tests used in the present enquiry with the reliability of the new Terman Revision of the Stanford Binet Scale.</u>

Terman and Merrill in the statistical introduction of "Measuring Intelligence" furnish the only available data on the reliability of the new Terman Revision of the Stanford Binet Scale. The methods used by these investigators of calculating reliability coefficients are similar to the methods used in the present enquiry. The two parallel forms of the Binet Scale, forms M and L, were given to the same group of children with a time interval of less than a week between the two testings. The children tested were classified into brightness categories of 20 points of I.Q. The average difference in I.Q. (mean absolute deviation of I.Q.) was calculated for each 20 point I.Q. category. The standard deviations of differences in I.Q. were calculated by dividing the average differences by .7979. Standard errors of I.Q. were calculated at each brightness level by dividing the standard deviations of differences in I.Q. by $\sqrt{2}$. Reliability coefficients were then found by substituting the values of the calculated standard errors of I.Q. in the formula for the standard error of a test score

<u>NOTE.</u> Some doubt exists as to whether the method outlined above is exactly that used by Terman and Merrill. Their figures check exactly with the method given above, although they may have used a slight variation of it.

and solving for r_{11}, using 16.5 as the standard deviation of I.Q.

The following table gives Terman and Merrill's values for average differences in I.Q., standard errors of I.Q., probably errors of I.Q., and reliability coefficients for the new Revision of the Binet Scale at different brightness levels

I.Q. Level	Ave. diff.	S.E.	P.E.	Reliability Coefficients	N
130 and over	5.92	5.24	3.54	.898	154
110-129	5.55	4.92	3.29	.910	872
90-109	5.09	4.51	3.04	.924	1291
70-89	4.35	3.85	2.60	.945	477
below 70	2.49	2.21	1.49	.982	57

An examination of the reliability coefficients given in the above table indicates that the New Stanford Binet is more reliable at the lower than at the upper levels of intelligence. Therefore no unique reliability coefficient exists for this test. This lack of uniqueness in the reliability coefficient is somewhat more pronounced in Terman and Merrill's data than in the data already presented for Moray House Tests.

Table 32 gives reliability coefficients and standard errors of I.Q. for Moray House Tests for categories corresponding to those used by Terman and Merrill in calculating reliability coefficients for the New Revision of the Binet Scale. These reliability coefficients for Moray House Tests are strictly comparable with those found for the New Revision of the Binet Scale.

(1) In each case parallel forms of the same test ~~was~~ were used in the estimation of reliability.

(2) The method of estimation is the same in each case.

(3) The time interval between the application of the two parallel forms is approximately the same. (In the case of the Binet less than one week, in the present enquiry exactly one week)

(4) Both sets of reliability coefficients are based on fairly large samples of the population.

Since in our enqiury into the reliability of the Moray House Tests, children with I.Q.'s above 130 and below 70 were deleted, a comparison of reliabilities can be made only for categories between these limits.

A comparison of the reliability coefficients for Moray House Tests with those for the New Revision of the Binet Scale indicates that there is little or no difference between the reliabilities of these two tests.

The only apparent difference is that Moray House Tests seem to be slightly more reliable at the upper levels of ability than the Binet Scale, and slightly less reliable at the lower levels of ability, that is the increase in reliability with decrease in ability is more pronounced for the Binet Scale than for Moray House Tests.

Educationists and psychologists have frequently made the tacit assumption that individual tests were more reliable instruments in the measurement of mental capacity than group tests. This assumption in favour of individual tests on grounds of their higher reliability is unwarranted, as this investigation has demonstrated that group tests of intelligence of the Moray House type are as reliable as the New Revision of the Binet Scale, generally recognised as the most reliable individual test of intelligence constructed thus far. Furthermore, there is some evidence to indicate that later Moray House Tests are more reliable than the tests used in this enquiry, and that with improved techniques of item selection employed in the construction of later tests the reliability may be still further INCREASED ~~invalid~~.

TABLE 32

Table of reliability coefficients for Moray House Tests at different levels of intelligence. Values of the standard deviation of variation in I.Q., and standard error of I.Q. are also given.

Test M.H.T.	I.Q. Level	Reliability Coefficient	S.D.$_d$	S.E. I.Q.	N
21-23	above 110	.916	6.1560	4.3529	379
21-23	90-110	.915	6.1938	4.3796	798
21-23	below 90	.933	5.4786	3.8739	358
21-26	above 110	.924	5.8488	4.1357	379
21-26	90-110	.930	5.6045	3.9629	798
21-26	below 90	.922	5.9075	4.1772	358
23-26	above 110	.921	5.9573	4.2124	379
23-26	90-110	.902	6.6390	4.6944	798
23-26	below 90	.954	4.5351	3.2068	358

THE CONSTANCY OF THE INTELLIGENCE QUOTIENT

The Constancy of the Intelligence Quotient.

The problem of the constancy of the Intelligence Quotient is closely associated with test reliability. Indeed, some difficulty exists in discriminating adequately between the two concepts. There exists, however, implicit in the idea of I.Q. constancy some conception of a time factor over which the abilities designated as intelligence, may, or may not, vary, which idea is not implicit in the usual definitions of reliability. Psychologists display a tendency to regard a reliability coefficient as a term purely descriptive of test efficiency, but as we have attempted to make clear elsewhere in this thesis we cannot dissociate altogether test reliability from trait reliability. It is true that we can estimate roughly what the reliability of a test would be had the trait tested been perfectly reliable, but a number of considerations render a convenient accurate estimate of reliability coefficients of this type difficult to attain. Since the majority of intelligence tests are prognostic in character, and are used as predictive indices of future behaviour, it is essential that some quantitative determination of the constancy or variability of the abilities measured by them be reached. Obviously if the I.Q. is seriously influenced by education and environmental conditions its value as a prognostic index will be considerably impaired.

Hitherto extensive research has been carried out to determine the constancy of the Stanford Binet I.Q. (old revision). These experiments have usually taken the form of testing a number of children twice with a time interval between the successive testings, and interpreting the results either by the correlation between test and retest (that is in terms of a reliability coefficient overlaid with trait unreliability) or by some measure of dispersion such as the mean absolute deviation or standard deviation applied to the I.Q. differences between initial and successive tests.

Unfortunately these investigations on the constancy of the Stanford Binet I.Q. were conducted by a miscellany of investigators, each investigator working with relatively small samples, and with different time intervals. Furthermore, the statistical interpretations of the results obtained is not in all cases admirable. Frequently, failure to correct obtained coefficients for selection, renders a comparison of the results of different investigators invalid.

Few investigators have occupied themselves with problems associated with the constancy of I.Q. as measured by Group tests of intelligence. The increasing large scale use of group tests by Education Authorities in selecting children for different types of secondary education, and indeed the increasing importance of the prognostic decisions based on the results of group tests indicates that the constancy of of the group I.Q. is a problem of considerably more practical importance and interest at the present time to the educationist than the problem of the constancy of the Stanford Binet I.Q.. Practical considerations render the use of individual tests for educational selection impossible.

Retests with Group Tests of Intelligence after a Time Interval of Seven Weeks.

Data for an investigation into the constancy of the group Intelligence Quotient was furnished by the Doncaster Education Authority. Doncaster as part of their procedure in selecting candidates for special places in secondary schools had administered two intelligence tests, Moray House Tests 24 and 26, to a complete year group of 11 year olds with a time interval between the testings of roughly seven weeks. M.H.T 24 was administered on February 3rd., 1939 and M.H.T. 26 on 31st. March, 1939.

The tests were standardised at Moray House by the usual method, care being taken to make the necessary allowance in the standardisation for those 11 year old children who had received special places during the 1938 examination as 10 year olds. This technique is known as replacing the cream.

The differences in I.Q. between the first and second testings were calculated for each child, and these differences grouped in 5 point I.Q. intervals as estimated by the first test, M.H.T.24 From these distributions of I.Q. differences at five point I.Q. levels of ability, standard deviations, reliability coefficients, and other parameters were calculated.

DISTRIBUTION OF I.Q. VARIATION.

The distributions of I.Q. variation at each 5 point I.Q. level are given in table 33. The distributions of variation in I.Q. for boys and girls separately, and for boys and girls combined, are given in Table 34. The two tests were given to 500 boys, and 530 girls, 1030 candidates in all. The standard deviation of variation in I.Q. for boys was found to be 5.325 (N=500), and for girls 5.330 (N=530). No significant difference exists between the I.Q. variability of boys and girls. The standard deviation of variation in I.Q. for boys and girls combined was 5.316 (N=1030). the reliability coefficients found over this seven weeks interval, calculated by the formula

$$r_{11'} = 1 - \frac{\sigma^2_{(1-1')}}{2\sigma^2}$$

when $\sigma = 15$ was found to be .9370 for boys, .9369 for girls, and .9372 for boys and girls combined. We may conclude from these calculations that the I.Q's calculated by the tests used have exhibited a very high degree of constancy over the time interval of seven weeks.

VARIATION IN I.Q. RELATIVE TO LEVEL OF ABILITY.

The standard deviation s of variations in I.Q. were calculated at each 5 point I.Q. level of ability. Standard errors of I.Q. were also calculated by dividing the standard deviation of variation in I.Q. obtained at each I.Q. level by 2. These standard deviations of variation and standard errors of I.Q. are given in Table 35, together with the number of cases upon which each parameter is based.

Reliability coefficients were calculated at each I.Q. level. These reliability coefficients with their standard errors are given in Table 36. Examination of these coefficients suggest that the I.Q. tends to be slightly more constant at the lower than at the upper ranges of intelligence. To test this hypothesis the coefficients attained were converted into z scores by Fisher's Tables. Each z score was given a weight equal to the reciprocal of its variance, that is (N-3). A least square line was fitted to the series of weighed points thus obtained. The slope of the best fitting least square line was found to be -.0421. This slope has a standard error of .0114. The equation of the best fitting least square line is

$$z = 1.7426 - .0421a$$

where a represents any given level of ability measured from the mean.

We may conclude from the above data that the tendency for the I.Q. to be more variable at the upper than at the lower ranges of ability is significant. Smoothed values of z were obtained, and the values of z converted into smoothed values of r. Values of z, smoothed values of z, and smoothed values of r are given in Table 36.

118

TABLE 33

Distributions of Variation in I.Q. at Different Levels of Ability, Doncaster Data, Interval Seven Weeks.

Int.	-70	70-	75-	80-	85-	90-	95-	100-	105 -	110-	115-	120-	125-	130-
20	-	-	-	-	-	-	-	-	-	-	-	-	-	-
19	-	-	-	-	-	-	-	1	-	-	-	-	-	-
18	-	-	-	-	-	-	-	-	-	-	-	-	-	-
17	-	-	-	-	-	1	1	-	-	1	-	-	-	-
16	-	-	-	-	-	-	-	1	-	-	-	-	-	-
15	-	1	-	-	1	-	1	-	-	-	1	-	-	-
14	1	-	-	-	2	-	1	-	1	2	1	-	-	-
13	-	-	-	-	-	1	-	-	1	2	-	-	-	1
12	-	-	-	-	-	1	1	1	-	2	1	-	-	1
11	1	-	-	-	1	-	-	3	1	-	1	1	1	-
10	1	-	1	1	1	3	4	2	1	4	1	2	1	0
9	-	1	3	-	2	2	3	3	3	4	2	2	1	-
8	-	-	3	1	6	6	5	3	6	2	-	5	-	-
7	1	-	1	3	3	2	2	3	6	6	3	1	2	-
6	1	3	3	1	1	9	6	7	4	2	1	2	-	-
5	-	2	7	1	4	4	8	7	13	6	6	2	-	-
4	1	-	4	3	8	7	6	4	9	7	4	5	5	1
3	3	1	5	6	5	13	12	8	7	7	1	6	3	-
2	2	2	5	5	7	11	15	14	8	7	7	2	2	-
1	2	-	4	5	4	7	8	7	2	10	5	1	1	1
0	3	5	6	8	8	17	17	10	10	9	2	4	6	1
-1	1	1	2	5	2	4	4	10	13	2	3	1	3	2
-2	1	1	1	6	7	9	10	14	3	4	8	3	3	1
-3	1	-	3	4	3	8	8	11	13	7	7	3	1	1
-4	-	1	-	-	2	6	8	6	3	13	2	3	-	3
-5	1	-	-	2	4	7	3	3	5	4	1	2	3	-
-6	-	1	1	-	-	1	2	4	3	5	4	3	-	-
-7	-	-	1	1	1	2	4	4	3	6	4	5	1	1
-8	-	-	-	1	2	1	1	4	3	6	1	1	1	-
-9	-	-	-	1	-	1	-	3	-	2	-	1	1	-
-10	-	-	-	-	1	-	-	-	1	-	1	2	-	1
-11	-	-	-	-	-	-	-	-	-	-	1	-	-	-
-12	-	-	-	-	-	1	-	-	-	-	4	-	-	-
-13	-	-	-	-	-	-	-	-	-	-	1	2	-	-
-14	-	-	-	-	1	-	-	-	-	-	-	-	-	-
-15	-	-	-	-	-	-	-	1	-	2	-	1	-	-
-16	-	-	-	-	-	-	-	-	-	-	-	-	-	-

TABLE 34

DISTRIBUTIONS OF DIFFERENCES IN I.Q.

Doncaster Data, M.H.T. 24/26.

I.Q. diff.	Girls	Boys	Total
19	1	0	1
18	2	0	2
17	0	1	1
16	0	2	2
15	0	3	3
14	4	4	8
13	3	2	5
12	4	3	7
11	4	5	9
10	9	13	22
9	11	15	26
8	18	19	37
7	15	18	33
6	19	21	40
5	32	28	60
4	31	33	64
3	34	43	77
2	39	48	87
1	31	26	57
0	53	53	106
-1	23	30	53
-2	36	35	71
-3	34	36	70
-4	28	19	47
-5	16	19	35
-6	8	16	24
-7	19	14	33
-8	11	10	21
-9	4	5	9
-10	4	2	6
-11	1	0	1
-12	3	2	5
-13	1	2	3
-14	0	1	1
-15	2	2	4
	500	530	1030
S.D.	5.325	5.330	5.316

TABLE 35

Table of Standard Deviations of Variations in I.Q. at Different Levels of Ability with Standard Errors of I.Q. Doncaster Data.

I.Q. Range	$S.D._d$	$S.E._{I.Q.}$	N
70-	4.6665	3.2997	20
70-74	4.8729	3.4456	19
75-79	3.7995	2.6868	50
80-84	3.8116	2.6952	54
85-89	5.4560	3.8579	76
90-94	4.7672	3.3709	124
95-99	4.7263	3.3420	130
100-104	5.3692	3.7966	134
105-109	4.8949	3.4612	119
110-114	6.2023	4.3856	122
115-119	6.3033	4.4571	73
120-124	6.4678	4.5734	60
125-129	4.7366	3.3492	35
130-	6.2002	4.3842	14

TABLE 36

Table Showing Decrease in Reliability with Increase in Ability. Doncaster Data. M.H.T. 24/26. Interval 7 Weeks.

I.Q. Level	r	$S.E._r$	Smoothed values of r	Values of z	Smoothed Values z	N
70-	.952	.0211	.965	1.852	2.016	20
70-74	.947	.0236	.962	1.702	1.974	19
75-79	.968	.0089	.959	2.060	1.932	50
80-84	.968	.0087	.955	2.060	1.890	54
85-89	.934	.0147	.951	1.689	1.848	76
90-94	.950	.0088	.947	1.831	1.806	124
95-99	.950	.0085	.943	1.831	1.764	130
100-104	.936	.0107	.938	1.705	1.722	134
105-109	.947	.0095	.933	1.702	1.679	119
110-114	.915	.0148	.927	1.559	1.637	122
115-119	.912	.0197	.921	1.540	1.595	73
120-124	.907	.0229	.914	1.476	1.553	60
125-129	.950	.0164	.907	1.831	1.511	35
130-	.915	.0229	.899	1.559	1.469	14

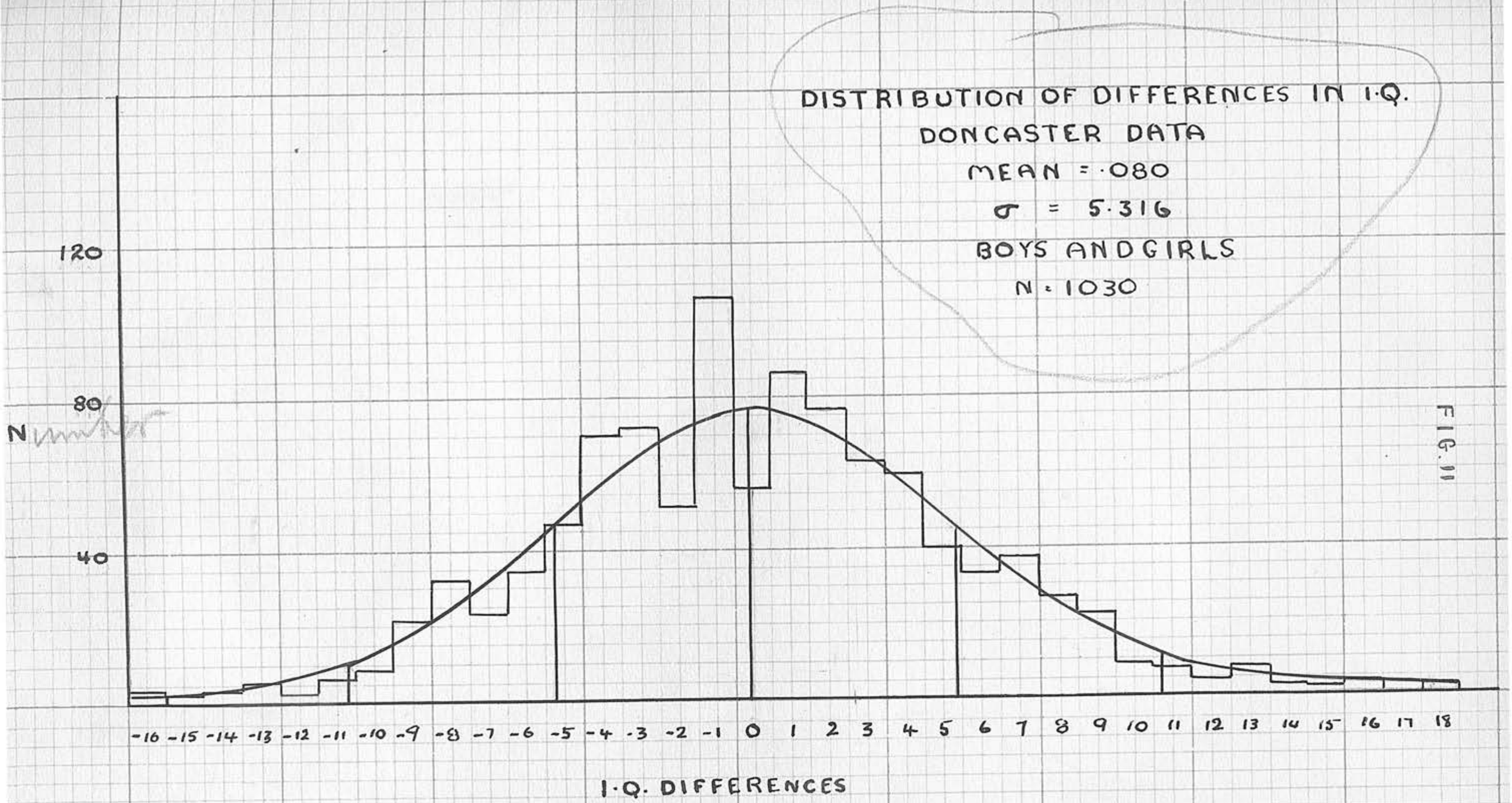

DISTRIBUTION OF DIFFERENCES IN I.Q.
DONCASTER DATA
MEAN = ·080
σ = 5·316
BOYS AND GIRLS
N = 1030
120
80
40
N
-16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18
I.Q. DIFFERENCES

FIG. 11

THE CONSTANCY OF THE GROUP I.Q. OVER LONGER TIME INTERVALS.

Some data are available relative to the constancy of Intelligence Quotients as measured by Group Tests of Intelligence over time intervals ranging from 15 to 38 months. These data have been studied and presented as a thesis for the Degree of Bachelor of Education at the University of Edinburgh. A brief summary of these results is given here to render the findings of the present enquiry more complete.

THE CONSTANCY OF THE GROUP I.Q. OVER LONGER TIME INTERVALS

Two Moray House Group Tests of Intelligence were administered to 952 children in Northumberland with varying time intervals between the successive testings. Three Groups took part in the experiment.

(1) 394 children who had been tested with a Moray House Test at 11+ in 1934, and who were retested at 14+ in 1937.

(2) 363 children who had been tested with a Moray House Test at 11+ in 1935, and who were retested at 13+ in 1937.

(3) 195 pupils who had been tested with a Moray House Test at 11+ in 1936, and who were retested at 12+ in 1937.

Differences in I.Q. between test and retest were calculated for each Group, and normal curves fitted to the distributions of differences thus obtained. Pearson's formulae with Sheppard's corrections were used in the estimations of values of B_1 and B_2. The results for the three groups are as follows:-

THE CONSTANCY OF THE GROUP I.Q. OVER LONGER TIME INTERVALS.

Some data are available relative to the constancy of Intelligence Quotients as measured by Group Tests of Intelligence over time intervals ranging from 15 to 38 months. These data have been studied and presented as a thesis for the Degree of Bachelor of Education at the University of Edinburgh. A brief summary of these results is given here to render the findings of the present enquiry more complete.

Two Moray House Group Tests of Intelligence were administered to 952 children in Northumberland with varying time intervals between the successive testings. Three Groups took part in the experiment.

(1) 394 children who had been tested with a Moray House Test at 11+ in 1934, and who were retested at 14+ in 1937.

(2) 363 children who had been tested with a Moray House Test at 11+ in 1935, and who were retested at 13+ in 1937.

(3) 195 pupils who had been tested with a Moray House Test at 11+ in 1936, and who were retested at 12+ in 1937.

Differences in I.Q. between test and retest were calculated for each Group, and normal curves fitted to the distributions of differences thus obtained. Pearson's formulae with Sheppard's corrections were used in the estimations of values of B_1 and B_2. The results for the three groups are as follows:-

	B_1	B_2	t	N
Group 1	.0000	3.044	15 months	394
Group 2	.0395	2.958	26 months	363
Group 3	.0000	2.337	38 months	195

In no case does B_1 differ significantly from zero, or B_2 from 3. Consequently we may conclude that the normal curve of errors describes with considerable accuracy variations in I.Q. from test to retest, and that no systematic factor is operating in causing the discrepancies between I.Q's as measured by these tests.

The standard deviations of the differences in I.Q. between test and retest were calculated for each group; also the correlation between test and retest. The standard deviation of the differences in I.Q. for each group, and the correlations between test and retest are as follows:-

	$S.D._d$	r_{11}	t	N
Group 1	5.42	.912	15 months	394
Group 2	5.69	.895	26 months	363
Group 3	6.90	.776	38 months	195

Examination of the above parameters indicates that the correlations between test and retest varies inversely with increase in the time interval separating the testings.

Since, however, the children to which the tests were administered did not represent a complete year group, but rather a selected sample, it was necessary to correct the above coefficients for selection. The coefficients corrected for selection may be obtained by using the formula

$$r_{11'} = 1 - \frac{\sigma^2_{(1-1)}}{2\sigma^2}$$

where $\sigma = 15$. The correlation coefficients after correction for selection are as follows:-

	r
Group 1	.935
Group 2	.929
Group 3	.895

Examination of the above coefficients reveals that Intelligence Quotients as estimated by Moray House Tests display an unusual degree of constancy even over relatively long time intervals.

Table 37 gives the distributions of differences in I.Q. for each group.

A Comparison of the Constancy of the Group I.Q. with the Stanford Binet I.Q. (Old Revision).

Numerous investigators have, in the past devoted considerable attention to the constancy of the Binet I.Q. These investigations have usually taken the form of administering the Binet Scale twice to the same group of children, allowing a more or less lengthly time interval to elapse between the testings. A miscellany of techniques has rendered a valid comparison of the results of investigators in this field unusually difficult. The greatest difficulty in making a comparison results from failure on the part of many investigators to correct their obtained coefficients for selection, or to furnish information indicative of the degree of selection characterized by the groups tested.

Examination of the work of investigators in this field discloses that the correlation between Binet test and retest varies as some inverse function of the time interval separating the successive testings. Table 38 gives some indication of the type of results obtained over varying time intervals. This table is reproduced from an article of Robert L. Thorndike, "The Effect of the Interval between Test and Retest on the "Constancy of the I.Q.". Thorndike converted the values of r given in this Table into z scores, and fitted a least square line to the series of points thus obtained, weighting each point by the reciprocal of its variance (N-3). The equation thus obtained for the best fitting least square line was

$$z = 1.415 - .00916t$$

x Thorndike, Robert L., (1933) "The Effect of the Interval Between Test and Retest on the Constancy of I.Q." J.Educ.Psychol. xxiv, pp. 543-549.

By converting values of z thus obtained back into values of r for different values of t we obtain values of r for varying time intervals as follows (t in months):-

t	r
0	.889
10	.868
20	.843
30	.814
40	.781
50	.743
60	.698

By interpolation we can find the correlation after an interval of 15 months, 26 months, and 38 months. A comparison of these correlations with the correlations between successive applications of Moray House Tests is given below.

t	Binet r	M.H.T. r	uncorrected
15	.856	.935	.912
26	.826	.929	.895
38	.788	.895	.776

These correlations imply that I.Q.'s as estimated by Moray House Tests exhibit greater constancy than I.Q's as measured by the Old Revision of the Binet Scale. page 226

Although the comparison made here seems to be greatly to the advantage of Moray House Tests, it is necessary in all fairness to the Binet Scale to bear in mind that this favourable comparison is to some extent at least invalidated by lack of information concerning the degree of ~~solution~~ selection of the groups tested by experimentees on the constancy of the Binet I.Q. Underselection, however, may in part be counteracted by the fact that certain investigators report in their experiments a variance of Binet I.Q. for the group tested greater than the known variance of Binet I.Q. in a represtative population.

TABLE 37

DISTRIBUTIONS OF DIFFERENCES IN I.Q. AT

VARIOUS TIME INTERVALS.

MORAY HOUSE TESTS.

I.Q. diff.	Group 1 Interval 15 months.	Group 2 Interval 26 months.	Group 3 Interval 38 months.
16.5	0	1	1
15.0	2	3	3
13.5	2	2	4
12.0	5	4	7
10.5	8	10	4
9.0	11	7	9
7.5	18	12	13
6.0	21	19	15
4.5	40	26	11
3.0	30	27	19
1.5	40	38	10
.0	41	47	18
-1.5	43	29	10
-3.0	41	38	20
-4.5	33	30	10
-6.0	27	26	9
-7.5	12	17	10
-9.0	6	10	9
-10.5	6	9	7
-12.0	3	3	3
-13.5	3	5	1
-15.0	0	0	2
-16.5	2	0	0
Total	394	363	195
S.D.	5.42	5.69	6.90
Mean	.194	-.28	.76

TABLE 38

RETESTS WITH THE STANFORD BINET.

Experimenter	N	t, months	r
Cuneo and Terman	25	0	.95
Lincoln	30	0	.95
Brown	221	0-12	.91
Cuneo and Terman	21	5-7	.942
Randall	103	0-18	.798
Rosenow	69	$7\frac{1}{2}$ or 11(mn.10.25)	.82
Berry	351	6-18(mn.11)	.74
Baldwin	173	12	.901
Garrison	298	12	.88
Garrison & Robinson	131	12	.88
Garrison & Robinson	131	12	.92
Gray & Marsden	100	12	.883
Gray & Marsden	42	12	.834
Rugg & Colloton	137	10-16	.84
Brown	149	14(av.)	.87
Brown	320	12-24	.87
Cuneo & Terman	31	20-24	.852
Berry	273	19-30(mn.23)	.67
Baldwin	139	24	.817
Garrison	127	24	.91
Garrison & Robinson	131	24	.91
Gray & Marsden	42	24	.839
Randall	37	19-30	.699
Brown	149	29(av.)	.70
Brown	99	24-36	.88
Gordon	44	30.7 (av.).	.84
Berry	82	31-48(av.35).	.56
Baldwin	105	36	.797
Gray & Marsden	42	36	.843
Randall	6	31-42	.793
Madsen	34	41	.85
Brown	41	36-48	.87
Baldwin	71	48	.786
Garrison	43	48	.83
Randall	6	43-66	.801
Baldwin	37	60	.812
	3840		

THE CONSTANCY OF ARITHMETIC QUOTIENTS.

DATA.

Data for an investigation into the constancy and reliability of Arithmetic Quotients as measured by Moray House Arithmetic Tests were made available by the Doncaster Education Authority. Doncaster as part of their annual examination in selecting candidates for special places in secondary schools had administered two Moray House Arithmetic Tests, M.H.A.11 and M.H.A.9, to a complete year group of over 1000 children with a time interval separating the two testings of roughly 7 weeks. M.H.A.11 was administered on 3rd. February, 1939, and M.H.A.9 on 31st. March, 1939.

THE CONSTANCY OF ARITHMETIC QUOTIENTS

TESTS USED.

The tests used in this enquiry, M.H.A.11 and M.H.A.9, are regarded as parallel forms, and have been used by many Education Authorities as part of their special places examination. Each test consists of 102 items. The first 42 items on each test are simple questions in addition, subtraction, multiplication and division. Of the first 42 items on M.H.A.11, 11 are addition, 10 subtraction, 11 multiplication and 10 division. Of the corresponding 42 items on M.H.A.9, 11 are addition, 10 subtraction, 10 multiplication and 10 division. The remaining 60 items

THE CONSTANCY OF ARITHMETIC QUOTIENTS.

DATA.

Data for an investigation into the constancy and reliability of Arithmetic Quotients as measured by Moray House Arithmetic Tests were made available by the Doncaster Education Authority. Doncaster as part of their annual examination in selecting candidates for special places in secondary schools had administered two Moray House Arithmetic Tests , M.H.A.11 and M.H.A.9, to a complete year group of over 1000 children with a time interval separating the two testings of roughly 7 weeks. M.H.A.11 was administered on 3rd. February, 1939, and M.H.A.9 on 31st. March, 1939.

TESTS USED.

The tests used in this enquiry, M.H.A.11 and M.H.A.9, are regarded as parallel forms, and have been used by many Education Authorities as part of their special places examination. Each test consists of 102 items. The first 42 items on each test are simple questions in addition, subtraction, multiplication and division. Of the first 42 items on M.H.A.11, 11 are addition, 10 subtraction, 11 multiplication and 10 division. Of the corresponding 42 items on M.H.A.9, 11 are addition, 10 subtraction, 10 multiplication and 10 division. The remaining 60 items on each test are of the problem type. The time of administration for each test is 30 minutes.

STANDARDISATION.

M.H.A.11 was standardised by Mr. W.G.Emmett at Moray House in the usual way by finding the 5th., 16th., 50th., 84th., and 95th., percentile points for each month of age separately and fitting least square lines to each set of 12 points thus found. The slopes of the percentile lines are as follows:-

	Slope
5%ile	.273
16%ile	.782
50%ile	1.680
84%ile	1.093
95%ile	1.016

The slopes of the 95th. and 50th. percentile lines appeared somewhat high when compared with corresponding slopes for the same test for Northumberland children. Consequently in the final standardisation 1.2 was used as the slope of the 95%ile line.

The second test M.H.A.9 had been obtained by the Doncaster Authority from the University of London Press, and in the determination of Arithmetic Quotients the norms furnished by the University of London Press had been used.

Consequently it was necessary for the purposes of the present investigation to restandardise the test on Doncaster children. This standardisation was carried out in the usual way. The scores of 31 candidates who, at 10+ had been awarded special places as a result of their performance in the 1938 examination were added to the final grid. The Arithmetic Quotients of these candidates on M.H.A.10 on 18th. March, 1938, were obtained. From these quotients it was possible to estimate the raw scores that would have been obtained by these candidates had they received the test at 11+ instead of 10+. The estimates thus found were used in the final standardisation.

The slopes of the appropriate percentile lines in this standardisation were found to be as follows:-

	Slope
5%ile	.364
16%ile	.532
50%ile	1.630
84%ile	.790
95%ile	.866

The 50%ile slope, 1.630, when compared with the corresponding slope for the same test for Northumberland, and also when compared with the slope used in the final standardisation of M.H.A.11A was found to be too high. Furthermore the slope for the 16%ile line appeared somewhat too small. Consequently in the final standardisation 1.2

was used as the slope of the 50%ile line, and 0.7 as the slope of the 16%ile line. The final standardisation was based on the scores of 1040 candidates, 1009 of 11+, and 31 'creamed' candidates.

MEAN CHANGE IN A.Q.

The process of standardisation is designed to eliminate any mean change in A.Q. from test to retest. Consequently we are concerned in this investigation with an examination of the variation in A.Q. from test to retest relative to the mean. The approximation of the mean change in A.Q. to zero is some indication of the efficiency of the standardisations of the two tests. The mean change in A.Q. for the total number of candidates taking both tests, 1030 in all, was found to be .187. The standard error of this mean is .127. The insignificance of this mean is one indication that the two standardisations were satisfactory. The mean change in A.Q. for boys was found to be .456 (N=500) and for girls -.066 (N=530). The ratio of the difference between these means to the standard error of the difference is 2.023.

If, however, we examine the mean difference in A.Q. from test to retest at each 5 point A.Q. level of ability we find that some of the means depart significantly from zero.

Means calculated at different levels of ability are given in Table 41 together with their standard errors, and

and the number of cases upon which each mean is based. The largest departure from zero is the mean difference at the 125-129 A.Q. level of ability, 3.829. This mean differs significantly from zero, the ratio of its departure from zero to its standard error being 4.768. The mean at the 120-124 level (A.Q) of ability also departs significantly from zero.. These departures in the mean change in A.Q. from zero must be attributed to faults in the standardisation. Departures of the mean from zero at the extreme levels of ability may be attributed to overestimation or underestimation in the extrapolation of the norms at these levels. Another source of discrepancy is the influence of sampling error upon the slopes of the percentile lines upon which the norms are based. On the whole, however, the slight departures of the means from zero at certain levels of ability is of no great importance, and does not invalidate the findings of this enquiry in any way.

PROCEDURE.

The difference in Arithmetic Quotient between the first and second testings was calculated for each child, and these differences, grouped in class interval of 1 point difference, were classified according to 5 point A.Q. levels of ability. From these distributions of A.Q. differences at 5 point A.Q. levels of ability, standard deviations, correlations and other parameters were calculated.

DISTRIBUTIONS OF A.Q. DIFFERENCES.

The distributions of A.Q. variations at each 5 point A.Q. level of ability are given in Table 40. The distributions of variation in A.Q. for boys and girls separately and for boys and girls combined are given in Table 39. The standard deviation of variation in A.Q. for boys was found to be 4.3795 (N=500) and for girls 3.8868 (N=530). The standard deviation of variation in A.Q. for the complete group (boys and girls combined) was found to be 4.1316.

The correlations found over the seven week interval calculated by the formula

$$r_{11'} = 1 - \frac{\sigma^2_{(1-1')}}{2\sigma^2}$$

where $\sigma = 15$, were found to be .9574 for boys, .9666 for girls and .9620 for the complete group.

VARIATION IN A.Q. RELATIVE TO LEVEL OF ABILITY.

The standard deviations of differences in Arithmetic Quotient were calculated at each 5 point A.Q. level of ability. Standard errors of A.Q. were calculated by dividing the standard deviation of difference in A.Q., obtained at each 5 point A.Q. level of ability, by $\sqrt{2}$. The standard deviations of differences between test and retest, and corresponding standard errors of A.Q., are given in Table 42. The number of cases upon which each parameter is based is also given.

Reliability coefficients were also calculated at each level of ability by the same method as used in calculating reliability coefficients for intelligence tests at different levels of ability. These coefficients range from .944 to .989. No reliance can be placed on this latter coefficient since it is based on only 19 cases. No general tendency can be said to exist for dull children to be more constant in their responses to the arithmetic tests used in this enquiry than bright children , no increase in test retest correlation with decrease in ability being observable.

SUMMARY.

In summary it is reasonable to conclude as a result of the above calculations that the abilities measured by Moray House Arithmetic Tests exhibit a very high degree of constancy over relatively short time intervals. Furthermore the high coefficients obtained indicate that Moray House Arithmetic Tests are very reliable.

137

TABLE 39

Distributions of Differences in Arithmetic Quotient between Test and Retest.

Diff.	Boys	Girls	Total
15			
14	1		1
13	3	1	4
12	1	0	1
11	2	0	2
10	9	3	12
9	4	5	9
8	7	6	13
7	17	13	30
6	20	11	31
5	25	23	48
4	33	31	64
3	35	40	75
2	33	49	82
1	41	43	84
0	54	57	111
-1	52	68	120
-2	36	42	78
-3	36	32	68
-4	34	45	79
-5	17	21	38
-6	14	14	28
-7	11	13	24
-8	9	5	14
-9	2	5	7
-10	1	1	2
-11	1	0	1
-12	2	1	3
	500		
mean	.456	-.066	.187
S.D.	4.3795	3.8868	4.3795

TABLE 40

Distributions of Differences in Arithmetic

Quotients at Various Levels of Ability.

Diff.	70-	70-74	75-79	80-84	85-89	90-94	95-100	100-104	105-109	110-114	115-119	120-124	125-129	130
14	-	-	-	-	-	-	-	1	-	-	-	-	-	-
13	-	-	-	2	-	-	-	0	-	-	-	1	1	-
12	-	-	-	0	-	-	-	0	-	-	-	-	1	-
11	-	-	-	0	-	-	-	0	-	-	-	-	2	-
10	-	-	-	1	4	1	2	0	-	-	-	2	2	-
9	-	-	-	1	1	0	0	2	1	2	-	1	1	-
8	-	-	-	0	1	0	2	1	1	1	2	4	1	-
7	-	-	-	1	7	1	4	2	2	5	2	3	1	2
6	-	-	-	4	2	2	3	3	4	3	2	2	3	3
5	-	-	4	1	6	3	8	6	4	4	3	3	4	2
4	-	1	2	6	5	6	8	7	8	7	10	1	2	1
3	1	2	4	1	10	12	6	8	8	9	3	7	3	1
2	2	1	2	2	10	4	11	14	11	8	10	2	3	2
1	1	0	1	5	17	11	5	14	8	8	9	5	0	0
0	5	1	4	2	13	15	15	17	14	12	4	5	2	2
-1	5	0	5	2	19	12	18	13	12	13	12	2	6	1
-2	1	1	3	6	6	10	8	17	12	8	2	1	2	1
-3	2	2	4	3	5	10	8	5	9	9	8	2	0	1
-4	0	1	4	5	9	9	9	11	12	9	9	1	0	0
-5	1	2	2	2	0	7	4	6	6	3	3	1	0	1
-6	1	1	2	3	3	4	2	3	4	3	1	0	0	1
-7	0	1	3	0	3	7	2	3	1	0	2	2	0	-
-8	-	2	0	0	0	1	1	2	1	2	3	1	1	-
-9	-	0	2	0	1	0	0	2	1	1	-	-	-	-
-10	-	0	-	1	0	0	1	0	-	-	-	-	-	-
-11	-	1	-	-	0	0	-	0	-	-	-	-	-	-
-12	-	-	-	-	1	2	-	0	-	-	-	-	-	-
-13	-	-	-	-	-	-	-	1	-	-	-	-	-	-
	19	16	42	48	123	117	117	138	119	107	85	46	35	18

TABLE 41

Table of Mean Change in A.Q. at different Levels of Ability.

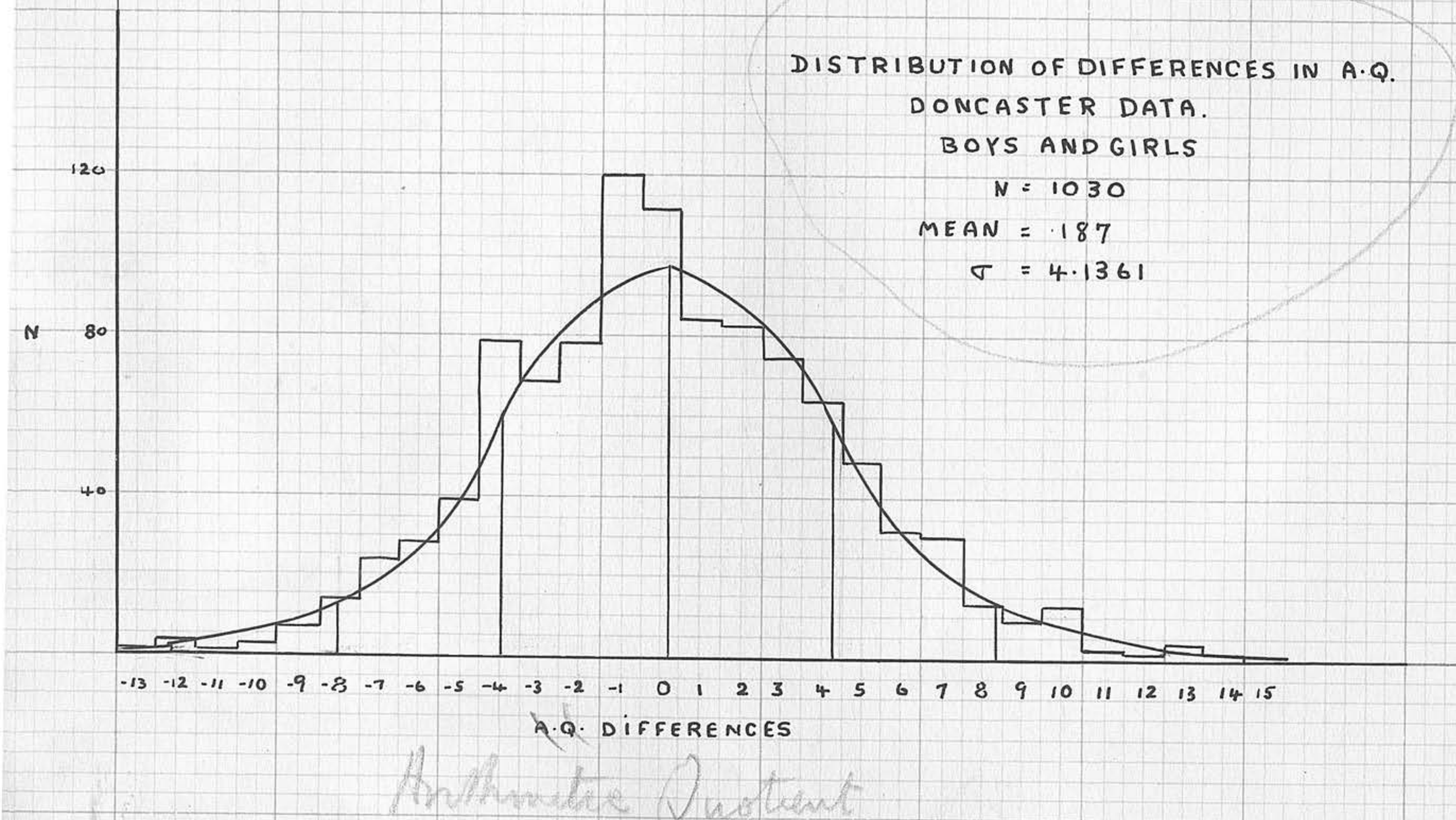
A.Q. Level.	Mean Change	S.E.$_m$	$\frac{D}{S.E._m}$	N
130+	2.000	.947	2.112	18
125-129	3.829	.803	4.768	35
120-124	2.500	.697	3.587	46
115-119	.000	.414	.000	85
110-114	.262	.369	.710	107
105-109	-.235	.330	.712	119
100-104	-.188	.338	.556	138
95-100	.350	.356	.983	117
90-94	-.983	.353	2.785	117
85-89	.756	.361	2.094	123
80-84	.750	.726	1.033	48
75-79	-.310	.607	2.158	42
70-74	-3.125	1.087	2.875	16
70-	-.842	.502	1.677	19

TABLE 42

Table of Standard Deviations of A.Q. Differences, Reliability Coefficients, and Standard Errors of A.Q. at Different Levels of Ability.

A.Q. Level.	$S.D._d$	r_{11}	S.E. E.Q.	N
130+	4.0173	.9641	2.8406	18
125-129	4.7511	.9498	3.3595	35
120-124	4.7265	.9504	3.3421	115
115-119	3.8177	.9676	2.6995	85
110-114	3.8156	.9676	2.6980	107
105-109	3.6014	.9712	2.5465	119
100-104	3.9758	.9649	2.8113	138
95-100	3.8552	.9670	2.7260	117
90-94	3.8164	.9676	2.6986	117
85-89	4.0004	.9644	2.8287	123
80-84	5.0312	.9437	3.5576	48
75-79	3.9324	.9656	2.7806	42
70-74	4.3475	.9580	3.0741	16
70-	2.1878	.9894	1.5470	19

FIG. 12


DISTRIBUTION OF DIFFERENCES IN A.Q.
DONCASTER DATA.
BOYS AND GIRLS
N = 1030
MEAN = ·187
σ = 4·1361
N
120
80
40
-13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15
A.Q. DIFFERENCES
Arithmetic Quotient

THE CONSTANCY OF ENGLISH QUOTIENTS.

DATA

The Doncaster Education Authority, while furnishing data for enquiries into the constancy and reliability of Intelligence and Arithmetic Quotients, made available additional data for an enquiry into the constancy of English Quotients. Doncaster had included as part of their special places examination, two Moray House English Attainment Tests,

THE CONSTANCY OF ENGLISH QUOTIENTS

M.H.E.11 and M.H.E.9. These two tests were administered to a complete year group of over 1000 candidates with a time interval between the two testings of roughly 7 weeks. M.H.E.11 was administered on 3rd. February, 1939, and M.H.E.9 on 31st. March, 1939.

TESTS USED.

The tests used in this investigation, M.H.E.11 and M.H.E.9, are regarded as parallel forms of the same test. Both tests have been widely used by many Educational Authorities in the selection of candidates for special places in secondary schools. M.H.E.11 consists of 150 items, M.H.E.9 of 151 items. The test items are similar in type. As with other Moray House Tests no reason exists to believe that these tests depart from a high degree of equivalence. The time of administration (40 minutes) was the same for both tests, and the method of administration the same.

THE CONSTANCY OF ENGLISH QUOTIENTS.

DATA

The Doncaster Education Authority, while furnishing data for enquiries into the constancy and reliability of Intelligence and Arithmetic Quotients, made available additional data for an enquiry into the constancy of English Quotients. Doncaster had included as part of their special places examination, two Moray House English Attainment Tests, M.H.E.11 and M.H.E.9. These two tests were administered to a complete year group of over 1000 candidates with a time interval between the two testings of roughly 7 weeks. M.H.E.11 was administered on 3rd. February, 1939, and M.H.E.9 on 31st. March, 1939.

TESTS USED.

The tests used in this investigation, M.H.E.11 and M.H.E.9, are regarded as parallel forms of the same test. Both tests have been widely used by many Educational Authorities in the selection of candidates for special places in secondary schools. M.H.E.11 consists of 150 items, M.H.E.9 of 151 items. The test items are similar in type. As with other Moray House Tests no reason exists to believe that these tests depart from a high degree of equivalence. The time of administration (40 minutes) was the same for both tests, and the method of administration the same.

STANDARDISATION.

M.H.E.11 was standardised in the customary way at Moray House by Mr. W.G.Emmett. The slopes of the 5th., 16th., 50th., 84th., and 95th. percentile lines are as follows:-

	slope
5%ile	0.775
16%ile	1.409
50%ile	1.774
84%ile	1.481
95%ile	1.440

The slopes are comparable with slopes found for the same test in other areas.

The Doncaster Authority had used norms furnished by the University of London Press in converting raw scores on M.H.E.9 into E.Q's. Consequently it was necessary to restandardise this test on Doncaster children. This was accomplished in the usual way. As in the restandardisation of M.H.A.9 the scores of 31 candidates, who, at 10+ had been awarded special places as a result of their performance in the 1938 examination were estimated, and added to the grid.

The slopes of the appropriate percentile lines on the restandardisation of M.H.E.9 were found to be as follows:-

	slope
5%ile	1.884
16%ile	1.236
50%ile	1.650
84%ile	.995
95%ile	.565

Since a comparison of these slopes with comparable slopes in other areas indicated that the slope of the 5%ile line was too high, and the 95%ile too low, due possibly to sampling error, 1.384 was used as the 5%ile slope, and .865 as the 95%ile slope, in the final standardisation.

PROCEDURE.

The procedure used in the present investigation was exactly similar to that used in studying the constancy of Intelligence and Arithmetic Quotients. The difference in English Quotient between test and retest were calculated for each child, and these differences, grouped in class intervals of 1 point E.Q. difference, were classified according to 5 point E.Q. levels of ability. From these distributions of E.Q. differences at each level of ability the necessary parameters were computed.

MEAN DIFFERENCE IN E.Q.

The process of standardisation is designed to eliminate any mean change in E.Q. from test to retest for the whole group. The mean change in E.Q. for the whole group was found to be .0184. This mean has a standard error of .127, and is obviously quite insignificant. If, however, the mean differences are calculated for each 5 point E.Q. level of ability separately, a few means are found which depart significantly from zero. Means calculated at different levels of ability are given in Table 43, together with their standard errors, and the number of cases upon which each mean is based. The largest departure from zero is the mean difference at the 125 to 129 E.Q. level of ability, 2.6181, and the next largest, -2.560, at the "below 70" E.Q. level of ability. These departures in mean difference from zero must be regarded as faults in the standardisation, the former being due either to overestimation in the extra: :polation of the norms at the upper level of ability in the second test, or underestimation at the same level in the first test, the latter figure, -2.560, being attributable to a similar fault. It would of course be possible to adjust one or other of the standardisations, or both, in order to make the mean differences more nearly zero, and therefore increase the correlation between the two tests by some very minute quantity. Such an increase, however, would seem to be spurious because (a) we cannot determine which of the standardisations is at fault (b) any estimation of test reliability must take into consideration sources of unreliability arising out of the process of standardisation itself, including faults in the norms due to sampling errors in the slopes of the different percentile lines upon which the norms are based. In a standardisation of the ordinary type no index exists whereby it may be determined whether the extrapolations of the norms furnish slight overestimates or slight underestimates of the capacity of the children tested. Furthermore, slight underestimates or overestimates in the norms at the extreme levels of ability are of little or no importance in the selection of candidates for secondary school places, the crucial level of ability being in the neighbourhood of 110 E.Q. Sampling errors in the slopes of the percentile lines, upon which the norms are based, may at times lead to quite considerable discrepancies. Errors of this type may be eliminated in some degree by a critical comparison of the obtained slopes with corresponding slopes for the same test in other areas.

DISTRIBUTIONS OF E.Q. DIFFERENCES.

The distributions of differences in E.Q. from test to retest for boys and girls separately, and for boys and girls combined are given in Table 44. The standard deviation of variation in E.Q. for boys is 4.7232 (N=500), for girls 4.3938 (N=530), and for the whole group 4.5915 (N=1030). The difference between the standard deviation for boys and that for girls is not significant, the ratio of the difference to the standard error of the difference being 1.631.

Although the mean change in E.Q. for the whole group is .0184, a figure which does not differ significantly from zero, the mean for the boys alone is .600 with a standard error of .211. The ratio of the departure of this mean from zero to its standard error is 2.84. The mean for the girls on the other hand is -.530 with a standard error of .191, the ratio of the departure of this mean from zero to its standard error being 2.77. The standard error of the difference between the means for boys and for girls is .2846. The difference between the means for boys and girls is significant, the ratio of the difference to the standard error of the difference being 3.970. If this statistic is to be relied upon we must conclude that over the seven week interval separating the application of the two English Tests the achievement in English for boys was significantly greater than for girls, a somewhat unusual conclusion. This result on the other hand may be merely a statistical curiosity.

From the standard deviation of the differences the correlations between test and retest were calculated using 15 as the standard deviation of E.Q. The correlation for boys thus calculated was found to be .9504 (N=500) and for girls .9571 (N=530). The correlation for the whole group between test and retest was found to be .9532 (N=1030). These figures adequately demonstrate that (a) English Quotients ~~as~~ estimated by Moray House Tests have exhibited a high degree of constancy over the seven week time interval; that is, the traits measured by these tests are highly reliable. (b) the tests themselves as instruments of mental measurements are highly reliable apart from the reliability of the traits measured.

VARIATION IN E.Q. RELATIVE TO BRIGHTNESS.

The standard deviation of difference in English Quotients were calculated at each 5 point E.Q. level of ability, in order to determine whether Moray House English Quotients exhibited varying constancy atvarying levels of ability. The distributions from which these standard deviations were calculated are given in Table 45. The standard deviations are given in Table 46. These standard deviations of variation in E.Q. range from 3.0599 at the "below 70" E.Q. level of ability to 4.9535 at the 85 to 89 E.Q. level of ability. No consistent increase in variability with increase in ability is apparent. Little weight can be attached to the standard deviations of variation given here for the extreme levels of ability due to the small number of cases upon which these particular parameters are based.

Correlation coefficients were calculated at each level of ability by methods used and described elsewhere in this research. These correlation coefficients range from .9455 (N=95) to .9792 (N=25). The difference between these two coefficients is not significant.

A column of standard errors of E.Q. is also given in Table 46. The standard error of a person's English Quotient is roughly 3 points of E.Q.

SUMMARY.

(1) The correlation between two Moray House English Tests, M.H.E.9 and M.H.E.11, after a time interval of seven weeks yielded the high coefficient of .9532 (N=1030) in a complete population. This correlation must be regarded as highly satisfactory, and is indicative that (a) Moray House English Quotients are remarkably constant over relatively short time intervals. (b) the tests used are themselves highly reliable.

(2) No uniform and general tendency is apparent, indicating that the abilities measured by these tests are less variable in dull than in bright children.

(3) The standard error of a person's English Quotient is approximately 3 points of E.Q.

169

TABLE 43

Table of Mean Change in E.Q. at Different Levels of Ability.

E.Q. Level	Mean Range	$S.E._m$	$\frac{D}{S.E._m}$	N
130+	.750	1.227	.611	16
125-129	2.168	.634	4.129	34
120-124	.407	.673	.605	54
115-119	1.250	.477	2.621	76
110-114	.074	.427	.173	121
105-109	1.0170	.448	2.270	117
100-104	-.2263	.369	.613	137
95-99	-.5203	.417	1.248	123
90-94	-.8140	.393	2.071	118
85-89	-.2840	.508	.559	95
80-84	-.5080	.523	.971	65
75-79	-.0513	.729	.070	39
70-74	-1.0000	1.172	.853	10
70-	-2.5600	.612	4.183	25

TABLE 44

Distributions of Differences in English Quotient

Between Test and Retest.

Diff.	Boys	Girls	Total
17	2		2
16			0
15	1		1
14	0		0
13	3		3
12	2		2
11	4	1	5
10	7	3	10
9	9	4	13
8	13	8	21
7	12	15	27
6	21	12	33
5	20	18	38
4	32	29	61
3	40	44	84
2	37	31	68
1	35	58	93
0	49	51	100
-1	48	47	95
-2	39	45	84
-3	28	43	71
-4	26	26	52
-5	25	30	55
-6	25	16	41
-7	8	8	16
-8	7	16	23
-9	3	12	15
-10	1	5	6
-11	1	4	5
-12	1	2	3
-13	0	1	1
-14	0	0	0
-15	0	0	0
-16	0	0	0
-17	1	0	1
-18		1	1
	500	530	1030
Mean	+.600	-.530	.0184
S.D.	4.7232	4.3943	4.5915

TABLE 45

DISTRIBUTIONS OF DIFFERENCES IN ENGLISH QUOTIENTS AT VARIOUS LEVELS OF ABILITY.

M.H.E. 11/9

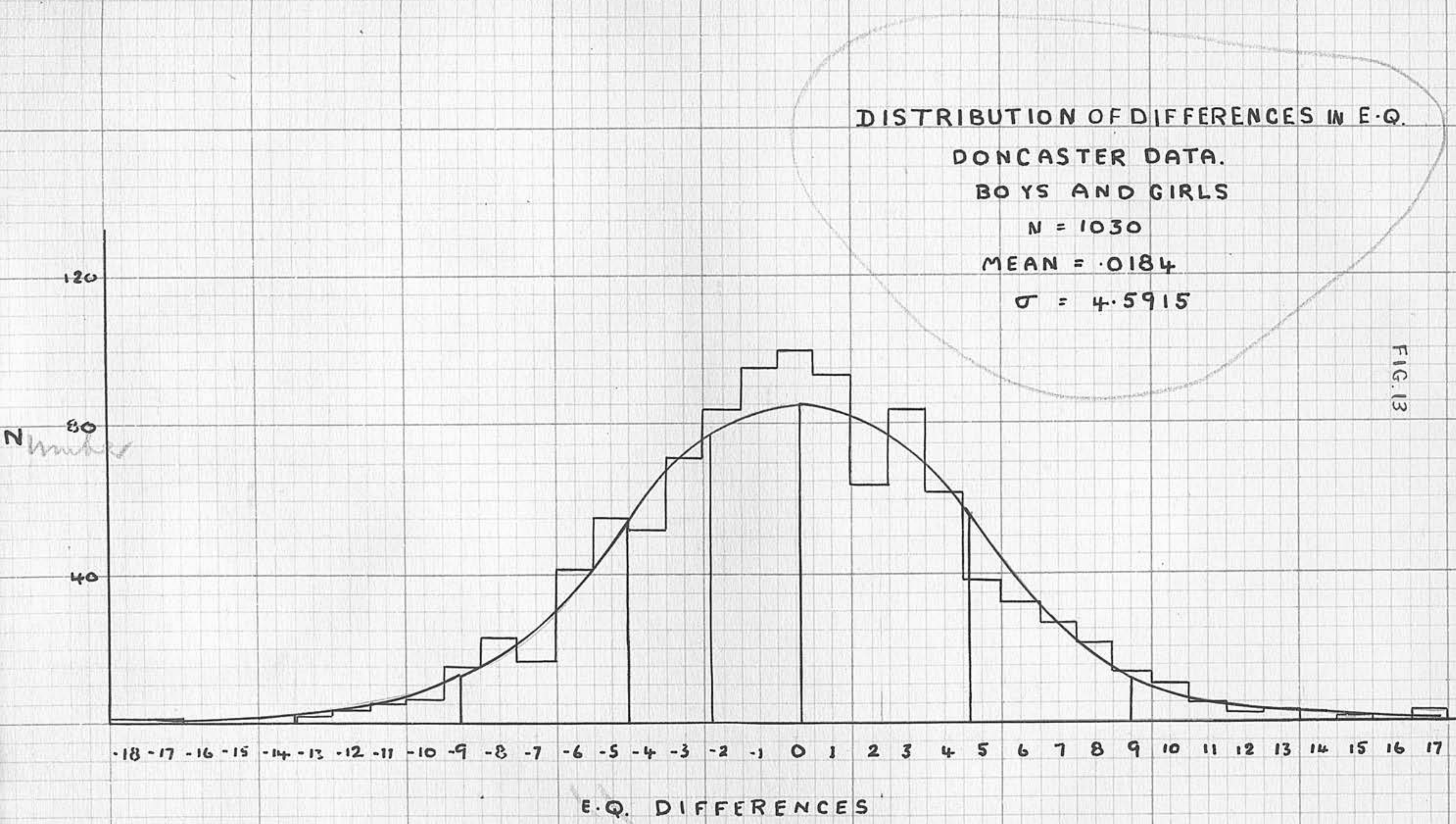
E.Q. diff.		130+	125 129	120 124	115 119	110 114	105 109	100 104	95 99	90 94	85 89	80 84	75 79	70 74	70- 74
17											1	1			
16											0	0			
15							1				0	0			
14							0				0	0			
13					1		0				0	0	2		
12					0		1	1			0	0	0		
11			1	1	0		1	1			1	0	0		
10			1	1	0		5	1		1	0	0	1		
9		1	1	1	2	5	2	0		1	0	0	0		
8		1	2	0	4	3	1	2	3	2	2	1	0		
7		1	1	4	4	4	4	2	2	3	2	0	0		
6		0	0	3	1	2	4	6	7	2	6	1	1		
5		1	1	3	4	6	2	2	8	4	2	3	1	1	
4		2	4	2	3	6	12	11	7	4	4	6	0	0	
3		0	7	3	7	11	12	13	11	5	8	3	1	2	1
2		1	3	3	10	9	6	5	7	12	4	2	4	0	2
1		0	4	6	7	10	10	13	12	12	9	4	4	1	1
0		2	3	4	6	13	13	15	8	8	10	8	6	0	4
-1		2	2	4	7	11	10	15	6	10	13	11	3	1	0
-2		0	0	4	8	10	6	14	11	12	5	5	4	1	4
-3		2	3	3	1	7	8	6	7	14	6	5	5	1	3
-4		0	0	2	6	5	6	6	4	7	4	5	2	1	4
-5		0	1	2	3	5	5	9	9	7	4	3	3	1	3
-6		2		3	0	4	3	5	7	5	8	3	0	0	1
-7		1		1	0	1	0	3	3	3	1	1	1	1	0
-8				2	1	5	1	2	4	1	2	3	1		1
-9				2	1	2	1	2	5	2	0				0
-10						0	1	2	1	0	1				1
-11						0	1	1	1	2	0				
-12						1	1			0	1				
-13						0				1	0				
-14						0					0				
-15						0					0				
-16						0					0				
-17						1					0				
		16	34	54	76	121	117	137	123	118		65	39	10	25

TABLE 46

Table of standard deviations of E.Q. differences, reliability coefficients, and standard errors of E.Q. at different levels of ability.

E.Q. Level.	$S.D._d$	r_{11}	S.E. E.Q.	N
130+	4.9096	.9464	3.4716	16
125-129	3.6990	.9696	2.6156	34
120-124	4.9451	.9457	3.4967	54
115-119	4.1611	.9615	2.9423	76
110-114	4.6994	.9509	3.3229	121
105-109	4.8518	.9477	3.4307	117
100-104	4.3131	.9587	3.0498	137
95-99	4.6262	.9524	3.2712	123
90-94	4.2626	.9596	3.0141	118
85-89	4.9539	.9455	3.5029	95
80-84	4.2187	.9605	2.9830	65
75-79	4.5513	.9539	3.2182	39
70-74	3.7036	.9695	2.6188	10
70-	3.0599	.9792	2.1637	25

FIG. 13


DISTRIBUTION OF DIFFERENCES IN E·Q.
DONCASTER DATA.
BOYS AND GIRLS
N = 1030
MEAN = ·0184
σ = 4·5915
N
120
80
40
-18 -17 -16 -15 -14 -13 -12 -11 -10 -9 -8 -7 -6 -5 -4 -3 -2 -1 0 1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17
E.Q. DIFFERENCES

A NOTE ON THE RELATIONSHIP BETWEEN THE RELIABILITY AND VALIDITY OF TESTS

Examination of the Doncaster Data dealing with the reliability and constancy of Moray House Intelligence, Arithmetic and English Quotients, brings to light the fact that of the three types of test those regarded as measures of Intelligence are least reliable. This fact requires explanation. The reliability coefficients found for the three types of test are repeated here for comparative purposes.

	r	N
Intelligence	.9370	1030
Arithmetic	.9620	1030
English	.9532	1030

These reliabilities are, in two respects, not directly comparable.

(1) The times of administration are different for each test, the Intelligence requiring 45 minutes, the Arithmetic 30 minutes, and the English 40 minutes.

(2) The number of items are different, the Intelligence Test having 100 items, the Arithmetic 102 items, and the English 150 items.

The figures given above show that the Arithmetic test is by far the most reliable of the three despite the fact that the time of its administration is only 30 minutes. The English test with its 150 items is less reliable than the Arithmetic and more reliable than the Intelligence tests. It is possible to estimate by the Spearman-Brown formula the reliability of the English test had it been constructed of only 100 items, but such a test would then require about 27 minutes to administer, and as such would not be directly comparable with the Intelligence test requiring 45 minutes to administer. None the less if some common ground of comparison could be reached the English test would in all likelihood be characterised by higher reliability than the Intelligence test. Since some measure of doubt, however small, exists, the observations developed below will be largely concerned with the comparative reliabilities of the Intelligence and Arithmetic tests.

The reliability of a test is not only dependent on the actual reliabilities of the items which it contains, but also on the intercorrelations of all the items in much the same way that the correlation between a battery of tests and another battery of tests, or between a battery of tests and a criterion, is dependent on all the intercorrelations between the several variables. The greater the number of items the greater the importance to be attached to the inter-item correlations, and the less the importance to be attached to the actual item reliabilities. With a test of 100 items there are only

100 item reliabilities whose influence on the reliability of the whole test is greatly outweighed by the influence of the 4950 different inter-item correlations.

A test whose inter-item correlations are high, tends to be more reliable than a test whose inter-item correlations are low, and by the selection of items to yield high inter-item correlations, we increase the reliability of the whole test. Thus the more homogeneous the items, the more closely they approximate to the measurement of a single trait rather than a composite of two or more traits, the more reliable the test tends to be. This implies that the higher the general factor variances of the items, and the smaller the group and specific factor variances, the more reliable the test. Thus it is possible, although the arithmetical labour involved is enormous, to purify a test by the elimination of those items that exhibit a low inter-item correlation, and thus attain a test characterised by high internal consistancy and high reliability. It will be understood that increasing the inter-item correlations will only make the test as a whole approximate more closely to the measurement of a unit trait when the items themselves may be regarded as measures of a unit trait. If each item measures a composite of traits selecting items that yield high inter-item correlations will imply that the test itself measures a composite of abilities. In such a case the composite of factors measured by each item behaves as a single factor.

The simple theory outlined above explains the difference between the reliabilities of different tests, which, if the number of items, the times of administration, and objectivity were the only factors influencing reliability, would be equally reliable. Although data are not at the moment available it is most probably that the intercorrelations of the items on Moray House Arithmetic Tests are on the whole higher than the intercorrelations of the items on Moray House Intelligence Tests; that is to say the Intelligence Test seems to measure a greater complexity of abilities than the Arithmetic test. The inter-item correlation matrix for the Arithmetic test is of a lower rank than the inter-item correlation matrix for the Intelligence test.

Increasing the inter-item correlation in order to approximate more closely to the measurement of a unit trait and to increase test reliability may, however, be inadvisable from the point of view of validity. An unfortunate incompatability exists between reliability and validity concepts which is as yet unresolved. By increasing the inter-item correlations, and thereby making the test more homogeneous in structure, one will usually, although not always, decrease the correlation of a test with an external criterion. The truth of the above statement depends on the nature of the criterion. Success in secondary school or in an occupation, or in fact any criterion of the usual type which we wish to predict, is not dependent on a single mental trait but upon a composite of traits, and the efficiency of the test or test battery in predicting such criteria depends on the adequacy of the test or test battery in sampling such traits. The test samples what the child can do. Thus it seems that by constructing a test approximating to the measurement of a single unit trait we decrease the correlation of each item with the criterion. By increasing the inter-item correlation we increase the reliability of the test at the expense of validity. By decreasing the inter-item correlation we increase the validity of the test at the expense of reliability.

In the case of Moray House Tests the superior reliability of the Arithmetic Tests over the Intelligence Tests indicates that the former is more homogeneous in structure, but as is known the Intelligence Tests correlate more highly with the later performance of the pupils than the Arithmetic Tests, and this despite their greater unreliability. The influence of the greater prevalence of random errors will depress the correlation of the Intelligence Test with a criterion more than the correlation of an Arithmetic Test with a criterion. Occasionally we attach a weight to the Intelligence Test equal to twice the weight of the Arithmetic Test. Thus the less reliable test is given the greater weight by virtue of its higher validity. How this incompatability between reliability and validity will be resolved is not at the moment apparent.

SPLIT-HALF RELIABILITY COEFFICIENTS.

A number of split-half reliability coefficients are available for Moray House Tests estimated from random samples of over 200 cases. These coefficients are invariably higher than coefficients obtained by correlating parallel forms after a time interval due either to the correlation of errors or to the absence of functional variability. The correlation between the odd and even items ($r_{\frac{11}{22}}$), the 'boosted' split-half reliability (r_{11}), the number of cases in the sample (N), the standard deviation of the sample ( ), and the standard deviation of the population ( ), for five samples of Moray House Intelligence Tests, and one Moray House English Test are as follows:-

Test	Sample	$r_{\frac{11}{22}}$	r_{11}	N		
M.H.T.21	W. Yorkshire	.9278	.9625	212	19.96	22.07
M.H.T.23	W. Yorkshire	.9393	.9687	212	17.95	20.08
M.H.T.23	Darlington	.9560	.9775	235	19.38	20.38
M.H.T.24	Northumberland	.9427	.9705	242	19.37	20.07
M.H.T.26	W. Yorkshire	.9457	.9721	212	17.35	19.47
M.H.E.11	Northumberland	.9661	.9828	222	31.27	31.77

135

SELECTED BIBLIOGRAPHY

Aitken, A.C., (1939). "Statistical Mathematics," Oliver and Boyd, pp. 125-127.

Brown, R.Ralph. (1933). "The Time Interval Between Test and Retest in its Relation to the Constancy of Intelligence," J.Educ. Psychol. XXIV pp.81-95.

Brown, W., (1910). "Some Experimental Results in the Correlation of Mental Abilities," B.J.P. pp. 296-322.

Brown, W., and Thomson, G.H., (1925), "The Essentials of Mental Measurement," pp.156-160.

Dunlap, Jack W., (1934). "Comparable Tests and Reliability," J.Educ.Psychol., pp.442-453.

Foran, I.C., (1926) "The Constancy of Intelligence," Catholic University of America Research Bulletin No.10.

Fisher, R.A., (1938) "Statistical Methods for Research Workers," Oliver and Boyd.

Flanagan, John C., (1939). "General Considerations in the Selection of Test Items and a Short Method of Estimating the Product-Moment Coefficient from Data at the tails of the Distribution," J.Educ. Psychol., XXX pp. 674-680.

Garrett, Henry E., (1937) " Statistics in Psychology and Education", pp.311-331.

Handy, Uvan, and Lenz, Theodore F., (1934) , "Item Value and Test Reliability," J.Educ.Psychol. pp.703-708.

Holzinger, Karl J., and Swineford, Frances, (1937), "The Bi-factor Method", Psychometrika, Vol.2, pp.41-54.

(1937), "Student Manual of Factor Analysis", Chicago.

(1928), "Statistical Methods for Students in Education", Boston.

(1938), "The Relationship Between Three Multiple Orthagonal Factors and Four Bi-factors", J.Educ. Psychol., Vol.xxlx, pp. 513-519.

Jeffreys, Harold, (1937), "Scientific Inference", pp.52-83.

Jordon, R.C., (1935), "Empirical Study of the Reliability Coefficient", J.Educ.Psychol. pp.416-426.

Kelly, T.L., (1921), "The Reliability of Test Scores", J.Educ. Research, Vol.3, pp.370-379.

(1923), "A New Method for Determining the Significance of Differences in Intelligence and Achievement Scores", J.Educ. Psychol. XIV pp.321-333.

(1923), "Statistical Method", New York.

(1925), "The Applicability of the Spearman-Brown Formula for the Measurement of Reliability", J.Educ.Psychol., XVI pp.300-303.

Kuder, G.F., and Richardson, M.W., (1937), "The Theory and Estimation of Test Reliability", Psychometrika, Vol. 2, pp. 151-160.

Long, John A., and Sandiford, Peter, (1935), "The Validation of Test Items", University of Toronto.

Merrill, Walter W., Jr., (1937), "Sampling Theory in Item Analysis", Psychometrika, Vol. 2, pp. 215-223.

Mosier, Charles I., (1936), "A Note on Item Analysis and the Criterion of Internal Consistency", Psychometrika, Vol. I, pp. 275-282.

Paulsen, G.B. (1931), "A Coefficient of Trait Variability", Psychol. Bulletin, XXVIII, 218.

Richardson, M.W., (1936), "The Relationship Between Difficulty and the Differential Validity of a Test", Psychometrika, Vol. 1 pp. 33-55.

Richardson, M.W., (1936), "Notes on the Rationale of Item Analysis", Psychometrika Vol. 1 pp. 69-76.

Richardson, M.W., and Kuder, G.F., (1939) "The Calculation of Test Reliability based on the Method of Rational Equivalence", J. Educ. Psychol. XXX, 681-687.

Rodgers, Allan G., (1936), "The Application of Six Group Intelligence Tests to the Same Children, and the Effects of Practice", B.J.E.P. Vol. VI, 291-305.

Roff, Merrill, (1937) "The Relationship Between Results Obtainablewith Raw and Corrected Correlation Coefficients in Multiple Factor Analysis", Psychometrika, Vol.2, pp. 35-39.

Ruch, Ackerson, Jackson, "An Empirical Study of the Spearman-Brown Formula as Applied to Educational Test Material", J. Educ.Psychol. vol.XVII, 1926, pp309-313.

Shen, E, (1924), "The Standard Error of Certain Estimated Coefficients of Correlation", J.Educ. Psychol vol. XV, pp. 462-465.

(1926), "A Note on the Standard Error of the Spearman Brown Formula", J.Educ.Psychol, XVII,pp93-94.

Spearman, C., (1907), "Demonstration of Formulae for True Measurement of Correlation", American J. of Psychol. vol 18, p 161, (1907).

(1910), "Correlation Calculated from Faulty Data", B.J.P., vol.3, p.281.

(1932) "Abilities of Man", London.

Swineford, Frances, (1936) "Bi-serial r versus Pearson r as Measures of Test-item Validity", J.Educ.Psychol. XXVII, pp.471-472.

Terman, Lewis.M., and Merrill, Maud A.,(1937) "Measuring Intelligence" pp.33-47.

Terman, Lewis M., (1916) "The Intelligence of School Children".

Thomson, Godfrey H. (1939) "The Factorial Analysis of Human Ability", University of London Press.

(1940) "Weighting for Battery Reliability and Prediction", B.J.P., vol XXX, pp.357-365.

Thorndike, Robert L.(1933)"The Effect of the Interval between Test and Retest on the Constancy of I.Q.", J.Educ.Psychol. vol.XXIV, pp.543-549.

Thouless, Robert H., (1936) "Test Unreliability and Function Fluctuation", B.J.P. XXVI p 325.

(1939)"The Effects of Errors of Measurement on Correlation Coefficients", B.J.P. XXIX, pp.383-403.

Thurstone L.L., (1932) (1939) "The Reliability and Validity of Tests" Ann Arbour, U.S.A.

(1935) "The Vectors of Mind" Chicago.

Walker, D.A., (1931), "Answer Pattern and Score-scatter in Tests and Examinations, B.J.P. XXII pp.73-86

(1936),"Answer Pattern and Score-scatter in Tests and Examinations", B.J.P.XXVI, pp.301-308.

(1940) "Answer Pattern and Score-Scatter in Tests and Examinations", B.J.P.XXX, pp.248-260.

Woodrow, H. (1932), "Quotidian Variability", Psychol. Review, XXXIX, pp.245-246.

Zubin, J., (1934) "The Method of Internal Consistency for Selecting Test Items" J.Educ.Psychol., vol. xxv, pp. 345-356.

APPENDIX

This appendix is a record of an empirical enquiry on the application of Sheppard's Correction for grouping. This enquiry bears no immediate relationship to the main subject of this thesis.

THE APPLICATION OF SHEPPARD'S CORRECTION FOR GROUPING.

THE APPLICATION OF SHEPPARD'S CORRECTION FOR GROUPING.

Sheppard's correction for grouping, although widely used by statisticians in certain fields, is apparantly not in general use among psychometricians. The majority of standard deviations and correlations reported in psychological and educational literature are calculated from grouped data, and are uncorrected for grouping. This paper attempts to show the influence of grouping on standard deviations and correlations, and advances empirical evidence to illustrate with what accuracy values corrected for grouping with Sheppard's correction approximate to values obtained from ungrouped data in a continuous distribution.

In the calculation of statistical measures from grouped data the values of each variate within a given class interval are assigned the value of the mid-point of that interval. Thus in the calculation of a correlation coefficient from such data we are not calculating the relationship between the continuous variates x and y, but rather the relationship between the mid-points of certain class intervals into which the variates x and y have been grouped. With a normal distribution, and many other types of distributions, the point of concentration of the variate is not the mid-point of the class interval but a point slightly nearer the mean. Thus statistical measures calculated from the odd moments remain uninfluenced by grouping, because the errors made by the assumption that the scores are concentrated at the mid-point of each interval will tend to balance on both sides of the mean, while with the even moments the errors will not balance but will add together.

Grouping error tends to increase the size of the uncorrected standard deviations, and to reduce the size of the uncorrected correlations. The usual formula for correcting a standard deviation for grouping is as follows:-

$$\sigma = \sqrt{\tilde{\sigma}^2 - \frac{i^2}{12}}$$

where σ, $\tilde{\sigma}$ are the corrected, and uncorrected estimates respectively of the standard deviation and is the class interval.

The correction to be applied to a correlation coefficient for grouping depends on the observation that with two normally distributed variates x and y the quantity $\tilde{r}_{xy}\tilde{\sigma}_x\tilde{\sigma}_y$ is independent of the class interval used. It immediately follows from this observation that

$$r_{xy} = \frac{\tilde{r}_{xy}\tilde{\sigma}_x\tilde{\sigma}_y}{\sigma_x\sigma_y}$$

where $\tilde{r}_{xy}$ and r_{xy} are the uncorrected and corrected values of the correlation between x and y . Since, however,

$\tilde{r}_{xy}\sigma_x\sigma_y$ the usual product-moment formula for a correlation coefficient corrected for grouping may be written as follows:-

$$r_{xy} = \frac{\sum xy}{N\sqrt{\left(\tilde{\sigma}_x^2 - \frac{i_x^2}{12}\right)\left(\tilde{\sigma}_y^2 - \frac{i_y^2}{12}\right)}}$$

where i_x and i_y represent the class intervals of x and y respectively. When correlation coefficients are calculated by the diagonal adding method the formula for a corrected coefficient becomes

$$r_{xy} = \frac{H + V - D}{2\sqrt{\left(H - \frac{N}{12}\right)\left(V - \frac{N}{12}\right)}}$$

where H, V, and D represent the sum of the squares of the deviations from the mean values of x, y, and x-y, respectively.

Fisher has pointed out that in averaging correlation coefficients the values of z should be obtained from uncorrected values of r, and a correction added to the resulting coefficient equivalent to the average correction of the averaged values of r.

The corrected value of r is always larger than the uncorrected value of r. The larger the value of r the larger the absolute value of the correction to be made for grouping. The relative value of the correction is constant, given constant values for the standard deviations of the variates correlated. The size of the correction is independent of N, the number of cases.

Errors introduced by using uncorrected values of r when r is large are much more significant than errors resulting from a corresponding group when r is small. Not only is the absolute discrepancy between the uncorrected and the corrected value of r greater when r is large, but small differences between large correlations represent a much greater difference in the degree of relationship between the variates correlated than equivalent differences between small coefficients, and for this reason are more important to the statistician.

EXPERIMENTAL.

To determine the influence of grouping on standard deviations and correlations, and to estimate the accuracy with which values corrected for grouping approximate to values obtained from ungrouped data in a continuous distribution, the I.Q's of 952 children on two Intelligence tests were plotted on a grid with a class interval of unity. This was a somewhat laborious procedure. The two distributions of scores were approximately normal. The standard deviations of the two variables, and the correlation between them were calculated. The class interval was then excessively increased by telescoping, as it were, the original grid, and further standard deviations and correlations were calculated with class intervals of 2, 3, 4, 5, 6, 7, 8, 9, 10, 12, 14, 16, 18, and 20.

Table 1 gives the uncorrected and corrected standard deviations for variable x at different units of class interval, and the number of arrays upon which each measure is based. The corrected standard deviation with a class interval of unity is taken as the standard, and the deviations from this standard of the uncorrected and corrected standard deviations, calculated at each step interval, are given in columns d_1 and d_2, respectively. It will be observed that the uncorrected standard deviation with a class interval of unity is the same as would have been obtained from ungrouped data. This value is, however, corrected on the basis of the assumption that the distribution is theoretically continuous.

Table 2 furnishes corresponding data for variable y. These data indicate clearly that grouping tends to influence the size of the uncorrected standard deviation, and when the class interval is large this influence is substantially marked. Furthermore the application of Sheppard's correction results in an estimate of the standard deviation closely approximating to the value that would have obtained from an ungrouped continuous variate. Certain substantial discrepancies in the corrected values occasionally appear. These are due to the purely arbitrary nature of the points fixed as the top of the last class interval and the bottom of the first.

TABLE 1

Class interval.	No. of arrays	$S.D._x$ uncorrected	$S.D._x$ corrected	d_1	d_2
1	60	12.1550	12.1516	.0034	.0000
2	30	12.1549	12.1412	.0033	-.0104
3	20	12.1634	12.1325	.0118	-.0191
4	15	12.1740	12.1191	.0224	-.0325
5	12	12.1175	12.0313	-.0341	-.1203
6	10	12.1836	12.0599	.0320	-.0917
7	9	12.3123	12.1452	.1607	-.0064
8	8	12.4592	12.2433	.3076	.0917
9	7	12.6432	12.3734	.4916	.2218
10	6	12.4806	12.1421	.3290	-.0095
12	5	12.4620	11.9708	.3104	-.1808
14	5	12.6512	11.9883	.4996	-.1633
16	4	12.8747	12.0177	.7231	-.1339
18	4	13.1897	12.1230	1.0381	-.0286
20	3	13.3611	12.0493	1.2095	-.1023

TABLE 2

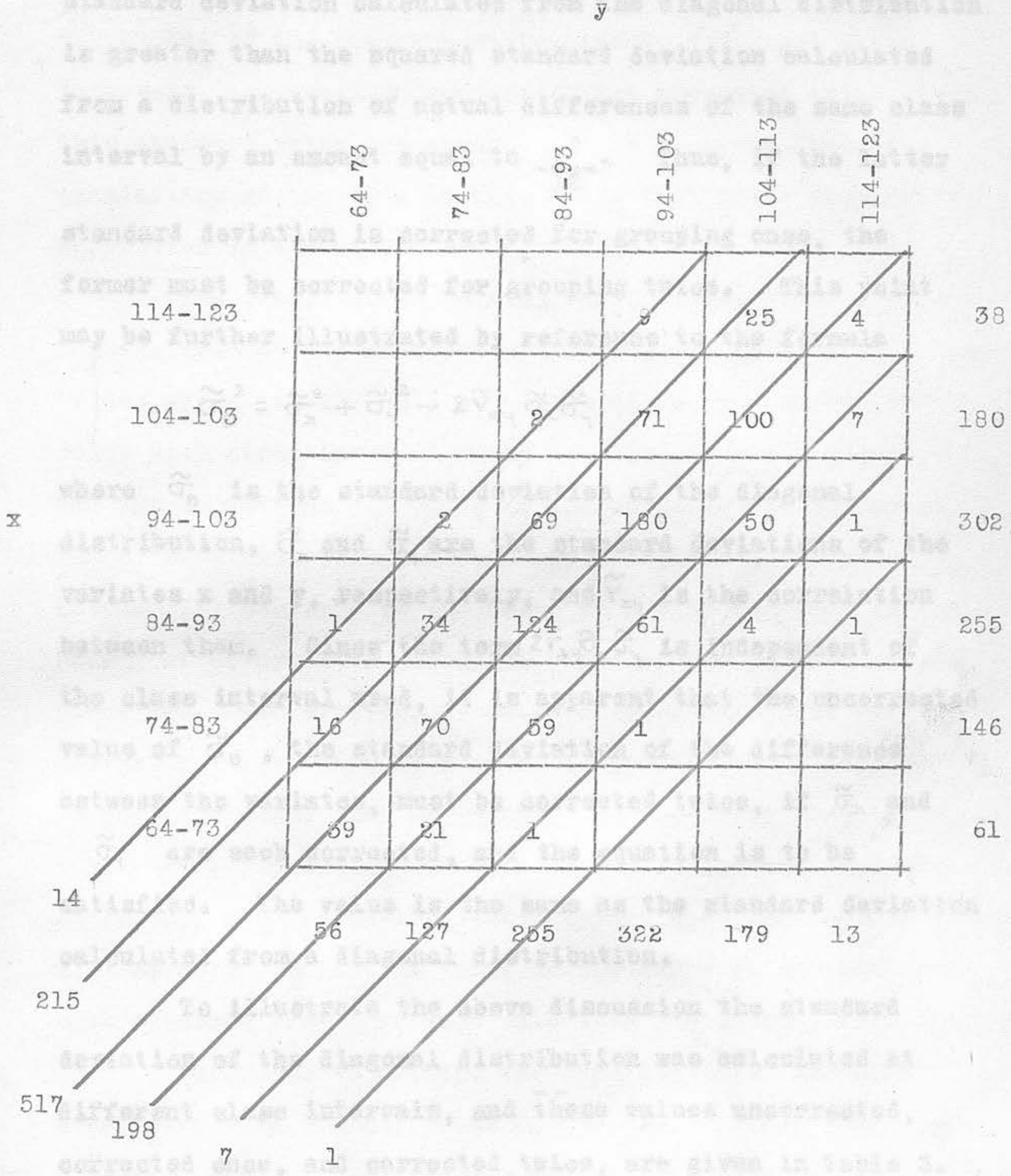
class interval.	No. of arrays.	S.D.$_y$ uncorrected	S.D.$_y$ corrected	d_1	d_2
1	55	11.2309	11.2272	.0037	.0000
2	28	11.2563	11.2416	.0291	.0144
3	19	11.2518	11.2184	.0246	-.0088
4	14	11.3123	11.2523	.0851	.0260
5	11	11.3570	11.2645	.1298	.0373
6	10	11.3988	11.2664	.1716	.0392
7	8	11.3421	11.1595	.1149	-.0677
8	7	11.4128	11.1768	.1856	-.0504
9	7	11.5848	11.2896	.3576	.0624
10	6	11.5273	11.1600	.3001	-.0872
12	5	11.8006	11.2807	.5634	.0535
14	4	11.5885	10.8608	.3613	-.3664
16	4	11.7920	10.8498	.5648	-.3774
18	4	12.5132	11.3834	1.2860	.1562
20	3	12.5510	11.1442	1.3238	-.0830

Table 3 gives the standard deviations of the difference between the variates x and y calculated from diagonal distributions at different class intervals. This procedure may be illustrated by reference to the correlation grid in Figure 1 with a class interval of 10 points of raw score. By adding this correlation grid diagonally from north-east to south-west we obtain a distribution of the differences between the variables x and y. By adding from north-west to south-east we obtain a distribution of the sum of the variables x and y. Thus, if we wish to calculate the standard devåation of variation in I.Q. between test and retest, instead of calculating the actual distance in I.Q. for every child, and making a distribution og these differences, it is possible to plot the I.Q's on a correlation grid, and to calculate the standard deviation of difference in I.Q. direct from the distribution found by diagonal adding. The diagonal distribution in Figure 1 is, however, not the same as the distribution that would have resulted by subtracting every child's score in variable x from his score in variable y, and grouping the differences thus obtained in a frequency distribution of class interval 10. Because a peculiarity in the grouping of the diagonal distribution exists, the standard deviation of x-y calculated from the diagonal distribution is greater than the standard deviation of x-y calculated from the distribution made by subtracting the appropriate values of y from x, and grouping the differences

TABLE 3

class interval.	No. of arrays.	$S.D._{x-y}$ uncorrected	$S.D._{x-y}$ corrected once	$S.D._{x-y}$ corrected twice	d_1	d_2	d_3
1	35	5.9401	5.9331	5.9261	.0140	.0070	.0000
2	18	5.9662	5.9382	5.9101	.0401	.0121	-.0160
3	12	6.0219	5.9592	5.8960	.0958	.0331	-.0301
4	10	6.1348	6.0251	5.9134	.2087	.0990	-.0157
5	10	6.2517	6.0828	5.9091	.3256	.1567	-.0170
6	7	6.2382	5.9929	5.7371	.3121	.0668	-.1890
7	7	6.5170	6.1958	5.8570	.5909	.2697	-.0691
8	6	6.6952	6.2843	5.8428	.7691	.3582	-.0833
9	5	7.0182	6.5196	5.9794	1.0921	.5935	.0533
10	6	7.2845	6.6836	6.0330	1.3584	.7575	.1069
12	5	7.4767	6.6258	5.6481	1.5506	.6997	-.2780
14	5	8.3318	7.2860	6.0624	2.4057	1.3599	.1363
16	3	8.5042	7.1406	5.4456	2.5781	1.2145	-.4805
18	3	9.1499	7.5313	5.4516	3.2238	1.6052	-.4745
20	3	9.9576	8.1130	5.6997	4.0315	2.1869	-.2264

FIGURE 1



y
64-73
74-83
84-93
94-103
104-113
114-123
x
114-123
9
25
4
38
104-103
2
71
100
7
180
94-103
2
69
180
50
1
302
84-93
1
34
124
61
4
1
255
74-83
16
70
59
1
146
64-73
39
21
1
61
56
127
255
322
179
13
14
215
517
198
7
1

with class interval equal to that of x and y. The squared standard deviation calculated from the diagonal distribution is greater than the squared standard deviation calculated from a distribution of actual differences of the same class interval by an amount equal to $\frac{i^2}{12}$. Thus, if the latter standard deviation is corrected for grouping once, the former must be corrected for grouping twice. This point may be further illustrated by reference to the formula

$$\tilde{\sigma}_D^2 = \tilde{\sigma}_x^2 + \tilde{\sigma}_y^2 - 2\tilde{r}_{xy}\tilde{\sigma}_x\tilde{\sigma}_y$$

where $\tilde{\sigma}_D$ is the standard deviation of the diagonal distribution, $\tilde{\sigma}_x$ and $\tilde{\sigma}_y$ are the standard deviations of the variates x and y, respectively, and $\tilde{r}_{xy}$ is the correlation between them. Since the term $2\tilde{r}_{xy}\tilde{\sigma}_x\tilde{\sigma}_y$ is independent of the class interval used, it is apparent that the uncorrected value of $\tilde{\sigma}_D$, the standard deviation of the difference between the variates, must be corrected twice, if $\tilde{\sigma}_x$ and $\tilde{\sigma}_y$ are each corrected, and the equation is to be satisfied. The value is the same as the standard deviation calculated from a diagonal distribution.

To illustrate the above discussion the standard deviation of the diagonal distribution was calculated at different class intervals, and these values uncorrected, corrected once, and corrected twice, are given in Table 3.

The standard deviation of the difference with class interval unity, is taken as the standard value, and the deviations d_1, d_2, d_3 of the standard deviations at different class intervals, uncorrected, corrected, and corrected twice, from this standard value are given. It is apparent from an examination of the data in this Table that twice Sheppard's correction is the correction required.

The correlations between the variates x and y were also calculated at different units of class interval. These values are given in Table 4. Here again, the corrected value with class interval unity is taken as the standard, and the deviations d_1 and d_2 of the obtained and corrected values of r from this standard are calculated. A very substantial decrease in the value of r with decrease in the number of arrays into which the variates are grouped can be observed. The discrepancy between the uncorrected and corrected values of r is such as to furnish sound support to the conclusion that correlation coefficients must be corrected for grouping if accurate statistics are desired. These data are indicative that Sheppard's correction furnishes a remarkably accurate estimate of the correlation that would have obtained from ungrouped data with continuous variates.

In order to examine the functioning of Sheppard's correction with a small value of r a new grid was drawn up with 1828 cases. Values of r were found as before at

TABLE 4

Class interval	No. of arrays x	No. of arrays y	r_{xy} uncorrected	r_{xy} corrected	d_1	d_2
1	60	55	.8739	.8744	.0005	.0000
2	30	28	.8729	.8750	.0015	.0006
3	20	19	.8706	.8754	.0038	.0010
4	15	14	.8661	.8746	.0083	.0002
5	12	11	.8601	.8733	.0143	-.0011
6	10	10	.8621	.8812	.0123	.0068
7	9	8	.8513	.8771	.0231	.0027
8	8	7	.8462	.8793	.0282	.0049
9	7	7	.8357	.8762	.0387	.0018
10	6	6	.8187	.8692	.0557	.0052
12	5	5	.8114	.8836	.0630	.0092
14	5	4	.7671	.8638	.1073	-.0106
16	4	4	.7656	.8914	.1088	.0170
18	4	4	.7478	.8943	.1266	.0199
20	3	3	.7063	.8821	.1681	.0077

successive class intervals. Table 5 gives values of r uncorrected and corrected for different class intervals. The deviations of the uncorrected and corrected values, respectively, from a standard value .3672 are given in columns d_1 and d_2. The number of arrays are given, in this case the number of arrays of the x variable being equal to the number of arrays of the y variable for each value of r.

It will be observed that the d_1 column of Table 4 is in every case greater than the d_1 column of Table 5, illustrating that the larger the value of r the larger the absolute value of Sheppard's correction, and emphasizing that correcting for grouping is of much more importance when r is large than when r is small. Examination of the d_2 columns of Tables 4 and 5 shows that Sheppard's correction furnishes a remarkably accurate estimate of the correlation that would have obtained from ungrouped data with continuous variates. Furthermore, if there is reason to believe that the distributions of the two correlated variables approximate normality some work can be avoided by using a coarse grouping with a small number of arrays and correcting for grouping. Tables 4 and 5 show that accurate results can be obtained with as few as six arrays, the error made by using only six arrays in Table 4 being .49 per cent, and in Table 5 .03 per cent. With less than six arrays the purely arbitrary position of the class intervals will in most cases lead to slight discrepancies in the corrected value of r.

TABLE 5

Class Interval.	No. of arrays	r_{xy} uncorrected	r_{xy} corrected.	d_1	d_2
1	60	.3670	.3672	-.0002	.0000
2	30	.3663	.3672	-.0009	.0000
3	20	.3648	.3668	-.0024	-.0004
4	15	.3632	.3667	-.0040	-.0005
5	12	.3613	.3667	-.0059	-.0005
6	10	.3581	.3660	-.0091	-.0012
7	9	.3548	.3653	-.0124	-.0019
8	8	.3520	.3658	-.0152	-.0014
9	7	.3514	.3685	-.0158	.0013
10	6	.3457	.3669	-.0215	-.0003
12	5	.3340	.3634	-.0332	-.0038
14	5	.3452	.3873	-.0220	.0101
16	4	.3134	.3616	-.0538	-.0056
18	4	.3112	.3758	-.0560	.0086
20	3	.2729	.3423	-.0943	-.0249

SUMMARY.

If the distributions of variates used in statistical work are approximately normal the use of Sheppard's correction furnishes accurate estimates of the standard deviations and correlations that would have resulted from the use of ungrouped data. Correcting a correlation coefficient for grouping is essential when the grouping is coarse and the number of arrays is large. Otherwise inaccurate statistics will result. The discrepancies found in small correlations due to failure to correct for grouping are of less importance. Reasonably accurate results can be attained with a small number of arrays if the distributions of variates are normal.