THE APPLICATION OF ANALYSIS OF VARIANCE IN PSYCHOMETRIC EXPERIMENTATION

by

A. E. C. PILLINER

Thesis presented for the Degree of Ph.D. University of Edinburgh

September, 1965.
I wish to acknowledge the debt I owe to the following:

(1) To Professor J. G. Pilley and Dr. D. N. Lawley, my supervisors for their encouragement and advice.

(2) To Miss E. G. Taylor and Dr. J. Sutherland for their ready agreement to the inclusion in this thesis of an account of a study with which they were associated.

(3) To the Ministry of Education and Culture of the State of Israel for permission to make use of the results of a study of part of the 'Bagrut' examination, conducted annually in Israel.

(4) To Mrs. Margaret Milne and Mrs. Barbara Muir who undertook the typing of this thesis - a truly formidable task.
## INDEX

### GENERAL INTRODUCTION

<table>
<thead>
<tr>
<th>Chapter One</th>
<th>- Introduction</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>1</td>
</tr>
</tbody>
</table>

### PART ONE - STUDIES OF TEST RELIABILITY

<table>
<thead>
<tr>
<th>Chapter Two</th>
<th>- Introduction</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>38</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Three</th>
<th>- The Coefficient of Equivalence</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>44</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Four</th>
<th>- The Coefficient of Internal Consistency</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>57</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Five</th>
<th>- Summary of Results of Chapters Three and Four</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>67</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Six</th>
<th>- The Relationship between Stability and Equivalence Coefficients</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>69</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Seven</th>
<th>- Numerical Example</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>88</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Eight</th>
<th>- Zero Error</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>102</td>
</tr>
</tbody>
</table>

### PART TWO - THE RESCALING OF TEACHERS' ASSESSMENTS

<table>
<thead>
<tr>
<th>Chapter Nine</th>
<th>- Introduction</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>125</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Ten</th>
<th>- Methods of Rescaling</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>130</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Eleven</th>
<th>- The Reliability of the Rescaling Procedure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>160</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Twelve</th>
<th>- Making Assessments and Choosing a Rescaling Test</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>228</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Thirteen</th>
<th>- Current Procedures for Rescaling Assessments</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>261</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Fourteen</th>
<th>- Final Remarks and Conclusions</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>270</td>
</tr>
</tbody>
</table>

### PART THREE - THE RELIABILITY OF MARKING IN A GENERAL EXAMINATION

<table>
<thead>
<tr>
<th>Chapter Fifteen</th>
<th>- The Bagrut Examination in Israel</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>273</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Sixteen</th>
<th>- Experimental Enquiry</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>277</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Seventeen</th>
<th>- Alternative Procedures for Distributing Scripts to Markers</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>295</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Chapter Eighteen</th>
<th>- Conclusions and Recommendations</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td>302</td>
</tr>
</tbody>
</table>

### REFERENCES

<table>
<thead>
<tr>
<th>REFERENCES</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(i)-(vii)</td>
</tr>
</tbody>
</table>
GENERAL INTRODUCTION

...
CHAPTER ONE - INTRODUCTION

A - MAIN PURPOSE OF THIS THESIS

The main purpose of this thesis is to use the statistical technique of analysis of variance in the study of methods of evaluating and improving the precision of measurement, that is, the reliability, of psychometric instruments. The term 'psychometric instruments' will be interpreted broadly to include objectively scorable tests; the assessments made by teachers of the abilities or potentialities of their pupils; and examinations, particularly the large-scale public examinations conducted by local or national organisations.

A.I - OBJECTIVE TESTS

The large majority of reliability studies reported in the literature relate to objective tests. This is not surprising; ever since the time of Binet and Simon, which may be taken roughly as the beginning of the objective testing era, the problems of test construction have attracted the attention of statisticians or of educationists and psychologists with a statistical bent. The names of Spearman, Thurstone, Thomson, Thorndyke and Burt come to mind, to mention only a few. It is now generally recognised that the test constructor must possess not only the qualities of educational insight and psychological competence, but also of statistical expertise, if his test is to be acceptable and efficient as a testing instrument. The test statistician is now one of the team. It has rightly become customary for the constructor of a reputable test to include, in the information he furnishes about his product, details of its statistical characteristics, including its reliability. To reinforce this custom, the
American Psychological Society (1954) and the British Psychological Society (1959) have laid down specifications concerning the reporting of information, including statistical, to which test constructors and publishers increasingly conform.

A.II - TEACHERS' ASSESSMENTS

Despite the fact that on the face of it the assessments made by teachers of their pupils' abilities should have high predictive validity in respect of future performance, it is only recently that these assessments have been extensively used. Many teachers were reluctant at first to accept responsibility for making judgements which amounted to decisions about their pupils' futures. In view of the tensions engendered by secondary school selection, they preferred that these decisions should be made on the basis of some external procedure. In the last few years, however, the reluctance of teachers thus to involve themselves has diminished, and the number of Local Authorities using teachers' assessments is increasing.

However accurately individual teachers can judge the relative abilities of their own pupils, there remains the problem of making these judgements comparable in standard from one teacher to another. This problem can be solved by using a procedure in which all teachers' assessments are brought to a common standard by rescaling on the results of a test or tests administered to all the pupils of all the teachers. Probably the first extensive use of a rescaling technique - certainly the most celebrated - was that of McClelland in his massive 1935-1936 Dundee study (54). McClelland, however, did not explore the theoretical foundations of the rescaling techniques he employed, and it is only in the last few years that any serious attempt to do so has been made (39), (65). In
particular, little is known about the reliability of teachers' assessments and but small attention has been given to the devising of methods appropriate to their study.

A.III - EXAMINATIONS

The publications relating to examination reliability are surprisingly few. The investigations of Hartog and Rhodes, and of their research associates, Thomson and Burt, gave reason for disquiet about the efficiency of the public examination. The results, published as long ago as 1935, failed to provoke the further research one might have expected. Presumably the examining bodies themselves conduct enquiries into the examinations they set, but they do not make their results generally known. In the few reports that are available, there is little to allay misgivings.

This thesis, then, is concerned with the further study of reliability within these boundaries.

A.IV - ANALYSIS OF VARIANCE

The statistical tool employed in the study is the method of analysis of variance devised by Fisher (27) during the 1920's and described recently by Green and Tukey (30) as 'perhaps our single most powerful methodological technique'.

It is a reasonable conjecture that the method of analysis of variance was first devised as a research tool for use in agriculture science. This, at least, is the impression one derives from the context of Fisher's original exposition (35). His illustrative examples relate to 'treatments', 'plots', 'yields', and 'blocks'.* The method itself, however, is a statistical

* It is interesting that the terminology first used by the Rothamsted group of workers with specific regard to agricultural experimentation has become current generally. The writers of text-books on educational statistics refer to 'randomised blocks', 'split plot' designs, etc.
device, applicable to any area in which quantitative information is to be handled, provided the assumptions underlying the method hold for the information.

The possibilities of variance analysis as a research tool were quickly seen by workers in fields other than the agricultural. Burt (6) in Britain, Hoyt (38) in the United States, and Jackson (41) in Canada were among the first to employ it in education and psychology, and its use in these areas has now become commonplace.

However, as in agricultural experimentation, the method of analysis of variance has been generally applied by educationists and psychologists to the type of research set up in order to test hypotheses concerning specific 'treatment' effects by way of appropriate experimental designs. Recently, for example, it has been employed to compare the results obtained with programmed and conventional methods of instruction given respectively to different groups of pupils randomly selected. The methods are 'treatments'; the groups of pupils are 'blocks'; the individual pupil is a 'replication within a block'; his score on the post-treatment test is a 'yield'; and the experimental design is that termed 'randomised blocks'. Central to researches of this sort is the null hypothesis (no difference in the mean 'yields' from the different 'treatments') which is tested by the appropriate F-test.

The importance of analysis of variance as a tool in practical research of the above type is well established. Probably its use, in studies of a more theoretical nature, is less generally recognised. The study of reliability theory is an example of the latter type, in which a small number of workers have brought to bear the full power of variance analysis only in the last decade.
though still important, are no longer central. The main interest lies in estimating population parameters, and in particular, components of variance. These are then used in different combinations of inclusions and exclusions to give estimates of different reliability coefficients. By this means, light is shed on the concept of reliability which even now is imperfectly understood by many educationists and psychologists.

One can only speculate about why analysis of variance, almost from its inception, was employed in education and psychology for testing hypotheses about 'treatment' effects but relatively neglected as a tool for studying reliability theory. Possibly its use in testing hypotheses about 'treatments' such as different methods of teaching was immediately apparent by analogy with its use in agriculture. Moreover, this kind of study needs little more than a rule-of-thumb acquaintance with the technique of variance analysis. However, the use of the technique to study reliability theory requires a greater degree of statistical sophistication, the more so since recently the development of the theory itself has gathered impetus in the hands of Cronbach, Lord, Ebel, and others.

**B - SUBSIDIARY AIM OF THIS THESIS**

The main aim of this thesis has been stated. Its subsidiary aim is to urge, and so far as possible to exemplify, the desirability of an increased understanding of the potentialities of variance analysis on the part of educationists and psychologists in general. The general 'impression from reading the literature, particularly the British publications, is that these potentialities are frequently not realised. Sometimes a modification of the experimental design would have led to a clearer and more general answer to the
question the investigator has asked. At other times, where there is opportunity to obtain useful supplementary information from the design, the opportunity is missed although the main question is answered. Occasionally there is uncertainty over the correct error term to use in a test of significance.

Within the framework of its main aim and of the areas chosen for study, this thesis generally will serve also the subsidiary aim. In particular, the remainder of this introductory chapter will be concerned with it. A brief statement of the principles of experimental design and a note on terminology, followed by an illustrative example, will obviate repetition at a later stage.

**B.I - EXPERIMENTAL DESIGN**

Kempthorne (44) has stated admirably the general principles involved in setting up a statistically designed investigation. So far as experimental design is concerned, these principles may be summed up as follows:-

(1) The devising of the experimental design must directly follow from an explicit statement of the problem and from the formulation of relevant hypotheses.

(2) The questions then to be asked are: What are the possible outcomes of this design? Are these outcomes in fact relevant to this problem and these hypotheses?

(3) These are followed by two further questions: Are these outcomes amenable to statistical analysis? If so, will the conclusions from this analysis be valid for the population to which they are intended to apply?

**B.II - STATISTICAL ANALYSIS**

Leaving Kempthorne, we turn to the statistical analysis which reflects the experimental design and which is of almost equal importance. It is obvious that the degree of command of statistical techniques which the
investigator has at his disposal imposes limits on the experimental design. There is thus a two-way relationship between experimental design and statistical analysis.

If the method of analysis of variance is to be used for the statistical treatment of the experimental results, it is important for the investigator to satisfy himself that the scale of the dependent variable (the 'yield') is such that the underlying assumptions of the method are not violated. If this danger exists, he should seek an appropriate transformation. Fortunately, in psychometric studies, in which the dependent variable is often marks or 'quotients', transformations are seldom necessary.

B.III - 'CROSSING'

Two sources of variation (classification) are 'crossed' if each instance of one classification is paired in turn with every instance of the other. For example, if n scripts are each assessed by m markers, 'Scripts' and 'Markers' are crossed, since each of the n scripts is paired with each of the m markers. When complete* crossing occurs it is possible to average over either classification independently of the other. Moreover, whenever it occurs, a corresponding interaction term (in this case, Scripts × Markers) figures in the analysis.

B.IV - 'NESTING'

'Nesting' of one classification in another occurs when one has meaning only

---

* Complete crossing will normally be built into the experimental design wherever possible since it facilitates both analysis and interpretation of the results. Green and Tukey (30) point out that crossing is sometimes incomplete with unselected populations. 'Such combinations as height 3 feet, weight 300 pounds, or height 7 feet, weight 30 pounds, are non-existent, or at least very unlikely.'
with reference to the other. For example, 'Persons', in an analysis including both persons and sexes, is obviously nested in 'Sexe'. Averaging over the outer (nesting) classification automatically averages over the inner (nested) classification. Interaction between a nesting classification and one nested in it is clearly impossible. Interaction however can occur between either the nesting or the nested classification and any other classification that crosses with it. In the illustrative example which follows, 'Children within Sexes' is crossed with 'Tests', giving rise to a Test x Children within Sexes interaction. Under these conditions, it is automatic that 'Sexe' also crosses with 'Tests', so that a Sex x Test interaction also occurs.

C - ILLUSTRATIVE EXAMPLE

The intention of the example which follows is to demonstrate generally the potentialities of the method of analysis of variance as a statistical tool in psychometric research.

The example starts with a fairly simple experimental design, chosen as typical of the class of 'mixed' designs which are often useful in this sort of research. It proceeds to the corresponding statistical analysis. The several intermediate stages are demonstrated by which the final analysis is reached and it is argued that this stage-by-stage build-up clarifies the meaning of the statistical analysis and the relationship between it and the experimental design. The selection of correct tests of significance is shown to depend on the assumptions made about the data classification in the experimental design. The example is also used to show how useful supplementary information can be derived from the analysis.
C.I - EXPERIMENTAL DESIGN

Consider the situation in which $m$ parallel tests are administered to each of $n$ boys and $n$ girls (though not strictly necessary, the numbers of boys and girls are kept the same to avoid irrelevant complications). This situation may be diagrammed as follows:

<table>
<thead>
<tr>
<th></th>
<th>Boys</th>
<th>Girls</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>$T_1$</td>
<td>$T_2$</td>
</tr>
<tr>
<td>$B_1$</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$B_2$</td>
<td>*</td>
<td>*</td>
</tr>
<tr>
<td>$B_n$</td>
<td>*</td>
<td>*</td>
</tr>
</tbody>
</table>

(The asterisks represent individual scores)

An experimental design such as this would be useful in providing answers to several rather different questions which will be considered later. For the moment it is enough to point to its obvious applicability to the study of sex differences in performance on a particular sort of test. Perhaps a better word than 'sort' would be 'family', with its implication of a general resemblance overlying specific differences. In the broadest terms, this is approximately what is meant by 'parallel' tests, a matter which will be more fully discussed in later chapters.

The use of a single test of this sort or family could produce misleading results since some characteristic specific to the test, not counter-balanced in others of the family, might be dominant. Specifically, sex difference is at issue, and a single test could produce a result different from the average of
several. Moreover, the use of several will give a clearer picture, both for the overall sex bias which is characteristic of the family as a whole, and of discrepancies from the general pattern displayed in this respect by individual members. In technical terms, the use in this situation of a random sample of tests makes possible the evaluation of both the sex differences and the Sex × Test interaction.

C.II - STAGE-BY-STAGE DEVELOPMENT OF STATISTICAL ANALYSIS

Much is to be learnt from a close study of this design. It may be looked at in several slightly different ways, all of them useful.

(1) Since Children are nested within Sexes, it may be regarded as leading to a simple one-way analysis of variance, Between and Within Sexes, with replication on the same individuals within the sexes.

(2) Since both Boys and Girls are crossed with Tests, it may be seen as giving rise to a pair of two-way analyses in which one source of variation (Tests) is common to both while the other (Boys or Girls) is not. Because of the crossing, there will be an interaction line in each analysis.

(3) It may be regarded as the starting-point for a degraded three-way analysis (Children, Tests, plus replication), the degradation arising from the fact that the children in one replication are different from those in the other.

These different ways of looking at the experimental design, when taken together, lead to the conclusion that the statistical design which follows from it will be most clearly understood and will extract the maximum information if it is built up in accordance with (1), (2), and (3) above, as follows.

(1) Between and Within Sexes, for which the degrees of freedom are:
(2) (i) Children, tests and their interaction, all within sexes. There are two distinct two-way analyses (a) and (b), one for each sex.

(a)  

**TABLE IIa**

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Boys</td>
<td>n - 1</td>
</tr>
<tr>
<td>Tests</td>
<td>m - 1</td>
</tr>
<tr>
<td>Boys x Tests</td>
<td>(n - 1)(m - 1)</td>
</tr>
<tr>
<td>Total (2a)</td>
<td>mn - 1</td>
</tr>
</tbody>
</table>

(b)  

**TABLE IIb**

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Girls</td>
<td>n - 1</td>
</tr>
<tr>
<td>Tests</td>
<td>m - 1</td>
</tr>
<tr>
<td>Girls x Tests</td>
<td>(n - 1)(m - 1)</td>
</tr>
<tr>
<td>Total (2b)</td>
<td>mn - 1</td>
</tr>
</tbody>
</table>

(2) (ii) The result of pooling these tables horizontally is as follows:

**TABLE III**

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children within Sexes</td>
<td>2(n - 1)</td>
</tr>
<tr>
<td>Tests within Sexes</td>
<td>2(m - 1)</td>
</tr>
<tr>
<td>Children x Tests within Sexes</td>
<td>2(n - 1)(m - 1)</td>
</tr>
<tr>
<td>Total (within Sexes)</td>
<td>2(mn - 1)</td>
</tr>
</tbody>
</table>
Note that the degrees of freedom for the second line in the analysis of Table I and for the fourth ('Total') line in the analysis of Table III are identical.

(2) (iii) The identity just noted leads to the combination of these two analyses to give:

**TABLE IV**

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Sexes</td>
<td>1</td>
</tr>
<tr>
<td>Between Children within Sexes</td>
<td>2(n - 1)</td>
</tr>
<tr>
<td>Between Tests within Sexes</td>
<td>2(m - 1)</td>
</tr>
<tr>
<td>Children x Tests within Sexes</td>
<td>2(n - 1)(m - 1)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>2mn - 1</td>
</tr>
</tbody>
</table>

(3) (i) The analysis in Table IV is not quite complete. A glance at the original diagram will show that the tests in Tables IIIa and IIIb are the same, although the children of course are different. It is therefore possible to proceed at least part of the way towards a full three-way analysis - that part of it involving sexes and tests which are, of course, crossed. This is done by partitioning 'Between Tests within Sexes' (the third row in Table IV) into 'Between Tests' and the interaction 'Sexes x Tests'. Evidently the first of these must have \((m - 1)\) degrees of freedom, and therefore the second, by difference, \((m - 1)\) also. That these two results are consonant with the general rules for writing down degrees of freedom may be shown as follows:
13.

**TABLE V**

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Sexes</td>
<td>1</td>
</tr>
<tr>
<td>Between Tests</td>
<td>m - 1</td>
</tr>
<tr>
<td>Sexes x Tests</td>
<td>m - 1</td>
</tr>
<tr>
<td></td>
<td>Total = 2(m - 1)</td>
</tr>
</tbody>
</table>

(3) (ii) From the analyses of Table IV and Table V, the final analysis can now be written down, summarising in the most efficient manner the information presented in the original diagram. Following custom the word 'Between' is omitted, and initial letters are used for the interactions.

**TABLE VI**

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sexes</td>
<td>1</td>
</tr>
<tr>
<td>Tests</td>
<td>m - 1</td>
</tr>
<tr>
<td>S x T</td>
<td>m - 1</td>
</tr>
<tr>
<td>Children within Sexes</td>
<td>2(n - 1)</td>
</tr>
<tr>
<td>C x T within Sexes</td>
<td>2(n - 1)(m - 1)</td>
</tr>
<tr>
<td>Total</td>
<td>2mn - 1</td>
</tr>
</tbody>
</table>

The final outcome (Table VI) is seen to be an example of a 'mixed' design, in which, as Lindquist says, 'some of the comparisons are inter-subject and some are intra-subject'. Here the inter-comparisons are 'Sexes' and 'Children within Sexes' and the intra-comparisons 'Tests, 'S x T' and 'C x T within sexes'. The final analysis is in fact an example of Lindquist's
Type 1 design (ibid. page 267 et seq). Lindquist distinguishes between two factors A and B which in the design above correspond to 'Tests' and 'Sexes' respectively. The stage-by-stage building up in the preceding pages of the final analysis is however different from Lindquist's and, it is believed, simpler to follow. It is part of the subsidiary aim of this thesis to urge that, starting from a design specific to the investigation being undertaken, a stage-by-stage procedure should be followed whenever possible in order to arrive at the final statistical analysis. If the investigator follows this procedure, it is likely (a) that his final analysis will mirror his original experimental design more accurately than if he relies on the already completed analytic designs presented in publications such as those of Lindquist (49) and Cox (8), excellent though they are; (b) that his understanding of the potentialities and the limitations of the analysis will be clearer and more complete; and hence (c) that his published report will be more clearly stated and so more intelligible to his readers, particularly if he reports briefly the stages leading to the analysis when this is of some complexity.

There is no need at this stage to go into all the details of the computational procedures leading to the sums of squares, which in any case follow logically from the steps by which the final analysis is built up and from the degrees of freedom. What should be noted, however, is the insight about these sums of squares given by the build-up from first principles. For instance, the sums of squares in the fourth and fifth rows of the final analysis in Table VI are conveniently arrived at by working through two separate two-way analyses, one for boys, the other for girls and then pooling (see Tables IIa and IIb). The sources of variation in these separate analyses are either boys or Girls; Tests; and the interaction either B x T or C x T. Before these analyses are pooled, they are worth studying separately, since each provides
useful information in its own right. The customary F-tests can be carried out within each analysis. Thus, if the mean-squares for boys, Tests and the interaction B x T are designated \( B, T \) and \( BT \) respectively, then

(i) a significant mean-square ratio \( F = \frac{B}{BT} \) indicates that the group of \( m \) tests is discriminating among the boys;

(ii) a significant \( F = \frac{T}{BT} \) indicates that for the boys the tests are not strictly parallel.

The following may also be noted.

(iii) The mean error variance of a boy's total over \( m \) tests is estimated by \( mBT \), and that of a boy's score on one test by \( BT \);

(iv) the variance of boys' totals over \( m \) tests is estimated by \( mB \), and the mean variance of boys' scores over one test by \( \frac{1}{m} (\overline{B} + \frac{m - 1}{m} BT) \);

(v) the variance of boys' 'true' totals over \( m \) tests is estimated by \( m (\overline{B} - BT) \), and the mean variance of boys' 'true' scores on one test by \( \frac{1}{m} (\overline{B} - BT) \);

(vi) the reliability coefficient of boys' totals over \( m \) tests \( (r_{mm}) \) is estimated by \( (\overline{B} - BT) / \overline{B} \), and that of boys' scores on one test \( (r_{11}) \) by \( (\overline{B} - BT) / (\overline{B} + \frac{m - 1}{m} BT) \).

Similar information is of course obtainable from the corresponding analysis for girls.

Naturally it is unnecessary to report all information of this sort in a particular study in which variance analysis is used. However, a full realisation of the several potentialities of a statistical analysis must enlarge the area of choice open to the investigator and help him to select at the outset the experimental design most appropriate to the study. Between the experimental design and the statistical analysis, used in conjunction with it, there is a two-way relationship. The statistical analysis must reflect faithfully the
experimental design; and the experimental design must be one capable of
efficient statistical evaluation.

A further advantage of the build-up by stages of a final analysis of some
complexity from simpler analyses is that it becomes easy to see whether the
requirements of the assumptions basic to the final analysis are being met.
In the present example, the two analyses just discussed, one for boys, the
other for girls, are to be pooled, line by line. However, tests of significance
based on the final analysis are strictly legitimate only if the variation is
homogeneous, within corresponding lines, in the two analyses to be pooled.
For example, pooling the interaction sums of squares followed by dividing their
total by the pooled degrees of freedom to obtain the Children x Test interaction
mean square \( \bar{CT} \) would be a theoretically unsound procedure were the original
mean-squares \( \bar{BT} \) and \( \bar{GT} \) to differ significantly. In this instance, a simple
F-test would resolve the issue. If the number of analyses to be pooled were
greater than two, it would be necessary to use some other test, such as Bartlett's.
Fortunately, in an analysis of this sort, trouble arising from heterogeneity of
the data is seldom encountered. Moreover, there is some evidence (see, for
example, Norton's study, described in (49), pages 78-86) that the method
of analysis of variance is more robust than the basic assumptions seem to allow;
and that heterogeneity of variance, if it does occur, must be considerable to
be of serious consequence. However, this empirical finding, comforting though
it is, affords no warrant for failure to check that the assumptions revealed by
the stage-by-stage method of arriving at the final analysis are not being unduly
strained.
C.III - TESTS OF SIGNIFICANCE

Once the statistical analysis has been arrived at, and the computational procedures which follow from it have been carried out, the way is open to make tests of significance relevant to the hypothesis at issue. It was pointed out earlier (page 9) that the experimental design with which this discussion is concerned had an obvious application in the study of sex differences. For the purpose of such a study, the appropriate hypothesis is that the mean sex difference is zero. The mean-square for sexes, which will be termed $S$, therefore constitutes the numerator in an $F$-test, the denominator of which is one of the other mean-squares in the analysis. But which?

The answer to this question depends on the assumptions made in setting up the experiment, and these in turn depend on the purpose of the study.

Four different situations are to be distinguished.

(1) Inferences are to be made about the sex difference outcome if another random sample of $n$ boys and another random sample of $n$ girls were tested with another random sample of $m$ tests.

This situation corresponds to a model in which 'Sexes' are a fixed effect and 'Children' and 'Tests' random.

(2) Inferences are to be made about the sex difference outcome if these $n$ boys and these $n$ girls were tested with another random sample of $m$ tests.

This situation corresponds to a model in which 'Sexes' and 'Children' are fixed effects and 'Tests' random.

(3) Inferences are to be made about the sex difference outcome if another random sample of $n$ boys and another random sample of $n$ girls were tested with these $m$ tests.

This situation corresponds to a model in which 'Sexes' and 'Tests' are fixed effects and 'Children' random.
(4) Inferences are to be made about the sex difference outcome if these 
n boys and these n girls were re-tested with these m tests.

This situation corresponds to the completely fixed model.

The four situations may be represented diagrammatically as follows:-

<table>
<thead>
<tr>
<th>Situation</th>
<th>Sexes</th>
<th>Chn.</th>
<th>Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>F</td>
<td>R</td>
<td>R</td>
</tr>
<tr>
<td>2</td>
<td>F</td>
<td>F</td>
<td>R</td>
</tr>
<tr>
<td>3</td>
<td>F</td>
<td>R</td>
<td>F</td>
</tr>
<tr>
<td>4</td>
<td>F</td>
<td>F</td>
<td>F</td>
</tr>
</tbody>
</table>

For convenience, situations (1) and (4) will be left until later, and the discussion will now relate to situations (2) and (3).

It is not difficult to imagine circumstances in which situation (2) or situation (3) would be realistic. Situation (2) would arise if the aim were to know how the mean sex difference might be affected for a particular group of boys and girls if the m tests used were replaced by a set of m parallel tests. Situation (3) would arise if the aim were to know how stable the mean sex difference is in face of a change in the identity of the testees. In both situations, it is assumed that all other conditions are held constant. In situation (2) the interest would lie mainly in the children tested; in situation (3) mainly in the characteristics of the tests used.

The safest, and at the same time the most illuminating way of deciding on the appropriate tests of significance is to set out the variance analysis including several variance components which the mean-squares in the analyses respectively estimate. The results are as follows:-
TABLE VII

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>Expectation of MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Sexes</td>
<td>1</td>
<td>S</td>
<td>S</td>
<td>$\sigma_{ct.s}^2 + \sigma_{st}^2 + \sigma_{s}^2$</td>
</tr>
<tr>
<td>Tests</td>
<td>m - 1</td>
<td>T</td>
<td>T</td>
<td>$\sigma_{ct.s}^2 + \sigma_{st}^2 + 2\sigma_{t}^2$</td>
</tr>
<tr>
<td>$S \times T$</td>
<td>m - 1</td>
<td>ST</td>
<td>ST</td>
<td>$\sigma_{ct.s}^2 + \sigma_{st}^2$</td>
</tr>
<tr>
<td>Children within Sexes</td>
<td>2(n - 1)</td>
<td>$C_W$</td>
<td>$C_W$</td>
<td>$\sigma_{ct.s}^2 + \sigma_{st}^2$</td>
</tr>
<tr>
<td>$C \times T$ within Sexes</td>
<td>2(n - 1)(m - 1)</td>
<td>$C_{TW}$</td>
<td>$C_{TW}$</td>
<td>$\sigma_{ct.s}^2$</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Total</th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td></td>
<td></td>
<td></td>
<td>Total $2mn - 1$</td>
</tr>
</tbody>
</table>

In order to clear up a possible difficulty of terminology, it should be stated that $\sigma_{ct.s}^2$ and $\sigma_{st}^2$ are used to indicate within-sex variance. This terminology is based on a loose analogy of these variance components with partial variances. In the same way that $\sigma_{2.1}^2$ represents the variance of variable 2 with 1 held constant, so $\sigma_{ct.s}^2$ represents the population interaction variance for children x tests with sex held constant.

The next step is to decide which of the three effects 'Sexes', 'Tests', and 'Children' are fixed, and which random.

(1) In situation (2), 'Sexes' and 'Children' are fixed effects and 'Tests' random. The application of Schultz's rule (74) leads to the deletion of the term in $\sigma_{cos}^2$ from the expectation of the mean-square for 'Sexes.' It is then clear from the expectations that the mean-square for 'Sexes' is to be tested against the mean-square for 'Sexes x Tests', that is:

$$F = \frac{S}{ST} \text{ with } 1 \text{ and } (m - 1) \text{ degrees of freedom.}$$

A significant F would lead to the rejection of the null hypothesis (zero sex difference) under the conditions of situation (2).
In situation (3), 'Sexes' and 'Tests' are fixed effects and 'Children' random. In this case the operation of the rule leads to the deletion of the term in $\sigma_{st}^2$ from the expectation of the mean square for 'Sexes'. It is evident that the latter is now to be tested against the mean-square for 'Children within Sexes', that is:

$$F = \frac{S}{\bar{\sigma}_w} \text{ with } 1 \text{ and } 2(n - 1) \text{ degrees of freedom.}$$

A significant F would lead to the rejection if the null hypothesis (zero sex difference) under the conditions of situation (3).

G.IV - DISCUSSION

The selection of the proper tests of significance brings to an end the first part of this study. The reader is reminded that one object of the study is to illustrate in some detail how to arrive by stages at the statistical analysis which best reflects an experimental design and which makes it possible to extract the maximum amount of information from the data. Any experimental design would have served the purpose in some degree. That used was chosen because it is an example of the class of 'mixed' designs which occur frequently in the pages that follow. These 'mixed' designs, with one effect nested within another, like a Chinese puzzle, are especially suitable for the kind of research in which practical considerations impose some limitations on the investigator. Children are 'within' classes or schools, and it is often easier to obtain the co-operation of teachers if the whole class or school is tested than if some children are selected for the study and others not. Subtests occur within tests, and it is frequently more sensible to study subtests in relation to the test which they comprise than to do so in isolation. Often both effects, schools and children within schools, or tests and subtests within tests, are of interest, each complementing the other. A study in which schools are the units can be quite
misleading if it ignores variation among children within schools. Equally, information is sacrificed in a study which involves a number of children, and which pays no regard to the fact that they come from different schools. In such cases as these a 'mixed' design is a most useful instrument of research.*

In building up the statistical design in the preceding pages, and in selecting the appropriate significance tests, much has been taken for granted. This was intentional: to have digressed from time to time to justify arguments or statements would have diluted and confused the main trend of the exposition. Little indeed needs justification, since most of the procedures employed are well-known - it is the linking up of these procedures in this particular manner which is perhaps less familiar.

**C.V - RATIONALE OF TESTS OF SIGNIFICANCE**

However, there is one matter which constantly gives rise to confusion and which therefore needs discussion in some detail. This is the rationale of selecting the significance test relevant to a specific situation. In the illustrative analysis above, the rule-of-thumb proposed by Schultz has been applied without comment or discussion. This rule is stated and its use illustrated (though it is barely explained or justified), in Edwards (18, page 303). Since the rule has wide applications which continually occur in the following page, it is advisable at this point to show that the application of Schultz's rule leads to results identical with those obtained from a more basic analysis. The discussion will again relate to tests of the null hypothesis (zero sex difference) outlined above.

The model basic to the analysis of Table VII is as follows:–

---
* See, for example, Chapter 11.
where

\[ M \] is a component common to all children;

\[ s_k \] is a component common to all children of sex \( k \);

\[ t_j \] is a component specific to test \( j \);

\[ st_{kj} \] is a component resulting from interaction between sex \( k \) and test \( j \);

\[ c_{ik} \] is a component specific to child \( i \) of sex \( k \);

\[ ct_{ijk} \] is a component resulting from interaction between child \( i \) of sex \( k \) and test \( j \);

and \( i \) runs from 1 to \( n \); \( j \) from 1 to \( m \); and \( k \) from 1 to 2.

Without loss of generality, \( M \) may be put equal to zero and henceforth ignored.

It is assumed that the expectations of

\[ s_k, t_j, st_{kj}, c_{ik}, \text{ and } ct_{ijk} \]

are all zero.

To obtain the observed sex totals \( S_{k(o)} \) over \( n \) children and \( m \) tests, sum in equation (1) over \( i \) and \( j \):—

\[ S_{k(o)} = \sum_{ij} x_{ijk} = n s_k + n \sum_j t_j + n \Sigma_{j} st_{kj} + m \Sigma_{i} c_{ik} + \Sigma_{ji} ct_{ijk} \quad \ldots \quad (2) \]
The observed total in equation (2) is to be thought of as made up of two parts, 'true' total and 'error of measurement'. The 'true' total is constant over all replications of the experiment; the 'error of measurement' varies from one replication to another.

The 'true' total in a particular situation is defined as the mean of the observed totals over an infinite number of replications of the experiment under the conditions specified for that situation.

The 'error of measurement' of an observed total in a single replication is the difference between that observed total and the 'true' total.

The conditions of situation (2) require that the tests change from one replication to another, while the children are the same in all replications. Therefore, as $m \to \infty$, we have asymptotically $S_k(T)(2)$, the 'true' sex total corresponding to situation (2):

$$S_k(T)(2) = mnS_k + mn\sum \epsilon_{ijk}$$

From equations (2) and (3), therefore,

$$e_k(2) = \sum_{j} n\epsilon_j + \sum_{j} \epsilon_{kj} + \sum_{ijk} \epsilon_{ijk}$$

From equation (3), the expectation of the variance of 'true' totals in a sex comparison, which will be called the 'true' variance, is

$$E(\text{Variance of } S_k(T)(2)) = m^2n^2\sigma_s^2 + m^2n\sigma_{st}^2$$

From equation (4), the expectation of the variance of errors of measurement, which will be called 'error' variance is

$$E(\text{Variance of } e_k(2)) = mn^2\sigma_{st}^2 + mn\sigma_{ot}^2$$
Hence, under the conditions of situation (2), the expectation of observed sex totals, which will be called the 'total' variance, is

\[ E(\text{Variance of } S_k(o)) = (m^2 n^2 \sigma_s^2 + m^2 n^2 \sigma_{cs}^2) + (mn \sigma_{st}^2 + mn \sigma_{ct}^2) \]  

i.e. 'Total' variance = 'True' variance + 'Error' variance

Since \( E(\text{Variance of } S_k(o)) = m n E(\text{Mean Square for Sexes}) \), the latter may be similarly divided into 'true' and 'error' mean square:

\[ E(\text{M S for Sexes}) = (m n \sigma_s^2 + \sigma_{cs}^2) + (n \sigma_{st}^2 + \sigma_{ct}^2) \]  

(Situation 2) 

\( 'True' \) \quad \( 'Error' \) \quad 'Total'

Clearly, in testing for significance of the sex difference in situation (2), the correct error term to use is \( \overline{ST} \) (see Table VII). This is the same error term as that selected by the Schultz rule-of-thumb.

Situation (3) requires that the children change from one replication to another while the tests are the same in all replications. Putting \( n = \infty \) therefore in equation (2), we obtain eventually:

\[ E(\text{M S for Sexes}) = (m n \sigma_s^2 + \sigma_{cs}^2) + (m \sigma_{st}^2 + \sigma_{ct}^2) \]  

(Situation 3) 

\( 'True' \) \quad \( 'Error' \) \quad 'Total'

In this case, the correct error term for testing sex difference is \( \overline{C_w} \), again in agreement with the Schultz rule.

An alternative general rule for selecting 'true' and 'error' portions of a mean-square expectation is as follows:

(1) A variance component in a line contributes to 'true' variance if, in addition to the subscript or subscripts defining the line's classification, it
possesses a subscript or subscripts corresponding to any one fixed classification.

(2) A variance component in a line contributes to 'error' variance if, in addition to the subscript or subscripts defining the line's classification, it possesses a subscript or subscripts corresponding to random classifications only.

The same rule applies whether the classification is crossed with, or nested in, other classifications.

Note that the type (fixed or random) of the line classification itself is irrelevant.

Situations (2) and (3) exemplify the class of 'mixed' models in which some of the variables are fixed and some random.

As a tail-piece to this discussion, the two other experimental situations, (1) and (4), will be touched on briefly.

In situation (1), replications are envisaged as occurring with other random samples of boys and girls and other random samples of tests. The partition into 'true' and 'error' variance is then as follows:

\[
E (\text{MS for Sexes}) = (m n \sigma_s^2) + (\sigma_{ct.s}^2 + m \sigma_{c,s}^2 + n \sigma_{sb}^2)
\]

\[
\begin{align*}
\text{'True' variance} & \quad \text{'Error' variance}
\end{align*}
\]

No mean-square expectation in the analysis of variance in Table VII corresponds to error terms in this equation. The statistical design is therefore inappropriate in this situation.

* A distinction is to be made between (1) the term 'mixed' models, which relates to the status of the variables as 'fixed' or random, whatever the design; and (2) the term 'mixed' design, which relates to a particular group of statistical designs (see page 13), and which may or may not be 'mixed' models.
Finally, situation (4), in which replications are envisaged with these boys and girls and these tests, is an example of a completely 'fixed' model. Replication under precisely the same conditions would of course produce identical results. All four terms in the expectation of the mean-square must therefore be in the 'true' variance category. The results of an experiment in which the assumptions made are those of situation (4) would possibly be of interest as a record of the outcome of a single testing event occurring with a particular group of children on a specific occasion, but of no interest otherwise. With situation (4) we have moved from the realm of 'sampling' to that of 'descriptive' statistics.

However, if the restriction of 'precisely the same conditions' is removed, there is more to be said about situation (4). It would be realistic to suppose that the administration of the succession of tests would not be free of disturbance. Conditions of administration might not be identical from one test to another; minor fluctuations in attention or health might affect the performances of at least some of the testees. Errors would therefore intrude which the experimental design does not cater for and which in consequence the statistical analysis is incapable of isolating. These errors, which it is reasonable to suppose are random, would inflate the variance component \( \sigma_{ct.s}^2 \) by some unknown amount. It may be argued that if the tests used in the experiment are truly parallel, then the real within-sex interaction between Children and Test is negligible, in which case the bulk of the variance represented in the analysis by \( \sigma_{ct.s}^2 \) results from unspecified random error. In this case, the use of \( \sigma_{ct.s}^2 \) (or, more correctly, \( \bar{CT}_w \)) to test the significance of the Sex effect would be a valid procedure. In any case, if \( \bar{S} \) is significant when tested against \( \bar{CT}_w \), which, in addition to random error
variance, may include interaction variance also, then $S$ would undoubtedly be significant if tested against a variance term incorporating random error variance alone. In all cases of completely fixed models which make no provision for the isolation of random error, it is perhaps advisable to play for safety in this way.

It should be noted that this qualification of the meaning of $CT_w$ does not affect the interpretations given to situations (1), (2) and (3).

C.VI - CHILD DIFFERENCES

So far the discussion has centred on sex differences. We now turn to a study of the additional information, relating to child differences, that can be extracted from the experimental and statistical design.

It was pointed out on page 14 that the pair of two-way analyses, one for Boys, the other for Girls, merit separate study before they are pooled in the final analysis. Each furnishes information in its own right.

The following discussion is concerned with significance tests for

(i) differences among boys;
(ii) differences among girls;
(iii) differences among children of different sex;
(iv) differences among children whose sex is not known.

The starting point for the discussion is the model stated in equation (1), page 22.

Sum over $j$ in equation (1) to obtain the total over all tests for child $i$ of sex $k$:

$$T_{ijk} = \sum_j x_{ijk} = m_{ik} + \sum_j t_{kj} + \sum_j s_{kj} + u_{ik} + \sum_j c_{t_{ijk}}$$  \hspace{1cm} (11)
For a different child \(i'\) of the same sex, the corresponding total is

\[ T_{i'jk} = \sum_j x_{i'jk} = m s_{ik} + \sum_j t_{ij} + \sum_j s t_{kj} + m c_{i'k} + \sum_j c t_{i'jk} \]

The difference between the child \(i\) and child \(i'\) totals is

\[ T_{ijk} - T_{i'jk} = m (c_{ijk} - c_{i'jk}) + \sum_j (c t_{ijk} - c t_{i'jk}) \quad \text{(12)} \]

The first term on the right is the 'true' difference and the second the error of measurement of this difference. The error variance of the difference is therefore

\[ ev(T_{ijk} - T_{i'jk}) = m (\sigma_{e}^2 + \sigma_{e'}^2) \]

in which the bracketed terms on the right denote the individual error variances of measurement on the same single test for child \(i\) and child \(i'\) respectively.

Assuming homogeneity of error variance over all children in sex \(k\), we may replace the last equation by

\[ ev(T_{ijk} - T_{i'jk}) = 2m \sigma_{ct.s}^2 \quad \text{(13)} \]

whence

\[ ev_{T_{ijk}} = ev_{T_{i'jk}} = m \sigma_{ct.s}^2 \quad \text{(14)} \]

Hence, provided that comparisons are restricted to children of one sex, the appropriate error variance of measurement of a child total over \(m\) tests is

\[ m \sigma_{ct.s}^2 \]

where \(m\) is the number of tests and \(\sigma_{ct.s}^2\) the expectation of the mean-square for the Child \(x\) Test interaction in the table for that sex. For boys, this interaction is \(BT\), and for girls \(GT\) (see Tables IIa and IIb, page 11).

Comparisons between individual children of the same sex may be based therefore on \(BT\) or \(GT\), and confidence limits for individual child totals, on these mean-squares also.
If $BT$ and $GT$ are found not to differ significantly, the pooled interaction $CT_W$ in the full analysis (Table VII, page 19) should be substituted for each in making comparisons or setting confidence limits as above.

So much for intra-sex comparisons. The matter is slightly less simple if the comparison is to be made between the scores of a boy and a girl.

Starting with equation (11) and writing $b$ and $g$ for $k$, we obtain two further equations, the first for a boy's total, the second for a girl's, over the same $m$ tests:--

$$T_{ijb} = ms_b + \sum_j s_{bj} + \sum_j mc_{ib} + \sum_j ct_{ijb}$$

$$T_{ijg} = ms_g + \sum_j s_{gj} + \sum_j mc_{ig} + \sum_j ct_{ijg}$$

The difference between these totals, with a slight re-arrangement of terms, is

$$T_{ijb} - T_{ijg} = m(s_b - s_g) + m(c_{ib} - c_{ig}) + \sum_j (s_{bj} - s_{gj}) + \sum_j (ct_{ijb} - ct_{ijg}) . \quad (15)$$

The first two terms on the right of equation (15) together make up the 'true' difference, and the third and fourth together comprise the error of measurement of this difference. Assuming homogeneity of variance in the error terms, we may write therefore:

$$ev(T_{ijb} - T_{ijg}) = 2m(\sigma_{st}^2 + \sigma_{ct}^2) \quad \cdots \quad \cdots \quad \cdots \quad (16)$$

and

$$ev_{T_{ijb}} = ev_{T_{ijg}} = m(\sigma_{st}^2 + \sigma_{ct}^2) \quad \cdots \quad \cdots \quad \cdots \quad (17)$$

Hence, for comparing the scores of two children of different sexes, the appropriate error variance of measurement depends on the number of tests, on the Sex x Test interaction, and on the Child x Test interaction within Sexes. That this conclusion is reasonable may be seen by considering what would happen if
either (i) $\sigma_{st}^2 = 0, \sigma_{ct,s}^2 > 0$ or (ii) $\sigma_{st}^2 > 0, \sigma_{ct,s}^2 = 0$.

In case (i) there would be perfect correlation between sex means on the $m$ tests employed in the experiment, and hence, by inference, on other sets of $m$ tests drawn from the same population of tests. Mean sex difference would not affect the error variance, although it would rightly appear as a part of the 'true' difference, constant for every boy-girl comparison (see equation (6)).

The error variance would be attributable entirely to imperfect within-sex correlations among test scores and would be the same as for intra-sex comparisons.

In case (ii) there would be perfect correlation among children's scores within sex, so that for intra-sex comparisons the error variance would be zero. Inter-sex comparisons, however, would still be affected by the variability of sex means on the tests; the difference between totals on the same $m$ tests for a particular boy and a particular girl would still depend on the particular sample of $m$ tests taken by both. The expectation of the variance attributable solely to this source is $\sigma_{st}^2$.

Comparisons of equations (14) and (17) lead to the conclusion that the precision of measurement of intra-sex differences is higher than that of inter-sex differences, a conclusion to be borne in mind when the sexes of the children to be compared is known.

But what if the sexes of the children are not known? The question is one of some theoretical interest, and although unlikely to arise in this context, it might well do so in others.

Intuitively, one would suppose that the error variance to use in comparisons among children whose sex is unknown is the mean of the two error variances given in equations (14) and (17), that is, $m(\sigma_{st}^2 / 2 + \sigma_{ct,s}^2)$. This supposition can be tested by considering what kind of analysis of variance would replace that developed in the previous pages if the sexes of the children were disregarded.
This new analysis would have to be of the simple two-way sort, the sources of variation being 'Children', 'Tests', and the interaction 'Children x Tests'.

Now it is possible to modify the analysis of Table VII, page 19, so as to obtain a simple two-way analysis in which the distinction between the sexes is lost, a situation corresponding to the condition of sex unknown. It is to be imagined that the same m tests and the same 2n children are involved, but that it is not known which children are boys and which are girls. The analysis would then be as shown below:

**TABLE VIII**

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
<td>2n - 1</td>
<td>C</td>
<td>C</td>
</tr>
<tr>
<td>Tests</td>
<td>m - 1</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>C x T</td>
<td>(2n - 1)(m - 1)</td>
<td>CT</td>
<td>CT</td>
</tr>
<tr>
<td>Total</td>
<td>2mn - 1</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Note that the degrees of freedom for 'Children', (2n - 1) in the new analysis, are the sum of the degrees of freedom for 'Sexes', (1) and 'Children within Sexes' (2n - 2), in the analysis in Table VII, page 19. Note also that the degrees of freedom for the C x T interaction, (2n - 1)(m - 1), in the new analysis, are the sum of the degrees of freedom for S x T, (m - 1), and C x T within Sexes, (2n - 2)(m - 1).

Pooling sums of squares follows the pattern set by pooling degrees of freedom. Thus

\[ C = S + C_w, \]

and \[ CT = ST + CT_w. \]
In terms of mean-squares:

\[(2n - 1)\bar{c} = \bar{s} + 2(n - 1)\bar{c}_w\]

\[(2n - 1)(m - 1)\bar{c}_T = (m - 1)\bar{s}_T + 2(n - 1)(m - 1)\bar{c}_T\]

From the expectations of mean-squares in Table VII.

\[E(\bar{c}) = \frac{1}{2n - 1} \int \sigma_{ct,s}^2 + m\sigma_{c,s}^2 + m\sigma_s^2 + 2(n - 1)\left(\sigma_{ct,s}^2 + m\sigma_{c,s}^2\right) \, \]

\[= \sigma_{ct,s}^2 + \frac{n}{2n - 1} \sigma_{st}^2 + m\sigma_{c,s}^2 + \frac{mn}{2n - 1} \sigma_s^2\]

Also \[E(\bar{c}_T) = \frac{1}{(2n-1)(m-1)} \int (m - 1)(\sigma_{ct,s}^2 + m\sigma_{c,s}^2) + 2(n - 1)(m - 1) \sigma_{ct,s}^2 \]

\[= \sigma_{ct,s}^2 + \frac{n}{2n - 1} \sigma_{st}^2\]

As \(n \to \infty\)

\[E(\bar{c}) = (\sigma_{ct,s}^2 + \frac{1}{2} \sigma_{st}^2) + m(\sigma_{c,s}^2 + \frac{1}{2} \sigma_s^2)\]

\[E(\bar{c}_T) = (\sigma_{ct,s}^2 + \frac{1}{2} \sigma_{st}^2)\]

But \[E(\bar{c}) = \sigma_{ct}^2 + m\sigma_c^2\]

and \[E(\bar{c}_T) = \sigma_{ct}^2\]

where \(\sigma_{ct}^2\) and \(\sigma_c^2\) are variance components estimated in this new 'telescoped' analysis.

Hence \(\sigma_c^2 = \sigma_{c,s}^2 + \frac{1}{2} \sigma_s^2\) . . . . . . . . . . (18)

and \(\sigma_{ct}^2 = (\sigma_{ct,s}^2 + \frac{1}{2} \sigma_{st}^2)\) . . . . . . . . . . (19)

Equation (19) gives the expectation of the error variance of a child's score on a single test when the sex of the child is unknown. Hence the expectation of the error variance of the child's total over m tests is
$$m \sigma_{ct}^2 = m(\sigma_{ct}^2 + \frac{1}{2} \sigma_{st}^2)$$

which confirms the supposition arrived at intuitively (see page 30).

Note that \( \sigma_{ct}^2 > \sigma_{st}^2 \) unless \( \sigma_{st}^2 = 0 \). It follows that we may expect \( \overline{CT} \) to be larger than \( \overline{CT}_W \) unless the Sex x Test interaction is zero.

The conclusion of the argument is this:—

If a simple two-way analysis is carried through on the data resulting from the administration of a set of parallel tests to a group of children without regard to their sex, the derived error variance over-estimates the correct error variance appropriate to comparisons among the scores of children of the same sex; and under-estimates the correct error variance appropriate to comparisons among the scores of children of different sexes.

This loss of precision is the price that must be paid for lack of information regarding the sexes of the children whose scores are compared. The seriousness of the loss depends on the magnitude of the Sex x Test interaction, that is, on the extent to which the difference in response of boys and girls to a test (as shown by the difference in the sex means for that test) varies from one test to another.

To end this discussion on a practical note, it should be pointed out that

\[ \frac{1}{n}(\overline{ST} - \overline{CT}_W) \]

in the analysis in Table VII estimates \( \sigma_{st}^2 \), which is required for equation (19).

C.VII — TEST RELIABILITY

The analysis is capable of providing still further information. We turn now to its use in estimating test reliability. The relationship between analysis of variance and reliability will be more fully discussed in a later chapter. At present it is enough to note that the basic concept of reliability is conveyed by the equation:
Reliability = \frac{\text{True Variance}}{\text{Observed Variance}} = \frac{\text{True Variance}}{\text{True Variance} + \text{Error Variance}}

(1) Consider first the reliability of boys' test scores.

The expectation of the variance of boys' means over p tests is \( (\sigma_b^2 + \sigma_{bt}^2/p) \), in which \( \sigma_b^2 \) is 'true', and \( \sigma_{bt}^2 \) 'error' variance. Hence the expectation of the reliability of boys' means (or totals) over p tests is

\[
\rho (\text{boys - p tests}) = \frac{\sigma_b^2}{(\sigma_b^2 + \sigma_{bt}^2/p)} = \frac{p \sigma_b^2}{p \sigma_b^2 + \sigma_{bt}^2}
\]

Equation (20) is general for all values of p, and, as might be expected, \( \rho \to 1 \) as \( p \to \infty \). The variance components \( \sigma_b^2 \) and \( \sigma_{bt}^2 \) can be estimated from the mean-squares in the analysis of variance for boys completed from that for which the degrees of freedom are indicated in Table IIa, page 11. Denoting the mean-square for Boys by \( \bar{B} \) and that for the interaction by \( \bar{E}T \), we have

\[
E(\bar{B}) = m \sigma_b^2 + \sigma_{bt}^2, \quad \text{and} \quad E(\bar{ET}) + \sigma_{bt}^2
\]

whence

\[
\sigma_b^2 = \frac{1}{m} E(\bar{B} - \bar{ET}).
\]

From equation (20), therefore,

\[
\rho (\text{boys - p tests}) = E \left\{ \frac{\frac{p}{m} (B - \bar{ET})}{\frac{p}{m} (B - \bar{ET}) + \bar{ET}} \right\} = E \left\{ \frac{B - \bar{ET}}{\bar{B} + \frac{1}{p} (m - p) \bar{ET}} \right\}
\]

Hence \( \rho (\text{boys - p tests}) \) is estimated by

\[
f(\text{boys - p tests}) = \frac{\bar{B} - \bar{ET}}{\bar{B} + \frac{1}{p} (m - p) \bar{ET}}
\]
If \( p = m \), the number of tests used in the analysis,

\[
r_{(\text{boys} - m \text{ tests})} = \frac{(\bar{B} - \bar{BT})}{\bar{B}},
\]

a well-known formula,

and if \( p = 1 \),

\[
r_{(\text{boys} - 1 \text{ test})} = \frac{(\bar{B} - \bar{BT})}{(\bar{B} + (m - 1) \bar{BT})},
\]

also familiar.

The reliability of equation (21) is an example of the class of reliabilities inferred from the results of one administration of a test or test battery. This class of inferred reliabilities will be treated more fully in a later chapter. In this case, the reliability should be interpreted as an estimate of the correlation that would have resulted, if two random samples of \( p \) tests from the same test population had both been administered to the same group of boys under identical conditions, and the totals (or means) over the two test samples correlated. Similar interpretations, mutatis mutandis, applies to the reliabilities discussed below.

(2) In the same way, estimates of the reliability of girls' scores can be obtained, starting from the analysis sketched in Table IIb, page 11. Of course, \( \bar{G} \) and \( \bar{GT} \) replace \( \bar{B} \) and \( \bar{BT} \) in the equations.

(3) If the assumption of homogeneity of variance is valid, pooled (and therefore more stable) estimates of the within-sex reliability are obtainable from the full analysis of Table VII on substituting \( \bar{G}_W \) and \( \bar{GT}_W \) for \( \bar{B} \) and \( \bar{BT} \).

(4) Finally, if the sexes of individual children are not known, the mean-squares to use are \( \bar{G} \) and \( \bar{GT} \) in the analysis of Table VIII, page 31.

A closer look at the reliabilities from (3) and (4) is informative. The reliability estimated in (3) above is
(Chn. within sexes - p tests) = \frac{p \sigma_{c.s}^2}{p \sigma_{c.s}^2 + \sigma_{ct.s}^2} \quad \cdot \quad \cdot \quad \cdot \quad (22)

(see Table VII for these population variances)

The reliability estimated in (4) is

(Chn. regardless of sex - p tests) = \frac{p \sigma_{c}^2}{p \sigma_{c}^2 + \sigma_{ct}^2} \quad \cdot \quad \cdot \quad \cdot \quad (23)

(see Table VIII)

In equation (22), the components relate to populations of children within sexes, and in equation (23) to a population of children regardless of sex. The conditions for these two reliabilities to be the same is that the sexes do not differ in their mean response to the tests. If this condition holds, segregation by sex should not affect the outcome, so that whether the analysis is conducted according to Table VII or Table VIII should be immaterial. In terms of variance components, the condition can be stated as \sigma_s^2 = \sigma_{st}^2 = 0.

The argument above can readily be tested by substituting from equation (9) and (10) in equation (14), whence we obtain

\begin{align*}
\rho \text{ (Chn. regardless of sex - p tests)} &= \frac{p \sigma_{c,s}^2 + \frac{1}{2} \sigma_{s}^2}{(p \sigma_{c,s}^2 + \frac{1}{2} \sigma_{s}^2) + (\sigma_{ct,s}^2 + \frac{1}{2} \sigma_{st}^2)} \quad \cdot \quad \cdot \quad \cdot \quad (24)
\end{align*}

Equation (24) reduces to equation (22) if \sigma_s^2 = \sigma_{st}^2 = 0. The argument above is therefore validated.
This introductory chapter has been concerned with the following:

(1) Statement and preliminary discussion of the main purpose of the thesis: the study of the reliability of measurement in three major psychometric areas.

(2) Brief outline of the uses to which the statistical technique known as analysis of variance has been put in educational and psychological research.

(3) Statement of the subsidiary aim of the thesis: to exemplify some of the advantages that would accrue from a more informed use of analysis of variance techniques in educational and psychological research.

(4) Discussion of the relationship between experimental design and the statistical analysis which follows from it.

(5) An example designed to illustrate both (3) and (4), which includes:
   (a) the stage-by-stage building up of the statistical analysis based on the experimental design;
   (b) the rationale of significance testing in analysis of variance;
   (c) the relationship between two analyses when one is a 'collapsed' version of the other;
   (d) an outline (to be filled in later) of the derivation of reliability formulae from components of variance;
   (e) an introduction to the methods by which the studies which follow will be conducted, with the dual purpose of providing a general framework for these studies and of integrating them.
PART ONE

STUDIES OF TEST RELIABILITY
CHAPTER TWO

A - INTRODUCTION

The early confusion which surrounded the idea of 'reliability' in psychometric work is apparent to anyone reading the literature on the subject. The confusion seems to have arisen from the concept of the reliability coefficient of a test, and its complement, the error of measurement of a test score. The various methods of estimating test reliability, test-retest with the same test, test-retest with different tests, splitting a test into two parts, more or less equivalent, and 'boosting' the correlation between the parts; all of these seem to have been regarded as devices for measuring the reliability coefficient of a test, and as doing so with varying degrees of efficiency.

If the reliability coefficient is conceived of as something fixed for a particular test, logical difficulties must arise in using any method for its estimation. If the method is that of test-retest with the same test, then the reliability coefficient may be over-estimated if the interval between testing is short, because of recall of the test content on the second testing. If the interval is long, then the trait measured may alter materially during that interval.

With the methods of test-retest with a parallel test, the difficulties of the first method are avoided, only to be replaced by others. The parallel test - and how parallel is it? - is not the same as the first test. How, then, can the correlation between the two be the reliability coefficient of either?

The 'boosted split-half' method, while possessing the advantage over the others that only a single testing is necessary, also gave rise to difficulties. Different 'splits', with equal claims to produce equivalent half-tests, yield
different correlations between halves, and hence, on 'boosting', to different estimates of the reliability coefficient. There is the added complication that the assumption of zero correlation among errors may not hold, in which case, however 'good' a particular split, the reliability coefficient is overestimated.

This confusion has now been largely dispelled. It is now recognised that there is no single reliability coefficient or error of measurement characterising a test in all circumstances. Instead, there are several, obtainable in different ways, and differing in size. Each is regarded simply as an index of the efficiency of the test as a measuring device when used under certain conditions, and as specific to these conditions. More briefly, a reliability coefficient is an index of the efficiency of a specific measuring procedure.

PREVIOUS USE OF ANALYSIS OF VARIANCE IN RELIABILITY STUDIES

The most general formulation of a test's reliability is

\[ \hat{\rho}_{TT} = 1 - \frac{\sigma_e^2}{\sigma_T^2} \]

in which \( \hat{\rho}_{TT} \) is the reliability coefficient; \( \sigma_T^2 \) is the variance of test scores, or 'total' variance; and \( \sigma_e^2 \) is the variance of uncontrolled measurement error, or 'error variance of measurement'.

The error variance is obviously subject to change as the conditions of testing change, and the reliability coefficient will alter accordingly.

The shape of the formula above makes it obvious that the method of analysis of variance is an appropriate tool for studying reliability. Hoyt (38) was among the first to use analysis of variance to generalise the Kuder-Richardson
Formula 20 (45) which relates specifically to tests with dichotomously scored items. Burt (7) and Mahmoud (53) showed that various reliability formulae could be expressed in terms of variance components estimated from analyses of variance. Finlayson (25) and Pilliner (64) used 3-way analyses to study the reliability of essay marking. They too used variance components to express reliability coefficients. Ebel (17) studied the reliability of ratings using the analysis of variance. Rajaratnam (69) discussed the problems arising in estimating the reliability of scores assigned to different persons in different tests. In a series of brilliant articles, Cronbach and his collaborators have clarified the distinctions between different kinds of reliability coefficient, and have raised the discussion of the issue of reliability to a new level of sophistication in their liberalization of reliability theory. (14), (70) They too have employed analysis of variance methods. Lindquist devotes a chapter in his treatise on experimental design to the estimation of variance components and their use in reliability studies (49).

A.I - PURPOSES OF THIS STUDY

The purposes of this study are as follows:

(1) To use the method of analysis of variance to show the relationship between the Coefficient of Equivalence, the Coefficient of Internal Consistency, the Coefficient of Stability, and the Coefficient of Stability and Equivalence.

(2) To suggest a practical routine procedure for the estimation of the Coefficient of Equivalence from test data.

(3) To discuss the difficulties arising when comparisons have to be made among scores obtained on different tests.

(4) To investigate the systematic standardisation error, or 'zero' error, of a series of Moray House Verbal Reasoning Tests.
Before embarking on the main study, it will be useful to discuss briefly the relationship between reliability and validity. It is recognised, of course, that the over-riding consideration is the validity of a test in respect of the criterion it is intended to predict. This has sometimes been taken to mean that provided the test is a valid predictor, its reliability is of secondary importance. One can only disagree with this statement. Broadly interpreted, it could mean that it does not matter if the ranking produced by Test A differs materially from that produced by Test B, provided that either ranking correlates equally well with the criterion.* But it does matter to the testees.

Moreover, the unreliability of a test sets a limit to its validity. As Spearman said: 'The relations of reliability and validity are one-sided. Low reliability necessarily involves low validity, but the converse is not true. Wherever we find bad agreement between different measurements, then we can safely say that the examination is bad. But when the measurements agree we cannot forthwith say that the examination is good.' (79) Spearman's comment exactly sums up the relationship between the two concepts.

More recently, Cronbach has commented on the relationship from a slightly different standpoint. 'Even those investigators who regard reliability as a pale shadow of the more vital matter of validity cannot avoid considering the reliability of their measures. No validity coefficient and no factor analysis can be interpreted without some appropriate estimate of the magnitude of the error of measurement.' (13), page 297

* Burt (7) points out the fallacy of Eysenck's argument that reliability must be adequate if predictive validity is satisfactory. Eysenck is referring (23) to a test with a validity coefficient of .7. As Burt says, a test with this validity might have a reliability coefficient of \(0.7^2 = 0.49\), which would be quite inadequate for most purposes.
The truth of the matter is as follows. If the criterion is to be predicted by a single test, then measures on it should be both as valid and as reliable as possible. If the criterion is to be predicted by a composite or battery of tests, then high validity may be achieved through deliberate choice of tests, which, though correlating well individually with the criterion, do not intercorrelate highly among themselves. However, within each test in the composite, high reliability is desirable. Whatever each part of the composite measures, it should measure as accurately as possible. This is a slightly different statement of the issue from that of Ferguson (24).

A.III - 'PARALLEL' TESTS

Basic to the definition of parallel tests is the concept of a universe or pool of questions or items from which those incorporated in the tests are to be drawn. Tests are parallel if they consist of the same number of items selected from the pool by identical sampling procedures. It is perhaps not possible to improve on Henrysson's formulation of this basic idea (37). He points out that if all the items in the pool measure 'one behaviour domain', then a random sampling procedure is adequate, unless the items differ in difficulty. In this case, a stratified sampling procedure, based on difficulty levels, is appropriate. If these are 'two or more domains of behaviour', if, that is, the pool consists of several item sub-pools, each testing a different facet of ability, then the stratified procedure should be extended to take account of these differences also.

The pool of items is of course conceptual. In practice, no test constructor proceeds in this manner. This, however, in no way diminishes the usefulness of the concept, which is, indeed, no more than a specific application of the general concept of a 'population' on which most of the theory of statistical inference is based. We need not be deterred from using the concept because in
fact the pool of items does not exist. To quote Cronbach again: 'There is no practical testing problem where the items in the test and only these items constitute the trait under examination. We may be unable to compose more items because of our limited skill as test-makers, but any group of items in a test of intelligence or knowledge or emotionality is regarded as a sample of items. If there weren't "plenty more where these came from," performance on the test would not represent performance on any more significant variable.' (13)

Accepting, then, the concept of the item pool, we go on to discuss the implications of the definition of parallel tests stated above. For the educationist or psychologist, it implies that parallel tests are similar in content. For the test statistician, it implies that parallel tests have the same expectations of statistical characteristics such as mean, standard deviation, shape of distribution, and correlation with other parallel tests (that is, reliability). For the test constructor, it points the way to the method of constructing parallel tests in practice. In general, the definition implies that one parallel test may legitimately be substituted for any other, and that the differences in outcome are the result of sampling error only. Note that the use of the word 'substituted' implies constancy of testing conditions. Note also that by a simple extension of the argument, the word 'test' can be replaced by 'group' or 'battery' of tests, it being understood that for groups to be parallel they must be assembled, at least conceptually, from a universe of individual tests by identical sampling procedures.
CHAPTER THREE

A - THE COEFFICIENT OF EQUIVALENCE

A.I - DEFINITION

The Coefficient of Equivalence is the correlation between scores on two parallel tests administered - as Cronbach puts it - 'simultaneously'. Stated otherwise, it is the correlation between observed scores on a test and hypothetical scores on a parallel test that might have been substituted for it on the single occasion of testing. The Coefficient of Equivalence therefore cannot be obtained directly, but must be inferred from the evidence afforded by the internal relationships among the part-scores on this one occasion of testing.

This formulation of the definition seems preferable to that sometimes seen (for example, Cronbach (13), page 298), that the Coefficient of Equivalence is the correlation between two parallel tests administered at 'virtually' the same time. The term 'virtually' does not sufficiently emphasise the essential idea of simultaneity of administration. If we must think in terms of two separate administrations, it is better to specify the two different occasions of testing as identical.

Ferguson preferred to define test reliability as 'the accuracy (not constancy) with which a test measures the abilities which it measures at the time it measures them' (24). At that time (1940) he had in mind such measures of reliability as those given by 'boosted split-half' methods or by the Kuder-Richardson formulae, but the sentence quoted is an apt definition of a test's Coefficient of Equivalence.
The test constructor normally assembles the items for his test in accordance with some principles. Often his objective is to construct the test so that scores over all the items in it, that is, total scores, have the validity characteristic he requires. With this kind of test, part-scores, that is, scores on particular groups of items, are of secondary interest. Even so, the constructor will use some educational or psychological principle in choosing items for his test. For example, a verbal reasoning ('intelligence') test will include item types such as analogies, series, 'odd man out', classification, and other types of item known to justify inclusion in the test. It is possible, at least notionally, to regard these groups of different types of item as 'subtests' or 'strata' within the test.

However, tests may be intentionally structured for more specific reasons. Situations are common in which the relations among scores on the various subtests are no less important, educationally or psychologically, than the global test score. One example is the use of a test battery* for counselling purposes; a second is its use for diagnosing areas of weakness; a third is its use in the preparation of profiles for record cards. In all these cases, differences among part-scores are to be interpreted.**

Finally, a test may be structured for the specific purpose of promoting validity. A test consisting of several subtests, within which inter-item correlations are high, but between which they are low, often possess high predictive validity.*** As will be shown, a test constructed in this way may also be highly reliable.

* The extension from subtest-within-test to test-within-battery is obvious.
** For discussion of the reliability of these differences see page 81.
*** See also page 42.
A. III - EXPERIMENTAL DESIGN AND MODEL

These considerations lead to the experimental design and model appropriate to the study of the Coefficient of Equivalence.

The model must make provision for the possibility of low inter-subtest correlation alongside high intra-subtest correlation. It must also reflect the complex structure of a conceptual universe from which parallel tests may be derived by a stratified sampling procedure (see page 42). Such tests will be called 'stratified-parallel' tests.

The experimental design is as follows:

<table>
<thead>
<tr>
<th>Item</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th></th>
<th>1'</th>
<th>2'</th>
<th>3'</th>
<th></th>
<th>1''</th>
<th>2''</th>
<th>3''</th>
</tr>
</thead>
<tbody>
<tr>
<td>Person 1</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x'</td>
<td>x'</td>
<td>x'</td>
<td></td>
<td>x''</td>
<td>x''</td>
<td>x''</td>
</tr>
<tr>
<td>2</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x'</td>
<td>x'</td>
<td>x'</td>
<td></td>
<td>x''</td>
<td>x''</td>
<td>x''</td>
</tr>
<tr>
<td>3</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td></td>
<td>x'</td>
<td>x'</td>
<td>x'</td>
<td></td>
<td>x''</td>
<td>x''</td>
<td>x''</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Each subtest is a random sample of items from a different sub-universe. All persons work all the items in all the subtests.

Although not diagrammed in the same way, the design is formally identical with that shown in Chapter 1, page 9. It will be seen that items are nested within subtests, and that both items and subtests are crossed with persons. Hence there will be two interactions: Persons x Subtests, and Persons x Items within Subtests. The design then leads to a model making provision for the separate
existence of the two different correlations referred to above. Since, within
each of a fixed number of sub-universes, the diagram may be thought of as
extending indefinitely in a horizontal direction, it also reflects the
conceptual complex universe which is to be the source of 'stratified-parallel'
tests.

The model giving expression to the design is

\[ x_{ijk} = M + p_i + s_k + ps_{ik} + q_{jk} + c_{ijk} \]  \hspace{1cm} (1)

\( x_{ijk} \) is the score of person \( i \) on item \( j \) in subtest \( k \).

\( M \) is the general mean over the universe of scores.

\( p_i \) is common to all scores made by person \( i \).

\( s_k \) is common to all scores on items in subtest \( k \).

\( ps_{ik} \) is the interaction for person \( i \) and subtest \( k \); it measures the
discrepancy between \( i \)'s level of performance on subtest \( j \) and his
general level of performance on the whole test.

\( q_{jk} \) is specific to item \( j \) nested in subtest \( k \).

\( c_{ijk} \) is the interaction for person \( i \) and item \( j \) in subtest \( k \); it measures
the discrepancy between \( i \)'s level of performance on item \( j \) in subtest \( k 
and his general level of performance on subtest \( k \).

For the test, \( i = 1, 2, ... , l \); \( j = 1, 2, ... , m \); \( k = 1, 2, ... , n \).

It is assumed that these components of score are independent of each other,
and that the sum of each over its subscript or subscripts in the universe is
zero.

A.IV - DERIVATION OF RELIABILITY COEFFICIENTS

Sum, first over \( j \) and then over \( k \) to obtain person \( i \)'s total score \( T_i \) on
the whole test:-
The expectation of the variance of test totals over persons is therefore

$$E(\text{Var. } T) = m^2 n \sigma_p^2 + m^2 n \sigma_{ps}^2 + mn^2 \sigma_{pq.s}^2$$  \hspace{1cm} (3)$$

With a different test, 'stratified-parallel' to the first, the new test totals $$T'_i$$ will have the same expectation of variance over the same persons as that given by equation (3).

The expected covariance over persons of $$T_i$$ and $$T'_i$$ will be

$$E(\text{Cov. } TT') = m^2 n^2 \sigma_p^2 + m^2 n \sigma_{ps}^2$$  \hspace{1cm} (4)$$

This is so because the subtests are the same for both tests, but the items different within the subtests.

The expected correlation between the test totals is therefore

$$\rho_{TT'} = \frac{mn \sigma_p^2 + m \sigma_{ps}^2}{mn \sigma_p^2 + m \sigma_{ps}^2 + \sigma_{pq.s}^2}$$  \hspace{1cm} (5)$$

$$\rho_{TT'}$$ is the Coefficient of Equivalence of the test. It is the inferred correlation between scores on the test in hand and scores on a conceptual test which is 'stratified-parallel' to it and which is administered 'simultaneously' with it, or, alternatively, administered under identical conditions.

A.V - INFERENCES FROM THE MODEL

(i) Put $$n = 1$$ in equation (5). We obtain the

Expected Coefficient of Equivalence for a Subtest of $$m$$ Items:
The subsequent development requires the correlations between persons' scores on different subtests. This is easily obtained from the model of equation (1).

Expected Correlation between Person Totals on Different Subtests:

\[
\rho_{SS'} = \frac{m \sigma_p^2 + m \sigma_{ps}^2}{m \sigma_p^2 + m \sigma_{ps}^2 + \sigma_{pq's}^2}
\]

\[
\rho_{S_1S_2} = \frac{m \sigma_p^2}{m \sigma_p^2 + m \sigma_{ps}^2 + \sigma_{pq's}^2}
\]

(iii) We now consider two extreme cases. These are tabulated for convenience below.

<table>
<thead>
<tr>
<th>Case</th>
<th>( \sigma_{ps}^2 \rightarrow 0 )</th>
<th>( \sigma_{pq's}^2 \rightarrow 0 )</th>
<th>( \sigma_{ps}^2 &gt; 0 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>(a)</td>
<td>( \sigma_{ps}^2 \rightarrow 0 )</td>
<td>( \frac{mn \sigma_p^2}{mn \sigma_p^2 + \sigma_{pq's}^2} )</td>
<td>1</td>
</tr>
<tr>
<td>(b)</td>
<td>( \sigma_{pq's}^2 \rightarrow 0 )</td>
<td>( \frac{m \sigma_p^2}{m \sigma_p^2 + \sigma_{pq's}^2} )</td>
<td>( &lt;1 )</td>
</tr>
<tr>
<td></td>
<td>( \sigma_{ps}^2 &gt; 0 )</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Discussion

Case (a)

Making \( \sigma_{ps}^2 \rightarrow 0 \) is equivalent to saying that there is no real differential response on the part of persons to the different subtests. Two conclusions emerge. The first is that a simpler model would have served to obtain \( \rho_{TT'} \). Anticipating subsequent discussion, we may say that in this case the Coefficient of

* See page 63.
Internal Consistency would have served as well as the Coefficient of Equivalence as a measure of the 'simultaneous' reliability of the test.

The second point is that $\rho_{SS} = \rho_{S_1S_2}$. This means that the correlation between parallel subtests is no greater than that between different subtests. In this sense, the test is homogeneous, and stratification has been ineffective from a statistical point of view.

The two conclusions are of course connected. If $\sigma_{ps}^2 = 0$, construction of the test by random sampling would have served as well as stratified sampling. But this could not have been known in advance. Commenting on the relation between stratified and random sampling in another connection, Tippett writes:

'A knowledge of the sources of variation and of their relative importance such as that given by an investigation using an analysis of variance and by a priori technical knowledge is of great assistance. When there are no real variations between the strata, sampling by strata is at its worst, and even then is as good as random sampling.' (89, page 210)

**Case (b)**

Making $\sigma_{pq,s}^2 \to 0$, and leaving $\sigma_{ps}^2 > 0$ is equivalent to saying that the item error variance within subtests approaches zero, while the error variance between subtests does not. All the items within a subtest are measuring the same trait, which is different from that measured by the others. The purpose of case (b) is to demonstrate the possibility of high overall reliability ($\rho_{TT} \to 1$) despite subtest heterogeneity ($\rho_{S_1S_2} < 1$).

These extreme cases have an important implication for the test constructor. He can structure, or stratify, his test in any manner which seems to him educationally or psychologically desirable without sacrificing reliability, provided that within each stratum he selects items which inter-correlate as
highly as possible. Items for a particular subtest should therefore be selected on the basis of their correlation with that subtest pool, not the aggregate over all subtests. In this way he will maximise the reliability of the test as measured by the Coefficient of Equivalence. He will also increase the likelihood of good predictive validity, provided that the several subtests each correlate well with the criterion he wishes to predict (see page 42).

A.VI - RELIABILITY OF DIFFERENCE SCORES

It was pointed out earlier (page 45) that a testee's part-scores and the differences between them are frequently as important as his total score when the test is to be used for counselling, diagnostic, or profile, purposes. Since the differences among part-scores are to be interpreted, and decisions based on them, their reliability is an important matter.

For individual subtests* 1 and 2, the reliability** of the person differences \(d(1 - 2)\) is given by

\[
\rho_{d(1-2)} = \frac{\rho_{11} + \rho_{22} - 2\rho_{12}}{2(1 - \rho_{12})}
\]

( (48), page 777; (33), page 353)

The analysis of variance provides an overall test of the reliability of differences among part-scores.

Sum over \(j\) in the model of equation (1), page 47, to obtain \(S_{ik}\) the total for person \(i\) on subtest \(k\):

\[
S_{ik} = \sum_{j} x_{ijk} = m_{i} + m_{k} + m_{ps_{ik}} + \sum_{j} g_{ijk} + \sum_{j} p_{qijk}
\]

* The extension from subtest-within-test to test-within-battery is obvious.

** This is easily proved using a pooling-square.
Similarly, the total $S_{ik}$ for person $i$ over subtest $k'$ is:

$$S_{ik} = \sum_{j} x_{ijk'} = m_{p} + m_{s_{k'}} + \sum_{j} q_{jk} + \sum_{j} p_{jq_{jk}}$$  \hspace{1cm} (9)$$

From equations (8) and (9), the difference between person $i$'s scores on subtests $k$ and $k'$ is:

$$S_{ik} - S_{ik'} = m(s_{k} - s_{k'}) + m(ps_{ik} - ps_{ik'}) + \sum_{j} (q_{jk} - q_{jk'}) + \sum_{j} (p_{ijk} - p_{q_{jk}})$$  \hspace{1cm} (10)$$

The expectation of the variance over persons of the difference in equation (10) is therefore:

$$E(\text{Var. } S_{ik} - S_{ik'}) = 2m^{2} \sigma^{2}_{pq.} + 2m \sigma^{2}_{pq.}$$  \hspace{1cm} (11)$$

The expectation of the variance of the corresponding difference in a 'stratified-parallel' test is also given by equation (11).

The expectation of the covariance is

$$E(\text{Cov. } S_{ik} - S_{ik'}) = 2m^{2} \sigma^{2}_{ps}$$  \hspace{1cm} (12)$$

The expectation of the mean reliability of differences among subtests is therefore

$$R(\sum_{ik} - \sum_{ik'}) = \frac{m \sigma^{2}_{ps}}{m \sigma^{2}_{ps} + \sigma^{2}_{pq.}}$$  \hspace{1cm} (13)$$

Equation (13) shows yet again the importance of minimising $\sigma_{pq.}^{2}$, the mean error variance of score on an item. As equations (11) and (13) show, the expected mean error variance of a person's subtest total is $2m \sigma_{pq.}^{2}$. If this error variance makes a substantial contribution to the total variance of difference scores (equation 11), then little reliance can be placed on the subtest differences and attempts to interpret them may be misleading.*

* For further discussion, see page 55.
B - ESTIMATION OF VARIANCE COMPONENTS

From the experimental design (page 46) and the model (page 47), the appropriate analysis of variance follows:

TABLE I - ANALYSIS OF VARIANCE FOR STRATIFIED TEST

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>E(MS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons</td>
<td>((\ell - 1))</td>
<td>(P)</td>
<td>(\bar{F})</td>
<td>(\sigma_{pq,s}^2 + m\sigma_{ps}^2 + mn\sigma_p^2)</td>
</tr>
<tr>
<td>Subtests</td>
<td>((n - 1))</td>
<td>(S)</td>
<td>(\bar{S})</td>
<td>(\sigma_{pq,s}^2 + \sigma_{qs}^2 + m\sigma_{ps}^2 + l_m\sigma_s^2)</td>
</tr>
<tr>
<td>(P \times S)</td>
<td>((\ell - 1)(n - 1))</td>
<td>(PS)</td>
<td>(\bar{PS})</td>
<td>(\sigma_{pq,s}^2 + m\sigma_{ps}^2)</td>
</tr>
<tr>
<td>Items within (S)</td>
<td>(n(m - 1))</td>
<td>(Q_W)</td>
<td>(\bar{Q}_W)</td>
<td>(\sigma_{pq,s}^2 + l\sigma_{qs}^2)</td>
</tr>
<tr>
<td>(P \times I) within (S)</td>
<td>(n(\ell - 1)(m - 1))</td>
<td>(PQ_W)</td>
<td>(\bar{PQ}_W)</td>
<td>(\sigma_{pq,s}^2)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td></td>
<td>(\ell mn - 1)</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

B.1 - THE COEFFICIENT OF EQUIVALENCE

'Persons' and 'subtests' are fixed effects and 'Items' random. The appropriate significance test for \(\bar{F}\) is therefore \(F = \bar{F} / \bar{Q}\). This test of significance is logically consistent with the argument leading to equation (5).

The variance components for equation (5) are estimated by

\[
\begin{align*}
\sigma_{pq,s}^2 &= \bar{PQ}_W \\
\sigma_{ps}^2 &= \frac{1}{m}(\bar{PS} - \bar{PS}_W) \\
\sigma_p^2 &= \frac{1}{mn}(\bar{F} - \bar{PS})
\end{align*}
\]

and the coefficient of equivalence is estimated by
\[ r_{TT'} = \frac{\text{ms}_p^2 + \text{ms}_p^2}{\text{ms}_p^2 + \text{ms}_p^2 + s_{pq,s}^2} \quad (14a) \]

\[ = \frac{\bar{F} - \bar{F}_W}{\bar{F}} \quad (14b) \]

Equation (14a) is more general than equation (14b). Once the variance components have been estimated, the Coefficient of Equivalence can be estimated for other values of m (that is, other sizes of subtest).

**B.II - MEAN CORRELATIONS BETWEEN PERSON TOTALS ON SUBTESTS**

From equation (7), the estimate of \( \rho_{s_1s_2} \) is

\[ r_{s_1s_2} = \frac{\text{ms}_p^2}{\text{ms}_p^2 + \text{ms}_p^2 + s_{pq,s}^2} \quad (15a) \]

\[ = \frac{\bar{F} - \bar{F}_S}{\bar{F} + (n - 1)\bar{F}_S} \quad (15b) \]

**B.III - MEAN RELIABILITY OF DIFFERENCE SCORES**

From equation (13), the reliability of difference scores is estimated by:

\[ r(S_{ik} - S_{ik'}) (S_{ik} - S_{ik'})' = \frac{\text{ms}_p^2}{\text{ms}_p^2 + s_{pq,s}^2} \quad (16a) \]

\[ = \frac{\bar{F}_S - \bar{F}_W}{\bar{F}_S} \quad (16b) \]

The reliability of difference scores is not significant if the test

\[ F = \frac{\bar{F}_S}{\bar{F}_W} \]

shows that \( \bar{F}_S \) is not significantly greater than \( \bar{F}_W \). A non-
significant result would cast doubt on the efficiency of the test in serving any of the purposes for which difference scores are relevant.

If there are two subtests only, which we may call 1 and 2, the Persons x Subtest mean-square is

\[ ms^2 \left( \bar{x}_{i1} - \bar{x}_{i2} \right) / 2, \]

in which \( \bar{x}_{i1} \) and \( \bar{x}_{i2} \) are person \( i \)'s means on subtests 1 and 2 respectively. If there are \( n \) subtests,

\[ FS = \frac{m}{n(n - 1)} \sum_{k,k'} \left( x_{ik} - x_{ik'} \right)^2 \]

\[ k = 1, 2, \ldots, n; \quad k' = 1, 2, \ldots, n; \quad k \neq k'. \]

The differences \( (\bar{x}_{ik} - \bar{x}_{ik'}) \) are the 'difference scores' which are important for such purposes as counselling, diagnosing of weakness, and profiles. The estimate of the mean error variance of these differences is \( 2 \frac{F_Q}{W} / n \), so that, for a particular person \( i \), the mean difference \( (\bar{x}_{ik} - \bar{x}_{ik'}) \) may be tested against its standard error \( \sqrt{2 \frac{F_Q}{W} / n} \), using Student's \( t \) test, with \( 2(m - 1) \) degrees of freedom. Considerable doubt attaches to the validity of the \( t \)-test, however, if the overall test, \( F = FS / \frac{F_Q}{W} \) (d.f. \( (\ell - 1)(n - 1), n(\ell - 1)(m - 1) \) ) produces a non-significant result. In this case, Lindquist would regard individual \( t \)-tests as 'improper', since 'all observed differences could be due to chance alone' ((49), page 96). Even if the requirements of overall significance is relaxed, and a procedure* such as Duncan's (16) is used to test individual differences for significance, it would be hazardous to attach much importance to the results. Difference scores are notoriously unstable, particularly when the correlation between the subtests on which the differences

* This procedure is illustrated in (20), Chapter 10.
are based does not differ greatly from the mean reliability of those subtests (see page 49). It seems probable that an analysis of variance used on the results obtained with a test or test battery intended for diagnostic purposes would be a useful preliminary step in assessing the efficiency of the test or battery for its intended purpose.
CHAPTER FOUR

A - THE COEFFICIENT OF INTERNAL CONSISTENCY

Much discussion has been devoted in the literature to the group of measures of test reliability now generally classed together as Coefficients of Internal Consistency. All these measures are inferred from the relationships among part-scores obtained by a group of testees on a single occasion of testing, and all were originally regarded as reliability coefficients.

With the 'split-half' method, the test is divided arbitrarily into two parts of equal length, the part-scores are correlated, and the correlation is 'boosted' using the Spearman-Brown formula (Brown (5), page 290, and Spearman (78), page 299). Variants of the method were devised by Rulon (72) and Flanagan (28).

The split-half methods of estimating reliability have been criticised on the ground that they do not give the same result as that obtained by test-retest procedures. This criticism is symptomatic of the semantic confusion which once existed and to which reference has already been made. There is, of course, no reason why the results of procedures so different should agree. Cronbach (13) has distinguished between measures of equivalence (typified by split-half coefficients) and measures of stability (typified by test-retest coefficients).

The split-half methods have also been criticised on the ground that the coefficients obtained are inconsistent from one split to another. Again, it is to be expected that the correlation between two half-tests (and hence that between whole-tests inferred from it) will be affected by the manner in which the test is divided. The correlation between two half-tests is a measure of their equivalence, and it should be obvious that different 'splits' will produce
half-tests exhibiting different degrees of equivalence.

A method statistically preferable would be to divide the test into more than two parts, to intercorrelate the parts, and finally to apply the generalised Spearman-Brown formula to the mean inter-correlation in order to infer the reliability of the whole test. In this way, chance inequalities would tend to cancel out.

The procedure above, taken to its logical extreme, is the method described by Kuder and Richardson. Their method amounts virtually to finding the mean inter-item correlation and 'boosting' by the number of items in the test. The best known of the several working formula devised by Kuder and Richardson is that known as KR20:-

\[
\rho_{tt} = \frac{n}{n - 1} \left(1 - \frac{\sum_{i=1}^{n} p_i q_i}{s_t^2}\right) \quad (i = 1, 2, \ldots, n) \quad . . \quad (17)
\]

\(p_i\) is the proportion of testees passing item \(i\); \(q_i = (1 - p_i)\); \(s_t^2\) is the variance of test scores; and \(n\) is the number of items in the test.

It is interesting that although Kuder and Richardson stated the underlying assumption that the test is uni-factor, they did not reach the obvious conclusion that whole-test measures inferred in this way are measures of internal consistency.

Following the methods suggested by Johnson and Neyman (43) and by Jackson (41), Hoyt (38) applied analysis of variance to a matrix of scores obtained by each of \(k\) persons on each of \(n\) items. His starting-point was the method devised by Rulon for estimating whole-test reliability from half-test scores:

\[
r_{tt} = 1 - \frac{s^2}{s^2(x + y)}
\]
where \( x \) and \( y \) are the scores obtained by same testee on the two half-tests \( X \) and \( Y \). Hoyt pointed out that \( s^2_{(x + y)} \), the variance of whole-test scores, is independent of the particular 'split', but that \( s^2_{(x - y)} \), the variance of the difference, is not. The variance \( s^2_{(x - y)} \) is a measure of the discrepancy between 'true' and 'observed' scores, but in practice is likely to be unstable since it depends on the particular split.

Hoyt suggested remedying this defect by using a two-way analysis of variance on a matrix of scores, leading to the reliability estimate

\[
r_{tt} = (\bar{P} - \bar{E}) / \bar{P}
\]

in which \( \bar{P} \) is the mean-square for Persons and \( \bar{E} \) the Persons x Items interaction mean-square. Hoyt shows that \( \bar{E} \) is the 'best' estimate of the discrepancy variance in the light of the least-squares criterion.

Hoyt also pointed out that when the test items are dichotomously scored, equation (18) and equation (17) are algebraically identical. It appears to have been left to Cronbach ((13), page 299) to state the general expression for 'Coefficient alpha':

\[
\alpha = \frac{m}{m-1} \left(1 - \frac{\sum_i s^2_i}{s^2_t}\right)
\]

Equation (19) follows directly from an equation couched in terms of mean-square from an analysis of variance:

\[
\alpha = \frac{m}{m-1} \left(1 - \frac{\bar{W}_I}{\bar{P}}\right)
\]

Equation (20) is readily derived from equation (18).

Cronbach also proved rigorously that 'Coefficient alpha' is the mean of all
possible split-half coefficients for a given test (ibid. pages 304 - 306).

B - BASIC SIMILARITY OF THOSE METHODS

B.I - GENERAL

These methods - 'boosted split-half' and Rulon's and Flanagan's variants of it; Kuder-Richardson's KR20 and Hoyt's analysis of variance - are all basically similar. The same model serves for them all:

\[ x_{ij} = M + p_i + q_j + p_{qij} \]  

in which \( x_{ij} \) is the score of person \( i \) on test \( j \); \( M \) is common to all scores; \( p_i \) is specific to person \( i \); \( q_j \) is specific to item \( j \); and \( p_{qij} \) is the discrepancy between 'true' and 'observed' scores. The usual assumptions apply, and \( i \) runs from 1 to \( Q \), and \( j \) from 1 to \( n \).

With all of the methods discussed, the expectation of the reliability of the whole test is

\[ \rho_{TT} = \frac{n \sigma_p^2}{n \sigma_p^2 + \sigma_{pq}^2} \]  

The proof of the above equation is as follows:-

Assume that a test is assembled from \( n \) items selected at random from a universe of items. Then the total \( T_i \) of person \( i \) on the whole test is

\[ T_i = n p_i + \sum_j q_j + \sum_j p_{qij} \]  

The expectation over \( i \) of the variance of the \( T \)'s is

\[ E(\text{Var. } T) = n^2 \sigma_p^2 + n \sigma_{pq}^2 \]
If a parallel test is assembled from \( n \) other items randomly chosen from the same universe of items, then equation (24) gives the expected variance of person totals on this test also.

The expectation of the covariance, or 'true' variance, is

\[
E(\text{Cov. \, TT}) = n^2 \sigma_p^2
\]

and the expectation of the discrepancy variance, or 'error' variance, is

\[
E(\text{Discrepancy Var.}) = n \sigma_{pq}^2
\]

The expectation of the test reliability is therefore given by equation (22).

B.II - 'BOOSTED SPLIT-HALF'

From equation (21), the expected total for person \( i \) on a half-test of \( n/2 \) items is

\[
T_i(1/2) = \frac{n}{2} p_i + \sum_j q_j + \sum_j p_{ij} j \quad (j = 1, 2, ..., n/2)
\]

The expectation of the correlation between two half-tests is therefore

\[
\rho_{(T/2)(T/2)} = \frac{n \sigma_p^2}{\left( n \sigma_p^2 + 2 \sigma_{pq}^2 \right)}
\]

On substituting from equation (27) in the Spearman-Brown formula

\[
\rho_{TT} = \frac{2 \rho_{(T/2)(T/2)}}{1 + \rho_{(T/2)(T/2)}}
\]

we obtain the expectation of equation (22).

B.III - KUDER-RICHARDSON

It was earlier stated that the Kuder-Richardson procedure is virtually a 'boosting' to full test-length of the mean item intercorrelation. It is easy to
see that the latter has the expectation

$$\rho_{ii} = \frac{\sigma_p^2}{(\sigma_p^2 + \sigma_{pq}^2)}$$

which when substituted in the general Spearman-Brown formula:

$$\rho_{TT} = \frac{n\rho_{ii}}{1 + (n - 1)\rho_{ii}}$$

at once gives the expectation of equation (22).

\[\text{B. IV - RULON}\]

From equation (21) the expected difference between two half-test totals for person \(i\) is

$$\frac{1}{2} T_i - \frac{1}{2} T_i' = \sum_j (q_j - q_j') + \sum (p_{qij} - p_{qij}') \quad (j = 1, 2, \ldots, n/2)$$

The expected variance over \(i\) of these differences is

$$E(\text{Var. diffs.}) = 2 \cdot \frac{n}{2} \cdot \sigma_{pq}^2$$

$$= n \sigma_{pq}^2$$

which, from equation (26), is the expectation of the discrepancy variance \(\sigma_T^2\) 'error' variance for the whole test. This result at once leads to the reliability expectation of equation (22).

(5) To sum up: The basic similarity of all the methods discussed above for inferring whole-test reliability from a single administration of the test is made clear by (i) stating a model (equation (21)); (ii) deriving from general principles the expectation of the inferred reliability in terms of variance.
components (equation (22)); (iii) following the procedures specific to the methods discussed; and (iv) arriving in all cases at the same expectation of whole-test reliability.

C.I - LIMITATION OF THE MODEL

The model of equation (21) assumes that a single factor accounts for all of person i's 'true' score and hence for all of the 'true' variance. If this assumption is valid, then the $P_{TT}$ of equation (22) is a valid measure, both of the internal consistency of the test, and of its reliability. If the assumption of unit rank in the score matrix is not valid, then $P_{TT}$ is not a completely valid measure of the reliability of the test, though it is still valid as a measure of its internal consistency.

These internal consistency measures are however still frequently used as estimates of reliability for tests of various kinds, whether or not their internal structure is adequately represented by the model of equation (21). Cronbach (13), Guttman (34) and others have shown that the reliability is under-estimated if the model is not adequate. Hitherto, the method of analysis of variance has not been used to demonstrate the extent of this under-estimate, or the conditions under which it occurs.

C.II - MODEL OF EQUATION (21) USED WITH A HETEROGENEOUS TEST

It will be recalled that Table I, page 53, related to a test consisting of n subtests, each containing m items, that is, mn items in all. If this test, were analysed as though it were a homogeneous test, the analysis would be based on the model of equation (21) with $j = 1, 2, \ldots, mn$, instead of on that of equation (1). This procedure is tantamount to ignoring the division by subtest. The analysis of variance corresponding to this situation is shown in Table II.
**TABLE II - ANALYSIS OF VARIANCE APPROPRIATE FOR A HOMOGENEOUS TEST**

<table>
<thead>
<tr>
<th>Source</th>
<th>DF</th>
<th>SS</th>
<th>MS</th>
<th>E(MS)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons</td>
<td>l-1</td>
<td>P</td>
<td>Π</td>
<td>$\frac{\sigma_{PQ}^2}{\sigma_{P}^2} + mn \sigma_{P}^2$</td>
</tr>
<tr>
<td>Items</td>
<td>mn-1</td>
<td>Q</td>
<td>Ψ</td>
<td>$\frac{\sigma_{PQ}^2}{\sigma_{Q}^2} + e \sigma_{Q}^2$</td>
</tr>
<tr>
<td>$P \times I$</td>
<td>$(l-1)(mn-1)$</td>
<td>$PQ$</td>
<td>$\bar{PQ}$</td>
<td>$\sigma_{PQ}^2$</td>
</tr>
<tr>
<td>Total</td>
<td>$\sum mn - 1$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table II is a 'collapsed' version of Table I. The 'Persons' line is unchanged. The 'Items' line in Table II is obtained by pooling the degrees of freedom and the sums of squares for the lines relating to 'Subtests' and 'Items within Subtests' in Table I. Finally, the $P \times I$ line in Table II is derived by pooling the degrees of freedom and the sums of squares for the $P \times S$ and $P \times I$ within $S$ lines in Table I. The new table thus analyses the test as though it is homogeneous; in fact, it may or may not be homogeneous.

In Table II, the mean-square expectations have been given capital letters as subscripts to distinguish them from the mean-square expectations in Table I.

Noting that $\text{SS}(PQ) = \text{SS}(PS) + \text{SS}(P_{QW})$

(Table II) (Table I)

we obtain

\[(mn - 1) \bar{PQ} = (n - 1) \bar{PS} + n(m - 1) \bar{PQ}\]

On substituting expectations we have
\[ \sigma_{PQ}^2 = \sigma_{pq.s}^2 + \frac{m(n-1)}{mn-1} \sigma_{ps}^2 \quad \cdots \quad (28) \]

\[ \sigma_{pq.s}^2 + \frac{n-1}{n} \sigma_{ps}^2 \quad \cdots \quad (28a) \]

if \( m \) is large.

Hence \( \sigma_{PQ}^2 > \sigma_{pq.s}^2 \) unless \( \sigma_{ps}^2 = 0 \) or unless \( n = 1 \).

\( \sigma_{pq.s}^2 \) is the error variance of score on a single item corresponding to the Coefficient of Equivalence (Table I, page 53), while \( \sigma_{PQ}^2 \) is the error variance of a single item corresponding to Coefficient alpha (Table II, page 64). Moreover, the test score data are the same whether analysed by the method of Table I or of Table II. Writing \( \sigma_T^2 \) for the expectation of the test variance, we have, for the Coefficient of Equivalence

\[ C. \ Eq. = 1 - \frac{\sigma_{pq.s}^2}{\sigma_T^2} \quad \cdots \quad (29) \]

and for Coefficient alpha

\[ \alpha = 1 - \frac{\sigma_{PQ}^2}{\sigma_T^2} \quad \cdots \quad (30) \]

Hence Coefficient alpha is an underestimate of the Coefficient of Equivalence unless \( \sigma_{ps}^2 = 0 \) or unless \( n = 1 \). If \( \sigma_{ps}^2 = 0 \), the Persons x Test interaction in Table II is not significant; from a statistical point of view the division of the tests into subtests is unnecessary. If \( n = 1 \), there is, of course, no division by subtest.

Since all the practical routines for estimating reliability which were discussed in Chapter Four are in effect estimates of varying efficiency of Coefficient alpha, it follows that these also may be expected to underestimate the Coefficient of Equivalence.
The expectation of coefficient alpha for a test consisting of \( n \) subtests, each with \( n \) items, may be written:

\[
E(\alpha) = E \left[ \frac{\sqrt{(\bar{P} - \bar{F})}}{\bar{F}} \right]
\]

\[
= \frac{mn \sigma_p^2}{(mn \sigma_p^2 + \sigma_{pq}^2)}
\]

In terms of the variance components in Table I, this may be written

\[
E(\alpha) = \frac{(\sigma_p^2 + \frac{m-1}{mn-1} \sigma_{ps}^2)}{(\sigma_p^2 + \frac{m-1}{mn-1} \sigma_{ps}^2) + \frac{n-1}{n(mn-1)} \sigma_{ps}^2 + \frac{1}{mn} \sigma_{pq,s}^2}
\] (31)

Equation (31) shows how coefficient alpha takes account of the two possible sources of low internal consistency. One of these sources is a large Person x subtest interaction resulting from inter-subtest heterogeneity. Coefficient alpha allows for this by including part of the variance component \( \sigma_{ps}^2 \) in the error variance. If the number of subtests \( (n) \) is large, this proportion is substantial.

The other source of low internal consistency is poor inter-item correlation. The expectation of the mean inter-item correlation is:

\[
\rho_{qq} = \frac{(\sigma_p^2 + \frac{m-1}{mn-1} \sigma_{ps}^2)}{(\sigma_p^2 + \sigma_{ps}^2 + \sigma_{pq,s}^2)}
\]

Suppose that \( \sigma_{ps}^2 = 0 \); that is, a single factor accounts for the whole of the 'true' variance. Even so, \( \rho_{qq} \), and hence Coefficient alpha, may be small if \( \sigma_p^2 \) is small. Equation (31) shows how coefficient alpha takes account of this source of low internal consistency also.

Thus, whatever may be its short-comings as a measure of reliability, coefficient alpha is an efficient measure of internal consistency.
(i) The expectation of the Coefficient of Equivalence is

\[ \rho_{TT} = \frac{\sigma_p^2 + \sigma_{ps}^2}{\sigma_p^2 + \sigma_{ps}^2/n + \sigma_{pq}^2/s_n} \]  

(see equation (5), page 48)

\[ \rho_{TT} \] is estimated by

\[ r_{TT} = \frac{(\bar{r} - \bar{r}_0)}{\bar{r}} \]  

(see equation (14b), page 54.)

(ii) The expectation of Coefficient alpha is

\[ E(\alpha) = \frac{\sigma_p^2 + (m-1)\sigma_{ps}^2}{\sigma_p^2 + \sigma_{ps}^2/n + \sigma_{pq}^2/s_n} \]  

(see equation (31), page 66)

\[ E(\alpha) \] is estimated by

\[ \alpha = \frac{(\bar{r} - \bar{r}_0)}{\bar{r}} \]  

(see equation (18), page 59, and Table II, page 64)

\[ \alpha = \frac{n}{n-1} \cdot \frac{\bar{r} - \bar{r}_1}{\bar{r}} \]  

(see equation (20), page 59)

\[ \alpha = \frac{n}{n-1} \cdot \frac{s_t^2 - \bar{r}^2}{s_i^2} \]  

(see equation (19), page 59)

(iii) The Coefficient of Equivalence is the best estimate of the 'simultaneous' reliability of a structured test.

(iv) Coefficient alpha underestimates the Coefficient of equivalence unless \( \sigma_{ps}^2 = 0 \), in which case the two coefficients are the same.
(v) As $m \to \infty$, Coefficient alpha $\to$ Coefficient of Equivalence, and both $\to$ unity.

(vi) For the reasons stated in (iv) and (v), Coefficient alpha is a lower bound to the Coefficient of Equivalence.

(vii) The use of any of the routine procedures for estimating the 'simultaneous' reliability of a test (split-half, Kuder-Richardson, Rulon, two-way analysis of variance) leads to an estimate of coefficient alpha.

(viii) The model of equation (1) and the analysis of variance of Table I (or working routines based on it*) will always lead to the most efficient estimate of the Coefficient of Equivalence of a test, whether or not there is a significant Persons x Subtest interaction. If the interaction is not significant, the resulting estimate of the Coefficient of Equivalence also estimates Coefficient alpha. The model of equation (21) and the analysis of Table II always gives an estimate of coefficient alpha, but this is only an estimate of the Coefficient of Equivalence under the conditions of (iv) and (v) above.

(ix) The model of equation (1) and the analysis of Table I make possible the estimation of the error variance and the reliability of difference scores, an important consideration if these differences are to be used for educational or psychological purposes.

---

* See Chapter Seven, page 99.
A - THE RELATIONSHIP BETWEEN STABILITY AND EQUIVALENCE COEFFICIENTS

A. I - DEFINITIONS

A test's Coefficient of Stability is the correlation between the sets of scores obtained when the same test is administered twice to the same testees, with a specified interval between the administrations.

A test's Coefficient of Stability and Equivalence* is the correlation between the sets of scores obtained when two parallel tests are administered to the same testees with a specified interval between the administrations.

The information provided by stability coefficients arising from test-retest procedures is obviously different from that furnished by equivalence coefficients inferred from the results of a single testing. Stability coefficients measure the persistence of a trait over a period of time. Equivalence coefficients estimate the accuracy of measurement of a trait on the single occasion of measurement.

This distinction is emphasised in the recommendations made to test constructors and publishers by both the American and the British Psychological Societies, (1954) and (1959). Both stress the importance of giving the test user full information concerning the types of reliability coefficient that may be quoted in the test manual.

A. II - PURPOSE OF THIS STUDY

The purpose of this study is to devise models which will help to clarify the distinction between stability and equivalence estimates of reliability. The

* This is an ambiguous name for what is essentially a stability coefficient.
method of analysis of variance will be used.

An experimental design which will serve this purpose must clearly make provision for estimating both equivalence and stability measures of reliability. It must therefore incorporate a procedure capable of providing internal test data on one occasion together with a replication of the procedure on at least one other occasion.

B.I - DESIGN A

The test is a unifactor test for which the two-way analysis of variance data obtained on a single occasion leads to estimates of both Coefficient alpha and the Coefficient of Equivalence. The design is as follows:

<table>
<thead>
<tr>
<th></th>
<th>$q_1$</th>
<th>$q_2$</th>
<th>\ldots</th>
<th>$q_j$</th>
<th></th>
<th>$q_1$</th>
<th>$q_2$</th>
<th>\ldots</th>
<th>$q_j$</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>$P_1$</td>
<td>$\times$</td>
<td>$\times$</td>
<td>\ldots</td>
<td>$\times$</td>
<td></td>
<td>$\times$</td>
<td>$\times$</td>
<td>\ldots</td>
<td>$\times$</td>
<td></td>
</tr>
<tr>
<td>$P_2$</td>
<td>$\times$</td>
<td>$\times$</td>
<td>\ldots</td>
<td>$\times$</td>
<td></td>
<td>$\times$</td>
<td>$\times$</td>
<td>\ldots</td>
<td>$\times$</td>
<td></td>
</tr>
<tr>
<td>$P_1$</td>
<td>$\times$</td>
<td>$\times$</td>
<td>\ldots</td>
<td>$\times$</td>
<td></td>
<td>$\times$</td>
<td>$\times$</td>
<td>\ldots</td>
<td>$\times$</td>
<td></td>
</tr>
</tbody>
</table>

The same test, consisting of items drawn at random from a universe of items, is administered on each of two occasions to the same testees.

In this design, all classifications are completely crossed, and no complication occurs arising from nesting. The model must make provision for all possible first and second order interactions among the classifications.

The model leading to a straightforward three-way analysis of variance is appropriate to this design.
The score $x_{ijk}$ of person $i$ on item $j$ on occasion $k$ is made up as follows:

$$x_{ijk} = M + p_i + q_j + c_k + pq_{ij} + pq_{ik} + qo_{jk} + pqo_{ijk}.$$  \hspace{1cm} (32)

in which

- $M$ is a component common to all scores and taken as zero;
- $p_i$ is a component specific to person $i$;
- $q_j$ is a component specific to item $j$;
- $c_k$ is a component specific to occasion $k$;
- $pq_{ij}$ is a component representing the interaction between person $i$ and item $j$;
- $pq_{ik}$ is a component representing the interaction between person $i$ and occasion $k$;
- $qo_{jk}$ is a component representing the interaction between item $j$ and occasion $k$;
- $pqo_{ijk}$ is a component representing the second order interaction between person $i$, item $j$, and occasion $k$.

The usual assumptions are made as to independence and summation to zero in the population of scores.

- $i = 1, 2, \ldots, \ell$; $j = 1, 2, \ldots, m$; $k = 1, 2, \ldots, n$.

From equation (32), summing over $j$ and $k$ to obtain person $i$'s total over all items and occasions, we obtain

$$T_i = \sum_{kj} x_{ijk} = \sum_{j} \sum_{k} x_{ijk} = mnp_i + n\sum_{j} q_j + m\sum_{k} c_k + n\sum_{j} \sum_{k} pq_{ij} + m\sum_{k} \sum_{j} pq_{ik} + \sum_{j} \sum_{k} qo_{jk} + \sum_{j} \sum_{k} \sum_{i} pqo_{ijk}.$$  \hspace{1cm} (33)

The expectation of the variance over $i$ of person totals is

$$E(Var. T) = m^2n\sigma_p^2 + mn^2\sigma_{pq}^2 + m^2n\sigma_{po}^2 + mn\sigma_{qo}^2.$$  \hspace{1cm} (34)

If the test is regarded as fixed and the occasions random, the expectation of the covariance or true variance is
E(Cov.) = mn \sigma_p^2 + mn \sigma_{pq}^2 \quad \ldots \ldots \quad (35)

The expectation of the 'error' variance is then

E(Error Var.) = m^2 \sigma_{po}^2 + mn \sigma_{pqo}^2 \quad \ldots \ldots \quad (36)

**B.11 - RELIABILITY COEFFICIENTS**

The same test of m items is administered on each of n random occasions, and then re-administered on n other random occasions. The expected correlation between the two overall totals is:

\[ \rho(T_1T_2) = \frac{mn \sigma_p^2 + n \sigma_{pq}^2}{mn \sigma_p^2 + n \sigma_{pq}^2 + m \sigma_{po}^2 + \sigma_{pqo}^2} \quad \ldots \ldots \quad (37) \]

If the totals are taken over one occasion only, the expectation of the error variance of test totals is obtained by putting n = 1 in equation (36).

E(Error Var.) = m^2 \sigma_{po}^2 + m \sigma_{pqo}^2 \quad \ldots \ldots \quad (38)

If the same test is administered, first on one random occasion, and then on another, the expectation of the correlation between the totals is:

\[ \rho(T_1T_2) = \frac{m \sigma_p^2 + \sigma_{po}^2}{m \sigma_p^2 + \sigma_{pq}^2 + m \sigma_{po}^2 + \sigma_{pqo}^2} \quad \ldots \ldots \quad (39) \]

\[ \rho(T_1T_2) \] is a Coefficient of Stability.

Using similar arguments it may be shown that if randomly parallel tests of m items are administered on the same occasions, the expectation of the correlation between the totals is
\[
\rho_{(T_{101})(T_{201})} = \frac{m \sigma_p^2 + m \sigma_{pq}^2}{m \sigma_p^2 + \sigma_{pq}^2 + m \sigma_{po}^2 + \sigma_{pqo}^2} \quad \ldots \ldots \quad (4.0)
\]

\[
\rho_{(T_{101})(T_{201})} \text{ is a Coefficient of Equivalence, and also, for this model, a Coefficient alpha.}
\]

Finally, if randomly parallel tests of \( m \) items are administered on two random occasions, the expectation of the correlation between the totals is

\[
\rho_{(T_{101})(T_{201})} = \frac{m \sigma_p^2}{m \sigma_p^2 + \sigma_{pq}^2 + m \sigma_{po}^2 + \sigma_{pqo}^2} \quad \ldots \ldots \quad (4.1)
\]

\[
\rho_{(T_{101})(T_{201})} \text{ is a Coefficient of Stability and Equivalence.}
\]

**B. III - Interpretation of Variance Components**

These results are not difficult to interpret. The variance component \( \sigma_p^2 \) measures the variation over persons of the constant part of each person's score on an item. It is independent both of the particular identity of the item and of the occasion on which it is administered. Therefore \( \sigma_p^2 \) always contributes to 'true' score.

The variance component \( \sigma_{pq}^2 \) measures the variation over persons of the differential responses of each person to specific items regardless of the occasion. In non-technical language, \( \sigma_{pq}^2 \) measures personal idiosyncrasies with respect to particular items, persisting over all occasions of testing. Therefore \( \sigma_{pq}^2 \) contributes to 'true' variance whenever the same test is repeated.
The variance component $\sigma_{po}^2$ measures the variation over persons of the differential response of each person to particular occasions of testing, regardless of the test content. Therefore $\sigma_{po}^2$ contributes to 'true' variance for every item (and therefore for the whole test) administered on that occasion.

Finally, the variance component $\sigma_{pqc}^2$, which is specific to a particular item and a particular occasion, is obviously always error variance if either item or occasion changes.

B.IV - COMPARISON OF RELIABILITY COEFFICIENTS

All that can be said with certainty from a study of equations (39), (40) and (41) is that the expectation of the Coefficient of Stability and Equivalence is the smallest of the three, if full significance of the variance components is established. No definite conclusion can be reached about the relative magnitudes of the Coefficient of Equivalence and the Coefficient of Stability. Cronbach ((13), page 308) reached a similar conclusion from rather different arguments.

C - ESTIMATION OF VARIANCE COMPONENTS

The variance components in equations (34) - (41) may be estimated from the three-way analysis of variance in Table III.
TABLE III - ANALYSIS OF VARIANCE FOR PERSONS, ITEMS AND OCCASIONS (TEST HOMOGENEOUS)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>$E(\text{MS})^*$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons</td>
<td>$(l - 1)$</td>
<td>$P$</td>
<td>$\bar{P}$</td>
<td>$\sigma^2_{pq} + m \sigma^2_{po} + n \sigma^2_{pq0} + mn \sigma^2_p$</td>
</tr>
<tr>
<td>Items</td>
<td>$(m - 1)$</td>
<td>$Q$</td>
<td>$\bar{Q}$</td>
<td></td>
</tr>
<tr>
<td>Occasions</td>
<td>$(n - 1)$</td>
<td>$0$</td>
<td>$\bar{0}$</td>
<td></td>
</tr>
<tr>
<td>$P \times I$</td>
<td>$(l - 1)(m - 1)$</td>
<td>$PQ$</td>
<td>$\bar{PQ}$</td>
<td>$\sigma^2_{pq} + n \sigma^2_{pq}$</td>
</tr>
<tr>
<td>$P \times O$</td>
<td>$(l - 1)(n - 1)$</td>
<td>$PO$</td>
<td>$\bar{PO}$</td>
<td>$\sigma^2_{pq} + m \sigma^2_{po}$</td>
</tr>
<tr>
<td>$I \times O$</td>
<td>$(m - 1)(n - 1)$</td>
<td>$QO$</td>
<td>$\bar{QO}$</td>
<td></td>
</tr>
<tr>
<td>$P \times I \times O$</td>
<td>$(l - 1)(m - 1)(n - 1)$</td>
<td>$PQO$</td>
<td>$\bar{PQO}$</td>
<td>$\sigma^2_{pq}$</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>$lmn - 1$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

*Only those required for the estimation are recorded.

From Table III:

- $E(\bar{PQO}) = \sigma^2_{pqo}$
- $\frac{1}{m} E(\bar{PO} - \bar{PQO}) = \sigma^2_{po}$
- $\frac{1}{n} E(\bar{PQ} - \bar{PQO}) = \sigma^2_{pq}$
- $\frac{1}{mn} E(\bar{P} - \bar{PQ} - \bar{PO} + \bar{PQO}) = \sigma^2_p$

Since equation (37) - (41) are general, the estimates above can be used in the equations inserting any desired value of $m$. Thus the reliabilities of tests of
any length can be estimated, provided that the variance component estimates derived from the analyses are relevant for the tests concerned.

D - INFORMATION FROM INTERMEDIATE STAGES

In the Introductory Chapter it was urged (pages 5 and 6) that whenever possible the investigator should take the opportunity of obtaining supplementary information inherent in the experimental design. In the present study the opportunity to do so does arise.

In order to arrive at the final three-way analysis of Table III, page 75, the routine procedure is to set up three two-way analyses: one for Persons and Occasions after summing over Items; one for Persons and Items after summing over Occasions; and one for Items and Occasions after summing over Persons. These three analyses give the sums of squares for the three main effects, Persons, Occasions and Items, and the three first-order interactions, \( P \times O \), \( P \times I \) and \( I \times O \). The sum of squares for the second-order interaction \( P \times I \times O \) is the balance between the total of the six sums of squares mentioned above and the corrected sum of squares for all the entries in the original table of scores.

If the investigator uses this routine procedure, he misses the opportunity afforded by the design (page 70) of obtaining separate estimates of the Coefficient of Equivalence (or Coefficient alpha) for each occasion separately. For each of the \( n \) occasions a two-way analysis may be derived like that of Table II (page 64). This procedure enables the investigator to judge the stability of the Coefficient of Equivalence of the test as the occasion changes.

Using the same notation as before to denote within-occasion variances, we may write for a single occasion \( k \):-
\[ \hat{p}_{kk} = \frac{m \sigma^2_{p.k}}{m \sigma^2_{p.k} + \sigma^2_{pq.k}} \]  

which is estimated by

\[ r_{kk} = \frac{(\bar{r}_k - \bar{r}_{Q,k})}{\bar{r}_k} \]

It is a matter of algebra to show that the mean of all the variance components \( \sigma^2_{p.k} \) is

\[ \frac{1}{n} \sum_k \sigma^2_{p.k} = \sigma^2_{p.k} = (\sigma^2_p + \sigma^2_{po}) \]  

(see Table III)

and that the mean of all the \( \sigma^2_{pq.k} \)'s is similarly

\[ \overline{\sigma^2_{pq.k}} = (\sigma^2_{pq} + \sigma^2_{pog}) \]  

(see Table III)

Equation (4.0), page 73, may therefore be written

\[ \hat{p} = \frac{m \sigma^2_{p.k}}{m \sigma^2_{p.k} + \sigma^2_{pq.k}} \]

This result indicates that the Coefficient of Equivalence obtained from the three-way analysis of Table III is the result of pooling over all occasions and thus relates to an 'average' occasion. Since it takes all the information into account, the \( \hat{p} \) of equation (4.0) or equation (4.3) is the best least square estimate available, but it is useful to know how the observed single-occasion estimates vary about this pooled estimate.

E.1I - DESIGN B

The test is structured, and contains \( n \) subtests, each of \( m \) items. For a single occasion, the design is that shown on page 46, the model is that of
The new design, incorporating occasions, is to be thought of as a series of replications of the design on page 46, each replication corresponding to one occasion. It is convenient to think of the layout on page 46 as repeated $n$ times in a horizontal direction. The test, consisting of $m$ subtests, is the same on each replication. Each subtest is a random sample of items from a different sub-universe. All persons work all items in all subtests and on all occasions.

Persons are crossed with Items, Subtests and Occasions; Subtests are crossed with Occasions; and Items are nested within Subtests. The main effects will therefore be: Persons, Subtests, Occasions, and Items within Subtests. There will be first-order interactions between: Persons and Subtests; Persons and Occasions; Subtests and Occasions; Persons and Items within Subtests; Occasions and Items within Subtests. There will be second-order interactions between: Persons, Subtests and Occasions, and between Persons, Items within Subtests, and Occasions.*

The model incorporating these various requirements is as follows:

$$x_{ijkh} = M + p_i + s_k + o_h + ps_{ik} + p^o_{ih} + s^o_{kh} + p^o s_{ikh} + q_{jk} + Pq_{ijk} + q^o_{jkh} + Pq^o_{ijkh}$$

$x_{ijkh}$ is the score of person $i$ on item $j$ within subtest $k$ on occasion $h$, and it is made up of the following independent components:

---

*The analysis may be regarded as a degraded four-way analysis, the degradation arising from the fact that the items in the subtests differ from one subtest to another (see Introductory Chapter, (3), page 10.)
\( M \) is common to every score, and as usual is taken as zero.

\( p_i \) is specific to person \( i \).

\( s_k \) is specific to subtest \( k \).

\( o_h \) is specific to occasion \( h \).

\( p_{sk} \) is the interaction between person \( i \) and subtest \( k \).

\( p_{ih} \) is the interaction between person \( i \) and occasion \( h \).

\( s_{kh} \) is the interaction between subtest \( k \) and occasion \( h \).

\( p_{skh} \) is the second-order interaction for person \( i \), subtest \( k \), and occasion \( h \).

\( q_{jk} \) is specific to item \( j \) in subtest \( k \).

\( p_{ijk} \) is the interaction between person \( i \) and item \( j \) in subtest \( k \).

\( s_{jkh} \) is the interaction between occasion \( h \) and item \( j \) in subtest \( k \).

\( p_{ijkh} \) is the second-order interaction between person \( i \), occasion \( h \), and item \( j \) in subtest \( k \).

\( i = 1, 2, \ldots, I; \quad j = 1, 2, \ldots, m; \quad k = 1, 2, \ldots, n; \quad h = 1, 2, \ldots, u, \) and the usual assumptions are made about summation to zero in the population.

Sum over \( j, k \) and \( h \) to obtain person \( i \)'s total score \( T_i \) over all items in all subtests replicated on all occasions:

\[
T_i = \sum_{j, k, h} \sum \left[ \mu \sum_{x, y} p_{ijkh} = \sum \left[ \mu \sum_{x, y} p_{ijkh} \right] + \sum \left[ \mu \sum_{x, y} p_{ijkh} \right] \right]
\]

The expectation of the variance of person totals is therefore:

\[
E(\text{Var. } T) = m^2 n^2 u^2 \sigma_p^2 + m^2 n^2 u^2 \sigma_{ps}^2 + m^2 n^2 u^2 \sigma_{po}^2 + m^2 n^2 u^2 \sigma_{ps}^2 + m^2 n^2 \sigma_p^2 \sigma_{po}^2 + m^2 n^2 \sigma_p^2 \sigma_{ps}^2 + m^2 n^2 \sigma_p^2 \sigma_{po}^2 \sigma_{ps}^2 (45)
\]

The expectation of the variance of similar totals over the same or 'stratified-parallel' tests on \( n \) other occasions is also given by equation (45).

The expectation of the covariance or 'true' variance depends on which effects are fixed and which are random.
(i) If the occasions and subtests are fixed, and the items within subtests are random, the expectation of the covariance is:

\[ E(\text{Cov.}) = \frac{2 m n u \sigma_p^2 + m n u \sigma_{ps}^2 + m n u \sigma_{po}^2 + 2 m n u \sigma_{ps}^2}{m n \sigma_p^2 + m n \sigma_{po}^2 + m \sigma_{ps}^2 + m \sigma_{ps}^2} \quad \ldots \quad (4.6) \]

From equations (4.5) and (4.6), putting \( u = 1 \), we obtain the expectation of the mean Coefficient of Equivalence for a test consisting of \( n \) subtests, each with \( m \) items, for an 'average' occasion:

\[ E(\text{C. Eq.}) = \frac{m n \sigma_p^2 + m \sigma_{ps}^2 + m \sigma_{po}^2}{m n \sigma_p^2 + m n \sigma_{po}^2 + m \sigma_{ps}^2 + m \sigma_{po}^2 + \sigma_{pq}^2 + \sigma_{pq}^2} \quad \ldots \quad (4.7) \]

(ii) If the subtests and items within subtests are fixed, and the occasions random, the expectation of the covariance is:

\[ E(\text{Cov.}) = \frac{2 m n u \sigma_p^2 + 2 m n \sigma_{ps}^2 + m n u \sigma_{po}^2}{m n \sigma_p^2 + m n \sigma_{po}^2 + m \sigma_{ps}^2 + m \sigma_{po}^2 + \sigma_{pq}^2 + \sigma_{pq}^2} \quad \ldots \quad (4.8) \]

From equations (4.5) and (4.7), putting \( u = 1 \), we obtain the expectation of the Coefficient of Stability:

\[ E(\text{C. Stab.}) = \frac{m n \sigma_p^2 + m \sigma_{ps}^2 + \sigma_{pq}^2}{m n \sigma_p^2 + m n \sigma_{po}^2 + m \sigma_{ps}^2 + m \sigma_{po}^2 + \sigma_{pq}^2 + \sigma_{pq}^2} \quad \ldots \quad (4.9) \]

(iii) If the subtests are fixed, and occasions and items within subtests are random, the expectation of the covariance is:

\[ E(\text{Cov.}) = \frac{2 m n u \sigma_p^2 + 2 m n u \sigma_{ps}^2}{m n \sigma_p^2 + m n \sigma_{po}^2 + m \sigma_{ps}^2 + m \sigma_{po}^2 + \sigma_{pq}^2 + \sigma_{pq}^2} \quad \ldots \quad (5.0) \]

From equations (4.5) and (5.0), putting \( u = 1 \), we obtain the expectation of the Coefficient of Stability and Equivalence:

* See discussion on page 77.
E(C. Stab. and Eq.) = \frac{mn \sigma_p^2 + m \sigma_{ps}^2}{mn \sigma_p^2 + mn \sigma_{pc}^2 + m \sigma_{ps}^2 + m \sigma_{pso}^2 + \sigma_{pq.s}^2 + \sigma_{pqo.s}^2} \quad (51)

### E.III - RELIABILITY OF DIFFERENCE SCORES

By procedures analogous to those employed on pages 51-2, the reliability of difference scores may be estimated from the model of equation (44).

Sum over j to obtain person i's total for subtest k:

\[ \Sigma x_{ijkh} = m_{p_i} + m_{ps_{ik}} + m_{p_{oih}} + m_{pso_{ikh}} + \sum_{j} p_{q_{ijk}} + \sum_{j} p_{q{o_i}_{jkh}} + \text{other components} \quad (52) \]

of score not containing i in the subscript and therefore irrelevant)

The same person's total on subtest k' on the same occasion is

\[ \Sigma x_{ij'kh} = m_{p_i} + m_{ps_{ik'}} + m_{p_{oih}} + m_{pso_{ik'h}} + \sum_{j} p_{q_{ijk'}} + \sum_{j} p_{q{o_i}_{jkh}} + \text{other components} \quad (53) \]

The difference between the two totals (relevant components only) is

\[ d_i = m(ps_{ik} - ps_{ik'}) + m(pso_{ikh} - pso_{ik'h}) + \sum_{j} (pq_{ijk} - pq_{ijk'}) + \sum_{j} (pq{o_i}_{jkh} - pq{o_i}_{jkh'}) \]

The expectation of the variance over persons of this difference is

\[ E(Var. S_k - S_{k'}) = 2(m \sigma_{ps}^2 + m \sigma_{pso}^2 + m \sigma_{pq.s}^2 + m \sigma_{pqo.s}^2) \quad (54) \]

1. The expectation of the correlation between these differences and those obtained from a 'stratified-parallel' test on the same occasion is
\[
\rho_{dd(1)} = \frac{m(\sigma_{ps}^2 + \sigma_{ps0}^2)}{m(\sigma_{ps}^2 + \sigma_{ps0}^2) + (\sigma_{pq,s}^2 + \sigma_{pq0,s}^2)}
\]  \quad \quad (55)

Equation (55) should be compared with equation (13), page 52, to which it corresponds.* It is a measure of the stability of subtest differences in the face of change in item content, other conditions (occasions) being held constant. This correlation is akin to the Coefficient of Equivalence (see equation (47)).

(2) It may be asked: How stable are subtest differences in the face of change in the occasion of testing? There are two possibilities: (a) the item content is the same on both occasions; (b) both the occasions and the item content change.

(2a) If the item content is the same but the occasions are different, the first and third terms in equation (54) are covariance and the others error variance. The expectation of the reliability of differences under these conditions is therefore

\[
\rho_{dd(2a)} = \frac{m \sigma_{ps}^2 + \sigma_{pq,s}^2}{m(\sigma_{ps}^2 + \sigma_{ps0}^2) + (\sigma_{pq,s}^2 + \sigma_{pq0,s}^2)}
\]  \quad \quad (56)

The reliability of equation (56) is a measure of the stability of subtest differences in the face of change of occasion, subtest content being held constant. This correlation is akin to the Coefficient of Stability (see equation (49)).

(2b) If both item content and occasion change, only the first term in equation (54) is covariance, and the three other terms are error variance. The expectation of the reliability of differences under these conditions is therefore

* Equation (55) relates to an 'average' occasion, and, following the discussion on pages 76 and 77, could have been written down from equation (13).
\[ \rho_{dd(2b)} = \frac{m \sigma_{ps}^2}{m(\sigma_{ps}^2 + \sigma_{pso}^2) + (\sigma_{pq\cdot s}^2 + \sigma_{pco\cdot s}^2)} \]  

(57)

The reliability of equation (57) is a measure of the stability of subtest differences in the face of change of occasion and of item content. This correlation is akin to the Coefficient of Stability and Equivalence (equation 51).

E.IV - COMPARISON OF RELIABILITY COEFFICIENTS

As with the previous comparisons, (page 74), based on Design A, no firm conclusion can be reached about the relative magnitudes of the Coefficient of Equivalence (equation (47)) and the Coefficient of Stability (equation (49)). All depends on how \( m (n \sigma_{po}^2 + \sigma_{pso}^2) \) compares in size with \( \sigma_{pq\cdot s}^2 \), and these are independent of each other. In principle, therefore, a test can have:

(i) a high Coefficient of Equivalence and a low Coefficient of Stability, in which case it must be concluded that the trait or traits measured accurately on a single occasion of testing are subject to change during the interval between occasions in the persons tested;

(ii) a low Coefficient of Equivalence and a high Coefficient of Stability, in which case it must be concluded that although difficult to measure accurately, the trait or traits measured are stable over the period between occasions of testing.

It should be pointed out that \( mn \sigma_p^2 \), which is common to both Coefficients, may be sufficiently large to dominate both.

Although the outcome is in doubt for the two Coefficients of Equivalence and of Stability, there is no doubt that the expectation of both is higher than that of the Coefficient of Stability and Equivalence (equation (51)).

Similar conclusions are reached from a comparison of the reliabilities of subtest differences (equations (55), (56) and (57)).
F - ESTIMATION OF VARIANCE COMPONENTS

The variance components in equations (45) - (57) may be estimated from the analysis of variance in Table IV.

(See page 85 for Table IV)
<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>$E(\text{MS})$ *</th>
</tr>
</thead>
<tbody>
<tr>
<td>Persons</td>
<td>$(\ell - 1)$</td>
<td>$P$</td>
<td>$P$</td>
<td>$\sigma_{pqo.s}^2 + u \sigma_{pq.s}^2 + m \sigma_{ps}^2 + mn \sigma_{po}^2 + mu \sigma_{ps}^2 + mnu \sigma_{p}^2$</td>
</tr>
<tr>
<td>Subtests</td>
<td>$(n - 1)$</td>
<td>$S$</td>
<td>$S$</td>
<td></td>
</tr>
<tr>
<td>Occasions</td>
<td>$(u - 1)$</td>
<td>0</td>
<td>0</td>
<td></td>
</tr>
<tr>
<td>$P \times S$</td>
<td>$(\ell - 1)(n - 1)$</td>
<td>$PS$</td>
<td>$PS$</td>
<td>$\sigma_{pqo.s}^2 + u \sigma_{pq.s}^2 + m \sigma_{ps}^2 + mu \sigma_{ps}^2$</td>
</tr>
<tr>
<td>$P \times O$</td>
<td>$(\ell - 1)(u - 1)$</td>
<td>$PO$</td>
<td>$PO$</td>
<td>$\sigma_{pqo.s}^2 + m \sigma_{ps}^2 + mn \sigma_{po}^2$</td>
</tr>
<tr>
<td>$S \times O$</td>
<td>$(n - 1)(u - 1)$</td>
<td>$SO$</td>
<td>$SO$</td>
<td>$\sigma_{pqo.s}^2 + m \sigma_{ps}^2$</td>
</tr>
<tr>
<td>$P \times S \times O$</td>
<td>$(\ell - 1)(n - 1)(u - 1)$</td>
<td>$PSO$</td>
<td>$PSO$</td>
<td>$\sigma_{pqo.s}^2 + m \sigma_{ps}^2$</td>
</tr>
<tr>
<td>Items within Subtests</td>
<td>$n(m - 1)$</td>
<td>$Q_w$</td>
<td>$Q_w$</td>
<td></td>
</tr>
<tr>
<td>$P \times I \times O \times S$</td>
<td>$n(\ell - 1)(m - 1)$</td>
<td>$PQ_w$</td>
<td>$PQ_w$</td>
<td>$\sigma_{pqo.s}^2 + u \sigma_{pq.s}^2$</td>
</tr>
<tr>
<td>$I \times O \times S$</td>
<td>$n(u - 1)(m - 1)$</td>
<td>$QO_w$</td>
<td>$QO_w$</td>
<td></td>
</tr>
<tr>
<td>$P \times I \times O \times S$</td>
<td>$n(\ell - 1)(u - 1)(m - 1)$</td>
<td>$PQO_w$</td>
<td>$PQO_w$</td>
<td>$\sigma_{pqo.s}^2$</td>
</tr>
<tr>
<td>Total</td>
<td>$\ell mn - 1$</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

* Only the $E(\text{MS})$ including $p$ in the subscript are included.
From Table IV:

\[ E(PQO_w) = \sigma_{pqo}^2 \]

\[ \frac{1}{u} \ E(PQ - PQO_w) = \sigma_{pqo}^2 \]

\[ \frac{1}{m} \ E(PSO - PQO_w) = \sigma_{ps}^2 \]

\[ \frac{1}{mn} \ E(PSO - PQO_w) = \sigma_{po}^2 \]

\[ \frac{1}{mn} \ E(PS - PSO - PQO_w + PQO_w) = \sigma_{ps}^2 \]

\[ \frac{1}{mn} \ E(PS - PQ + PSO) = \sigma_p^2 \]

Since equations (47) - (55) are general, the estimates above can be inserted in them with any desired value of \( m \), the number of items in a subtest.

For the particular values of \( m \) in Table IV itself, the labour of computing estimates of the variance components can be avoided. Estimates of the reliabilities in equations (47) - (55) can be obtained directly from the mean-square. The relationships are shown in Table V.
<table>
<thead>
<tr>
<th>Coefficient</th>
<th>Estimate</th>
</tr>
</thead>
<tbody>
<tr>
<td>Equivalence (Equation (47))</td>
<td>( (\bar{P} + \bar{F}_0) - (\bar{P}_w + \bar{P}_0\bar{Q}_w) ) / ( \bar{P} + \bar{F}_0 )</td>
</tr>
<tr>
<td>Stability (Equation (49))</td>
<td>( \frac{\bar{P} - \bar{F}_0}{\bar{P} + \bar{F}_0} )</td>
</tr>
<tr>
<td>Stab. and Equiv. (Equation (51))</td>
<td>( (\bar{P} - \bar{F}_0) - (\bar{P}_w - \bar{P}_0\bar{Q}_w) ) / ( \bar{P} + \bar{F}_0 )</td>
</tr>
<tr>
<td>( r_{dd}(1) ) (Equation (55))</td>
<td>( \frac{(\bar{P}_S + \bar{P}_S\bar{O}) - (\bar{P}_w + \bar{P}_0\bar{Q}_w)}{\bar{P}_S + \bar{P}_S\bar{O}} )</td>
</tr>
<tr>
<td>( r_{dd}(2) ) (Equation (56))</td>
<td>( \frac{\bar{P}_S - \bar{P}_S\bar{O}}{\bar{P}_S + \bar{P}_S\bar{O}} )</td>
</tr>
<tr>
<td>( r_{dd}(3) ) (Equation (57))</td>
<td>( \frac{(\bar{P}_S - \bar{P}_S\bar{O}) - (\bar{P}_w - \bar{P}_0\bar{Q}_w)}{\bar{P}_S + \bar{P}_S\bar{O}} )</td>
</tr>
</tbody>
</table>
CHAPTER SEVEN

A - NUMERICAL EXAMPLE

We turn now to a specific application of the theoretical principles discussed in the preceding sections.

The data consist of scores on an arithmetic test constructed for experimental use in a study of the relative performances of two matched groups, one taught with, and the other without, the help of Cuisenaire rods. Since we are now concerned only with the effect of test structure on reliability, details of the study will not be reported.

The original test consisted of 95 items unequally divided among three subtests. The 40 items in the first subtest were all small problems so chosen that the use of fractions was minimised. The 25 items in the second subtest were all 'mechanical' fractions. The 30 items in the third subtest were all small problems involving fractions. The test was administered without time limit, and the items were dichotomously scored. The two occasions of testing were a fortnight apart.

For the present purpose, data are needed for one group only, and that taught with Cuisenaire rods was chosen. Also it is convenient* to work with subtests all of the same length. The first and third subtests were therefore reduced to 25 items (to be the same length as the second) by random elimination of excess items, and the subtests rescored on this basis.

With tests differing in content as described above, there is an a priori probability of greater item homogeneity within subtests than between subtests.

* Convenient, but not necessary. The design is orthogonal even if the number of items differs from one subtest to another.
The methods of Table I (page 53) and Table IV (page 85) were therefore employed in the analysis of the data. Intermediate stages of the analysis are shown on the next pages. Table I, which relates to the first administration, shows three two-way analyses, one for each subtest, and the pool of them all. Table 2 shows the corresponding analyses for the same subtests on the second administration. Table 3 shows the analysis for the whole test on the first occasion in accordance with the model of equation (1), page 47, and the analysis of Table I, page 53. Table 4 shows the corresponding analysis for the same test on the second occasion. Table 5 shows the final analysis in accordance with the model of equation (44), page 78 and the analysis of Table IV, page 85. Finally, the reliability estimates derived from these several analyses are collected together in Tables 6 and 7.
### TABLE 1 - SUBTESTS (OCCASION 1)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
<td>27</td>
<td>39.7733</td>
<td>1.4360</td>
<td>27</td>
<td>33.0167</td>
<td>1.2228</td>
<td>27</td>
<td>40.8945</td>
<td>1.5146</td>
<td>81</td>
<td>112.6845</td>
<td>1.3912</td>
</tr>
<tr>
<td>Items</td>
<td>24</td>
<td>31.3432</td>
<td>1.3268</td>
<td>24</td>
<td>29.9257</td>
<td>1.2469</td>
<td>24</td>
<td>35.7896</td>
<td>1.4912</td>
<td>72</td>
<td>97.5585</td>
<td>1.3550</td>
</tr>
<tr>
<td>C x I</td>
<td>648</td>
<td>99.0274</td>
<td>1.5282</td>
<td>648</td>
<td>101.2694</td>
<td>1.5628</td>
<td>648</td>
<td>105.3843</td>
<td>1.6263</td>
<td>1944</td>
<td>305.5811</td>
<td>1.5724</td>
</tr>
<tr>
<td>Total</td>
<td>699</td>
<td>169.6439</td>
<td>-</td>
<td>699</td>
<td>164.2118</td>
<td>-</td>
<td>699</td>
<td>182.0694</td>
<td>-</td>
<td>2097</td>
<td>515.9241</td>
<td>-</td>
</tr>
</tbody>
</table>

Coefficient Alpha*  

0.8936  

0.8722  

0.8926  

0.8870

(Mean within $O_1$)

### TABLE 2 - SUBTESTS (OCCASION 2)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
<td>27</td>
<td>43.1131</td>
<td>1.5968</td>
<td>27</td>
<td>35.7402</td>
<td>1.3237</td>
<td>27</td>
<td>40.7215</td>
<td>1.5082</td>
<td>81</td>
<td>119.5748</td>
<td>1.4762</td>
</tr>
<tr>
<td>Items</td>
<td>24</td>
<td>36.0284</td>
<td>1.5612</td>
<td>24</td>
<td>31.5343</td>
<td>1.3139</td>
<td>24</td>
<td>36.9846</td>
<td>1.5410</td>
<td>72</td>
<td>104.3473</td>
<td>1.4520</td>
</tr>
<tr>
<td>C x I</td>
<td>648</td>
<td>103.6670</td>
<td>1.5998</td>
<td>648</td>
<td>93.8676</td>
<td>1.4486</td>
<td>648</td>
<td>95.1764</td>
<td>1.4683</td>
<td>1944</td>
<td>292.6822</td>
<td>1.5056</td>
</tr>
<tr>
<td>Total</td>
<td>699</td>
<td>182.3085</td>
<td>-</td>
<td>699</td>
<td>161.1421</td>
<td>-</td>
<td>699</td>
<td>172.3537</td>
<td>-</td>
<td>2097</td>
<td>516.8043</td>
<td>-</td>
</tr>
</tbody>
</table>

Coefficient Alpha  

0.8998  

0.8906  

0.9206  

0.8980

(Mean within $O_2$)

* $\bar{u} = (\bar{C} - \bar{CI}) / \bar{c}$
TABLE 3 - TEST (OCCASION 1)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
<td>27</td>
<td>100.1471</td>
<td>3.7092</td>
</tr>
<tr>
<td>Subtests</td>
<td>2</td>
<td>4.2137</td>
<td>2.1066</td>
</tr>
<tr>
<td>C × S</td>
<td>54</td>
<td>12.5374</td>
<td>.2322</td>
</tr>
<tr>
<td>C × W S</td>
<td>72</td>
<td>97.5585</td>
<td>1.3550</td>
</tr>
<tr>
<td>I × W S</td>
<td>72</td>
<td>305.6811</td>
<td>.16724</td>
</tr>
</tbody>
</table>

Total     2099  520.1378  -

Coefficient of Equivalence = .9576;  \( \alpha = .9571 \)
Mean Reliability of Subtest Differences = .3228

TABLE 4 - TEST (OCCASION 2)

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
<td>27</td>
<td>106.3381</td>
<td>3.9384</td>
</tr>
<tr>
<td>Subtests</td>
<td>2</td>
<td>3.7806</td>
<td>1.8903</td>
</tr>
<tr>
<td>C × S</td>
<td>54</td>
<td>12.2367</td>
<td>.2451</td>
</tr>
<tr>
<td>I × W S</td>
<td>72</td>
<td>104.5473</td>
<td>1.4520</td>
</tr>
<tr>
<td>C × W S</td>
<td>1944</td>
<td>292.6822</td>
<td>.15056</td>
</tr>
</tbody>
</table>

Total     2099  520.5849  -

Coefficient of Equivalence = .9618;  \( \alpha = .9611 \)
Mean Reliability of Subtest Differences = .3857
### TABLE 5 - FINAL ANALYSIS

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Children</td>
<td>27</td>
<td>201.4301</td>
<td>7.4604</td>
</tr>
<tr>
<td>Subtests</td>
<td>2</td>
<td>7.4173</td>
<td>3.7036</td>
</tr>
<tr>
<td>Occasions</td>
<td>1</td>
<td>2.9561</td>
<td>2.9561</td>
</tr>
<tr>
<td>C x S</td>
<td>54</td>
<td>21.4029</td>
<td>.3964</td>
</tr>
<tr>
<td>C x O</td>
<td>27</td>
<td>5.0551</td>
<td>.1872</td>
</tr>
<tr>
<td>S x O</td>
<td>2</td>
<td>.5770</td>
<td>.2885</td>
</tr>
<tr>
<td>C x S x O</td>
<td>54</td>
<td>4.3712</td>
<td>.08095</td>
</tr>
<tr>
<td>Items within Subtests</td>
<td>72</td>
<td>190.3676</td>
<td>2.6440</td>
</tr>
<tr>
<td>CI w S</td>
<td>1244</td>
<td>388.1196</td>
<td>.19965</td>
</tr>
<tr>
<td>IO w S</td>
<td>72</td>
<td>11.7382</td>
<td>.16303</td>
</tr>
<tr>
<td>CIO w S</td>
<td>1244</td>
<td>210.2437</td>
<td>.10815</td>
</tr>
</tbody>
</table>

| Total                   | 4199| 1043.6788| -   |

### TABLE 6 - ESTIMATES OF RELIABILITY COEFFICIENTS FROM INTERMEDIATE STAGES

<table>
<thead>
<tr>
<th>Subtest</th>
<th>Occasion 1</th>
<th>Occasion 2</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Alpha</td>
<td>C.Eq.</td>
</tr>
<tr>
<td>1</td>
<td>.8936</td>
<td>-</td>
</tr>
<tr>
<td>2</td>
<td>.3722</td>
<td>-</td>
</tr>
<tr>
<td>3</td>
<td>.3926</td>
<td>-</td>
</tr>
<tr>
<td>Mean within Subtests</td>
<td>.8870</td>
<td>-</td>
</tr>
<tr>
<td>Whole Test</td>
<td>.9571</td>
<td>.9576</td>
</tr>
</tbody>
</table>
TABLE 7 - ESTIMATES OF RELIABILITY COEFFICIENTS FROM FINAL ANALYSIS

<table>
<thead>
<tr>
<th></th>
<th>C.Eq.</th>
<th>C.Stab.</th>
<th>C.Stab. Eq.</th>
<th>r_{dd(1)}</th>
<th>r_{dd(2)}</th>
<th>r_{dd(3)}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Whole Test</td>
<td>.9598</td>
<td>.9510</td>
<td>.9391</td>
<td>.3552</td>
<td>.6608</td>
<td>.4692</td>
</tr>
</tbody>
</table>

The reliabilities in Table 7 have been computed from the mean-squares in Table 5 using the formulae recorded in Table V, page 87.

B - DISCUSSION OF RESULTS

(1) The alphas for subtests in Table 1 and 2 may be regarded as Coefficients of Equivalence for individual subtests. They reflect the error variance of measurement of item scores, estimated in each sub-analysis by the mean-square for C x I. These are remarkably consistent, and the pooled results on the right of Tables 1 and 2 are acceptable as estimates of the general level of these error variances. For whole subtests, therefore, the error variance of measurement may be estimated as

\[ 25 \times 0.15724 = 3.931 \quad \text{for Occasion 1} \]

and

\[ 25 \times 0.15056 = 3.764 \quad \text{for Occasion 2} \]

with corresponding standard errors of measurement of 1.93 and 1.94 respectively.

The same results are of course obtainable from the mean-squares in the CI w S lines of Tables 3 and 4.

(2) For the whole test of 75 items, the estimated error variance of measurement is
with corresponding standard errors of measurement of 3.434 and 3.360 respectively.

(3) The mean error variance over both occasions for individual items, for subtests, or for whole tests, may be obtained from Table 5 by pooling the lines for Cl w S and ClO w S:

Mean error variance over both occasions for a single item

\[
\frac{388.1196 + 210.2437}{2(1944)} = 0.15390
\]

whence the mean error variance over both occasions of a subtest is

\[
25 \times 0.15390 = 3.848 \quad (se_{\text{meas.}} = 1.962)
\]

and the mean error variance over both occasions of the whole test is

\[
75 \times 0.15390 = 11.543 \quad (se_{\text{meas.}} = 3.397)
\]

Because of the consistency of the error of measurement from subtest to subtest, the overall estimates derived from Table 5 satisfactorily represent the error variances for single subtests and for the whole test on both occasions. This outcome, however, was not inevitable. The results might have fluctuated from one subtest to another and from one occasion to another. Had they done so, the fluctuation would have been masked by averaging and accepting the result, from Table 5 alone. The fluctuation would have been revealed by the figures obtained in the intermediate analyses of Tables 1-4.

(4) From Tables 3 and 4, the dispersions of child totals on the test may be obtained. The variances of these test scores are

\[
75 \times 3.7092 = 278.19 \quad \text{for Occasion 1}
\]

and

\[
75 \times 3.9384 = 295.38 \quad \text{for Occasion 2}.
\]

The corresponding s.d.'s are 16.68 and 17.19 respectively. Again, the mean
variance of test scores can be obtained from Table 5. On pooling the lines for Children and C x 0 we obtain

\[
\frac{201.4301 + 5.0551}{2 \times 27} = 3.8238
\]

and \(75 \times 3.8238 = 286.79\)

which estimates the mean variance of test scores on a single occasion. The corresponding s.d. is 16.94.

Again, the pooled estimate is satisfactory, though, as argued above, it was not necessarily so.

(5) To tie in the estimate of the Coefficient of Equivalence in Table 7 with the mean error variance and mean test variance estimated from Table 5 in (3) and (4) above, it should be noted that

\[
C_{\text{Eq}} = 1 - \frac{\text{Error Variance}}{\text{Test Variance}}
\]

\[
= 1 - \frac{11.543}{286.790}
\]

\[
= .9598 \quad \text{(see Table 7, page 93)}
\]

(6) The Coefficient of Stability could not have been estimated without the analysis of Table 5. For these data, the Coefficient of Stability is slightly less than the Coefficient of Equivalence. As pointed out on page 83, this was not necessarily so.

The Coefficient of Stability is in fact quite high. It seems that the interval of a fortnight between testings has not unduly altered the rank order of the testees. An estimate of the error variance of measurement corresponding to this Coefficient is

\[
286.79 \times (1 - .9510) = 14.053
\]

* See (4) above.
and the corresponding standard error of measurement is 3.749 (compare the mean standard error of measurement of 3.397 obtained in (3) above which corresponds to the Coefficient of Equivalence.)

(7) The Coefficient of Stability and Equivalence in Table 7 is an estimate of the correlation that would have been obtained if a 'stratified-parallel' test had been substituted on the second occasion of testing. As expected, it is lower than either of the other two coefficients quoted. Even so, it is quite high, and the corresponding standard error of measurement is

$$\sqrt{286.79 \times (1 - .9391)} = 4.179$$

(8) The whole-test alphas in Table 6 were obtained from Tables 3 and 4 by pooling line 2 with line 4, and line 3 with line 5 to obtain the 'collapsed' analysis of Table II. This is tantamount to assuming the non-existence of inter-subtest heterogeneity. The test of this assumption is $F = \frac{CS}{CIwS}$ in Tables 3 and 4, giving respectively 1.477 and 1.628, with (54, 1944) degrees of freedom in both cases. Significance is established at about the 5 per cent level.

A significance test, which is rather more sensitive since it is based on all the data from two occasions, is available from Table 5. The numerator is again $CS$, and the denominator appropriate to the test must be decided on. For this purpose, reference is made to the variance components in Table IV, page 85. Since we are concerned with occurrences within an occasion, it is assumed that Occasion is a fixed effect. The application of Schultz's rule then leads to the deletion of $m \sigma^2_{pso}$ in the expected mean-square for $P \times S$. The denominator for the test is therefore the mean-square for $PI \times S$. For the numerical data of Table 5, $F = 1.985 \times (54, 1944)$ which indicates significance beyond the 5 per cent level.

Even so, the alphas of Table 6 differ only trivially from the corresponding
Coefficients of Equivalence. Since the test was dichotomously scored, alphas obtained from the analysis of variance of Tables 3 and 4 are identical with the results of applying the KR 20 formula to the data. It must be concluded that the application of the routine procedure in this case would have given, not only an estimate of the internal consistency of the test, but also a satisfactory estimate of its Coefficient of Equivalence.

(9) The conclusion just reached is bound up with the relatively low reliabilities of subtest differences reported in Table 6. These are a measure of the stability of the differences among child totals on subtests in the face of change of item content in the subtests the occasion of testing being regarded as fixed. More specifically, these reliabilities are the expected mean correlations between differences among child-subtest totals on the test in hand and the corresponding differences with a 'stratified-parallel' test that might have been used in its place. The reliability of difference scores will be high if the within-subtest error variance (bottom line of Tables 3 or 4) is a relatively small proportion of the between-subtest error variance (third line of Table 3 or 4). For these data, this is not so, as a glance at the figures will show.

Further light is shed on the matter by considering the mean intercorrelation of child totals on different subtests within the test. This mean intercorrelation is estimated by

\[
\bar{r} = \frac{\bar{C} - \bar{CS}}{\bar{C} + (n - 1) \bar{CS}}
\]

in which \(n\) is the number of subtests, here 3.

From Table 3, \(\bar{r} = .8331\), and from Table 4, \(\bar{r} = .8340\). These correlations are not far short of the subtest alphas in Table 6. This means that the average correlation between child totals on different subtests does not differ greatly from the correlation between child totals on parallel forms of the same subtest. Obviously, therefore, the test would not be efficient for diagnostic purposes.
In Table 7, the correlation \( r_{dd}(1) \), derived from the formula in Table V, page 87, or from equation (55), page 82, is in the same category as those just discussed. It is, in fact, an 'average' of the two \( r_{dd} \)'s in Table 6. However, \( r_{dd}(2) \) is relatively high. This measures the stability of child totals on a subtest in the face of a change of occasion. Specifically, it is the expected mean correlation between differences among child totals on the subtests administered on one occasion and the corresponding differences with the same subtests administered on another occasion. Over a period of a fortnight, these differences appear to be relatively stable.

The value of .4692 for \( r_{dd}(3) \) in Table 7 is anomalous, since \( r_{dd}(3) \) has an expectation below that of either \( r_{dd}(1) \) or \( r_{dd}(2) \). The reason for this anomaly is that the second order interaction mean-square for \( C \times S \times 0 \) is below expectation, and this particularly affects the estimation of \( \rho_{dd}(3) \).

(10) Finally, had the analysis of Tables 3 and 4 shown the \( C \times S \) interaction to be non-significant, it would have been legitimate to dispense with the lines relating specifically to subtests in both these analyses and that of Table 5. For Tables 3 and 4, this results would be achieved by pooling the line for Subtests with that for \( I \times S \), and the line for \( C \times S \) with that for \( C \times w \). Similarly, for Table 5, the pooling would be slightly more complicated. Subtests would join \( I \times S \) to give a single line for Items; \( C \times S \) would be pooled with \( C \times w \) to give \( C \); \( SO \) and \( IO \) would together give \( IO \); and \( CSO \) and \( CIO \) would give \( CIO \). The result would be the straightforward three-way analysis of Table III, page 75.
C - PRACTICAL ROUTINE FOR COMPUTATION OF COEFFICIENT OF EQUIVALENCE*

It seems probable that in general Coefficient alpha will be a sufficiently close approximation to the Coefficient of Equivalence for most practical purposes. However, situations may arise in which the under-estimation is sufficiently serious to warrant a special calculation of the Coefficient of Equivalence. This would probably be the case with a test designed for diagnostic purposes in which the subtests, while internally homogeneous in item content, differed very much from each other. In terms of the analyses of Tables 3 and 4, this would correspond to a very high $CS / CE_wS$ ratio.

The procedure suggested below for estimating the Coefficient of Equivalence is a simple extension of the routine procedures for obtaining Coefficient alpha, which most test constructors use as a matter of course. It consists of making a separate estimate of the error variance of person totals for each subtest, summing the error variances, and then using the relation:

\[
\text{Coefficient of Equivalence} = 1 - \frac{\sum (ev)}{\text{Test score variance}}
\]

The derivation of the procedure is as follows:

(1) For a single subtest $k$

\[
\alpha_k = \frac{\bar{P}_k - \bar{P}_k}{\bar{P}_k} = \frac{m_k}{m_k - 1} \cdot \frac{\bar{P}_k - \bar{W}_k}{\bar{P}_k}
\]

(see equation (20), page 59)

from which it follows that

* Since this routine was devised, a similar procedure has been suggested by Rajaratnam, Cronbach and Gleser ((70), page 46). Their equation (20) is identical, except for nomenclature, with equation (60).
\[ \alpha_k = \frac{m_k}{m_k - 1} \cdot \frac{V_k - \sum v_{jk}}{V_k} \]  

Equation (58) is a restatement using different nomenclatures of equation (19), page 59.

The error variance of a total on subtest \( k \) is

\[ ev_k = V_k(1 - \alpha_k) \]

From equation (58) it follows that

\[ ev_k = \frac{m_k \sum v_{jk} - V_k}{m_k - 1} \]  

Then:

\[ \text{Coefficient of Equivalence} = 1 - \frac{\sum (m_k \sum v_{jk} - V_k)}{V_T} \]  

In equation (60), \( m_k \) is the number of items in subtest \( k \); \( v_{jk} \) is the within-item variance of item \( j \) in subtest \( k \); \( V_k \) is the variance of scores on subtest \( k \); and \( V_T \) is the variance of whole-test scores.

For a dichotomously scored test, it is already a routine procedure to compute the sum of the within-item variance for the whole test using the relation

\[ v_i = p_i q_i = p_i(1 - p_i) \]

in which \( p_i \) is the proportion of testees passing item \( i \) and \( q_i = 1 - p_i \).

The same procedure is suggested in deriving the Coefficient of Equivalence, but the summation will be over single subtests instead of the whole test. If the test is not dichotomously scored, the within-item variance must be obtained by some other convenient method.
The suggested routine is laid out below, with illustrative figures obtained in an item analysis of the first-occasion results with the test discussed in the previous pages (see especially Table 3, page 91).

**SUGGESTED ROUTINE FOR ESTIMATING COEFFICIENT OF EQUIVALENCE**

<table>
<thead>
<tr>
<th>Subtest</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>m</td>
<td>25</td>
<td>25</td>
<td>25</td>
</tr>
<tr>
<td>$\Sigma v_j^*$</td>
<td>5.104</td>
<td>4.974</td>
<td>5.418</td>
</tr>
<tr>
<td>$m \Sigma v_j^j$</td>
<td>127.60</td>
<td>124.35</td>
<td>135.45</td>
</tr>
<tr>
<td>$V$ (subtest)</td>
<td>35.91</td>
<td>30.57</td>
<td>37.37</td>
</tr>
<tr>
<td>$m \Sigma v_j^j - V$</td>
<td>91.69</td>
<td>93.78</td>
<td>97.58</td>
</tr>
<tr>
<td>$\frac{1}{m - 1} (m \Sigma v_j^j - V)$</td>
<td>3.820</td>
<td>3.908</td>
<td>4.066</td>
</tr>
</tbody>
</table>

Total (= whole-test error variance) = 11.794

$V_T$ ( = whole-test variance ) = 278.19

Coefficient of Equivalence = .9576**

$\Sigma p_j q_j = \Sigma p_j^2 - \Sigma p_j^2$

** See C.Eq. reported at foot of Table 3, page 91.
NOTE - The following Chapter is an expansion of an article published in the British Journal of Educational Psychology (1960. Vol. XXX, pages 53-62) under the joint authorship of A. E. G. Pilliner, J. Sutherland and E.G. Taylor.

The first-named author designed the experiment and wrote the article; the second arranged with the Local Authorities for the administration of the tests; and the third was responsible for most of the computational work.
CHAPTER EIGHT

A - ZERO ERROR

A.I - INTRODUCTION

The purpose of this chapter is to draw attention to a particular source of error which is likely to be present to some degree in all standardised scores or "quotients" obtained when one or other of a series of standardised "objective" tests, all bearing the same label, is administered. Examples of such series are the Moray House verbal reasoning tests or the National Foundation English attainment tests, and there are many others. The experimental part of this study is concerned only with the Moray House verbal reasoning tests, but there is little doubt that parallel studies with other series of tests would produce similar results.*

Users of tests such as these are apt to assume rather lightly that the verbal reasoning (or whatever) performance level of two children is much the same if they obtain the same standardised score, the first child on test A in the series, the second on test B. Few nowadays would interpret this common standardised score as indicating exactly the same performance level for these children, since the concept of 'standard error of measurement' is a familiar one (see, for example, (32), page 479). Many, however, are insufficiently aware of the possibility of a bias attaching to all the standardised scores obtained on one of the tests relative to those obtained on the other; that children scoring, say, 80, 100, and 120 on one test in the series might have scored, 84, 105, and 123 had a different test been chosen.

* In a personal communication, Mr. D. A. Pidgeon, Senior Tests Officer of the National Foundation for Educational Research, tells us that the zero error studied here with Moray House verbal reasoning tests is observable with National Foundation tests also.
The occurrence of a bias in standardized scores obtained from one test relative to those from another in the same series may be misleading; for example, when a local authority uses one of a series of tests for the 'main' examination and another for the 'supplementary' examination for absentees; or when an allocation procedure involves a cut at some predetermined standardized score, regardless of the test used. Such procedures, too rigidly applied, are to be deprecated. They imply a faith in the consistency of calibration from test to test, which those who construct and standardise them are unable to share.

In the experimental investigation to be described, the mean standardisation bias of each of a number of Moray House verbal reasoning tests, relative to one chosen arbitrarily as standard, is evaluated. Since each bias reported is based on an experiment in which the subjects were a random sample from a larger group with a mean standard score of approximately 100 which corresponds to zero on the standard score scale; and since the bias under discussion is similar in some respects to the systematic errors occurring in physical measurement; it is proposed that this standardisation mean bias be termed 'zero' error.

A.II - OTHER RELEVANT WORK

Ebel (17) seems to have been aware that comparisons made using one test for person A and another for person B are less precise than comparisons in which the same test is used for both. Rajaratnam (69) discussed the problems arising when a different randomly selected set of judges is assigned to each individual. The formulae which she proposed (and which are extended in subsequent articles in collaboration with Cronbach and Gleser (14), (70)) obscure the essentially simple nature of the statistical problem that arises.
A. III - SOME THEORETICAL CONSIDERATIONS

Suppose a large number of tests are all administered to the same persons. The score of person \( i \) on test \( j \) may be regarded as made up of independent components as follows:

\[
x_{ij} = M + p_i + t_j + e_{ij}
\]

in which \( M \) is common to all scores and is assumed to be zero; \( p_i \) is specific to person \( i \); \( t_j \) is specific to test \( j \); and \( e_{ij} \) is random error associated with person \( i \) and test \( j \). All score components sum to zero in the population of scores.

Person \( i \)'s total on a random selection of \( m \) tests is

\[
T_i = \sum_{j=1}^{m} x_{ij} = mp_i + \sum_{j=1}^{m} t_j + \sum_{j=1}^{m} e_{ij} \quad (61)
\]

\((j = 1, 2, \ldots, m)\)

(1) If the same random sample of tests is administered to all persons, the expectation of the variance of person totals is

\[
E(\text{Var.})(1) = m^2 \sigma_p^2 + m \sigma_e^2 \quad (62)
\]

in which the first term is 'true' variance and the second 'error' variance.

(2) If a different random sample of tests is selected for each person, the expectation of the variance of person totals is

\[
E(\text{Var.})(2) = m^2 \sigma_p^2 + (m \sigma_t^2 + m \sigma_e^2) \quad (63)
\]

in which the first term is 'true' variance and the second and third are 'error' variance.

Clearly, \( E(\text{Var.})(2) \geq E(\text{Var.})(1) \), unless \( \sigma_t^2 = 0 \).
The first situation, in which the test procedure* is the same for all testees, is the usual one. However, the second, in which the test procedure differs from one testee to another, also occurs in a number of different contexts. The comparison of scores obtained by different children on different tests has already been mentioned. Other examples are: examinations in which the candidates have a choice of questions; selection for university on the results of examinations compiled by different Boards; assessment by different teachers of the potentialities of particular groups of children;** administration of individual tests such as the Terman-Merrill to different children by different testers (15); clinical studies on different subjects by different psychologists. The list could be extended indefinitely.

In all these cases, unless appropriate corrective measures can be taken, the error variance is inflated as shown in equation (63). In addition to the random error variance (the term in $\sigma_e^2$), there is now the error represented by $\sigma_t^2$, which results from differences in the testing procedure itself.

The distinction between influences producing $'\sigma_e^2'$ error and $'\sigma_t^2'$ error is sometimes difficult to draw. Doubtless $'\sigma_e^2'$ error frequently includes variation resulting from influences similar to those cited above as examples of sources of $'\sigma_t^2'$ error. However, $'\sigma_e^2'$ errors are usually undefined, or, if definable, they are accepted as too troublesome or too difficult to remove. Broadly speaking, however, the distinction is that $'\sigma_t^2'$ error can in principle be eliminated, while $'\sigma_e^2'$ error cannot.

Error of the $'\sigma_t^2$ type can be eliminated, or at least reduced, if it is possible to keep track of the specific testing procedures used. Suppose person a,

* The useful term 'test procedure' is due to Rajaratnam (69).

** This is more fully treated in Chapter 9.
tested by procedure 1, and person b, tested by procedure 2, are to be compared.

Their scores are as follows:

\[ x_{a1} = p_a + t_1 + e_{a1} \]
\[ x_{b2} = p_b + t_2 + e_{b2} \]

The difference is

\[ d_{(a1-b2)} = x_{a1} - x_{b2} = (p_a - p_b) + (t_1 - t_2) + (e_{a1} - e_{b2}) \quad (64) \]

If the comparison is replicated, with the same testing procedures, the expectation of the mean over replication is

\[ E(\bar{x}_{a1} - \bar{x}_{b2}) = (p_a - p_b) + (t_1 - t_2) \]

The error \((t_1 - t_2)\) is not eliminated by replication as is the random error, and is therefore a bias or 'zero' error. If the magnitude of the bias can be determined experimentally, it can be allowed for. The present study is concerned with the determination of the biases, or 'zero' errors, of a series of Moray House tests relative to a particular one taken as standard. Knowledge of these 'zero' errors makes it possible to allow for the bias if one test is used instead of another, provided that the identities of the tests are known. Normally this will be the case, but situations may arise in which it is not known which test was used. If so, the measurements obtained must be regarded as subject to error in accordance with equation (63).

In anticipation of subsequent and fuller discussion, it may here be pointed out that if persons a and b, tested previously by procedures 1 and 2 respectively, are now tested by procedures 2 and 1 respectively, the differences between their scores is:
\[ d_{(a_2-b_1)} = x_{a_2} - x_{b_1} = (p_a - p_b) + (t_2 - t_1) + (e_{a_2} - e_{b_1}) \quad (65) \]

From equations (64) and (65), we obtain

\[ d_{(a_1-b_2)} + d_{(a_2-b_1)} = 2(p_a - p_b) + \text{random error} \quad \ldots \quad (66) \]

and

\[ d_{(a_1-b_2)} - d_{(a_2-b_1)} = 2(t_1 - t_2) + \text{random error} \quad \ldots \quad (67) \]

Equation (66) shows how an unbiased comparison can be made between persons. Equation (67) shows how the magnitude of the bias can be evaluated.

A. IV - ORIGIN OF ZERO ERROR

The standardisation of a test initially involves its administration to a large sample of children (the nature of this sample will be discussed later), of the appropriate age range. This sample, called the 'standardisation' sample, is sub-divided by month of age, and within each month group the children's original or 'raw' scores are used to obtain their percentile ranks. These ranks are then transformed to a 'normal' distribution of standardised scores or 'quotients' usually with a mean of 100 and a standard deviation of 15. After repeating this procedure for all month groups in the sample a conversion table or table of norms is prepared which relates the raw scores and ages of the children in this sample to the corresponding standardised scores. Similar conversion tables are prepared for each test in the series, using each time a different standardisation sample. Each table of norms thus prepared makes it possible to match the performance level of any children subsequently taking the corresponding test with that of some children in the standardisation sample. For example, if
entering the table with some testee's age and raw score assigns him a standardised score of 100, his performance level on the test is equal to that of all the children in the standardisation sample whose several raw scores place them at the 50th percentile in their several month groups (86), (46).

Suppose that on some particular occasion, a test was chosen, that it was used with a number of children, and that the mean of their standardised scores was 100. Now suppose that a different test in the series had been chosen and used on this one occasion with the same children and that their mean standardised score on this different test was 106. Note that, in fact, only one test has been administered and only one mean is available; either 100 or 106; and that the test user accepts whichever turns up in ignorance that the other might have done so. It is not inconceivable that some important decision may hinge on the result.

The obstacles in the way of achieving the ideal that any child tested should obtain the same standardised score, whichever test in the series happens to be the one administered to him are:

(1) Differences in content from test to test, resulting in slightly different rank orders for the same children.

(2) Non-equivalence in the standardisation samples.

With regard to (1), the imperfect correlation results in unbiased errors of measurement which can be estimated with some precision if the correlation is known. It is worth pointing out that it is, in fact, unknown; the ordinary test-retest correlation based on the administration of different tests to the same testees on different occasions* under-estimates the correlation which presumably should be employed, namely, that between the children's scores on the test used and

* The Coefficient of Stability and Equivalence (see page 69).
the scores they would have obtained with a different test on this one occasion.* An under-estimate of the correlation means an over-estimate of the standard error of measurement. In estimating the standard error of measurement we shall err, therefore, on the safe side by basing it on the usual test-retest correlation.

With regard to (2), by 'equivalence' would be meant that we can legitimately assume each standardisation sample to be representative of the same specifiable 'population' of children; a population against which it makes sense to measure the prospective testees in test performance. If obtainable, equivalent samples would be inter-changeable in respect of the tests, without affecting the tables of norms. Testees would be pitted virtually against the same population, whatever the test, and any discrepancies in their standardised scores if one test were used rather than another would be due entirely to differences in test content, as discussed in (1). With equivalent samples as here defined, there would be no systematic bias associated with all the standardised scores obtained with a particular test. 'Non-equivalence' would imply either that the standardisation samples are not representative samples from the same population or that they are samples from different populations. In either case the possibility would exist of systematic as well as unsystematic errors occurring.

The population to be represented by the standardisation sample consists of a large number of local authority year-groups, the mean standardised scores and the allocation procedures of which show remarkable diversity. A number of these authorities prefer to use tests already fully standardised and so must be excluded from consideration. In practice, therefore, it is impossible to achieve completely representative sampling and some approximate procedure must be used, such as the following. The use of newly constructed tests not yet fully standardised is restricted to those other authorities which are prepared

* The Coefficient of Equivalence (see page 44).
to co-operate by making available the raw scores and ages of the children in year-groups working these new tests; in return these authorities receive specially constructed tables of norms based on the data supplied. From among the authorities in this sub-population which have chosen a particular new test a sample is selected which, on the basis of previous experience and current performance, is judged to be as nearly representative as possible of the general population of authorities. Finally, a 'national' conversion table for the new test is prepared from the pooled data of the authorities thus selected from the sub-population.

In spite of these difficulties, the zero error of any one Moray House verbal reasoning test relative to its neighbour in the list, is usually remarkably small; both, it should be noted, will have been standardised at about the same time. However, local authorities using new and old tests with consecutive year-groups (which may be expected to be about equal in mean performance), do experience a rise in mean standardised score from the new to the old test which indicates a non-equivalence in the samples used for standardising these more widely separated tests. This non-equivalence results less from the selection of poor samples from a given population than from a slow change in the performance characteristics of the population itself, particularly since the early 1950's. It appears to be associated mainly with the observed wide-spread increase during recent years in 'test-sophistication' which will be the subject of later discussion.

An experimental study designed specifically to estimate as precisely as possible the zero errors of commonly used tests was undertaken for the following reasons. First, such a study might contribute information useful in a general survey of test standards. Second, the results might be helpful to investigators using these tests for the study of population trends, earlier maturation, the effects of practice and coaching, and the like, including also the re-evaluation
of the results of previous studies in these fields, in which the zero errors of the tests were not taken into account. Third, the publication of the figures obtained might not only contribute to a more general awareness among test users of the existence of zero error, but also give those responsible for allocation procedures very practical assistance in the interpretation of the results obtained.

B.I - THE EXPERIMENTAL INVESTIGATION

The tests chosen were the verbal reasoning tests M.H.T.'s 33 and 39-58, for which 'national' norms had been prepared over the period 1945 to 1957. For six of these tests, namely M.H.T.'s 39-44, data were already available from Michael F. Moore's previous study (56) of the practice effects resulting from the administration of six verbal reasoning tests in succession to the same group of children. Moore's study, while primarily concerned with practice effect, differed from previous studies of this kind in that his experimental design enabled him to estimate the zero errors of the tests used and to disentangle these errors from the practice effect. Each of six groups of children worked six tests in six different orders, the whole conforming to a latin square design, the reciprocal nature of which enabled Moore to isolate both practice effect and zero error simultaneously.* The only drawback to Moore's elegant design was that the final computations had to be restricted to the scores of six numerically equal groups of children all of whom completed all six tests. Illness during the administration period resulted in considerable loss. It was therefore decided to employ, for the remaining tests, a method which minimised wastage while retaining the advantages of the latin square.

* The desirability of eliminating the effects of zero error in the tests used for practice and coaching experiments was previously noted by Vernon (90), page 59.
The simplest way of carrying out the experiment would have been to use a number of large independent random samples, that is, equivalent samples. Sample 1 would have test A, sample 2 test B, and so on ((49), page 47). With proper randomisation, the expected mean performance level is the same for all samples, so that any observed differences among the standardised score means would be ascribable to zero error. This method was rejected because of the practical difficulties in the way of efficient randomisation and the lack of precision inherent in the design.

Instead, separate comparisons were carried out with two tests at a time. In each comparison, both test A and test B were administered to the same group of children using the 'cross-over' method. The children were divided at random into two numerically equal groups. Group 1 took first test A, then test B; Group II took first test B, and then test A. The difference between the overall mean score on A in its two positions and the corresponding overall mean for B gave the mean zero error, with practice effect eliminated.

Test A was always M.H.T. 40, chosen as standard. Test B was one of the other tests, the experiment being repeated, each time with different children, for all the tests except for Moore's six. The variable was standardised score ('quotient') on the 'national' norms in the test manuals.

The children and the tests were allocated to the groups as follows: Primary schools were chosen at random from the Edinburgh and Midlothian lists. Six of these schools, again chosen at random, were allotted to each comparison AB₁, AB₂, AB₃, etc. Within each school the available children were allocated at random to two groups of the same size for each school, but not necessarily so from one school to another. Finally, within each school the group designation, i.e., I or II, was decided by the toss of a coin. Group I's testing order was always AB and group II's BA.
The main advantages of this design are: first, the effect of absences is less catastrophic than in the larger latin square. Second, the replications in six schools for each pair of tests provides information about the 'interaction' between school and zero error, that is, the extent of variation of zero error from school to school, in addition to information about the 'main' effect, the mean zero error overall. Third, since a random half of the random sample of children engaged in the experiment took M.H.T. 40 first, valid inferences can be drawn about the performance on this test of the population sampled, and this information is useful in judging the validity of regarding the zero errors obtained in the experiment as generally applicable.

B.11 - EXPERIMENTAL DESIGN AND STATISTICAL ANALYSIS

The design is slightly more complex than any hitherto encountered. It is most simply conceived of as a replication in randomly selected schools of the following crossover arrangement:–

<table>
<thead>
<tr>
<th>Group</th>
<th>Occasion 1</th>
<th>Occasion 2</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>A</td>
<td>x...x...x</td>
</tr>
<tr>
<td></td>
<td>x...x...x</td>
<td>x...x...x</td>
</tr>
<tr>
<td>B</td>
<td>x...x...x</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>A</td>
<td></td>
</tr>
</tbody>
</table>

Group 1 children are shown as taking Test 'A' on Occasion 1 and Test 'B' on Occasion 2. For Group 2 children the order of testing is reversed. The asterisks represent the scores of individual children. The dotted lines indicate that each child has two scores, one on Test 'A', the other on Test 'B', that is, children are crossed with tests.
Within each school, the appropriate variance analysis is as follows (only the degrees of freedom are shown):

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tests</td>
<td>1</td>
</tr>
<tr>
<td>Groups</td>
<td>1</td>
</tr>
<tr>
<td>Occasions</td>
<td>1</td>
</tr>
<tr>
<td>Children within Groups</td>
<td>2(c - 1)</td>
</tr>
<tr>
<td>C x T within Groups</td>
<td>2(c - 1)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4c - 1</strong></td>
</tr>
</tbody>
</table>

The scope of the experiment is increased by replicating in a number of schools, so that a further stage of the analysis may be set out as follows:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
</tr>
</thead>
<tbody>
<tr>
<td>Schools</td>
<td>s - 1</td>
</tr>
<tr>
<td>Groups within Schools</td>
<td>s</td>
</tr>
<tr>
<td>Tests within Schools</td>
<td>s</td>
</tr>
<tr>
<td>Occasions within Schools</td>
<td>s</td>
</tr>
<tr>
<td>Children within Groups within Schools</td>
<td>2(Σc - s)</td>
</tr>
<tr>
<td>C x T within Groups within Schools</td>
<td>2(Σc - s)</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td><strong>4Σc - 1</strong></td>
</tr>
</tbody>
</table>

Since Tests and Occasions are each crossed with Schools, the final analysis is as shown below:
The repeated \( W \) subscripts in the seventh and eighth lines reflect the fact that both Children and \( C \times T \) are doubly nested within Groups within Schools.

The only mean squares of interest for the present study are those for Tests, \( S \times T \), and \( C \times T \) within Groups within Schools. Their expectations are:

\[
E(\overline{T}) = \sigma^2_{ct.gs} + 2c \sigma^2_{st} + 2cs \sigma^2_t
\]

\[
E(\overline{ST}) = \sigma^2_{ct.gs} + 2c \sigma^2_{st}
\]

\[
E(\overline{CT}_{WW}) = \sigma^2_{ct.gs}
\]

The subscripts are self-explanatory, and the same letters are used for coefficients also. Since Schools are random, the appropriate F-tests are as follows:
For Tests: \[ F = \frac{T}{ST} \quad (d.f. \; l, (s - 1)) \]

For \( S \times T \): \[ F = \frac{ST}{ST_{WW}} \quad (d.f. \; (s - 1), 2(e - s)) \]

In (i), the null hypothesis tested is that the overall zero error = 0.

In (ii), the null hypothesis tested is that the differential zero error among schools is zero.

**B.III - EXPERIMENTAL RESULTS**

Table 1 summarises the main results of the enquiry:

**TABLE 1**

**ZERO ERRORS OF MORAY HOUSE VERBAL REASONING TESTS RELATIVE TO M.H.T. 40**

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>186</td>
<td>132</td>
<td>132</td>
<td>132</td>
<td>132</td>
<td>132</td>
<td>132</td>
<td>162</td>
<td>162</td>
<td>162</td>
<td>156</td>
</tr>
<tr>
<td>M.H.T.</td>
<td>33</td>
<td>39.1</td>
<td>40.1</td>
<td>41.1</td>
<td>42.1</td>
<td>43.1</td>
<td>44.1</td>
<td>45</td>
<td>46</td>
<td>47</td>
<td>48</td>
</tr>
<tr>
<td>Zero Error</td>
<td>-0.17</td>
<td>-0.92</td>
<td>0</td>
<td>-0.41</td>
<td>0.28</td>
<td>-0.42</td>
<td>-0.14</td>
<td>-0.31</td>
<td>-0.85</td>
<td>-1.09</td>
<td>0.20</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>170</td>
<td>122</td>
<td>192</td>
<td>186</td>
<td>174</td>
<td>182</td>
<td>182</td>
<td>210</td>
<td>168</td>
<td>188</td>
<td></td>
</tr>
<tr>
<td>M.H.T.</td>
<td>49</td>
<td>50</td>
<td>51</td>
<td>52</td>
<td>53</td>
<td>54</td>
<td>55</td>
<td>57</td>
<td>57</td>
<td>58</td>
<td></td>
</tr>
<tr>
<td>Zero Error</td>
<td>-4.38</td>
<td>-2.79</td>
<td>-4.82</td>
<td>-4.31</td>
<td>-5.76</td>
<td>-4.72</td>
<td>-6.70</td>
<td>-5.98</td>
<td>-6.11</td>
<td>-6.14</td>
<td></td>
</tr>
</tbody>
</table>

\[ N = \text{number of children in each experiment} \]

1 These tests were used by Mr. Moore, and the zero errors are those obtained by him. Permission to quote these results is gratefully acknowledged.
Mean standard error of zero errors relative to M.H.T. 40 = ± .57. Mean standard error of zero errors derived as in (2) below = .80.

(1) A negative sign means that the mean standardised score on the test indicated is expected to be less by the zero error recorded than that obtained by the same group if M.H.T. 40 had been used instead.

(2) Zero errors relative to tests other than M.H.T. 40 may be readily obtained from Table 1. Thus, the zero error of M.H.T. 58, relative to M.H.T. 56 is -6.14 - (-5.98) = -0.16 ± S.E..80.

Table 2 presents a typical example of the fifteen analyses of variance.* That in Table 2 results from the comparison of two tests, here M.H.T. 40 (test A) and M.H.T. 56 (test B). The analysis is the outcome of a design modified from the replicated 2 x 2 latin square ('cross-over') design of Cochran and Cox ((8), page 112-116). The numbers of children in the six schools concerned were 32, 40, 40, 40, 36, 30, and 32; within each school the children were allocated to groups I and II, in equal numbers. Group I's test order was AB; group II's, BA.

TABLE 2

ANALYSIS OF VARIANCE OF STANDARDISED SCORES OBTAINED BY CHILDREN IN SIX SCHOOLS ON M.H.T. 40 AND M.H.T. 56

(see next page)

* The scores from M.H.T. 39-44 were analysed by Moore in a single table.
The F's recorded indicate that the overall mean zero error is significant and that the zero error varies significantly among schools.

In Table 3 are reported the standardised score means for the six schools used for the M.H.T. 40 - M.H.T. 56 comparison. Each school mean is based on the sum of all scores obtained by the children in the school, half of the scores being in AB order, the other half in BA order. Table 3 gives also the mean zero error for each of the six schools. The regression of zero error on school mean should be noted.

**TABLE 3**

<table>
<thead>
<tr>
<th>School</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>N</td>
<td>32</td>
<td>40</td>
<td>40</td>
<td>36</td>
<td>30</td>
<td>32</td>
<td>210</td>
</tr>
<tr>
<td>Mean</td>
<td>110.23</td>
<td>108.19</td>
<td>108.14</td>
<td>100.28</td>
<td>96.07</td>
<td>92.38</td>
<td>102.99</td>
</tr>
</tbody>
</table>
B.IV - DISCUSSION AND IMPLICATION OF RESULTS

Each of the zero errors in Table 1 is based on a comparison between the arbitrary standard M.H.T. 4.0 and one other test administered to the same sample of children. The care taken over sampling procedure makes it reasonable to accept these observed zero errors as estimates of the population values, the population consisting of the pool of two year-groups of children of appropriate age. A sample of 1,291 children drawn from these two year-groups—those who took M.H.T. 4.0 first—had a mean on this test of 105.5, which we may take as a good estimate of the population mean on M.H.T. 4.0. Had this population of two year-groups taken one of the later tests with a zero error of 5-6 points, the expected mean would have been close to 100, this expectation tallying with our knowledge from other sources of the mean performance levels on later tests of these year-groups. The weight of the statistical evidence, namely: the adjusted mean of about 100; the observed standard deviation of 13.3; the coefficients of skewness ($\frac{\beta_1}{\sqrt{1}} = -0.029$) and kurtosis ($\frac{\beta_2}{s^4} = 2.52$); all this makes it probable that the Table 1 zero errors are applicable to most current year-groups.

These zero errors are themselves subject to error of measurement. Omitting Moore's six tests, there remain fifteen comparisons with M.H.T. 4.0 in most of which the Schools $\times$ Tests interactions were significant (the mean square of 50.0 in Table 2 is rather larger than average). The significance of these interactions has important consequences for the precision of the zero error estimates, since it implies a significant heterogeneity of zero error from school to school in each comparison which must be inferred in the population sampled; the appropriate errors of measurement, must, therefore, be based on the mean square for the Schools $\times$ Tests interactions. Thus, in the M.H.T. 4.0 - M.H.T. 56 comparison, the estimated zero error of 5.98 has a standard error of $\sqrt{2 \times 50/210} = 0.690$. 
The average of all such standard errors is 0.570, which is appropriate for zero errors as recorded in Table 1 (i.e., relative to M.H.T. 40). For zero errors obtained by taking the difference between two recorded errors (i.e., relative to a test other than M.H.T. 40), the mean standard error is \(0.570 \sqrt{2} = 0.80\) approximately.

A study of the regression of zero error on school mean in the fifteen comparisons reveals in some, but not all, a tendency for zero error to increase with school mean. The data of Table 3 provide a striking example; the regression is highly significant. In other comparisons the evidence is less clear, but the general trend is in the same direction. Fortunately, the observed regressions are insufficiently large to necessitate adjustment of observed zero error over the range of local authority mean quotients likely to be met with.

Table 2 illustrates a further point of some importance. The final mean square in the table (12.1) measures the mean error variance of measurement of children's individual standardised scores. This error variance implies a correlation between M.H.T. 40 and M.H.T. 56 scores of .95, referred to a population standard deviation of 15.* Correlations, similarly estimated, between M.H.T. 40 and each of the other tests, range from .92 to .96 with average .94. This uniformity of correlation is evidence that all the tests will place the same set of testees in much the same rank order as M.H.T. 40, and hence as each other. For the validities of the tests in the series with respect to a common criterion, this uniformity is important. There is no evidence of that cumulative error, producing a 'drift' from the criterion, against which Heim ([36], page 109) rightly warns us.

The zero errors in Table 1 show the effect on the mean standardised score of

* \(r = (15^2 - 12.1) / 15^2 = .95.\) This follows from: error variance = \(s^2(1 - r).\)
the same group of testees if one test, rather than another, happens to be chosen. For example, the mean on M.H.T. 58 is estimated as 6.14 (± SE .57) points lower than that on M.H.T. 40; and, similarly, the mean on M.H.T. 54 as 3.63 (± SE .80) points lower than that on M.H.T. 47. These apparent differences in 'difficulty' indicate a rise in the performance level of the samples available for standardisation from the earlier to the later tests. The same group of testees, pitted against a later instead of an earlier standardisation sample, will do relatively less well and so obtain lower standardised scores on the later test.

There are indications in Table I of a very slight upward trend in the zero errors over the first eleven tests listed, and a quite sharp discontinuity between that of M.H.T. 48 and that of M.H.T. 49. The upward trend then persists, but appears to be levelling out over the last four tests in the list.

There can be little doubt that this rise in performance level reflects mainly the effect of test-sophistication, increasing from earlier to later standardisation samples. Children in later samples have become more familiar than their predecessors with the type of material included in tests such as these through more extensive coaching or practice on similar material. It is significant that the discontinuity in zero error between M.H.T. 48 and M.H.T. 49 occurred at a time (1953) when many local authorities (including those used in the standardisation samples) were introducing full-scale practice tests as a preliminary to their allocation procedures; the time of the public controversy over coaching and practice (98). The further upward trend and the flattening towards the end of the series suggests that saturation may now have been reached, though we cannot be certain of this.

We can only speculate about other influences contributing to the rise in performance level reflected in increasing zero errors. Tanner (81), (82) has pointed out that as a result of better nutrition and perhaps because the world is getting
children in Western Europe and America are maturing physically at an earlier age than their predecessors; according to him the presumptions is that there is a similar trend in mental development. If this is so, the effect on performance and hence on zero error will reinforce that of increasing test-sophistication.

We turn now to the practical uses to which a knowledge of these zero errors can be put. In the first place they should be useful to those responsible for testing groups of children when comparisons have to be made between two groups of children tested with different tests. A knowledge of the zero error of one test relative to the other will enable the tester to adjust one of the observed group means so that the comparison between them is unbiased. The step from group means to individual scores is a little hazardous; this study is concerned with biasses in mean scores only, and provides no evidence that the bias is constant throughout the range of individual scores; indeed, some slight evidence is presented that zero error, like practice effect, is positively correlated with score. However, if the step from mean scores to individual scores is made, it will be in the right direction even if its magnitude is not entirely correct. Thus, in the case of a child who moved from one area where he obtained a standardised score on M.H.T. 47 of 114 to another where the test used was M.H.T. 57, it would be fairer to adjust his standardised score to 119, than to leave it at 114 in placing him in relation to the children in his new year-group. Again, in the case previously mentioned, of a supplementary examination administered to absentees from the main examination, it would not be unreasonable to diminish the score of each child taking the supplementary examination by 4 points if the supplementary test were M.H.T. 50 and the main test M.H.T. 55.

Zero error may need to be taken into account in interpreting the results of research investigations; for example, the study by Wiseman and Wrigley (94) of
the effects of coaching and practice based on test booklets containing material alleged to be similar to that incorporated in verbal reasoning tests. They used a number of Moray House tests to evaluate the practice and coaching effects at various stages in their experiment, and had reason to suspect, from the general pattern of group mean scores, standardisation errors in M.H.T. 39 and M.H.T. 40. If we adjust (where we have the data to do so) the mean scores obtained by Wiseman and Wrigley in the test run for the children who had simple practice we obtain the following:

<table>
<thead>
<tr>
<th></th>
<th>Initial Test</th>
<th>Practice Tests</th>
<th>Final Tests</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>M.H.T. 43</td>
<td>37*</td>
<td>38*</td>
</tr>
<tr>
<td>W. and W's Means</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Zero Errors</td>
<td>-0.42</td>
<td>-</td>
<td>-</td>
</tr>
<tr>
<td>Adjusted Means</td>
<td>100.04</td>
<td>(104.07)</td>
<td>(108.88)</td>
</tr>
</tbody>
</table>

The adjustments have had the effect of bringing the M.H.T. 39 mean more into line with the general trend, and the flattening out from M.H.T. 40 is slightly more regular than with the unadjusted array of means. The author's main conclusions remain unaltered, but it is easy to see how misleading the results might have been had tests with widely discrepant zero errors been chosen for the experiment.

Finally, the results of the present study have implications for the problem of test standards generally. The upward trend of zero errors is evidence that the level of performance on tests of this kind has risen generally and may continue to rise. Test standardisations must of necessity be based on the performance of

* M.H.T. 37 and M.H.T. 38, two very old tests, were out of print and could not be included in the range of tests studied.
contemporary samples of children; and these, it is argued here, are not equivalent to earlier samples. Are those responsible for the standardisations to continue to assign to each new test an arbitrary mean standardised score of 100 in spite of this known non-equivalence? Or are they to relate the test means to that of some previous test, standardised on a sample of children of an earlier 'vintage', assigning to that test only - as in this study - an arbitrary mean of 100? To adopt the first course is to fail to take account of, and, indeed, to obscure the undoubted rise in performance level the country over; to adopt the second is to accept the necessity of embarking on an extension of the work reported here, to co-relate all present and future tests in this and all other series of tests from whatever source - a formidable task. The second course, though perhaps preferable, presents difficulties, particularly with regard to tests of attainment. Test constructors are continually trying to improve their tests, to incorporate in them the results of recent research findings, to increase their validity, to make them more liberal in conception and to minimise undesirable 'backwash' effects on school curricula. To relate performance on some future English attainment test, say, to that on an early member of the series, when the tests have little in common except their name, would seem unrealistic.
PART TWO

THE RESCALING OF TEACHERS' ASSESSMENTS
CHAPTER NINE - INTRODUCTION

I - GENERAL

The problem of rescaling assessments for different groups (different schools, different age-groups or the like) arises in many different contexts; for example, in the rescaling of essay marks assigned by different markers to different groups of children; of marks obtained in Teachers' Certificate Examinations by students in different Training Colleges; and, particularly since the passing of the 1944 Education Act, of teachers' assessments used in the allocation procedures of local education authorities. This last application is of special importance at the present time owing to the increasing use now being made, after a slow start, of teachers' rescaled assessments in those allocation procedures.

II - THE NECESSITY FOR RESCALING ASSESSMENTS

The necessity for rescaling arises from the familiar fact that the means and the dispersions arbitrarily adopted by different teachers in different schools are not directly comparable. Each teacher making assessments for his own group (school or class) does so in isolation. If his assessments take the form of marks, the level and the scatter he assigns are necessarily arbitrary and unlikely to be comparable as to scale with the marks of other teachers with other groups. Attempts to achieve a common scale by preliminary consultations among teachers have not been successful. If the assessments take the form of ranks, as is quite common practice, they are obviously unuseable as they stand. Hence a single overall test which can be administered to all the testees is needed to determine the most probable means and group standard deviations.
III - PREVIOUS WORK

There is already a considerable body of literature on the subject of rescaling. McClelland (54) studied the relative validities of a number of predictors, including assessments rescaled in various ways. A number of handbooks have been published, advocating particular methods of rescaling, and giving advice on computational procedures and their practical application. Notable among these are the works of Sutcliffe and Canham (80), Edwards (19) and McIntosh et al (55). More recently, Vernon et al (91) have summarised the methods advocated by previous writers, and have not only made some evaluation of these various procedures, but have also attempted a statistical analysis of the errors arising if one test, rather than another, is used for rescaling. Yates and Pidgeon used extensively a simplified rescaling procedure in their Twickenham researches, and have shown that assessments thus rescaled have validity as high as, or higher than scores on objective tests in respect of their criterion (96).

Most of these writers appear to accept without question the customary basic procedure, which they use, either as it stands, or in slightly modified form. Little attempt seems to have been made to state the underlying assumptions, still less to study their implications. Sandon (73) makes some attempt, but restricts himself to one specialised aspect of the problem. As already stated, Vernon et al make use of statistical procedures in their discussion of rescaling, but do not enquire into their rationale.

A more recent writer, Miss Howard (39, pages 199-207), must be absolved from the charge of accepting uncritically a rule-of-thumb procedure. It is not until after a considerable discussion that she employs the generally accepted procedure in a validity study similar to that following. (See also (65), pp. 191-197)
Perhaps the most searching recent contribution* to a study of underlying principles is that of Greenall (31), but his article has not received the attention it merits, perhaps because its relevance to resealing was not immediately apparent at the time he was writing. Greenall's terminology and algebraic proof are not altogether easy to follow; and there seems to be a need for a fuller and simpler discussion of the chief alternative rescaling procedures available, and, above all, for an empirical comparison of them by actual application to a representative composite group.

IV - PURPOSES OF THIS STUDY

The purposes of this study are as follows:

(1) To state a general rescaling equation which implements assumptions also to be stated.

(2) Using this equation, to derive alternative rescaling procedures (a) when the criterion is known, (b) when it is not known; and to evaluate these procedures in the light of empirical evidence. This part of the enquiry seems necessary since one particular procedure has come to be unquestioningly used although there are others with a claim to recognition on the assumptions made.

(3) To study the reliability of the rescaling procedure. Except for a contribution (of doubtful value) from Vernon et al (91), no work has been done on this aspect of the matter. Rescaled assessments are coming increasingly into use. It would seem to be a matter of urgency that those using them should know, firstly, that these assessments would alter, and secondly, by how much they

* There are several purely theoretical discussions of the problem of estimating scores on a fallible dependent variable from scores on a fallible independent variable, notably those of Pearson (63), Cramér (11) and Wald (92). An early discussion of the same issue in a psychological context is that of Otis (60) and Thomson (87) reached similar conclusions, though from different arguments.
would probably alter, if, for the rescaling test used, another were substituted. The theoretical aspects will first be studied, and a practical method of estimating the reliability of the rescaling procedure will be devised. This method will be employed in two different experiments in which the conditions of administration of the rescaling tests differ.

(4) To make use of the theoretical foundations laid in (3) in questioning the belief, generally accepted though untested, that for small school groups rescaling is dangerously unreliable.

(5) To provide empirical evidence that test scores differing from teachers' expectations occur about equally frequently in both directions. This evidence is opposed to the findings of McIntosh et al (55) who state that 'flop' scores (i.e. test scores much lower than the corresponding teachers' assessments duly rescaled) occur with considerably greater frequency than their opposite.

(6) To consider the manner in which assessments should be made and the choice of appropriate rescaling tests. In these connections the 'forward-looking' and 'backward-looking' aspects of test scores and assessments will be discussed.

(7) To develop the statistical specification of the 'ideal' rescaling test; to devise equations which throw light on the sources of error arising when tests used for rescaling fail to conform to this specification; and to suggest lines of further research on the choice of rescaling tests based on the concept of the 'ideal' test.

(8) To note and comment briefly on some of the labour-saving variants, currently in use, of the accepted basic procedure of equating means and standard deviations; and to suggest a practical routine which takes account of differences in children's ages in a manner which appears to be less open to objection than other proposed routines.
V - NOTE ON TERMINOLOGY

Assessments adjusted to achieve comparability are commonly referred to as 'scaled' assessments, which seems to imply that before adjustment they were not scaled at all. However, the original unadjusted assessments were scaled, though the scales used by the several assessors were not comparable. It is preferable, therefore, to use the adjective 'rescaled' to describe the assessments after all have been placed on a common scale.
CHAPTER TEN *

METHODS OF RESCALING

A - DERIVATION OF THE GENERAL RESCALING EQUATION

A.1 - RATIONALE AND UNDERLYING ASSUMPTIONS FOR RESCALING

From a statistical standpoint, the primary question is how to construct a single scale which will preserve the order, and possibly the relative spacings, obtained from different teachers, when originally each had his own standard and his own way of spreading the marks.

The starting-point is an assumption that after adjustment the marks should approximate as closely as possible to those that would have been assigned by a single hypothetical teacher or judge who was in a position to estimate the achievements or potentialities of the children in all the schools. All the marks would then be on a common scale - that of the hypothetical teacher.

It is next assumed that these hypothetical marks would be positively correlated with scores on a test administered to all the children. If these marks and scores were put on the same scale - that of the scores - there would be some degree of correspondence between the group means and the group s.d.'s for the marks and scores. The group means and s.d.'s derived from the test scores are therefore accepted as starting points in furnishing estimates of the corresponding statistics for the hypothetical teacher. It would not do to assume without question that the group means and s.d.'s from the test are themselves the best estimates (though, in the event, appropriate enquiry may show them to be so). We do assume, however, the existence of some simple relationship between the test means and s.d.'s for the groups and those of the hypothetical teacher, and that this relationship can be brought to light by statistical investigation.

* This Chapter is a considerable expansion of an article originally published in 1968.
It is assumed finally that whatever adjustments must be made to the diverse arbitrary scales used initially by the teachers, the rank orders in which they place their own groups of children and the relative spacing of the children within these groups conform more closely to the rank orders and relative spacings of the hypothetical teacher than do those of the test scores.

Using these assumptions we obtain an overall array of assessments in which each teacher's rank order, and perhaps his relative spacings, are retained. The arbitrary scale he has used is rejected in favour of one common to all.

A.II - THE GENERAL EQUATION

We obtain first a general resealing equation which implements these assumptions. Let \( z \) denote the original assessment, \( x \) the test score, and \( y \) the adjusted assessment. For both \( x \) and \( y \) we may fix the mean for the entire sample at zero without loss of generality. Taking first some particular school with \( n \) children, let \( s_z \) and \( s_x \) denote the original standard deviations, and \( s_y \) that of the adjusted assessments. It is assumed (i) that within each school, \( s_y = s_x \), and (ii) that the adjusted means \( \bar{y} \) of the several schools are proportional to the test means \( \bar{x} \).

The first of these assumptions implies that within each school the arbitrarily assigned standard deviation \( s_z \) of the original assessments \( z \) must be altered so that the adjusted assessments \( y \) have the standard deviation \( s_x \). This is achieved by putting

\[
\frac{(z - \bar{z})}{s_z} = \frac{(y - \bar{y})}{s_x}
\]

or

\[
y = \bar{y} + \frac{(z - \bar{z}) s_x}{s_z}.
\]

The second assumption gives

\[
\bar{y} = k \bar{x}
\]

in which \( k \) is a constant to be determined.
The general resealing equation is therefore

\[ y = k\bar{x} + (z - \bar{z})s_x / s_z \]

that is,

\[ y = k\bar{x} + (y - \bar{y}) \]

(1)

This equation will be used extensively in the development which follows.

**B - ALTERNATIVE RESCALING PROCEDURES**

**B.I - GENERAL RELATIONS BETWEEN ASSESSMENTS, TEST SCORES, AND CRITERION SCORES**

For convenience, the general resealing equation just derived is repeated below:

\[ y = k\bar{x} + (z - \bar{z})s_x / s_z \]

\[ y = k\bar{x} + (y - \bar{y}) \]

(1)

The corresponding equation for the test is

\[ x = \bar{x} + (x - \bar{x}) \]

(2)

Squaring and summing, first within and then over all schools, we obtain, from equations (1) and (2) respectively:

\[ \Sigma \Sigma y^2 = k^2 \Sigma n\bar{x}^2 + \Sigma \Sigma (x - \bar{x})^2 \]

(3)

and

\[ \Sigma \Sigma x^2 = \Sigma n\bar{x}^2 + \Sigma \Sigma (x - \bar{x})^2 \]

(4)

The cross-products from equations (1) and (2) give

\[ \Sigma \Sigma xy = k \Sigma n\bar{x}^2 + \Sigma \sqrt{r_{xz}} \Sigma (x - \bar{x})^2 \]

(5)

or, alternatively,
\[ \Sigma xy = k \Sigma x^2 + r_{xy} \Sigma (x - \bar{x})^2 \]  \hspace{1cm} (6)

In equation (5), \( r_{xz} \) is the correlation between \( x \) and \( y \) for a particular school,* and in equation (6), \( r_{xy} \) is the within-schools correlation obtained by pooling sums of squares and sums of products after the \( z \)'s have been transformed into \( y \)'s. Thus \( r_{xy} \) is the weighted mean of the \( r_{xz} \)'s, the weights being the within-school sums of squares or variances.

Denoting the criterion by \( c \), we have

\[ c = \bar{c} + (c - \bar{c}) \]  \hspace{1cm} (7)

so that

\[ \Sigma \Sigma c^2 = \Sigma \Sigma c^2 + \Sigma \Sigma (c - \bar{c})^2 \]  \hspace{1cm} (8)

\[ \Sigma \Sigma xc = \Sigma \Sigma x\bar{c} + \Sigma \Sigma (x - \bar{x})(c - \bar{c}) \]  \hspace{1cm} (9)

\[ \Sigma \Sigma yc = \Sigma \Sigma y\bar{c} + \Sigma \Sigma (y - \bar{y})(c - \bar{c}) \]

\[ = k \Sigma \Sigma x\bar{c} + (r_{yc} / r_{xc}) \Sigma \Sigma (x - \bar{x})(c - \bar{c}) \]  \hspace{1cm} (10)

where \( r_{yc} \) and \( r_{xc} \) are within-schools correlations obtained by pooling sums of squares and sums of products.

The notation above is too cumbrous to be suitable for the subsequent development. For sums of squares, the single capital letters \( X \), \( Y \) and \( C \) will be used, and, for sums of products, pairs of capitals (\( XY \)), (\( XC \)), and (\( YC \)). The subscripts \( B \), \( W \) and \( T \) will be used to denote sums of squares or sums of products between schools, within schools, and over the total batch of schools respectively. Equations (3), (4), (6), (8), (9) and (10) may then be abridged as follows:

---

* A linear change of scale from \( z \) to \( y \) does not alter the correlation with \( x \).
\[
Y_{T} = k^{2}X_{B} + X_{W} \tag{3a}
\]
\[
X_{T} = X_{B} + X_{W} \tag{4a}
\]
\[
(XY)_{T} = kX_{B} + r_{XY}X_{W} \tag{6a}
\]
\[
C_{T} = C_{B} + C_{W} \tag{8a}
\]
\[
(XC)_{T} = (XC)_{B} + (XC)_{W} \tag{9a}
\]
\[
(YC)_{T} = (YC)_{B} + (YC)_{W}
\]
\[
= k(XC)_{B} + (r_{yc}/r_{xc})(XC)_{W} \tag{10a}
\]

The problem before us is that of assigning a value to \(k\) in the general rescaling equation (1). Four methods of doing so will be considered; in each it is assumed that regression is linear.

**B.II - Rescaling with Reference to the Criterion**

For two of the methods to be considered it is assumed that the criterion is known. **Method 1** is appropriate when test scores and assessments are to be used together to predict a known criterion; the method gives the optimum value of \(k\) in the rescaling equation as well as the weights for test score and rescaled assessments which maximize prediction of the criterion by the composite of the two. **Method 2** is used to obtain the optimum value of \(k\) (not necessarily the same as the previous \(k\)) in the rescaling equation when the assessments are used alone to predict a known criterion.

**Method 1**

If test scores and assessments are to be used together to predict the criterion, the variables used in prediction, as equations (1) and (2) show, are
(i) the test means \( \bar{x} \); (ii) the deviations \((x - \bar{x})\) of the test scores about the test school means; and (iii) the corresponding deviations \((y - \bar{y})\) of the rescaled assessments. Denoting these predictors by \(m\), \(d\) and \(f\), and the predicted criterion score by \(\hat{c}_s\), all being expressed in standard measure, we have

\[
\hat{c}_s = b_1m + b_2d + b_3f
\]

where the \(b\)'s are partial regression coefficients \((b_1 = b_{cm}d; b_2 = b_{cd}m; b_3 = b_{cf}md)\) to be evaluated. The assumption implicit in developing equations (3a) - (10a), that the schools means are independent of the deviations within schools, simplifies the work. The starting point in the evaluation of \(b_1, b_2\) and \(b_3\) is the following pooling square:

<table>
<thead>
<tr>
<th></th>
<th>(b_1)</th>
<th>(b_2)</th>
<th>(b_3)</th>
</tr>
</thead>
<tbody>
<tr>
<td>(b_1)</td>
<td>(r_{cm})</td>
<td>1</td>
<td>0</td>
</tr>
<tr>
<td>(b_2)</td>
<td>0</td>
<td>(r_{cd})</td>
<td>(r_{df})</td>
</tr>
<tr>
<td>(b_3)</td>
<td>0</td>
<td>(r_{df})</td>
<td>1</td>
</tr>
</tbody>
</table>

Any one of the several methods of arriving at the \(b\)'s (e.g. Aitken's method of pivotal condensation (Thomson(88)), gives:

\[
b_1 = r_{cm}; \quad b_2 = (r_{cd} - r_{cf}r_{df}) / (1 - r_{df}^2); \quad b_3 = (r_{cf} - r_{cd}r_{df}) / (1 - r_{df}^2).
\]

The first of these, \(r_{cm}\), is the correlation between the test school means and the criterion scores; every child in a particular school is assigned, on the one
hand, the mean of his school on the test, and, on the other, his own criterion score. Denoting the test mean in the \(i\)th school by \(\bar{x}_i\), and the criterion score of the \(j\)th child in this school by \(c_{ij}\); and remembering that the mean for the entire sample over all schools is zero for both predictors and criterion, we have for the \(i\)th school:

\[\text{Sum of products} = \bar{x}_i \sum_j c_{ij} = n_i \bar{x}_i \bar{c}_i;\]
\[\text{Sum of squares for predictor} = n_i \bar{x}_i^2;\]
\[\text{Sum of squares for criterion} = \sum_j (c_{ij} - 0)^2 = \sum_j c_{ij}^2.\]

Summing over all schools we obtain:

\[\text{Total sum of products} = \sum_i n_i \bar{x}_i \bar{c}_i = (XC)_B.\]
\[\text{Total sum of squares for predictor} = \sum_i n_i \bar{x}_i^2 = X_B.\]
\[\text{Total sum of squares for criterion} = \sum_{ij} c_{ij}^2 = C_T.\]

Hence \(r_{cm} = (XC)_B / \sqrt{(X_B C_T)}.\)

The second correlation \(r_{cd}\) is that between test score deviations from the school means and criterion scores. For the \(j\)th child in the \(i\)th school these are respectively \((x_{ij} - \bar{x}_i)\) and \(c_{ij}\). Using equation (7), multiplying, and summing first within school \(i\) and then over all schools we obtain:

\[\text{Total sum of products} = \sum_{ij} (x_{ij} - \bar{x}_i)(c_{ij} - \bar{c}_i) = (XC)_W.\]

(see equations (9) and (9a)).
Also, the total sums of squares for predictor and criterion are clearly \( X_w \) and \( C_T \) respectively, so that

\[
\hat{r}_{cd} = \frac{(XC)_w}{\sqrt{(X_wC_m)}}.
\]

The third correlation, \( r_{cf} \), that between rescaled assessments deviations from the school means and criterion scores is obtained similarly, so that

\[
r_{cf} = \frac{(YC)_w}{\sqrt{(Y_wC_m)}} = \frac{(YC)_w}{\sqrt{(X_wC_m)}}.
\]

The fourth and final correlation, \( r_{df} \), is that between test score deviations and rescaled assessment deviations, each from the several school means; clearly it is given by

\[
r_{df} = \frac{(XY)_w}{\sqrt{(X_wY_w)}} = \frac{(XY)_w}{X_w} = r_{xy}.
\]

An alternative, and in some ways more illuminating, method of deriving the correlations \( r_{cm} \) and \( r_{cd} \) is first to derive from equations (4a), (3a) and (9a) the equation

\[
\hat{r}_{xc} = \frac{\sum(XC)_B + (XC)_W}{\sqrt{\sum C_T (X_B + X_W)}}
\]

Assigning every child in a particular school the test mean for that school is equivalent to putting \((XC)_B\) and \((X_B)\) equal to zero in the equation above. We then obtain

\[
r_{cm} = \frac{(XC)_B}{\sqrt{(X_BC_m)}} \text{ as before.}
\]

Similarly the correlation between test score deviations and criterion scores is obtained by equating \((XC)_B\) and \(X_B\) to zero, thus obtaining

\[
r_{cd} = \frac{(XC)_w}{\sqrt{(X_wC_m)}} \text{ as before.}
\]

Finally, the third correlation \( r_{df} \) can be obtained from equations (3a), (4a) and (6a). We write
\[
\frac{r_{xy}}{r_{df}} = \frac{(kX_B + r_{xy}X_w)}{\sqrt{\sum X_B + X_w}(k^2 X_B + X_w)}
\]

(total)

and then put \(X_B = 0\), obtaining

\[
\frac{r_{df}}{r_{xy}} = \frac{r_{xy}}{r_{df}} 
\]
as before.

The partial regression coefficients derived from these correlations are the weights which give the multiple correlation with the criterion when the predictors \(x\), \((x - \bar{x})\), and \((y - \bar{y})\) are expressed in standard measure. To obtain weights in working units, we 'de-standardise' and use

\[
\begin{align*}
  w_1 &= b_1 \sqrt{C_T/X_B} ; \\
  w_2 &= b_2 \sqrt{C_T/X_w} ; \\
  w_3 &= b_3 \sqrt{C_T/X_w}.
\end{align*}
\]

(\(76\), page 343)

These weights now maximise prediction in the equation (all variables in working units):

\[
\hat{\sigma} = w_1 \bar{x} + w_2 (x - \bar{x}) + w_3 (y - \bar{y})
\]

\(\text{(11)}\)

It remains to evaluate \(k\) (\(= k_1\) for Method 1) in equation (1). Multiplying each term in equation (2) by \(w_2\) and each term in equation (1) by \(w_3\), and summing, we obtain

\[
w_2 x + w_3 y = (w_2 + k_1 w_3) \bar{x} + w_2 (x - \bar{x}) + w_3 (y - \bar{y})
\]

\(\text{(12)}\)

whence, from equation (11), the optimum value of \(k_1\) is given by

\[
k_1 = \frac{(w_1 - w_2)}{w_3}
\]

\(\text{(13)}\)

Equation (12) indicates also the practical working method. First rescale the original assessments using equations (1) and (13); then sum the resulting \(y\)'s, weighted by \(w_3\), and the corresponding \(x\)'s, weighted by \(w_2\); the weighted sum is the estimate \(\hat{\sigma}\), which correlates best with the criterion.
Method 2

If the rescaled assessments alone are to be used to predict the criterion, the variables used in prediction are (i) the test means $\bar{x}$, and (ii) the deviations $(y - \bar{y})$ of the rescaled assessments about the school means. The prediction equation (in standard measure) may be written:

$$\hat{c}'_s = b_4 m + b_5 f$$

where $m$ and $f$ are as before, and $b_4$ and $b_5$ are the partial regression coefficients $r_{cm.f}$ and $r_{cf_m}$ respectively. Since $m$ and $f$ are independent of each other, these regression coefficients become

$$b_4 = r_{cm} = b_1; \quad b_5 = r_{cf}.$$

The working weights are $w_1 = b_1 \sqrt{(C_T / X_B)}; \quad w_5 = b_5 \sqrt{(C_T / X_B)}$; and prediction by Method 2 is maximised in the equation

$$\hat{c}' = w_1 \bar{x} + w_5 (y - \bar{y});$$

whence, by arguments similar to those used previously, the constant $k ( = k_2$ for Method 2) is given by

$$k_2 = \frac{w_1}{w_5}.$$

B.III - RESCALING WITHOUT REFERENCE TO THE CRITERION

When the predictive validities of the test scores and estimates are unknown, any method used to rescale the estimates must of necessity be somewhat arbitrary. The variables predicting the criterion are still the school means $\bar{x}$, the deviations $(x - \bar{x})$ of the test scores about these means, and the corresponding deviations $(y - \bar{y})$ of the estimates, but how best to weight them is not known.
In the absence of any information about the criterion validity of either test scores or assessments, one way of rescaling assessments on test scores is based on the reasonable assumption that both are fallible measures of the same underlying trait - a trait valid (in an unknown degree) for the criterion. If test scores and assessments are assumed to be equally fallible, the regression line expressing the relationship between the corresponding 'true' measures is the principal axis of the ellipsoidal distribution. When test scores and assessments are on the same scale, this regression line is at 45° to the orthogonal reference axes, and the expectations of a particular child's test score and assessment are the same - a consequence of considerable practical convenience.

To obtain this convenience the arbitrary scale of the assessments is adjusted so that for each school the mean and the standard deviation are the same as for the test scores. The principal axis of the ellipsoidal distribution over all schools is then at 45° to each reference axis, and the assessments and test scores equivalent in Greenall's sense (31); the best numerical estimate of a child's rescaled assessment is his test score; and conversely, the best numerical estimate of his test score is his rescaled assessment.

The error variance of estimation (in either direction) is readily seen to be the error variance of the difference between corresponding scores on the two measures, that is, twice the error variance of measurement of the underlying trait. Thus the mean standard error in estimating a rescaled assessment from a test score or (vice versa) is \( s \sqrt{2(1 - r)} \), in which \( s \) is the overall

* 'True' has its usual statistical connotation: \( \lim_{n \to \infty} \frac{1}{n} \sum (\text{observed measures}) \).
standard deviation common to both measures and $r$ the overall correlation between them.*

Method 3 is that customarily used in one form or other. It has been discussed at some length because not infrequently it is employed without understanding of the underlying theory. It was noted earlier (see footnote, page 127), that although the statistical problem of estimating one fallible measure from another has engaged the attention of a number of writers none appears to have dealt specifically with its implications for the rescaling of teachers' assessments, particularly by Method 3.

Clearly in Method 3, since $\bar{y} = \bar{x}$ for each school, the constant is

$$k_3 = 1 \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad \ldots \quad (15)$$

Frequently the test scores and assessments rescaled by Method 3, when used together, are given equal weight; a commonsense procedure if information about the criterion is lacking. The composite is then

$$x + y = \sum \bar{x} + (x - \bar{x}) \frac{1}{\sum} + \sum \bar{y} + (y - \bar{y}) \frac{1}{\sum} = 2\bar{x} + (x - \bar{x}) + (y - \bar{y}),$$

the two sets of deviations therefore being assigned the same weight, and the test means twice this weight. Supposing the criterion to be known, however, possibly better prediction might be obtained by differentially weighting test scores and Method 3 assessments. It may be asked, whether a composite of test scores plus assessments rescaled by Method 3 can ever be as efficient in predicting the criterion as the corresponding composite in which the assessments are rescaled by Method 1; and if so under what conditions. The question may be answered as follows.

Firstly, in Method 3 the weight assigned to the school means is always equal to the sum of the weights assigned to the deviations. Secondly, the best weights

* For a fuller discussion of the standard error of a rescaled assessment (from a rather different point of view) see Chapter 11, page 16.
in Method 1 are the 'destandardised' partial regression coefficients \( w_1, w_2 \)
and \( w_3 \). It follows that a composite including assessments rescaled by
Method 3 can be as efficient a predictor as a composite including assessments
rescaled by Method 1 if it happens that \( w_1 = (w_2 + w_3) \). The same conclusion
follows on noting that \( k_1 = (w_1 - w_2) / w_3 \) which, for \( w_1 = w_2 + w_3 \), gives
\( k_1 = k_3 = 1 \). Clearly, however, such an occurrence will be a chance one;
rescaling by Method 1, in which the weights assigned to \( x \) is not predetermined
by those assigned to the deviations, is in principle the better.*

**Method 4**

A further method with a claim on our attention is one in which the rescaling
procedure is based on the ordinary regression line, test scores being the
independent, and assessments the dependent variable. The standard deviation of
the assessments is first equated, school by school, to that of the test scores;
this fixes the magnitude of each deviation \( (y - \bar{y}) = (z - \bar{z}) \frac{s_x}{s_z} \), as in
the three methods so far considered. Change of scale within schools leaves
unaltered the correlations \( r_{xz} \), that is, between test scores and estimates
for a particular school, but fixes the regression of rescaled assessments \( y \) on test
\( x \) as for \( r_{xz} \) for that school. This in turn leads to the estimate \( \bar{y} = \bar{r}_{xz} \bar{x} \),
thus completing the rescaling by fixing the level, for that school, of the
rescaled assessments which are \( y = \bar{r}_{xz} \bar{x} + (z - \bar{z}) \frac{s_x}{s_z} \). For the school
concerned, the regression line with slope \( b_{yx} = r_{xy} \), obviously minimises the
error variance of prediction of \( y \) by \( x \).

* A correspondent who read the original article (65), suggested that by suitable
weighting, the composite of test scores and assessments rescaled by Method 3 could
always be made as efficient a predictor of a given criterion as the composite of
test scores and assessments rescaled by Method 1. This is clearly not the case.
The result of repetition of the process over all schools is a series of y-means more or less collinear; completely so if $r_{xz}$ is the same for all schools, that is if all the teachers concerned assign sets of assessments in the first place which correlate equally well with the corresponding sets of test scores; but, in practice, only approximately collinear, since the teachers do in fact vary in this respect, so that $r_{xy}$ will not be constant for all schools. Clearly such a procedure minimises the total error variance of estimate of $y$ by $x$ over all schools.

A serious objection to the piece-meal procedure suggested above is that it might be unfair to all children in one or even several schools. It is unlikely that the assessments made by some particular teacher will have zero correlation with the test scores. If, however, this were to occur, that is, $b_{yx} = 0$, for some school, so that $\bar{y} = b_{xy} = \bar{x} = 0$, then that school's assigned mean for rescaled assessments would have to be the general average test score over all schools: and if that school's test mean $\bar{x}$ deviated appreciably from this overall mean one way or the other, one would strongly suspect that to assign the overall mean as the school mean for rescaled assessments would introduce a bias affecting all the children in the school one way or the other.

On grounds of fairness, therefore, it seems reasonable to use a single 'average' regression line over all the schools instead of a separate regression line for each. On statistical grounds also this course is preferable; the single regression obtained by pooling data from all schools is the best available estimate of the population value of the within-schools regression. This 'average' regression coefficient is clearly $r_{xy}$ in equation (6), that is, the weighted mean of the $r_{xz}$'s, the weights being the within-school sums of squares.
It follows that in Method 4, the constant is

\[ k_4 = r_{xy} \quad \ldots \quad \ldots \quad (16) \]

where \( r_{xy} = \frac{\sum (x - \bar{x})(y - \bar{y})}{\sum (x - \bar{x})^2} \)

Method (4) is that advocated by Bloom (2), page 137.

**B. IV - COMPARISON OF RESCALING PROCEDURES**

The four rescaling equations are:

With Method 1:

\[ y_1 = \frac{x(w_1 - w_2)}{w_3} + (z - \bar{z}) \frac{s_x}{s_z} \]

With Method 2:

\[ y_2 = \frac{xw_1}{w_5} + (z - \bar{z}) \frac{s_x}{s_z} \]

With Method 3:

\[ y_3 = \bar{x} + (z - \bar{z}) \frac{s_x}{s_z} \]

With Method 4:

\[ y_4 = r_{xy} \bar{x} + (z - \bar{z}) \frac{s_x}{s_z} \]

Methods 1 and 2, while obviously the best, must be ruled out when rescaling must be undertaken immediately after the examination in which tests and assessments are used. Once the criterion has become available, however, follow-back studies may show, for Method 1, the optimum value of \( k_1 \) and the weights for test scores and rescaled assessments which maximise prediction; and, for Method 2, in which the rescaled assessments are used alone, the optimum value of \( k_2 \). The information thus obtained may be used, with due caution, for similar examinations in the future. We may sum up for these two methods by saying that of the two, Method 1 is likely a priori to be the better in predicting the criterion since it incorporates one extra variable (the deviations of test scores from their means), provided that this extra variable itself correlates positively with the criterion, and not perfectly with the corresponding deviations of the assessments.
We turn to a discussion and comparison of Methods 3 and 4, and begin by examining the correlations of assessments rescaled by each method with the test scores. From equations (3a), (4a) and (6a), noting that, for Method 3, 
\[ k = k_3 = 1, \]
we have the total correlation \( r_{xy(T)} \) given by

\[
r_{xy(T)} = \frac{X_B + r_{xy}X_W}{X_T}
\]

in which, it will be recalled, \( r_{xy} \) is the average within-schools correlation between test scores and assessments.

It is important to note the effect on \( r_{xy(T)} \) of changes in \( r_{xy} \). If, for example, \( r_{xy} = 1 \), \( r_{xy(T)} = 1 \) also. Each teacher's original assessments would correlate perfectly with the corresponding test scores, school by school, the only difference between them being one of scale; and this difference would be removed by adjusting the within-schools standard deviations. In this case, the rescaling problem would be of academic interest, since there would be no need for the assessments. At the other extreme, if \( r_{xy} = 0 \), then \( r_{xy(T)} = X_B / X_T \), which is always positive unless the school means are all the same. Even though each teacher's assessments bore no relation whatever to the test scores, the overall correlation between assessments and scores would be positive; and this positive correlation would result solely from the equating of school means in the rescaling procedure. Method 3 is that commonly adopted; and the positive correlation just noted is one indication of the fact, often overlooked in discussions of teachers' assessments, that rescaling by Method 3 introduces information from the test used in scaling.

For Method 4, the total correlation between test scores and assessments (see equations (3a), (4a) and (6a), with \( k = k_4 = r_{xy} \)) is given by
The effect of putting $r_{xy} = 1$ is to make $r_{xy}(T) = 1$, as before. However, by putting $r_{xy} = 0$, we obtain $r_{xy}(T) = 0$ also. Assessments rescaled on the ordinary regression line will therefore exhibit a correlation overall with the test scores only if the original assessments bear some correspondence to the test scores, school by school. If there is no such correspondence, the rescaled estimates will reflect the fact.

Since in Method 4 the usual regression line is used; and since within each school, and hence over all schools, the error variance of prediction of $y$ by $x$ is minimised, there is here an interesting and instructive parallel with the procedure used in correcting a correlation $r_{xy}$ for selection, in the case where the selection is based on $x$, and the variances of $x$ are known in both the restricted and the whole group. The equation may be written

$$R^2_{xy} = \frac{r^2_{xy} S^2}{s^2 + r^2_{xy} (S^2 - s^2)}$$

where $R_{xy}$ is the correlation after correcting for selection and $r_{xy}$ the correlation in the restricted group; $S^2$ is the total variance in the whole group and $s^2$ the variance in the restricted group.

If the within-groups quantities $X_w, Y_w, (XY)_w$ etc. are thought of as referring to a conceptual school which is the 'average' of all the schools in the group, and if this 'average' school is regarded as a restricted sample (with variance $s^2$) from a larger group (with variance $S^2$), which is the whole group of schools; then the relationships that have been developed in terms of
analysis of variance and covariance should lead to the equation quoted above. That this is the case can be seen by rewriting the equation for \( r_{xy(T)} \) (see page 146) in the form

\[
 r_{xy(T)}^2 = \frac{r_{xy}^2}{X_T^2 + r_{xy}^2 (X_T - X_w)}
\]

If also we write

\[
 X_T = (N - 1)S^2
\]

\[
 X_w = (N - m)s^2
\]

where \( N \) is the total number of children and \( m \) the number of schools, and note that \( (N - 1) \div (N - m) \), the identity of \( R_{xy} \) and \( r_{xy(T)} \) becomes obvious.

**Validities of Method 3 and Method 4 Assessments**

The validities of Method 3 and Method 4 assessments in respect of the criterion are next examined. From equations (3a), (8a) and (10a) we may write the general equation for the validity coefficient with any of the methods of rescaling as

\[
 r_{yc(T)} = \frac{k (x_c)_B + (x_c)_w \frac{r_{yc}}{r_{xc}}}{\sqrt{k^2 (x_c)_B^2 + (x_c)_w^2 (C_m)_T}}
\]

For Method 3, \( k = k_3 = 1 \), so that

\[
 r_{yc(T)} = \frac{(x_c)_B + (x_c)_w \frac{r_{yc}}{r_{xc}}}{\sqrt{(x_c C_Y)}}
\]

For Method 4, \( k = k_4 = r_{xy} \), so that
These equations do not enable us to say in general which of these two validities will be the greater; all will depend on the particular situation. However, a further correlation (that between sets of assessments, one rescaled by Method 3, \( y(3) \), the other by Method 4, \( y(4) \)) helps to decide which of the two methods is the more suitable in practice. The correlation may be written

\[
r_{y(3)y(4)} = \frac{r_{xy} (x_B) + (x_C)_w r_{yc}/r_{xc}}{\sqrt{(r_{xy}^2 x_B + x_w)(x_T)}}
\]

For \( 0 < r_{xy} < 1 \), the equation shows that

\[
\sqrt{(x_w / x_T)} < r_{y(3)y(4)} < 1
\]

The minimum correlation, \( \sqrt{(x_w / x_T)} \), is bound to be high, since at the worst \( x_w \) will not differ greatly from \( x_T \) in practice. For likely values of \( r_{xy} \) the correlation between the two sets of assessments will be even higher. Thus the overall order of merit of the children on the rescaled assessments is necessarily similar whichever method is used. Hence, simplicity in use will decide the choice, which obviously falls on Method 3.

We may now complete this theoretical development by comparing the validity of assessments rescaled by this method with that of scores on the rescaling test itself. The validity of the assessments has been shown to be

\[
\sqrt{(x_C)_B + (x_C)_w r_{yc}/r_{xc}} / \sqrt{(x_C)_T}
\]

while that of the test scores is

\[
\sqrt{(x_C)_B + (x_C)_w} / \sqrt{(x_T)_T}
\]

Thus the magnitude of the ratio \( r_{yc}/r_{xc} \) alone determines the relative validities of Method 3 estimates and test scores.*

*The relative validities of rescaled assessments and test scores is the subject of further discussion in a later section (see page 251).
One final point must be made: the contribution \((XC)_B\) to the total validity of Method 3 assessments is typical of all four. This contribution, made by the re-scaling test, is independent of the original assessments. These validities must therefore be interpreted carefully; they must be attributed to the composite test-cum-assessments and not to the assessments alone. The phrase 'teachers' assessments' - so frequently used - is misleading, since it ignores the contribution of the test scores, without which the value of the teacher's original assessments would be limited.

**C - NUMERICAL ILLUSTRATION AND APPLICATION**

Data from Darlington are used to illustrate the preceding theoretical arguments. The predicting data consist of the test scores and teachers' assessments for an almost complete year-group of 935 children, divided unequally among 18 primary schools.

The test scores \((x)\) were the aggregate \(\frac{1}{2}(2VR + E + A)\), in which \(VR\), \(E\) and \(A\) are a child's standardised scores on 11+ Moray House tests of verbal reasoning, English and arithmetic respectively. Each test had been standardised separately by Lawley's method on the Darlington group itself.

Primary School teachers' assessments \((z)\), made before the 11+ examination, were rescaled (to give \(y\)) within schools by Method 3 on the aggregate above.

The criterion scores \((c)\) were the aggregate \(\frac{1}{2}(VR + E + A)\) on a 13+ Moray House battery administered to the same children after they had spent two years in various secondary schools.

Table I shows the 18 school means and s.d.'s for each of the three variables a, y and c, and the intercorrelations within schools.
TABLE I: SCHOOL MEANS, STANDARD DEVIATIONS AND INTERCORRELATIONS; VARIABLES x, y AND z

x = rescaling test;  y = rescaled assessments;  c = criterion

<table>
<thead>
<tr>
<th>School</th>
<th>n</th>
<th>(\bar{x} = \bar{y}(11+))</th>
<th>c (13+)</th>
<th>(s_x = s_y)</th>
<th>(s_c)</th>
<th>(r_{xy})</th>
<th>(r_{xc})</th>
<th>(r_{yc})</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11</td>
<td>92.27</td>
<td>93.91</td>
<td>11.63</td>
<td>11.89</td>
<td>.8878</td>
<td>.9704</td>
<td>.8583</td>
</tr>
<tr>
<td>2</td>
<td>17</td>
<td>98.41</td>
<td>97.29</td>
<td>12.22</td>
<td>13.43</td>
<td>.9344</td>
<td>.9702</td>
<td>.9474</td>
</tr>
<tr>
<td>3</td>
<td>35</td>
<td>110.06</td>
<td>111.29</td>
<td>11.72</td>
<td>12.45</td>
<td>.8078</td>
<td>.9600</td>
<td>.9172</td>
</tr>
<tr>
<td>4</td>
<td>23</td>
<td>89.78</td>
<td>92.52</td>
<td>10.82</td>
<td>10.72</td>
<td>.8856</td>
<td>.9679</td>
<td>.8671</td>
</tr>
<tr>
<td>5</td>
<td>31</td>
<td>94.84</td>
<td>93.94</td>
<td>9.01</td>
<td>9.17</td>
<td>.8505</td>
<td>.9416</td>
<td>.7888</td>
</tr>
<tr>
<td>6</td>
<td>28</td>
<td>101.14</td>
<td>103.36</td>
<td>11.90</td>
<td>11.62</td>
<td>.9255</td>
<td>.9655</td>
<td>.9084</td>
</tr>
<tr>
<td>7</td>
<td>37</td>
<td>93.68</td>
<td>92.73</td>
<td>10.44</td>
<td>10.55</td>
<td>.8735</td>
<td>.9408</td>
<td>.7729</td>
</tr>
<tr>
<td>8</td>
<td>33</td>
<td>98.03</td>
<td>98.15</td>
<td>11.61</td>
<td>12.11</td>
<td>.9165</td>
<td>.9645</td>
<td>.9109</td>
</tr>
<tr>
<td>9</td>
<td>43</td>
<td>92.74</td>
<td>94.21</td>
<td>13.21</td>
<td>12.46</td>
<td>.8890</td>
<td>.9253</td>
<td>.9300</td>
</tr>
<tr>
<td>10</td>
<td>51</td>
<td>95.88</td>
<td>96.51</td>
<td>14.04</td>
<td>15.05</td>
<td>.9020</td>
<td>.9642</td>
<td>.9022</td>
</tr>
<tr>
<td>11</td>
<td>54</td>
<td>100.02</td>
<td>101.54</td>
<td>14.02</td>
<td>13.49</td>
<td>.9402</td>
<td>.9632</td>
<td>.8745</td>
</tr>
<tr>
<td>12</td>
<td>44</td>
<td>98.20</td>
<td>96.73</td>
<td>14.05</td>
<td>14.50</td>
<td>.9527</td>
<td>.9650</td>
<td>.9357</td>
</tr>
<tr>
<td>13</td>
<td>55</td>
<td>94.51</td>
<td>97.51</td>
<td>10.41</td>
<td>11.55</td>
<td>.8966</td>
<td>.9470</td>
<td>.8897</td>
</tr>
<tr>
<td>14</td>
<td>73</td>
<td>105.58</td>
<td>107.44</td>
<td>12.74</td>
<td>12.70</td>
<td>.8765</td>
<td>.9070</td>
<td>.8875</td>
</tr>
<tr>
<td>15</td>
<td>80</td>
<td>105.60</td>
<td>106.24</td>
<td>13.76</td>
<td>13.15</td>
<td>.8901</td>
<td>.9289</td>
<td>.9174</td>
</tr>
<tr>
<td>16</td>
<td>80</td>
<td>98.78</td>
<td>98.06</td>
<td>14.65</td>
<td>13.14</td>
<td>.9433</td>
<td>.9477</td>
<td>.8761</td>
</tr>
<tr>
<td>17</td>
<td>89</td>
<td>99.08</td>
<td>100.87</td>
<td>11.92</td>
<td>13.02</td>
<td>.8825</td>
<td>.9616</td>
<td>.9485</td>
</tr>
<tr>
<td>18</td>
<td>151</td>
<td>102.22</td>
<td>102.74</td>
<td>13.44</td>
<td>13.24</td>
<td>.8968</td>
<td>.9575</td>
<td>.9114</td>
</tr>
</tbody>
</table>

Mean Within Schools - .9022 .9003

Since Method 3 rescaling was used, the school means \(\bar{y}\) and the school s.d.'s \(s_y\) are the same as those of the rescaling test, \(\bar{x}\) and \(s_x\).
Although the overall rescaling test mean is close to 100, the individual school means range from a maximum of 110.06 (school 3) to a minimum of 89.78 (school 4). The criterion means are obviously highly correlated with the test means (see Table III) and the criterion s.d.'s with the test s.d.'s \( r_{s_{c}x} = .8794 \). The 11+ predicting test and the 13+ criterion are similar, and some degree of correspondence is to be expected. It should be remembered, however, that after taking the 11+ test the 935 children were dispersed from their 18 primary schools among different secondary schools, grammar and modern. In view of the diverse experiences of the children from any one primary school during the two years between original test and final criterion, the correspondence between the means in the \( \bar{x} \) and \( \bar{c} \) columns is remarkable; so too is the correspondence between the 11+ and the 13+ s.d.'s, \( s_{x} \) and \( s_{c} \).

Equally striking are the high within-school correlations between the teachers' assessments (made before the 11+ test) and the 13+ criterion, as shown in the column headed \( r_{yc} \). It is important to note that these correlations owe nothing to rescaling. Each primary teacher's rank order at 11+ corresponds closely to that in which the aggregate \((VR + E + A)\) places the same children at 13+, despite the subsequent diversity of educational experience of the children he ranked. Furthermore, the teachers' long-term predictions of the criterion \( c \) are virtually as good as their short-term predictions of the aggregate scores \( x \).

In 3 of the 18 schools the correlation \( r_{yc} \) is higher than the correlation \( r_{xy} \), and the difference between the average within-school correlations at the foot of the table is trivial.

It was stated previously that the assessments were made before the 11+ test was administered. It must now be added that in Darlington junior tests had not preceded the 11+ test itself. Teachers were therefore not influenced, in making their assessments, by knowledge of earlier test performances from the same
children. Finally, it should be stated that the primary teachers were asked to assess their children for potential success in an academic course, and that the intentions of the 13+ testing were (i) to rectify errors made in allocation two years earlier, and (ii) to allocate children to technical courses. The possibility cannot be ruled out that in making their assessments for these children the teachers remember, and were influenced by, the correspondence between assessments and 11+ test scores for other children in previous years. They are unlikely, however, to have been similarly informed by the 13+ performances of other children which, indeed, they would not normally know; and yet their prediction of 13+ test performance is as good as that of 11+ test performance.

It would not do to assume, on this evidence alone, that the 11+ test and rescaled assessments are valid predictors of a criterion other than that used here. One would not expect them, however, to differ, in predictive validity for the more usually employed criteria, from similar data used in follow-up studies elsewhere (21).

Predictive validity for such criteria, though undeniably of over-riding importance, is not the main issue at this point. More will be said later about the conditions governing the choice of a rescaling test and the sources of error arising from the rescaling procedure. At present we need only note the necessary condition (though, in the light of what has just been said about validity, not a sufficient one): that the outcome of the rescaling procedure, in terms of rescaled assessments, should be independent of the test used for rescaling; that is, the test itself should have a high reliability coefficient of equivalence and stability. In assessing this reliability one would normally study the results of rescaling separately the same original assessments on each of a pair of tests drawn from the same range; the two resulting sets of rescaled assessments would then be compared.* In this instance we can make an

* For studies of this sort, see Chapter 11.
even more exacting comparison between assessments rescaled on an 11+ and on a 13+ battery of tests.

Take, for example, school 1. A child whose original assessment places him at the mean for his school would have, after rescaling on the 11+ battery, a rescaled assessment of 92 (to the nearest whole number). Had the 13+ test been used his rescaled assessment would have been 94. Or again, in school 17, a child one s.d. above his school mean would have, after rescaling on the 11+ test, and assessment of 111; and after rescaling on the 13+ test, one of 113.

There is no need to multiply instances; clearly the two arrays of rescaled assessments would be very highly correlated - with rescaling tests separated by two years. It is reasonable to expect an even closer correspondence with a pair of rescaling tests from the same 11+ range.

The 13+ test is now restored to its original role of criterion and analysis of variance and covariance carried out between and within primary school* groups, giving rise to the following:

(i) Sums of squares: 'between', 'within', and 'total', for x, y, and c;
(ii) Sums of products: 'between', 'within', and 'total', for x and y, x and c, y and c.
(iii) Intercorrelations: 'between', 'within', and 'total', between x and y, x and c, and y and c.

The sums of squares and sums of products are reported in Table II; the intercorrelations are shown in Table III.

* It should be noted that the 13+ data, from children dispersed over various secondary schools, have been classified, for the purposes of these analyses, in the children's original 11+ grouping by primary schools.
### TABLE II: SUMS OF SQUARES AND SUM OF PRODUCTS

<table>
<thead>
<tr>
<th>Source</th>
<th>(1) SS(x)</th>
<th>(2) SS(y)</th>
<th>(3) SS(c)</th>
<th>(4) SP(xy)</th>
<th>(5) SP(xc)</th>
<th>(6) SP(yc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Schools</td>
<td>20,334</td>
<td>20,334</td>
<td>21,450</td>
<td>20,334</td>
<td>20,223</td>
<td>20,223</td>
</tr>
<tr>
<td>Within Schools</td>
<td>150,587</td>
<td>150,587</td>
<td>150,810</td>
<td>135,862</td>
<td>135,681</td>
<td>135,681</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>170,921</td>
<td>170,921</td>
<td>172,260</td>
<td>156,196</td>
<td>162,869</td>
<td>155,904</td>
</tr>
</tbody>
</table>

### TABLE III: INTERCORRELATIONS OF x, y, AND C

<table>
<thead>
<tr>
<th>Source</th>
<th>r_{xy}</th>
<th>r_{xc}</th>
<th>r_{yc}</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Schools</td>
<td>1.0000</td>
<td>0.9683</td>
<td>0.9683</td>
</tr>
<tr>
<td>Within Schools</td>
<td>0.9022</td>
<td>0.9466</td>
<td>0.9003</td>
</tr>
<tr>
<td><strong>Total</strong></td>
<td>0.9138</td>
<td>0.9492</td>
<td>0.9086</td>
</tr>
</tbody>
</table>

In Table II, the identity of corresponding entries under (1) and (2) follows from the rescaling procedure employed (see also Table I, in which \( \bar{x} = \bar{y} \) and \( s_x = s_y \)). For the same reason, the top entries under (1), (2) and (4) are the same; and also the top entries under (6) and (5).
The intercorrelations in Table III are obtained from the entries in Table II by the usual product moment formula.

The weighted correlation between 11+ and 13+ test means after re-classifying the testees in the original primary school groups reaches the remarkable value of nearly .97. This is the numerical expression of the correspondence noted earlier between school means.

Within schools, the correlation between rescaled assessments and 13+ test scores is practically the same as that between these assessments and 11+ test scores. As might be expected, the within-school correlation between 11+ and 13+ test scores is higher still.

Conclusions from the overall ('total') correlations are similar to those from the within-school correlations. Again \( r_{yc} \) is almost as high as \( r_{xy} \), despite the interval of two years between assessment and criterion; and again, as expected, \( r_{xc} \) is even higher.

The four methods of rescaling developed previously will next be illustrated. All four employ all the entries in columns (1), (3) and (5) and the within-school entries in the other columns of Table II.

**Method 1 Rescaling**

The partial regression coefficients for Method 1 are found as on pages 135 and 138, from the data of Table II as follows:

\[
(i) \quad r_{om} = \frac{(X)B}{\sqrt{(S_B C_T)}} = 0.3417
\]

\[
(ii) \quad r_{od} = \frac{(X)W}{\sqrt{(X_W C_T)}} = 0.3857
\]
(iii) \( r_{of} = \frac{(YC)_W}{\sqrt{(X_W C_T)}} \)
\[ = 0.3424 \]

(iv) \( r_{df} = \frac{(XY)_W}{X_W} \)
\[ = 0.9022 \]

The relations on page 135 give the partial regression coefficients:

\[ b_1 = 0.3417; \quad b_2 = 0.6754; \quad b_3 = 0.2331 \]

From these b's and r's we obtain \( R_m = 0.9546 \) as the multiple correlation of the best weighted composite of m, d and f (see page 135) with the criterion.

Replacing the b's by \( w' \) to return to the original units, we obtain

\[ w_1 = 0.9945; \quad w_2 = 0.7223; \quad w_3 = 0.2493 \]

Hence \( k_1 = 1.092 \) (see page 138).

The Method 1 rescaling equation is therefore:

\[ y_1 = 1.092 \bar{x} + (z - \bar{z}) \frac{s_x}{s_x} \]

Maximum prediction (\( R_m = 0.9546 \)) is obtained when the \( y_1 \)'s thus calculated are combined with the x's in the proportion 0.2493 to 0.7223, or approximately 1 to 3.

Method 2 Rescaling

For Method 2, similar calculations (see page 139) lead to

\[ k_2 = 1.104 \]

The method 2 rescaling equation is therefore:

\[ y_2 = 1.104 \bar{x} + (z - \bar{z}) \frac{s_x}{s_y} \]

The validity of assessments rescaled by Method 2 and used alone is 0.9091.
Method 3 Rescaling

The results of rescaling by Method 3 (see page 140) are directly available from Tables II and III. For this method,

\[ k_3 = 1 \]

The Method 3 rescaling equation is therefore:

\[ y_3 = \bar{x} + \frac{(z - \bar{z}) s_x}{s_z} \]

The validity of Method 3 assessments used alone is 0.9086; this should be compared with the trivially higher validity (0.9091) from Method 2 assessments.

With a pooling square the validity of the straight sum of test scores and Method 3 assessments is found to be 0.9496, only slightly below the validity 0.9546, obtained with the best weighted composite of test scores and Method 1 assessments.

Method 4 Rescaling

For Method 4, \( k = r_{xy} \)

\[ k_4 = 0.9022 \quad (\text{see Table III}) \]

The Method 4 rescaling equation is therefore:

\[ y_4 = 0.9022 \bar{x} + \frac{(z - \bar{z}) s_x}{s_z} \]

With \( k_4 = 0.9022 \), equations (3a), (7a) and (10a) give the validity of Method 4 assessments as 0.9071, only slightly lower than the 0.9086 obtained with the Method 3 assessments.

For convenience, the final results are tabulated below.
TABLE IV: RESCALING EQUATION CONSTANTS

<table>
<thead>
<tr>
<th>Method</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
</tr>
</thead>
<tbody>
<tr>
<td>k</td>
<td>1.092</td>
<td>1.104</td>
<td>1.000</td>
<td>.902</td>
</tr>
</tbody>
</table>

TABLE V: VALIDITIES

<table>
<thead>
<tr>
<th>Predictor</th>
<th>Validity</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>0.9492</td>
</tr>
<tr>
<td>2.9x + y_1</td>
<td>0.9546</td>
</tr>
<tr>
<td>y_2</td>
<td>0.9091</td>
</tr>
<tr>
<td>y_3</td>
<td>0.9086</td>
</tr>
<tr>
<td>x + y_3</td>
<td>0.9496</td>
</tr>
<tr>
<td>y_4</td>
<td>0.9071</td>
</tr>
</tbody>
</table>

With large correlations such as these it matters little which method of rescaling is used. In each case the proportionality constant is close to 1; and the overall order of the children will be almost the same for all four methods. For example, the correlation between $y_3$ and $y_4$ assessments, calculated as on page 148, is 0.9995.

Method 3 is the customary rescaling procedure. If the evidence presented is to be trusted, Method 3 is nearly as good as the methods based on known validities.
However, while the above illustration serves its purpose of illustrating resealing procedures, the criterion employed is unusual. We must await the results of further empirical studies, carried out along similar lines, but with different criteria, before inferring that the customary method is generally comparable in efficiency with those methods in which weights derived from observed validities are utilised.
CHAPTER ELEVEN - THE RELIABILITY OF THE RESCALING PROCEDURE

A - PRELIMINARY DISCUSSION

The equation basic to all four methods of rescaling (see page 132) is

\[ y = k \bar{x} + (z - \bar{z}) \frac{s_x}{s_z} \]

In the following discussion it is assumed that Method 3 is employed, so that \( k = 1 \). As a result of the rescaling procedure the school means and within-school standard deviations for the rescaled assessments are then the same as the corresponding school means and within-school standard deviations for the rescaling test.

No test is infallible, so that the school test means, and hence the rescaled assessment means, are unreliable in some degree; and this would clearly be true even if within schools the original unscaled or 'raw' assessments were perfectly reliable, suffering only from the defect of arbitrary scale. If, as a result of imperfect test reliability, the position of a school test mean is in error by some amount, that error will be reflected in the corresponding rescaled assessment mean and will manifest itself as a constant bias affecting all the individual assessments for that school.

Similarly, if, because of imperfect test reliability, the standard deviation of the test scores in a school is in error by some amount, this error also will be reflected in the standard deviation of rescaled assessments which, for that school, will be more widely or more narrowly dispersed than they should.

Previous writers have recognised these facts, though they have tended to ignore the effects of test unreliability on the dispersion of rescaled assessments within schools, concentrating rather on the implications of test unreliability for rescaled school assessments means, particularly for small schools. The
problem presented by the small school has indeed been long regarded as the chief obstacle in the way of making use of rescaled assessments. McClelland (54) concludes a discussion of the effect of school size by saying that "no scaling system can work satisfactorily with small numbers". McIntosh et al (55) state that "the scaling methods described yield trustworthy results in general only when the number of pupils in the group is greater than sixteen, or, better still, twenty". In Yates and Pidgeon (96) we find: 'Schools offering a very small number of candidates present difficulties no matter what kind of scaling procedure is adopted'. Vernon et al attempt a quantitative evaluation of the errors involved which will be discussed later.

Yet, despite this recognition that teachers' rescaled assessments must be subject to errors because of unreliability in the rescaling test, remarkably little attempt has been made to study these errors in the light of any theoretical principles or to estimate their magnitude empirically. The discussion in Vernon et al, referred to above, appears to be the only reported attempted at such a study and, for reasons to be presented shortly, it seems basically unsound.

In the theoretical and empirical studies which follow, the errors arising out of test unreliability will be investigated from a different point of view which seems preferable on all counts.

B = DEFINITIONS

If a single set of original or 'raw' assessments for a number of children in different schools is rescaled, first on one test administered to all children, and again ab initio on an approximately parallel and equivalently standardised test, then two sets of rescaled assessments result. The overall reliability of the rescaling procedure is the correlation between the two sets of rescaled assessments;
and the mean standard error of an individual rescaled assessment is \( \frac{1}{\sqrt{2}} \) times the standard deviation, over all children, of the differences within pairs of rescaled assessments for individual children.

These definitions, which seem not only reasonable, but also the only ones possible, are closely analogous to the corresponding definitions of test reliability and test error of measurement, well-known in psychometric measurement. As will be shown, they make possible useful comparisons between, on the one hand, the overall reliability and mean standard error of rescaled assessments, and, on the other, the overall reliability and mean standard error of measurement of the test scores on which the rescaling was based.

To avoid the possibility of misunderstanding, it should be clearly understood that the standard error of a rescaled assessment is a standard error of measurement, and not a standard error of prediction. Like all other standard errors of measurement, it relates fundamentally to the dispersion of a large number of fallible measures of some characteristic of a single individual about the 'true' measure, the mean of them all. In this case the fallible measures are the 'scores' obtained on repeatedly rescaling the same original or 'raw' assessments on a large number of tests, so far as possible parallel in content and equivalently standardised. It will be noted that the definition initially stated is operational in that it makes use of the customary method of circumventing, in psychometric measurement, the practical difficulties of implementing the more fundamental definition which is in terms of large numbers of fallible measures and one 'true' measure for a single individual.

One further possible source of misunderstanding should be mentioned. We are concerned here with the reliability of the rescaling procedure, and not with that of the original assessments, which we are accepting as they stand. Whether they are reliable or not is irrelevant to the present discussion.
Vernon et al begin with this question ((91)page 139):

"If the children from a typical school score 100* on the first test or battery, and the averaged teachers' estimate is converted to this figure, how widely might the mean score vary on another test?"

These writers clearly see that the essence of reliability estimation is replication. Nevertheless, their approach to the problem they set themselves is both defective and incomplete. It is defective in that they make use of a statistical technique inappropriate in the context of their problem. It is incomplete in that they confine their attention to a single school in a situation where a study of the variability of several schools is essential.

In answer to their question Vernon et al employ the standard error of prediction, \( s\sqrt{(1 - r^2)} \), in which \( s \) is the standard deviation of the \( m \) test scores for the school concerned, and \( r \) the correlation, also for that school, between scores on the test which was used for rescaling and scores on the test which might have been used in its stead. The standard error of the mean on the second test, predicted from the mean on the first, is given as \( s\sqrt{(1 - r^2)} / \sqrt{m} \). Thus score on the test actually used is treated as an independent variable. In the numerical example the writers give to illustrate the method (ibid. page 190), they take the observed mean for both tests to be 100, the general mean over all scores. By doing so they ignore the regression towards the mean on the second test which, with their approach, they would need to take into account with a school whose mean on the first test was not 100. For example, with a school having a first test mean of 110, and with a correlation between tests of .9, the mean on the

* Presumably 'score an average of 100' is meant.
second test estimated in this way would be 109, and the standard error of this estimate $6.54 / \sqrt{m}$.

Even were the omission remedied, no confidence could be placed in estimates of standard error arrived at in this way. The whole approach is erroneous. Firstly, it wrongly assumes that the problem is one of prediction, assigning to the test actually administered the role of independent variable and to the test that might have been used instead that of dependent variable. In fact, however, the tests are equivalent as to purpose (rescaling) and hence equal in status; it is fortuitous that the one, and not the other, was chosen. Secondly, they are both fallible, and, since they are approximately parallel and equivalently standardised tests, approximately equally fallible. Thirdly, even were the tests equivalently standardised overall, so that for the whole group the mean, standard deviation, and constants of shape are the same for both tests, there is no guarantee whatever that for the schools which together make up the whole group equivalence would exist in these respects.

D - DIFFERENTIAL ZERO ERROR

As a simple illustration of this third point, using means only, consider two schools, each presenting the same number of candidates:

<table>
<thead>
<tr>
<th>School</th>
<th>Test X</th>
<th>Test Y</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>101</td>
<td>98</td>
</tr>
<tr>
<td>2</td>
<td>99</td>
<td>102</td>
</tr>
<tr>
<td>Overall</td>
<td>100</td>
<td>100</td>
</tr>
</tbody>
</table>
The schools have equal and opposite 'zero errors' which together produce a nil 'zero error' for the two schools taken together.

It is neglect of the possibility of differential zero error which mainly vitiates the approach of Vernon et al.

If, with a nil 'zero error' overall (or after making allowance for the observed 'zero error' overall), there remain differential school 'zero errors', (that is, a school-test interaction), the magnitude of this differential effect or interaction will determine the reliability of the school means on a test used for rescaling, and hence affect that of the rescaled assessments.

Differential school zero errors are a priori not unlikely. At best, 'parallel' tests are only approximately parallel, and the content of the first test may favour some schools, and that of the second, others. If there is an interval between the administrations of the two tests, then what happens in the schools during that interval may affect the school means differentially. For reasons such as illness or anxiety, some children may depress their school means by doing unduly badly on one occasion of testing, and not on the other.

Clearly, if a school-test interaction does exist, it implies, for the rescaled assessments, a bias in favour of all children in each of some schools, and against all children in each of others. Further, such biases are independent of, and superimposed on, the unbiased between-test variation (unreliability) for individual children within schools. It follows that knowledge of the latter alone affords no clue to the magnitude or even to the existence of differential school 'zero errors'. It also follows that data from the administration of a single test in a single school cannot possibly provide sufficient information.

Since this is the procedure followed by Vernon et al, their results are necessarily invalid.

We are forced to the conclusion that the only way in which differential

* See Chapter 8, page 102.
school zero means can be detected and their magnitude assessed is to study the results of administering two tests (at least) to children in each of a number of schools. The remainder of this chapter is devoted to such a study.

E - PLAN OF THIS STUDY

The plan in making this study is twofold. We shall first examine theoretically the conditions determining the reliability of the rescaling procedure, that reliability being defined as the correlation between assessments rescaled on one test and the same assessments rescaled on another. In this theoretical study, the reliability of the rescaling procedure will be compared with that of the tests used in rescaling. The second part of the plan is to design experimental procedures (suggested by the results of the theoretical study) and to use them with numerical data in order to estimate practically the reliability of the rescaling procedure.

F - THEORETICAL STUDY OF THE RELIABILITY OF A RESCALING PROCEDURE

In this theoretical study, the implications are first examined of inter-school heterogeneity for the correlation between tests X and Y, standardised overall to the same mean and standard deviation. This first correlation we take to be the reliability of either test (the kind of reliability is irrelevant). We then study the implications of inter-school heterogeneity for the correlation between assessments rescaled on test X and the same assessments rescaled on test Y (henceforth called X- and Y-assessments), in both cases using the linear transformation of Method 3. This second correlation is taken to be the reliability of the rescaling procedure. Finally the reliability of the tests is compared with that of the rescaling procedure.
F.I - RELIABILITY OF TEST SCORES

\( \bar{x} \) and \( \bar{y} \), \( s_x \) and \( s_y \), are the means and s.d.'s for a particular school of m children on tests X and Y, which, for all scores over all schools have the same mean (zero) and the same s.d. A particular child in this one school obtains scores \( x \) and \( y \) on the tests. Then:

\[
\begin{align*}
    x &= \bar{x} + (x - \bar{x}) \\
    y &= \bar{y} + (y - \bar{y})
\end{align*}
\]

For simplicity we shall suppose \( m \) constant over all schools.

Then, if \( r_{xy} \) is the correlation between X and Y scores in a particular school ('within-school' reliability), it is easy to show that the 'Total' correlation between X and Y scores (overall test reliability) is

\[
F_{xy(T)} = \frac{m}{m-1} \frac{\Sigma xy}{\sqrt{\left( \frac{m}{m-1} \Sigma x^2 + \Sigma s_x^2 \right) \left( \frac{m}{m-1} \Sigma y^2 + \Sigma s_y^2 \right)}}
\]

F.II - RELIABILITY OF RESCALED ASSESSMENTS

\( z \) is the 'raw' assessment of the same child as in F.I and \( s_z \) the s.d. of 'raw' assessments in his school. \( a \) and \( b \) are his X- and Y-assessments, both derived from \( z \). Then:

\[
\begin{align*}
    a &= \bar{x} + \frac{s_x}{s_z} (z - \bar{z}) \\
    b &= \bar{y} + \frac{s_y}{s_z} (z - \bar{z})
\end{align*}
\]

and it is easy to show that the 'Total' correlation \( r_{ab(T)} \) between X- and Y-assessments (overall reliability of the rescaling procedure) is
\[ r_{ab}(T) = \frac{\frac{m}{m-1} \Sigma x y + \Sigma s_x s_y}{\sqrt{\left(\frac{m}{m-1} \Sigma x^2 + \Sigma s_x^2 \right) \left(\frac{m}{m-1} \Sigma y^2 + \Sigma s_y^2 \right)}} \]  \quad (2)

Much can be learnt from equations (1) and (2).

Firstly, the reliability of the rescaling procedure is obviously higher than that of the tests used in rescaling (except in the unlikely event that the test scores within schools are perfectly correlated).

Secondly, the 'between schools' reliability \( r_{ab(B)} \) is the same for both test scores and rescaled assessments:­

\[ r_{ab(B)} = \Sigma xy / \sqrt{(\Sigma x^2 \Sigma y^2)} \]  \quad (3)

Equation (3) gives the correlation between school X- and Y-means, which is one of the fundamental determiners of the rescaling procedure's reliability.

Thirdly, the 'within-school' reliabilities for test scores, \( r_{xy(W)} \), and assessments, \( r_{ab(W)} \), based on these scores, are different. For the scores it is

\[ r_{xy(W)} = \Sigma r_{xy} s_x s_y / \sqrt{(\Sigma s_x^2 \Sigma s_y^2)} \]  \quad (4)

and for the estimates:

\[ r_{ab(W)} = \Sigma s_x s_y / \sqrt{(\Sigma s_x^2 \Sigma s_y^2)} \]  \quad (5)

the second correlation being higher than the first.

Equation (5) shows also that although for each school individually the correlation between X- and Y- assessments must be unity for whatever the values of \( s_x \) and \( s_y \) (an obvious but apparently neglected fact), nevertheless \( r_{ab(W)} \) cannot be unity unless for each school \( s_x = s_y \).
To sum up:

(i) Equation (2) is based on the hypothesis that tests standardised to have the same mean and s.d. over a number of schools do not necessarily have the same means and s.d.'s for individual schools.

(ii) The reliability of a rescaling procedure depends on the extent to which the school means and within-school standard deviations alter with change of rescaling test. Equation (2) shows exactly how this comes about.

(iii) The reliability of a rescaling procedure is higher than that of the test on which it is based.

It remains to design experiments to study whether changes in school means and school standard deviations occur if the rescaling test is changed, and so test the above hypothesis.

G - EXPERIMENTAL DESIGN I

G.I - THE DATA

This design is used to study the changes which occur in school rescaled means and standard deviations when first one, and then the other, of two tests, both administered to the same children, is used as the rescaling instrument; and to study also the implications of these changes for the reliability of the rescaling procedure.

The basic requirements are simple: scores on each of two tests administered to the same children from several schools, and the teachers' original or 'raw' assessments for the same children.
The data chosen for the study are a sample drawn from the scores of the complete year-group of eleven year old children tested in Darlington in 1958. In this year, Darlington used two batteries, each of verbal reasoning, English and arithmetic, with an interval of 40 days between the administrations. Each of the six tests was standardised by Lawley’s method on the Darlington year group itself, so that the mean and standard deviation imposed on the year-group for each test were 100 and 15 respectively, with all distributions normal. It is therefore reasonable to conclude that for the year-group the battery totals on three tests are also normally distributed about a mean of 300 with a standard deviation depending on the intercorrelation of scores on the three tests comprising the battery, probably about 42.

Darlington makes use also of teachers’ assessments rescaled on the first of the batteries.

For the purposes of this pilot study, a sample of six schools was randomly selected, and within each of these schools, 20 children, also chosen at random. For each child there are two battery totals and two rescaled assessments derived from his teacher’s original or ‘raw’ assessments by rescaling first on one battery total, then on the other. Rescaling was done by Method 3, using the battery mean and s.d.’s for the samples of 20 children henceforth referred to as ‘schools’. The two battery totals and the two rescaled assessments for each child are the units employed in the study.

G.II - SCHOOL MEANS AND STANDARD DEVIATIONS

The two batteries will henceforth be termed ‘tests’ X and Y, and the child totals on the ‘tests’ will be called ‘x- and y- scores’.
Table I shows the six school means (each over 20 children) on each of the tests X and Y and the within-school mean differences. Table II shows the six school standard deviations for tests and the within-school differences. These means and standard deviations are crucial to the rescaling procedure by Method 3.

**TABLE I - SCHOOL MEANS AND WITHIN-SCHOOL MEAN DIFFERENCES ON EACH OF TWO TESTS X AND Y**

<table>
<thead>
<tr>
<th>School</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test X</td>
<td>310.95</td>
<td>308.10</td>
<td>294.85</td>
<td>325.95</td>
<td>278.10</td>
<td>304.45</td>
<td>303.73</td>
</tr>
<tr>
<td>Test Y</td>
<td>306.50</td>
<td>304.55</td>
<td>297.10</td>
<td>331.90</td>
<td>280.75</td>
<td>299.70</td>
<td>303.42</td>
</tr>
<tr>
<td>Diffs. (Y-X)</td>
<td>-4.45</td>
<td>-3.55</td>
<td>+2.25</td>
<td>+5.95</td>
<td>+2.65</td>
<td>-4.75</td>
<td>-0.31</td>
</tr>
</tbody>
</table>

**TABLE II - WITHIN-SCHOOL STANDARD DEVIATIONS ON TWO TESTS AND WITHIN-SCHOOL DIFFERENCES**

<table>
<thead>
<tr>
<th>School</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>Test X</td>
<td>38.36</td>
<td>49.63</td>
<td>41.89</td>
<td>38.40</td>
<td>31.46</td>
<td>51.18</td>
<td>41.82</td>
</tr>
<tr>
<td>Test Y</td>
<td>36.99</td>
<td>45.73</td>
<td>43.12</td>
<td>43.40</td>
<td>29.68</td>
<td>48.32</td>
<td>41.21</td>
</tr>
<tr>
<td>Diffs. (Y-X)</td>
<td>-1.37</td>
<td>-3.91</td>
<td>+1.23</td>
<td>+5.00</td>
<td>-1.78</td>
<td>-2.84</td>
<td>-0.61</td>
</tr>
</tbody>
</table>
Two points are at once apparent from the figures in Table I under 1-6. The first is the quite high correlation between corresponding means in the X- and Y-arrays (rank correlation = 1; product moment correlation = .9626). The second point is the presence, despite this high correlation, of appreciable school zero errors ranging from -4.75 to +5.95.

The third point, apparent from the entries under 'All', is the close correspondence between the general means over all schools. For this sample, as for the parent population, the overall 'zero' error is negligible.

The figures in Table I strikingly resemble in pattern the fictitious figures used to illustrate, on page 164, the possible co-existence of school zero errors and an overall nil zero error. If statistical test establishes significant inter-school heterogeneity, we must accept its implications, one of which is that the argument used by Vernon et al. is invalid.

In Table II the agreement between corresponding s.d.'s is less pronounced than for means (ρ = .943; r = .901; schools 3 and 4 have exchanged ranks). Differential errors exist for s.d.'s no less than for means. Thus for school 4, test Y's s.d. is higher by 5 points than test X's, while for school 2 test Y's s.d. is nearly 4 points lower than test X's.

G.III - STUDY OF THE DATA USING ANALYSES OF VARIANCE

G.III.(1) - MODEL FOR TEST SCORES

The basic model for the analysis of test scores is:

\[ x_{ijk} = M + s_k + t_j + st_{kj} + c_{ik} + ct_{ijk} \]  

(6)
In this equation:

\[ x_{ijk} \] is the score of child \( i \) (\( i = 1, 2, \ldots, m \)) in school \( k \) (\( k = 1, 2, \ldots, l \)) on test \( j \) (\( j = 1, 2, \ldots, n \)). Note that \( i \) is 'nested' in \( k \), that is, child \( ik \) is a member of school \( k \) only.

\( M \) is a component common to all children; without loss of generality it may be put equal to zero and henceforth ignored.

\( s_k \) is a component common to all children in school \( k \).

\( t_j \) is a component specific to test \( j \).

\( s_{kj} \) is a component resulting from interaction between school \( k \) and test \( j \); a measure of the extent to which the school total or means depends on the identity of the test.

\( c_{ik} \) is a component specific to child \( i \) in school \( k \).

\( c_{tijk} \) is a component resulting from interaction between child \( i \) in school \( k \) and test \( j \); a measure of the extent to which the child's score depends on the identity of the test.

The usual assumptions are made about the independence of the components and their zero expectations.

Note that while estimates of the variances relating to \( c_{ik} \) and \( c_{tijk} \) may be obtained using data from a single school, estimation of the variance of the other components in the above equation requires data from more than one school. Note further, should the variance of \( s_{kj} \) in particular turn out to be significant, the impossibility of evaluating, from data for a single school, the effect on a
school mean of changing the test. The variance of $\sigma_{k,j}$, if significant, is that portion of the differential school zero error that cannot be ascribed to the within-school variance of $\sigma_{ij,k}$.

**C.III.(2) - ANALYSIS BASED ON THE MODEL**

An analysis of variance of the 120 test scores in six schools based on this model is presented below in Table III.

**TABLE III - ANALYSIS OF VARIANCE OF TEST SCORES ON TWO TESTS X AND Y**

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Schools</td>
<td>5</td>
<td>52,797.600</td>
<td>10,559.520</td>
<td>$\bar{S}/ST = 51.837 (5,5)$</td>
</tr>
<tr>
<td>(2) Tests</td>
<td>1</td>
<td>6.017</td>
<td>6.017 (T)</td>
<td></td>
</tr>
<tr>
<td>(3) Schools x Tests</td>
<td>5</td>
<td>1,018.533</td>
<td>203.707 (ST)</td>
<td>$ST/CT_w = 6.758 (5,114)$</td>
</tr>
<tr>
<td>(4) Children Within Schools</td>
<td>114</td>
<td>399,268.050</td>
<td>3,502.351 (CT)</td>
<td>$CT_w/CT_w = 116.187 (114,114)$</td>
</tr>
<tr>
<td>(5) Children x Tests Within Schools</td>
<td>114</td>
<td>3,436.450</td>
<td>30.144 (CT)</td>
<td></td>
</tr>
<tr>
<td>(6) Total</td>
<td>239</td>
<td>456,526.650</td>
<td></td>
<td>-</td>
</tr>
</tbody>
</table>

The expectations of the population variance components derived from the equations below are useful in deciding on the appropriate significance tests, and essential for other purposes. In writing these expectations the notation is similar to that used in the Introductory Chapter (see page 19).
The expectations are:

\[
E(\bar{S}) = \sigma^2_{ct.s} + n\sigma^2_{c.s} + m\sigma^2_{st} + mn\sigma^2_s
\]

\[
E(\bar{T}) = \sigma^2_{ct.s} + m\sigma^2_{st} + lm\sigma^2_t
\]

\[
E(\bar{ST}) = \sigma^2_{ct.s} + m\sigma^2_{st}
\]

\[
E(\bar{w}) = \sigma^2_{ct.s} + n\sigma^2_{c.s}
\]

\[
E(\bar{T}_w) = \sigma^2_{ct.s}
\]

From the mean squares in Table III, putting \(l = 6\), \(m = 20\), \(n = 2\), and substituting \(s\) for \(\sigma\), we obtain the following estimates of the variance components above:

\[
s^2_{ct.s} = 30.144; \quad s^2_{c.s} = 1736.104; \quad s^2_{st} = 8.678; \quad s^2_t = -1.617(?)
\]

and \(s^2_s = 172.090\).

Since the tests were both standardised on the year-group to the same mean, the expectation of the mean square for tests is zero. The small negative value is fortuitous.

The interpretation of the mean squares in Table II are as follows:

In line (1) of this table the comparison is among the six school totals (or means) over tests X and Y together. In line (2) it is between test totals (or means) over all six schools together. In line (3) the S \( \times \) T interaction mean square is a measure of the variance of the difference between school means on the two tests, that is, of the differential zero error. This is apparent on noting that
* \[ \overline{ST} = \frac{1}{5} \sum (\frac{\Sigma x - \Sigma y}{2 \times 20})^2 \]

where \(\Sigma x\) and \(\Sigma y\) are the two test totals for the same school. From this expression it is obvious that a significantly large \(\overline{ST}\) reflects important discrepancies in both directions between these totals (and hence means), that is, differential school zero error. Line (4) provides a comparison, within schools, of child totals (or means) over tests X and Y together; and line (5) is a measure (also within schools) of the variation of differences between child scores on tests X and Y, and thus of the reliability of measurement for individual children within a typical school. The latter is readily seen on noting that

\[ \overline{CT} = \frac{1}{14} \cdot \frac{1}{2} \sum (x - y)^2 \]

where \(x\) and \(y\) are the two scores for one child.

The appropriate error terms in testing the mean squares for significance must next be identified. It is assumed that 'Schools' and 'Children within Schools' are fixed effects, and 'Tests' random (see page 17). The application of Schultz's rule (see page 19) therefore leads to the F-tests indicated in Table III (page 174). All the mean-squares tested in that table are significant beyond the 1 per cent level.

The significant mean square for 'Schools' shows that the six school means differ significantly among themselves when the variate is summed scores (or the corresponding means) over two tests. These means, derived from the six pairs of means in Table I, are listed below:

<table>
<thead>
<tr>
<th>School</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>All</th>
</tr>
</thead>
<tbody>
<tr>
<td>Mean ((X + Y))</td>
<td>308.72</td>
<td>306.33</td>
<td>296.02</td>
<td>328.92</td>
<td>279.42</td>
<td>302.08</td>
<td>303.58</td>
</tr>
</tbody>
</table>

* *' denotes summation of deviations from the mean.
Such results are quite usual, though the magnitudes of the between-schools differences may cause some surprise. If the assumptions underlying the rescaling procedure are sound, the size of these differences — which reflect fairly faithfully the differences among the means in either the X-line or the Y-line in Table I — points to the necessity of the procedure.

The trivial mean square for 'Tests' has been discussed already. It corresponds to the mean difference of about 0.3 points listed under 'All' in Table I.

The significant 'School x Test' interaction is perhaps the most important result of the analysis. It establishes the presence of zero errors for individual schools despite the approximately nil zero error overall (see bottom line of Table I). For some schools the X-means are significantly larger than the Y-means; for others the reverse is the case. The consequences of a significant $ST$, both for test scores and for rescaled assessment, will be discussed later.

The mean-squares for 'Children within Schools' and for 'Children x Tests within Schools' are still to be interpreted. Each is the mean of six mean-squares derived from six two-way analyses of variance, one for each school. In each analysis the sources of variation are 'Tests', 'Children', and 'Children x Tests', and pooling the degrees of freedom and the sums of squares across all schools for the second and third sources mentioned leads to the mean-squares now discussed. Thus the significant $Cw$ indicates that the children in an 'average' school differ significantly among themselves when the variate is the sum of scores on tests X and Y. $CTw$ is interpreted as the mean residual error variance of a single X- or Y-score.*

---

* The composite nature of terms such as $CT_w$ has already been discussed (see page 26).
G.III.(3) - CORRELATIONS BETWEEN TEST SCORES

The correlation between school means for tests X and Y is readily derived from lines (1) and (3) in Table III. It is:

\[ r_{xy} = \frac{\bar{S} - \bar{ST}}{\bar{S} + \bar{ST}} \]  

\[ = .9621, \]

a result differing but slightly from the .9626 obtained by direct product-moment calculation and reported on page 172.

The mean within-school correlation between individual test scores is derived similarly from lines (4) and (5):

\[ r_{xy} \text{ (within schools)} = \frac{\bar{C}_w - \bar{CT}_w}{\bar{C}_w + \bar{CT}_w} \]

\[ = .9829. \]

G.III.(4) - RESCALED ASSESSMENTS - ANALYSIS OF VARIANCE

We turn now to the corresponding analysis of variance for rescaled assessments. The reader is reminded that there is one set of 120 original or 'raw' assessments, 20 from each of six schools, rescaled both on test X and test Y. These rescaled assessments will henceforth be called X- and Y-assessments. It should be clear that there is one X- and one Y-assessment for each child, and that the pattern of the data is exactly the same as for test scores. The basic model is therefore parallel to that for test scores (page 172).

It would be possible, though laborious, to analyse the 240 rescaled assessments by the same computational procedure as that used for test scores.
However, it is simpler and more illuminating to derive the analysis for rescaled assessments directly, line by line, from that for test scores (see Table III) which formally it resembles.

First note that the degrees of freedom in this new analysis must be the same as in the previous one. Next, since in rescaling the X- and Y-assessments school means and s.d.'s are equated to the X- and Y-score school means and s.d.'s respectively, it follows that the new analysis must have lines (1), (2), (3) and (6) identical with these lines in Table III.

Lines (4) and (5) are computed by re-partitioning their pooled sums of squares. The rationale is as follows.

In line (5) the sum of squares for children X tests within schools is easily shown to be

\[ CT_w = \frac{1}{2} \sum (x - y)^2 * \]

in which \( \sum' \) denotes summation with correction for mean over individuals within a school and \( \Sigma \) further summation over all schools. Expanding, and noting there are \( m \) children per school we obtain

\[ CT_w = \frac{m-1}{2} \sum (s_x^2 + s_y^2 - 2r_{xy} s_x s_y) \]

in which \( s_x^2 \) and \( s_y^2 \) are the variances and \( r_{xy} \) the correlation for X- and Y-scores in a particular school.

Since for any one school the X- and Y-assessments are merely linear transformations of the same original 'raw' assessments, the correlation between X- and Y-assessments is necessarily unity. Further, because of the rescaling procedure, \( s_x^2 \) and \( s_y^2 \) are the variances, not only for test scores, but also for

* See, for example, Jackson (41).
rescaled assessments. Hence to obtain the sum of squares (which will be denoted by \( CA_w \)) for children \( x \) assessments within schools, we have merely to put \( r_{xy} = 1 \) in the equation (9) for \( CT \). We obtain

\[
CA_w = \frac{m - 1}{2} \sum (s_x - s_y)^2
\]  

(10)

The differences \((s_x - s_y)\) are those already reported in the 'Difference' line of Table II (page 171). Substituting numerical values, we obtain

\[
CA_w = \frac{19}{2} (54.9802)
\]

\[
= 522.312
\]

This is the sum of squares in line (5), and line (4) is obtained by difference.

To sum up: by taking advantage of the forced equality of school means and s.d.'s for test scores and rescaled assessments; and by making use of the necessarily perfect within-school correlation between \( X \)- and \( Y \)-assessments; we are able to infer from the earlier analysis of test scores reported in Table III all the details of the new analysis of rescaled assessments. The new analysis is the same as that which would have been obtained directly from the rescaled assessments. However, this short cut not only avoids the necessity for rescaling, but also highlights the effect of the necessarily perfect correlation within schools between \( X \)- and \( Y \)-assessments. The new analysis is presented in Table IV.
TABLE IV - ANALYSIS OF VARIANCE OF ASSESSMENTS
AFTER RESCALING ON TEST X AND ON TEST Y

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Schools</td>
<td>5</td>
<td>52,797.600</td>
<td>10,559.520</td>
<td>(S)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>(S)</td>
<td></td>
<td>$\overline{S}/\overline{SA} = 51.837$ (5,5)</td>
</tr>
<tr>
<td>(2) Asst. X v Asst. Y</td>
<td>1</td>
<td>6.017</td>
<td>6.017</td>
<td>(A)</td>
</tr>
<tr>
<td>(3) $S \times A$</td>
<td>5</td>
<td>1.018.533</td>
<td>203.707</td>
<td>$\overline{SA}/\overline{CA_w} = 44.458$ (5,174)</td>
</tr>
<tr>
<td>Children within Schools</td>
<td>114</td>
<td>402,182.188</td>
<td>3,527.914</td>
<td>$\overline{C_w}/\overline{CA_w} = 769.951$ (114,114)</td>
</tr>
<tr>
<td>(5) $C \times A$ Within Schools</td>
<td>114</td>
<td>522.312</td>
<td>4.582</td>
<td>$\overline{CA_w}$</td>
</tr>
<tr>
<td>(6) Total</td>
<td>239</td>
<td>4,565,526.650</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

This table gives final expression to the earlier arguments. The identity of lines (1), (2), (3) and (6) in Tables III and IV displays the outcome of the rescaling procedure which equates school means and within-school standard deviations for the rescaled assessments to those for the test scores.

The mean-square $\overline{CA_w}$ in line (5) of Table IV is small; indeed, were it not for the inequalities in the within-school s.d.'s reported in Table II it would be zero.

For tests of significance, the pattern is the same as that of Table III; and the only estimate of a population variance component from Table IV needed later is an estimate of $\sigma^2_{ca,s}$, the within-school interaction variance between children and assessments; that is,

$$s^2_{ca,s} = \overline{CA_w}$$

$$= 4.582.$$
G.III.(5) - CORRELATIONS BETWEEN RESCALED ASSESSMENTS

In the same way as for the corresponding correlations from Table III, we obtain from Table IV,

(i) the correlation between school means for X- and Y-assessments:

\[ r_a(x) a(y) = \frac{\bar{S} - \bar{SA}}{\bar{S} + \bar{SA}} \]

\[ = .9621 \]

the same result, of course, as for school test means (see equation (7), page 178) and

(ii) the mean within-school correlation between individual X- and Y-assessments:

\[ r^w_a(x) a(y) = \frac{\bar{C} - \bar{CA}}{\bar{C} + \bar{CA}} \]

\[ = .9974, \]

much higher than the corresponding correlations for X- and Y-scores (see equation (8), page 178), failing, indeed, to reach unity only because of differences in s.d. between tests within schools.

G.III.(6) - FURTHER CORRELATIONS

(i) Test Scores

The correlation between the whole set of 120 X-scores and the whole set of Y-scores, ignoring the division by schools, is the reliability of either test as normally reported. This correlation is readily derived from an analysis of variance (Table V) obtained by pooling lines (1) and (4), and lines (3) and (5),
in Table III.* The subscript 'w', attached to \( C_w \) and \( CT_w \), and \( \overline{C}_w \) and \( \overline{CT}_w \) in the previous analysis, may now be dropped.

### Table V - Analysis of Variance of Test Scores Ignoring Division by Schools

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Tests X and Y</td>
<td>1</td>
<td>6.017</td>
<td>6.017</td>
<td>( (F) )</td>
</tr>
<tr>
<td>Children</td>
<td>119</td>
<td>452,065.650</td>
<td>3,798.871</td>
<td>( 101.474 ) (119,119)</td>
</tr>
<tr>
<td>Residual ((C \times T))</td>
<td>119</td>
<td>4,454.983</td>
<td>37.437</td>
<td>( CT )</td>
</tr>
<tr>
<td>Total</td>
<td>239</td>
<td>456,526.650</td>
<td>-</td>
<td>-</td>
</tr>
</tbody>
</table>

The reliability of either test, as normally reported, is given by

\[
 r_{xy} \text{ (overall)} = \frac{\overline{C} - \overline{CT}}{\overline{C} + \overline{CT}} \quad \ldots \quad (13)
\]

\[
 = .9805.
\]

Also,

(i) Mean standard error of a test** score = \( \sqrt{\overline{CT}} \)

\[
 = 6.119.
\]

(ii) Average test standard deviation** over the whole group

\[
 = \sqrt{\frac{1}{2}(\overline{C} + \overline{CT})}
\]

\[
 = 43.80.*
\]

* This procedure is similar to that used in the Introductory Chapter (see page 31).

** The reader is reminded that a 'test' is a battery consisting of \( (VR + E + A) \).
Thus the mean standard error of measurement by either of the tests, as usually reported, is \( \frac{1}{7} \) th of the test standard deviation, or about \( \frac{1}{40} \) th of the probable range of test scores. This degree of precision is high in psychometric measurement.

(ii) Rescaled Assessments

In the same way, from Table IV, we obtain the analysis of variance presented in Table VI, for rescaled assessments.

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>X- and Y-assessments</td>
<td>1</td>
<td>6.017</td>
<td>6.017</td>
<td></td>
</tr>
<tr>
<td>Children</td>
<td>119</td>
<td>454,979.788</td>
<td>3,823.360</td>
<td>295.286 (119,119)</td>
</tr>
<tr>
<td>Residual (C x A)</td>
<td>119</td>
<td>1,540.845</td>
<td>12.948</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>239</td>
<td>456,526.650</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Using the same methods as for tests (see page 183), we obtain:-

Overall correlation between rescaled assessments

\[ r(X\text{-asst.})(Y\text{-asst.}) \text{ (overall)} = .9933. \]

This is the reliability of the rescaling procedure as it would normally be reported. Finally,

Mean standard error of a rescaled assessment

\[ = 3.593, \]

and average standard deviation over all rescaled assessments

\[ = 43.80, \]

the same as for the tests - as it should be.
### G.III.(7) - COMPARISON OF RELIABILITIES AND STANDARD ERRORS FOR TESTS AND RESCALED ASSESSMENTS

For convenience, the correlations, standard deviations and errors of measurement are listed in Table VII below.

#### TABLE VII - CORRELATIONS, STANDARD DEVIATIONS, AND ERROR VARIANCES

<table>
<thead>
<tr>
<th></th>
<th>Tests</th>
<th>Rescaled Assessments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Correlations between school means</td>
<td>.9621</td>
<td>.9621</td>
</tr>
<tr>
<td>Within-school correlations</td>
<td>.9829</td>
<td>.9974</td>
</tr>
<tr>
<td>Overall correlations as normally reported</td>
<td>.9805</td>
<td>.9933</td>
</tr>
<tr>
<td>Mean s.d. overall</td>
<td>43.80</td>
<td>43.80</td>
</tr>
<tr>
<td>Mean s.e. of measurement as normally reported</td>
<td>6.119</td>
<td>3.593</td>
</tr>
</tbody>
</table>

(i) The correlations between means are necessarily the same in the two columns since the school means for tests and rescaled assessments are the same.

(ii) It was shown on page 168 that the reliability of rescaled assessments is necessarily higher than that of the test on which it is based. The empirical results, both within-schools and overall, and the s.e.'s of measurement, demonstrate the magnitude of the anticipated difference for these data.

Care must be taken, however, in interpreting these results. The test reliability and corresponding standard error of measurement are based on the
correlation between two different, though approximately parallel tests. The reliability of the resealed assessments, and the corresponding standard error of measurement, are based on the correlation between two sets of assessments obtained by twice rescaling the same original assessments.

It should be borne in mind constantly that this study is concerned with the reliability of the rescaling procedure, and not with that of the original or 'raw' assessments to which this procedure is applied. Had we started with two different sets of original assessments, and rescaled one set on X and the other on Y, the final correlations would certainly have been lower.

The correlations reported above, and the corresponding standard error, conclude one phase of this study. On pages 161 and 162 are defined the reliability of the rescaled procedure. The overall correlation .9933 between rescaled assessments, reported in Table VII, corresponds exactly to that definition. It is identical with the result obtained by correlating the two arrays, each of 120 assessments by the product-moment method. Moreover, the corresponding standard error of measurement, 3.593, is the same as that obtained by using the customary formula:

\[ s.e. (\text{meas.}) = \sqrt{\frac{s^2}{(1 - r(X-asst.)(Y-asst.))} \cdot (114) = 43.8(1 - .9933) = 3.593. \]

**G.III.(8) - DIFFICULTIES OF INTERPRETATION OF OVERALL RELIABILITY AS NORMALLY REPORTED**

There remain, however, unresolved difficulties which will be examined first in relation to the tests.
(a) **TESTS - THE CONVENTIONAL MODEL**

The simple assumptions underlying the conventional development of the standard error of measurement (see, for example (33)) are such that the standard error so derived is equally appropriate in comparing either (i) scores obtained by different children on the same test, or (ii) scores obtained by the same child on different (parallel) tests; that is, homogeneity of error variance is assumed, down the array of scores for different children, and along the (hypothetical) array of scores for a single child. This may be readily demonstrated as follows.

The conventional development assumes a model for the score of the $i^{\text{th}}$ child on the $j^{\text{th}}$ test of the form:

$$x_{ij} = M + c_i + t_j + ct_{ij}$$

(15)

in which $M =$ component common to all observations;

$c_i =$ component specific to child $i$, constant over all tests;

$t_j =$ component specific to test $j$, constant over all children;

$ct_{ij} =$ interaction for child $i$ and test $j$; the response specific to child $i$ whenever he encounters test $j$.

The usual assumptions are made as to independence of components and their summations to zero, and $M$ is put equal to zero.

(i) On retesting the same child repeatedly with different tests - the familiar hypothetical situation in which it is assumed that $c_i$ is unaffected - the components $t$ and $e$ change from test to test, so that from equation (15) the expectation of the variance of scores for this one child over the tests (error variance) is
\[ E(\text{ev}_{(i)}) = \sigma_t^2 + \sigma_{ct}^2 \]  

If, however, as is generally assumed, the tests are equivalently standardised, \( t_j \) is constant for all tests. Hence \( \sigma_t^2 = 0 \), and

\[ E(\text{ev}_{(i)}) = \sigma_{ct}^2 \]  

(ii) On testing different children with the same test, the components \( c \) and \( ct \) alter from child to child, so that the expectation of the variance of child scores is \( (\sigma_c^2 + \sigma_{ct}^2) \). Of this, \( s_c^2 \) is the expectation of 'true' variance, and that of the error variance is

\[ E(\text{ev}_{(ii)}) = \sigma_{ct}^2 \]  

Thus the expected error variance when comparing scores on different tests from the same child is the same as when comparing scores from different children on the same test (compare equations (17) and (18)).

It is this error variance which is estimated in the usual two-way analysis of variance in the manner of Hoyt (38) and others, and which corresponds to the usual form \( s^2(1 - r) \).

(b) TESTS - THE PRESENT MODEL

In the present study the data do not conform to the simple model above. Because of the division of the year-group into schools, and of the discrepancies among schools means and s.d.'s reported in Tables I and II, the hypothesis was adopted of a school x test interaction superimposed on, and independent of, the child x tests interaction, and the analysis supported this hypothesis. The model for the tests was
\[ x_{ijk} = M + s_k + t_j + st_{kj} + c_{ik} + ct_{ijk} \]  \hspace{1cm} (6)

(iii) On retesting the same child repeatedly with different tests, the components \( t_j, ct_{ijk} \) and \( st_{kj} \) (i.e. those with \( j \) in the subscript) all change from test to test, and the expectation of the error variance of scores for this one child over the tests is

\[ E(ev_{(iii)}) = \sigma^2_t + \sigma^2_{st} + \sigma^2_{ct.s} \]  \hspace{1cm} (19)

If the tests have been equivalently standardised over the whole group, \( t \) is constant over all tests, so that

\[ E(ev_{(iii)}) = \sigma^2_{st} + \sigma^2_{ct.s} \]  \hspace{1cm} (20)

(iv) On testing different children with the same test, there are now two possibilities.

(a) Children in the same school.

From the model it is easy to see that

\[ E(ev_{(iv)(a)}) = \sigma^2_{ct.s} \]  \hspace{1cm} (21)

(b) Children from different schools.

From the model again:

\[ E(ev_{(iv)(b)}) = \sigma^2_{st} + \sigma^2_{ct.s} \]  \hspace{1cm} (22)

Thus, strictly speaking, no one error variance is appropriate for these several comparisons as it was in the conventional development.

Let us sum up briefly. The conventional model assumes homogeneity of variance throughout both sets of test scores equivalently standardised. The error variance derived from this model is related to the overall correlation between
the tests and to their common overall standard deviation. It is appropriate for comparisons among the same child's scores on different tests. It is appropriate also for comparisons among scores of different children on the same test.

The more complex model which fits the data studied takes account of the observed heterogeneity of differences among school means as well as of differences among children within schools. No one error variance is now appropriate for all comparisons. Instead, components must be picked out from the model and incorporated in different error variances in accordance with the particular comparison to be made. Among these comparisons is one not envisaged at all in the simpler model — that between children in different schools.

Consider next the residual or error variance \( \bar{\sigma^2} \) in Table V, in which the division by schools was ignored, in accordance with the conventional model. This error variance is an average which, strictly speaking, is not appropriate for any of the above comparisons. The residual line in Table V was constructed by summing the \( S \times T \) and the \( C \times T \) lines in Table III. From the relations on page 182 it is a matter of simple algebra to show that

\[
E (\bar{\sigma^2}) \text{ (in Table V)} = \frac{m(l - 1)}{m \cdot \ell - 1} \sigma^2_{st} + \sigma^2_{ct.s} \text{ (in Table III)} \quad (23)
\]

\( \bar{\sigma^2} \) above corresponds to the error variance of measurement \( \sigma^2_{ct} \) in the conventional development on page 187. According to that development, \( \bar{\sigma^2} \) should be used in all comparisons; horizontal, between scores on different tests from the same child; and vertical, between scores on the same test from different children, regardless of their school. As we have just seen, however, where inter-school heterogeneity is present, no single error variance is appropriate for all comparisons we may wish to make.
Now any complete year-group must consist of a number of separate school-groups. Unless our present sample is unique in its behaviour, inter-school heterogeneity may be expected in most year-groups we encounter. It is therefore important to know how seriously wrong we are likely to be in using the error variance $\bar{C_T}$ derived from the conventional development typified by the analysis of Table V.

Note first that if $s^{2}_{st} = 0$, we are not wrong at all, for then not only does $\bar{C_T}$ (above) estimate $\sigma^2_{ct}$, but also the error variances on pages 188 and 189, $ev'(iii)$, $ev(iv)(a)$ and $ev(iv)(b)$, all estimate $\sigma^2_{ct}$ (see equations (20), (21) and (22)). Hence if there is no significant school x test interaction, the conventionally obtained error variance is appropriate for all comparisons.

It is much more probable, however, that $\sigma^2_{st}$ is not zero. Consider the effect of school size on the conventionally derived error variance $\bar{C_T}$. It is easy to see, from the expectation in equation (23), page 190), that the limiting values, for $m = 1$ and for $m = \infty$ respectively, are

$$E(\bar{C_T}) = \sigma^2_{st} + \sigma^2_{ct.s} \quad \text{for } m = 1 \quad (24)$$

$$E(\bar{C_T}) = \frac{L-1}{L} \sigma^2_{st} + \sigma^2_{ct.s} \quad \text{for } m = \infty \quad (25)$$

These limits differ by $\sigma^2_{st} / L$, and since $L$, the number of schools, is usually quite large, the effect of school size on $\bar{C_T}$ is likely to be small.
We may therefore conclude that

\[ ev'(iii) = \frac{C_T}{\bar{m}} = ev(iv)(b) \]

and that we shall not be far wrong in using \( C_T \), which estimates \( \sigma_{st}^2 + \sigma_{ct.s}^2 \), in the comparisons for which \( ev'(iii) \) and \( ev(iv)(b) \) are appropriate. It will be recalled that these comparisons were: between a single child's scores on different tests, \( ev'(iii) \) (see equation (20)); and between the scores on the same test of children in different schools, \( ev(iv)(b) \) (see equation (22)).

The above discussion which, the reader is reminded, relates to tests, is given numerical illustration in Table VIII below. The estimates used for \( \sigma_{st}^2 \) and \( \sigma_{ct.s}^2 \) are those derived and reported on page 175.

### TABLE VIII - ERROR VARIANCES FOR TESTS

<table>
<thead>
<tr>
<th>Comparison</th>
<th>Error Variance</th>
<th>Expectations</th>
<th>Numerical Estimates</th>
<th>S.E.</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Conventional (( m = 20 ))</td>
<td>( \frac{m(l - 1)}{m - 1} \sigma_{st}^2 + \sigma_{ct.s}^2 )</td>
<td>37.437*</td>
<td>6.119</td>
<td></td>
</tr>
<tr>
<td>(2) Conventional (( m = 1 ))</td>
<td>( \sigma_{st}^2 + \sigma_{ct.s}^2 )</td>
<td>38.822</td>
<td>6.231</td>
<td></td>
</tr>
<tr>
<td>(3) Conventional (( m = \infty ))</td>
<td>( \frac{l - 1}{l} \sigma_{st}^2 + \sigma_{ct.s}^2 )</td>
<td>37.376</td>
<td>6.113</td>
<td></td>
</tr>
<tr>
<td>(4) Same child, different tests</td>
<td>( \sigma_{st}^2 + \sigma_{ct.s}^2 )</td>
<td>38.822</td>
<td>6.231</td>
<td></td>
</tr>
<tr>
<td>(5) Children in different schools, same test</td>
<td>( \sigma_{st}^2 + \sigma_{ct.s}^2 )</td>
<td>38.822</td>
<td>6.231</td>
<td></td>
</tr>
<tr>
<td>(6) Children in same school, same test</td>
<td>( \sigma_{ct.s}^2 )</td>
<td>30.144</td>
<td>5.489</td>
<td></td>
</tr>
</tbody>
</table>

* See Table V
It seems safe to conclude that the test error variance derived from the conventional model which takes no account of inter-school heterogeneity, is a sufficiently close approximation for all practical purposes, even when inter-school heterogeneity is present, so that the assumptions underlying the conventional model are not met.

(c) **RESCALED ASSESSMENTS - THE CONVENTIONAL MODEL**

The development of the conventional model for rescaled assessments is exactly analogous to that for tests (see page 187), and need not be repeated. It is sufficient to note that the analysis of variance in Table VI follows from this model, and that the error variance of a rescaled assessment, (the mean-square CE in Table VI), is 12.91 - the square of the standard error 3.593 reported in Table VII.

(d) **RESCALED ASSESSMENTS - THE PRESENT MODEL**

Writing $\sigma^2_{sa}$ for the variance component of interaction between schools and assessments, and $\sigma^2_{ca,s}$ for that between children and assessments within schools, we have, in the same manner as for tests:

(i) **The expectation of the error variance of assessments obtained by repeatedly rescaling the same child's 'raw' assessment on different parallel tests, equivalently standardised overall is**

$$E(\text{ev}_{(v)}) = \sigma^2_{sa} + \sigma^2_{ca,s} \quad \cdots \cdots \cdots \cdots \cdots \cdots (26)$$

$\text{ev}_{(v)}$ is thus analogous to $\text{ev}'(\text{iii})$ on page 189 (see equation (20)).

(ii) **The expectation of the error variance appropriate for comparing assessments for different children in the same school, the assessments being rescaled on the same test is**
E (\text{ev}(vi)(a)) = \sigma_{ca.s}^2 \quad \text{... (27)}

ev(vi)(a)\text{ is analogous to } ev(iv)(a)\text{ on page } 189\text{ (see equation } (21)).

(iii) \text{ The expectation of the error variance appropriate for comparing assessments for children in different schools, the assessments being rescaled on the same test is }

E (\text{ev}(vi)(b)) = \sigma_{sa}^2 + \sigma_{ca.s}^2 \quad \text{... (28)}

ev(vi)(b)\text{ is analogous to } ev(iv)(b)\text{ on page } 189\text{ (see equation } (22)).

The variance components \( \sigma_{sa}^2 \) and \( \sigma_{ca.s}^2 \) may be estimated from the analysis of Table IV (we noted earlier on page 181 that \( \bar{CA}_w = 4.582 \)). However, \( \sigma_{sa}^2 \) need not be estimated, since we already have everything needed to develop alternative expressions for \( \text{ev(v)} \) and \( \text{ev(vi)(b)}. \) These alternatives are equivalent to those presented in the preceding paragraphs and strikingly illuminate the relationship between test scores and the assessments rescaled on them.

It was earlier noted (page 179) that because of the equating of school assessment means and s.d.'s to the corresponding test means and s.d.'s, certain lines in the analyses of Tables III and IV must be identical. Referring in particular to line 3, we have

Mean Square for S \times A = Mean Square for S \times T.

Substituting variance components we obtain

\[ m \sigma_{sa}^2 + \sigma_{ca.s}^2 = m \sigma_{st}^2 + \sigma_{ct.s}^2 \]

whence
\[ \sigma_{sa}^2 = \frac{1}{m} (m \sigma_{st}^2 + \sigma_{ot.s}^2 - \sigma_{ca.s}^2) \]

and finally,

\[ \sigma_{sa}^2 + \sigma_{ca.s}^2 = (\sigma_{st}^2 + \frac{1}{m} \sigma_{ot.s}^2) + \frac{m - 1}{m} \sigma_{ca.s}^2 \]  \hspace{1cm} (29)

This important equation gives the expectation of the error variances \( ev(y) \) and \( ev(vi)(b) \) in terms of \( \sigma_{st}^2 \) and \( \sigma_{ot.s}^2 \), both of which can be estimated from the test data (Table III and page 175) and \( \sigma_{ca.s}^2 \), which can be estimated from the assessments data (Table IV and page 181).

The equation above is illuminating because it shows exactly how the assessment error variances depend on the error variance of school test means.

A rescaled assessment is to be thought of as a school test mean plus an independent deviation related to the school s.d.. The error variance of a rescaled assessment is therefore the sum of the error variance of school means and that of the deviations, and this is exactly what equation (29) shows.

The first two terms on the right together make up the expected error variance of a school mean; and it follows directly from equation (10) (page 180) that \( \sigma_{ca.s}^2 \) in the third term is estimated by

\[ \overline{e_{A_w}} = \frac{1}{2L} \sum (s_x - s_y)^2 \]  \hspace{1cm} (30)

an expression involving the school s.d.'s.

It will be recalled, moreover, that the reliability of the rescaling procedure was shown to be necessarily higher than that of the test on which it is based. Therefore the error variance of a rescaled assessment must be less than that of a test score.

It may now be shown by how much these error variances differ. The difference between their expectations is
\[
\left( \sigma_{st}^2 + \sigma_{ct.s}^2 \right) - \left( \sigma_{st}^2 + \frac{1}{m} \sigma_{ct.s}^2 - \frac{m-1}{m} \sigma_{ca.s}^2 \right)
\]

\[
= \frac{m-1}{m} \left( \sigma_{ct.s}^2 - \sigma_{ca.s}^2 \right)
\]

This difference is estimated by

\[
\frac{m-1}{2lm} \sum \left[ s_x - s_y \right]^2 - \left( s_x - s_y \right)^2
\]

\[
= \frac{m-1}{8m} \prod \sum \left[ s_x s_y \left( 1 - r_{xy} \right) \right]
\]

where the summation is over all schools.

Clearly, the smaller the within-school test correlations, the larger is the difference between the test and assessment error variances.

The numerical estimates of the error variances for the rescaled assessments are presented in Table IX.
The standard errors in Table IX should be compared with those in Table VIII. The new errors are all smaller - slightly more than half for (1) - (5) and less than half for (6). Once more the average standard error in (1) obtained by correlating overall, ignoring schools, is a reasonable approximation to most of the other standard errors quoted, though (6), which is considerably overestimated by (1), is an exception.
G. IV. - EFFECT OF SIZE OF SCHOOL ON RESCALING ERROR

Reference has been made already (page 161) to the opinions expressed by several writers as to the effect of school size on errors of rescaling. To use the word 'opinion' seems fair since, with the single exception of Vernon et al, already referred to, there appears to have been no serious attempt to develop a theoretical relationship between size of school and size of rescaling error. All that we appear to have are statements about minimum school sizes below which rescaling is not considered feasible because of the inverse relationship between school and error size.

However, if the arguments of the previous sections are sound, it is now possible to provide a theoretical basis, hitherto lacking, on which to study the relationship between school size and the error variance of rescaled assessments. The theory will be developed in two stages.

G. IV. (1) - (STAGE 1). EFFECT OF SCHOOL SIZE ON ERROR VARIANCE OF A SCHOOL MEAN

It has been pointed out (pages 164 and 172) that one of the two sources of unreliability in the rescaling procedure is the differential variability of school means from one possible rescaling test to another. The expression of this variability is the Schools x Tests interaction in a table of analysis of variance such as that presented in Table III (page 174). This becomes apparent on noting that

\[
\overline{ST} = \frac{m}{2(\ell - 1)} \sum (\bar{x} - \bar{y})^2 \quad \text{ (see also page 176)}
\]

\[
= \frac{m}{2} s^2 (\bar{x} - \bar{y})
\]

(33)
\[ = \frac{m}{2} \text{ (error variance of the difference between school means on two possible rescaling tests).} \]

A concrete illustration of how the school means differ in an actual experiment is provided by the entries in Table I, page 171.

Now the expectation of \( \overline{ST} \) is

\[ E(\overline{ST}) = n \sigma_{st}^2 + \sigma_{ct.s}^2 \]

Hence the expectation of the error variance of a school mean is

\[ E(\text{ev of a school mean}) = \sigma_{st}^2 + \frac{\sigma_{ct.s}^2}{m} \quad \cdots \quad \cdots \quad \cdots \quad (34) \]

The error variance of a school mean thus consists of two parts, of which only the second is affected by school size, the first being independent of it. Both \( \sigma_{st}^2 \) and \( \sigma_{ct.s}^2 \) can be estimated from the mean-squares in Table III, page 174:

\[ E \left( \frac{\overline{ST} - \overline{CT}}{m} \right) = \sigma_{st}^2 \text{ and } E \left( \overline{CT} \right) = \sigma_{ct.s}^2 \]

Hence equation (34) leads at once to estimates of the error variance of school means for different sizes of school. Table X shows the results derived from these Darlington data.
TABLE X - ESTIMATED ERROR VARIANCES OF TEST MEANS
FOR SCHOOLS OF DIFFERENT SIZES

<table>
<thead>
<tr>
<th>School Size (m)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_v$(school mean)</td>
<td>38.82</td>
<td>25.75</td>
<td>18.73</td>
<td>16.21</td>
<td>14.71</td>
<td>11.69</td>
</tr>
<tr>
<td>$e_s$(school mean)</td>
<td>6.23</td>
<td>4.87</td>
<td>4.33</td>
<td>4.03</td>
<td>3.83</td>
<td>3.42</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>School Size (m)</th>
<th>15</th>
<th>20</th>
<th>25</th>
<th>30</th>
<th>40</th>
<th>$\infty$</th>
</tr>
</thead>
<tbody>
<tr>
<td>$e_v$(school mean)</td>
<td>10.69</td>
<td>10.18</td>
<td>9.88</td>
<td>9.68</td>
<td>9.43</td>
<td>8.68</td>
</tr>
<tr>
<td>$e_s$(school mean)</td>
<td>3.27</td>
<td>3.19</td>
<td>3.14</td>
<td>3.11</td>
<td>3.07</td>
<td>2.95</td>
</tr>
</tbody>
</table>

The figures show the regression of error variance and standard error on school size. Little need be said about them except that they lend little support to the prevailing belief that rescaling is hazardous with small schools.

G.IV.(2) - (STAGE 2). EFFECT OF SCHOOL SIZE ON ERROR VARIANCE AND STANDARD ERROR OF INDIVIDUAL RESCALED ASSESSMENTS

Interesting though they are, the error variances and standard errors in the preceding section provide only partial information. It is more important in practice to estimate the effect of the size of school on the error variance and standard error of rescaled assessment of an individual child.

The second of the two sources of unreliability in the rescaling procedure is the variability of school standard deviations from one possible rescaling test to another. It has already been shown (page 168) that the correlation
within any one school is necessarily perfect between sets of assessments obtained by rescaling the same original or 'raw' assessments on each of two different rescaling tests. However, rescaling error is introduced if, for that school, the standard deviations of scores on the two rescaling tests differ. This becomes apparent on noting from equation (10), page 180 that the interaction mean-square for Assessments and Children within Schools is

\[ \overline{CA}_{w} = \frac{1}{2} \sum (s_x - s_y)^2. \]

A concrete illustration of how these standard deviations differ in an actual experiment is provided by the data of Table II, page 171.

Now it was shown in equation (29), page 195, that the expectation of the error variance of a single assessment is

\[ E(\text{ev of a single assessment}) = \sigma_{sa}^2 + \sigma_{ca.s}^2 \]

\[ = (\sigma_{st}^2 + \frac{1}{m} \sigma_{ct.s}^2) + \frac{m - 1}{m} \sigma_{ca.s}^2. \]

(29)

All these variance components can be estimated from the mean-squares of Table III, page 174 and Table IV, page 181. Hence equation (29) leads directly to estimates of the error variances of single assessments in schools of different sizes. Table XI presents the results derived from the Darlington data.
TABLE XI - EFFECT OF SCHOOL SIZE ON ERROR VARIANCE AND STANDARD ERROR OF A RESCALED ASSESSMENT

<table>
<thead>
<tr>
<th>School Size (m)</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>10</th>
</tr>
</thead>
<tbody>
<tr>
<td>( ev(\text{resc. asst.}) )</td>
<td>38.82</td>
<td>26.04</td>
<td>21.78</td>
<td>19.65</td>
<td>18.37</td>
<td>15.82</td>
</tr>
<tr>
<td>( se(\text{resc. asst.}) )</td>
<td>6.23</td>
<td>5.10</td>
<td>4.67</td>
<td>4.43</td>
<td>4.29</td>
<td>3.98</td>
</tr>
<tr>
<td>School Size (m)</td>
<td>15</td>
<td>20</td>
<td>25</td>
<td>30</td>
<td>40</td>
<td>( \infty )</td>
</tr>
<tr>
<td>( se(\text{resc. asst.}) )</td>
<td>3.87</td>
<td>3.91</td>
<td>3.78</td>
<td>3.76</td>
<td>3.73</td>
<td>3.64</td>
</tr>
</tbody>
</table>

The regression of standard error on school size is even lower in Table XI than it was in Table X. The maximum standard error is less than twice the minimum and near-stability is reached at a remarkably small school size. It seems that there is little theoretical ground for the current pessimism over rescaling in small schools.

G.V - A NOTE ON 'FLOP' SCORES

McIntosh et al (55) page 60) discuss in detail the difficulties arising when 'flop' scores occur. A 'flop' score is a child's score on the test used for rescaling which is 'much less than would be expected of him, while the others have scored in accordance with expectation'. These writers point out that if a 'flop' score occurs and it is not balanced by its opposite, all rescaled assessments in the school concerned will be depressed, since the school mean on the rescaling test will be unduly low. If this occurs, the depression will be greater in small schools than in large.
McIntosh et al. also state (ibid. page 65) that 'cases are much rarer...

of the other type of deviation where a pupil scores much higher than expected'.

The statements of these writers that the frequency of below-expectation test scores is greater than that of above-expectation test scores must be accepted as a record of their not inconsiderable experience. Nevertheless, it has no obvious warrant from theoretical principles. There is no apparent reason why some children should not do better than expected on the test any less frequently than others do worse.

The effect of an unbalanced 'flop' score may be simply illustrated as follows:

<table>
<thead>
<tr>
<th>Child</th>
<th>a</th>
<th>b</th>
<th>c</th>
<th>d</th>
<th>e</th>
<th>Mean</th>
<th>s.d.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Original Assessment</td>
<td>10</td>
<td>8</td>
<td>6</td>
<td>4</td>
<td>2</td>
<td>7</td>
<td>√10</td>
</tr>
<tr>
<td>Test score (x)</td>
<td>4</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>6</td>
<td>√10/2</td>
</tr>
<tr>
<td>Rescaled Assessment (a)</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>6</td>
<td>√10/2</td>
</tr>
<tr>
<td>Difference (x - a)</td>
<td>-4</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>-</td>
</tr>
</tbody>
</table>

Child a's test score of 4 is a 'flop' score which places him last on the test. On rescaling, child a recovers the top position assigned him by his teachers, and all rescaled estimates are 1 point below what they would have been had child a scored as well on the test as expected.

The differences (x - a) should be noted. That for the child making the 'flop' score is much greater than the others, and opposite in sign, and the distribution of differences (x - a) is asymmetrical.

In an actual case, of course, the effect of an unbalanced 'flop' score would not be so clear-cut. Nevertheless, if unbalanced 'flop' scores do
occur in practice, the resulting lack of symmetry in the distribution of within-child differences between test scores and rescaled assessments should be detectable.

The matter is easily put to the test for the Darlington data. For each of the 6 schools of 20 children there is one set of 20 'raw' assessments, two sets of 20 test scores x and y, and two sets of 20 rescaled assessments, a and b. For each child there are 2 differences, \( (x - a) \) and \( (y - b) \).

Table XII shows these pairs of differences for the 20 children in each of six schools.
TABLE XII - INTRA-CHILD DIFFERENCES

<table>
<thead>
<tr>
<th>Child</th>
<th>School 1</th>
<th>School 2</th>
<th>School 3</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(x - a)</td>
<td>(y - a)</td>
<td>(x - a)</td>
</tr>
<tr>
<td>1</td>
<td>+</td>
<td>+</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>28</td>
<td>16</td>
<td>8</td>
</tr>
<tr>
<td>3</td>
<td>11</td>
<td>3</td>
<td>8</td>
</tr>
<tr>
<td>4</td>
<td>1</td>
<td>3</td>
<td>3</td>
</tr>
<tr>
<td>5</td>
<td>27</td>
<td>16</td>
<td>13</td>
</tr>
<tr>
<td>6</td>
<td>12</td>
<td>10</td>
<td>21</td>
</tr>
<tr>
<td>7</td>
<td>16</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>9</td>
<td>15</td>
<td>5</td>
</tr>
<tr>
<td>9</td>
<td>20</td>
<td>20</td>
<td>18</td>
</tr>
<tr>
<td>10</td>
<td>1</td>
<td>2</td>
<td>22</td>
</tr>
<tr>
<td>11</td>
<td>21</td>
<td>17</td>
<td>31</td>
</tr>
<tr>
<td>12</td>
<td>7</td>
<td>3</td>
<td>4</td>
</tr>
<tr>
<td>13</td>
<td>11</td>
<td>15</td>
<td>9</td>
</tr>
<tr>
<td>14</td>
<td>3</td>
<td>3</td>
<td>12</td>
</tr>
<tr>
<td>15</td>
<td>2</td>
<td>1</td>
<td>24</td>
</tr>
<tr>
<td>16</td>
<td>17</td>
<td>15</td>
<td>0</td>
</tr>
<tr>
<td>17</td>
<td>6</td>
<td>11</td>
<td>0</td>
</tr>
<tr>
<td>18</td>
<td>24</td>
<td>11</td>
<td>28</td>
</tr>
<tr>
<td>19</td>
<td>6</td>
<td>17</td>
<td>46</td>
</tr>
<tr>
<td>20</td>
<td>4</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td></td>
<td>16</td>
<td>19</td>
<td>17</td>
</tr>
</tbody>
</table>

|       | 123      | 119      | 108      | 109      | 119      | 121      |

Schools 4 - 6 on next page
Table XII. - INTRA-CHILD DIFFERENCES (contd.)

<table>
<thead>
<tr>
<th>Child</th>
<th>School 4</th>
<th>School 5</th>
<th>School 6</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>(x - a)</td>
<td>(y - b)</td>
<td>(x - a)</td>
</tr>
<tr>
<td>1</td>
<td>+</td>
<td>-</td>
<td>+</td>
</tr>
<tr>
<td>2</td>
<td>5</td>
<td>12</td>
<td>21</td>
</tr>
<tr>
<td>3</td>
<td>13</td>
<td>8</td>
<td>6</td>
</tr>
<tr>
<td>4</td>
<td>4</td>
<td>0</td>
<td>2</td>
</tr>
<tr>
<td>5</td>
<td>19</td>
<td>20</td>
<td>9</td>
</tr>
<tr>
<td>6</td>
<td>13</td>
<td>20</td>
<td>12</td>
</tr>
<tr>
<td>7</td>
<td>10</td>
<td>11</td>
<td>17</td>
</tr>
<tr>
<td>8</td>
<td>25</td>
<td>18</td>
<td>12</td>
</tr>
<tr>
<td>9</td>
<td>1</td>
<td>10</td>
<td>17</td>
</tr>
<tr>
<td>10</td>
<td>49</td>
<td>56</td>
<td>8</td>
</tr>
<tr>
<td>11</td>
<td>17</td>
<td>13</td>
<td>7</td>
</tr>
<tr>
<td>12</td>
<td>15</td>
<td>7</td>
<td>11</td>
</tr>
<tr>
<td>13</td>
<td>10</td>
<td>6</td>
<td>10</td>
</tr>
<tr>
<td>14</td>
<td>17</td>
<td>7</td>
<td>10</td>
</tr>
<tr>
<td>15</td>
<td>1</td>
<td>3</td>
<td>1</td>
</tr>
<tr>
<td>16</td>
<td>5</td>
<td>6</td>
<td>17</td>
</tr>
<tr>
<td>17</td>
<td>8</td>
<td>21</td>
<td>9</td>
</tr>
<tr>
<td>18</td>
<td>12</td>
<td>11</td>
<td>1</td>
</tr>
<tr>
<td>19</td>
<td>2</td>
<td>9</td>
<td>21</td>
</tr>
<tr>
<td>20</td>
<td>10</td>
<td>5</td>
<td>13</td>
</tr>
<tr>
<td></td>
<td>124</td>
<td>126</td>
<td>132</td>
</tr>
</tbody>
</table>

Table XIII, derived from the figures in Table XII, summarises the frequency distributions of \((x - a)\) and \((y - b)\) over all 120 children.
In Table XII there are 12 sets of 20 differences, six of them \((x - a)\) and the others \((y - b)\). None of these sets shows evidence of the asymmetry to be expected if some children had done 'much less well than would be expected' while the others had 'scored in accordance with expectation'. In each set of 20 differences the number of positive and negative differences is similar.

Table XIII shows the \((x - a)\) and \((y - b)\) differences pooled over all schools. Again there is no evidence of asymmetry. Tests are scarcely necessary, but for the record, \(\beta_1\) has been calculated for each. For the \((x - a)\)'s it is \(-.1341\), and for the \((y - b)\)'s \(-.1298\). Neither value is significant at the 5 per cent level.
Some children have undoubtedly done less well on the test than their teachers expected. But their scores are not 'flop' scores as McIntosh et al define them, since a similar number of children have done better on the tests than their teachers expected, and by about the same amount.

Since for each child there are two differences in Table XII, a further test for 'flop' scores is now available which was not so to McIntosh et al. Inspection of the data shows the existence of a high positive correlation for each school between the \((x - a)\) and \((y - b)\) arrays. These correlations are listed below in Table XIV.

**TABLE XIV - CORRELATIONS BETWEEN DIFFERENCES \((x - a)\) AND \((y - b)\)**

<table>
<thead>
<tr>
<th>School</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
</tr>
</thead>
<tbody>
<tr>
<td>(r(x-a)(y-b))</td>
<td>.907</td>
<td>.868</td>
<td>.844</td>
<td>.922</td>
<td>.845</td>
<td>.930</td>
</tr>
</tbody>
</table>

Pooled \(r(x-a)(y-b)\) within schools = .992 (d.f. 114)

The consistency of these correlations is noteworthy. It is the outcome of the high reliability of the tests X and Y and of the consistency, school by school, of the correlations between tests and assessments (see Table XV, page 209).

These correlations mean that the child who does unexpectedly badly or, conversely, unexpectedly well on the first test probably does unexpectedly badly or, conversely, unexpectedly well on the second, nearly six weeks later. Child 10, in school 4, for example, is 49 points below expectation on test X, and 56 points below expectation on test Y. Child 18 in school 2 is 46 points above expectation on test X and 39 points above on test Y. With a few exceptions in both directions (which on the whole balance out) the pattern throughout is similar.
These results may mean that some children are consistently poor examinees and that others are consistently good; or it may mean that the teachers' original judgements for some children were wrong. The figures do not support the belief that some children make 'flop' scores - or their opposite - on one occasion, and score according to their teachers' expectation on the other.

It must be concluded that the experience of McIntosh et al - that 'flop' scores occur more frequently than their opposite - is not borne out by the Darlington results. The crux of the matter is of course the high correlation between the tests scores on X and Y, which means that the performances of children on these two tests is similar, whatever the teachers' assessments were.

G.VI - CORRELATION BETWEEN TEST SCORES AND ASSESSMENTS

As a footnote to this enquiry, the correlations between test scores and estimates are reported in Table XV below.

<table>
<thead>
<tr>
<th>TABLE XV - CORRELATIONS BETWEEN TEST SCORES AND TEACHERS' ASSESSMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>School</td>
</tr>
<tr>
<td>--------</td>
</tr>
<tr>
<td>r(X)(Asst.)</td>
</tr>
<tr>
<td>r(Y)(Asst.)</td>
</tr>
</tbody>
</table>

Pooled correlations within schools:

*r(X)(Asst.) = .930;  r(Y)(Asst.) = .937

Total correlation:

r(X)(Asst.) = .938;  r(Y)(Asst.) = .945
The pooled within-school correlations in Table XV were obtained by pooling the sums of squares and sums of products which, taken individually, lead to the separate within-school correlations listed in the second and third lines of the table. The 'total' correlations include the between-means correlations which are of course very nearly unity.

The correlations are fairly consistent for all schools, and generally high — indeed, as high as some test-retest correlations. Although the correlations were calculated between test scores and rescaled assessments, those within schools (the first 12 in the table) are of course identical with the correlations between test scores and original assessments. Since the assessments were made before the tests were administered, it cannot be said that, in making their assessments, the teachers were influenced by these test results. The correlations are so high that in general the validities of tests and assessments for a common criterion will be rather similar, though in particular cases this might not be so. More specifically, if the test-assessment correlation is .93, and the test-criterion correlation .85 (the value frequently found in follow-up studies where the criterion is academic success), then the estimate-criterion correlation must lie between .98 and .60.

A further point to note is that with one exception all the correlations between tests Y and the assessments are slightly higher than the corresponding correlations between test X and the assessments. Individually the differences are slight and non-significant but their consistency of direction in five out of the six schools suggests a non-random effect. Y was the second test to be administered, and hence reflects the effect of practice obtained on X. These results lend some support to the hypothesis of Ortar (59), that the effect on test scores of limited coaching or practice is to increase their validity for
In the preceding enquiry, the 'tests' (in fact batteries VR + E + A) on
which the rescaling was based were administered 40 days apart. The tests
included in the first battery differed from those included in the second.
Hence the overall battery reliability coefficient of .9805 (see equation (13),
page 183) is a coefficient of equivalence and stability. The reliability of
the rescaling procedure, .9933 (see Table VII, page 185), is also a coefficient
of equivalence and stability. This is so because it is the result of
correlating overall the two sets of assessments obtained by rescaling the same
original set from several schools, first on one test battery, and then on
another administered six weeks later.

If, however, there is only one occasion of testing, and only one test on
that occasion, the method employed in the previous study is inappropriate.
It is important to have some estimate of the variation to be expected in a
child's rescaled assessment, if, in place of the test actually administered and
used for rescaling, some other test had been administered on that occasion and
used for rescaling instead. The correlation between the tests would be a
coefficient of test equivalence, and the corresponding correlation between the
assessments the coefficient of equivalence for the rescaling procedure.

* A similar enquiry was conducted by Dr. Ortar in 1961 while I was working with
her in Israel. Her subjects were Israeli children who were relatively 'test-
unsophisticated', and the increase in validity for teachers' estimates of
scholastic ability was much greater. With British children, so much more
'test-sophisticated', a much smaller increase is to be expected.
Single-occasion testing and assessments rescaling based on that testing, are not uncommon in allocation procedures. Frequently a choice of tests is made from several that are available and equally appropriate. So far as is known, there has been no serious study hitherto of the differences made to rescaled assessments if one test, rather than another, is chosen for the one occasion of testing.

Exact measures of coefficients of equivalence are, of course, unattainable in practice. It is impossible to know what changes would have occurred in testees' test order if for the test used another had been substituted. Nevertheless, we can come somewhere near the coefficient of equivalence by employing two tests with as short a time interval as possible between them.

Practice effect is an obvious difficulty which can be overcome, at least in part, by using the cross-over design described and employed earlier (pages 113-118) in estimating test 'zero error'. Indeed, the data used in that enquiry are now employed again in this.

The reader is reminded that, in the 'zero error' experiment, test X, and then test Y, were administered to a random half of the children in each of several schools, while test Y, and then test X, were administered to the other half.* On the assumption that the practice effect from first to second test is the same, no matter which is first, it is easy to allow for the practice effect and hence to estimate the overall test 'zero error' untrammelled by practice. The 'zero errors' for individual schools are also available.

We do not have, nor do we need to obtain, teachers' assessments corresponding to the test scores in the 'zero error' study. Our concern here is to obtain

* The data from Moore's 6 x 6 latin square are not made use of in this enquiry.
the standard error of a rescaled assessment; and, as we have seen (pages 178-180), thus can be done without calculating the rescaled assessments at all, provided that certain data are available. Fortunately, they are available from the 'zero error' experiments.

It is easy to see that if we did have assessments rescaled by Method 3 on tests X and Y in any one of the 'zero error' experiments, the overall mean 'zero error' for tests would be also the overall mean 'zero error' for assessments; and that the school mean 'zero errors' for tests would be also the school mean 'zero errors' for assessments. It has been pointed out already (pages 164-166) that knowledge of these school mean 'zero errors' is an important prerequisite for the calculations of the standard errors of individual children's rescaled assessments. Since we do have these school mean 'zero errors' our problem is already half solved.

II - THE 'ZERO ERROR' MODEL

II (1) Tests

The independent components* of a child's test score $x$ are:

$$x = M + c + g + s + t + o + st + so + ct$$

in which

$M$ is common to all children; $c$ is specific to the child; $g$ to his group; $s$ to his school; $t$ to a test; $o$ to the occasion of testing; $st$ and $so$ are school x test and school x occasion interactions; and $ct$ is the child x test interaction.

There are two groups, two tests, and two occasions of testing, and the design may be diagrammed for each of the several schools as follows:

---

* Subscripts are omitted since the equation will not be summed. It is stated here for illustration only.
The analysis of variance corresponding to this design and model is as follows:-

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Schools</td>
<td>( \ell - 1 )</td>
<td>S</td>
<td>S</td>
</tr>
<tr>
<td>(2) Tests</td>
<td>1</td>
<td>T</td>
<td>T</td>
</tr>
<tr>
<td>(3) S x T</td>
<td>( \ell - 1 )</td>
<td>ST</td>
<td>ST</td>
</tr>
<tr>
<td>(4) Occasion</td>
<td>1</td>
<td>0</td>
<td>0</td>
</tr>
<tr>
<td>(5) S x C</td>
<td>( \ell - 1 )</td>
<td>SC</td>
<td>SC</td>
</tr>
<tr>
<td>(6) Groups within Schools</td>
<td>( \ell )</td>
<td>( \bar{G}_W )</td>
<td>( \bar{G}_W )</td>
</tr>
<tr>
<td>(7) Children within Groups within Schools</td>
<td>( 2(\bar{m} - \ell) )</td>
<td>( \bar{C}_{WW} )</td>
<td>( \bar{C}_{WW} )</td>
</tr>
<tr>
<td>(8) C x T within Groups within Schools</td>
<td>( 2(\bar{m} - \ell) )</td>
<td>( \bar{C}_{ST,WW} )</td>
<td>( \bar{C}_{ST,WW} )</td>
</tr>
<tr>
<td>(9) Total</td>
<td>( 4\bar{m} - 1 )</td>
<td>Total</td>
<td></td>
</tr>
</tbody>
</table>

(i) \( \ell \) = number of schools.

(ii) \( m \) = number of children in a group within a school.

(iii) \( m \) differs from one school to another, and \( m' \) in the variance components is derived from

\[
m' = \frac{1}{\ell - 1} \left( \bar{m} m' - \frac{\sum m^2}{\sum m} \right)
\]

\( (44b), \) page 105. \]
H.11 (2) Application to Rescaled Assessments

Let us suppose that Method 3 has been used to rescale the teachers' original assessments on test X so that the rescaled assessment means and s.d.'s are the same as the test X means and s.d.'s for the groups within schools; and similarly, ab initio, on test Y. An analysis of variance could be derived from these assessments of the same sort as that just described for tests. The degrees of freedom would then be the same, line by line, as those in the test analysis. Also, the sums of squares, and hence the mean squares, would be numerically the same in both analyses for lines (1) - (6), and (9). The remaining lines, (7) and (8), could be obtained in a manner similar to that used in the previous design (page 179 et seq).

Such an analysis need not be carried out; it is mentioned merely to show that the arguments used in the preceding section are relevant here also. Instead, we shall extract from the analysis for tests, and from the inferred analysis for rescaled assessments, only what we need to obtain the standard error of a rescaled assessment under the conditions of the cross-over design.

Let us denote a particular child's assessment, rescaled on test X, by a, and, rescaled on test Y, by b. Let the mean zero error of test Y, relative to test X, calculated as in Chapter 8, be z. Then b is given by:

\[ b = a + z \]  

(standard error of the rescaled assessment).

We propose to estimate this standard error.

The arguments used in deriving the error variance (and hence the standard error) of a rescaled assessment are similar to those used in the preceding section (page 195). A rescaled assessment is to be thought of as a school test mean plus an independent deviation related to the school s.d. Hence the error variance of a rescaled assessment is the sum of the error variances of the school mean and of the deviation.
As shown in equation (29), page 195, the expectation of the error variance of a rescaled assessment for a school of size \( m \) is given by

\[
E(ev) = \left( \sigma^2_{st} + \frac{1}{m} \sigma^2_{ct.s} \right) + \frac{m - 1}{m} \sigma^2_{ca.s}
\]

If the schools are randomly chosen, and the children in the schools randomly assigned to the groups within schools, then \( \sigma^2_{st} \) and \( \sigma^2_{ct.s} \), derived from the analysis for tests, are estimates of the population variances \( \sigma^2_{st} \) and \( \sigma^2_{ct.s} \) for the pair of tests analysed.

\( \sigma^2_{ca.s} \), which estimates \( \sigma^2_{ca.s} \), is obtained as follows:-

In a particular school the scores on tests X and Y of a child in group I are \( x_1 \) and \( y_2 \), the subscripts indicating the occasion of testing. Similarly, the scores on the same tests of a child in group II in the same school are \( x_2 \) and \( y_1 \).

The interaction sum of squares \( C \times T \) for this school is

\[
CT = \frac{1}{2} \sum (x_1 - y_2)^2 + \sum (x_2 - y_1)^2
\]

(within a school)

and if there are \( m \) children in a group, the degrees of freedom are \( 2(m - 1) \).

We may therefore write:

\[
CT = \frac{m - 1}{2} \left( \sum s^2_{x_1 - y_2} + \sum s^2_{x_2 - y_1} \right)
\]

(within a school)

Noting, as in the previous section, that, within each school, assessments rescaled on X and on Y are perfectly correlated, we may now write, for the interaction sum of squares \( CA_W(1) \) between children and assessments within this school:
\[ \text{CA}_{W}(1) = \frac{m - 1}{2} \sum (s_{x_1} - s_{y_2})^2 + (s_{x_2} - s_{y_1})^2 \]

(one school)

Summing over all schools and dividing by the sum of the degrees of freedom over all schools we obtain the interaction mean-square for Assessments and Children within Schools:

\[ \bar{\text{CA}}_{WW} = \frac{\frac{1}{2} \sum(m - 1) \sum (s_{x_1} - s_{y_2})^2 + (s_{x_2} - s_{y_1})^2}{2 \sum(m - 1)} \]

\[ = \frac{s^2_{(ca)} \text{, estimating } \sigma^2_{(ca)}}{s_{(ca)}}, \text{estimating } \sigma^2_{(ca)} \]

Note that the summation is over all m's and all s.d.'s.

It is necessary, therefore, to calculate the within-group s.d.'s, \( s_{x_1}, s_{y_2}, s_{x_2}, s_{y_1} \), and \( s_{x}, s_{y} \), for each of the schools, and then to use the equation above to obtain \( s^2_{(ca)} \).

The conclusions from this discussion may now be summarised as follows:

a = teacher's assessment, after rescaling on test X, for a particular child;

b = the same teacher's assessment after rescaling on test Y, for the same child;

z = mean zero error for tests X and Y obtained as in Chapter 8;

\( \text{se}(\text{res. asst.}) = \text{standard error of a rescaled assessment} \);

m = school size;

\( s^2_{st}, s^2_{ct.s}, s^2_{ca.s} \) are respectively estimates of the population variance components \( \sigma^2_{st}, \sigma^2_{ct.s}, \text{and } \sigma^2_{ca.s} \).

Then the relation between the rescaled assessments a and b is

\[ b = a + z + \text{se}(\text{resc. asst.}) \]

in which
\[ s_e^2 = s_{st}^2 + \frac{1}{m} s_{ct.s}^2 + \frac{m - 1}{m} s_{ca.s}^2 \]  

(34)

The mean zero error is derived from a cross-over analysis appropriate to tests X and Y, as in Chapter 8; for equation (34), \( s_{st}^2 \) and \( s_{ct.s}^2 \) are obtained from the corresponding analyses of variance, and \( s_{ca.s}^2 \) is derived from the test standard deviations within groups within schools in the manner indicated by equation (33), page 217.

The numerical data required to derive the zero error \( z \) for tests X and Y and to complete the corresponding analyses of variance include the data necessary to derive \( s_e^2 \). The computations are laborious; it is therefore fortunate that their outcomes are remarkably similar (as will be shown) when M.H.T. 40 is paired in turn with each of a number of other tests. This similarity of outcome makes it reasonable to infer a fair degree of stability in the estimate of \( s_e^2 \), no matter which Moray House verbal reasoning test is used for rescaling. These computations will now be illustrated.

H.III - NUMERICAL ILLUSTRATION

The data for the comparison between M.H.T. 40 and M.H.T. 33 will be used to illustrate the procedure.

In the six schools the numbers of children were:

32, 30, 34, 30, 34, and 26, total = 186.

Within each school the children were divided at random into two numerically equal groups, so that the group sizes are each one-half of the numbers above.

The 'average' group size is

\[ m' = \frac{32 + 30 + 34 + 30 + 34 + 26}{6} = \frac{186}{6} = 31 \]

(see page 214)

\[ m' = 31 \]

\[ = 15.475 \]
The final analysis of variance for these tests is as follows:

<table>
<thead>
<tr>
<th>Source</th>
<th>df</th>
<th>SS</th>
<th>MS</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Schools</td>
<td>5</td>
<td>32,354.4</td>
<td>6,460.9</td>
</tr>
<tr>
<td>(2) Tests (M.H.T. 40 vs. M.H.T. 33)</td>
<td>1</td>
<td>2.8</td>
<td>2.8</td>
</tr>
<tr>
<td>(3) Schools x Tests</td>
<td>5</td>
<td>166.1</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>= 2 x 15.475 + s^2_{ct.s}</td>
<td></td>
</tr>
<tr>
<td>(4) 1st vs. 2nd occasion</td>
<td>1</td>
<td>190.2</td>
<td>190.2</td>
</tr>
<tr>
<td>(5) Schools x Occasions</td>
<td>5</td>
<td>139.2</td>
<td>27.8</td>
</tr>
<tr>
<td>(6) Groups within Schools</td>
<td>6</td>
<td>629.1</td>
<td>104.8</td>
</tr>
<tr>
<td>(7) Child Totals within Groups within Schools</td>
<td>174</td>
<td>31,769.7</td>
<td>182.6</td>
</tr>
<tr>
<td>(8) Child Scores x Tests (C x T) within Groups within Schools</td>
<td>174</td>
<td>2,444.5</td>
<td>14.049</td>
</tr>
</tbody>
</table>

The analysis is based on the model stated on page 213, and follows the pattern set on page 214.

All that we require from the analysis are the estimates $s^2_{st}$ and $s^2_{ct.s}$ of the variance components $\sigma^2_{st}$ and $\sigma^2_{ct.s}$. Solving the equation in lines (3) and (8) we obtain:

$$s^2_{st} = .620; \quad s^2_{ct.s} = 14.049$$

We next obtain $s^2_{cs.5}$ in the manner shown below. All the s's are within group s.d. s; test X = M.H.T. 40, test Y = M.H.T. 33; 1 indicates first occasion and 2 second occasion.
<table>
<thead>
<tr>
<th>School</th>
<th>$s_{x_1}$</th>
<th>$s_{y_2}$</th>
<th>$s_{x_2}$</th>
<th>$s_{y_1}$</th>
<th>$(s_{x_1} - s_{y_2})^2$</th>
<th>$(s_{y_1} - s_{x_2})^2$</th>
<th>$(m-1)$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>5.465</td>
<td>5.766</td>
<td>6.217</td>
<td>6.735</td>
<td>0.0967</td>
<td>0.2683</td>
<td>15</td>
</tr>
<tr>
<td>2</td>
<td>11.733</td>
<td>13.479</td>
<td>10.131</td>
<td>8.895</td>
<td>3.0485</td>
<td>1.5277</td>
<td>15</td>
</tr>
<tr>
<td>3</td>
<td>10.964</td>
<td>10.412</td>
<td>9.724</td>
<td>10.659</td>
<td>0.3047</td>
<td>0.8742</td>
<td>16</td>
</tr>
<tr>
<td>5</td>
<td>10.051</td>
<td>10.514</td>
<td>9.184</td>
<td>10.141</td>
<td>0.2144</td>
<td>0.1069</td>
<td>16</td>
</tr>
<tr>
<td>6</td>
<td>8.006</td>
<td>7.018</td>
<td>13.432</td>
<td>9.867</td>
<td>0.9761</td>
<td>12.7092</td>
<td>12</td>
</tr>
</tbody>
</table>

Substituting in the equation

$$s_{ca.s}^2 = \frac{1}{2} \sum (m-1) \left( (s_{x_1} - s_{y_2})^2 + (s_{x_2} - s_{y_1})^2 \right) / 2 \sum (m-1)$$

we obtain

$$s_{ca.s}^2 = 1.085$$

Hence the error variance of a rescaled assessment is

$$ev = 0.620 + 14.049 / m + 1.085 (m-1) / m,$$

in which $m$ is the number of children in the school.

Thus the estimates of the population variance components obtained from the analysis of variance can be used to obtain estimates of the standard error a child's rescaled assessment for any size of school.

The computations have been worked through for nine of the fifteen pairs of tests used in the zero error study, and the results are reported in Table XVI, M.H.T. 40 is paired in turn with each of the tests listed.

* For purposes of computation, the equation can be simplified to:

$$ev = 1.705 + 12.964 / m$$
### TABLE XVI - STANDARD ERRORS OF RESCALED ASSESSMENTS FOR VARIOUS SCHOOL SIZES

<table>
<thead>
<tr>
<th>m</th>
<th>33</th>
<th>34</th>
<th>46</th>
<th>M.H.T. 40 and M.H.T. 47</th>
<th>M.H.T. 49</th>
<th>56</th>
<th>57</th>
<th>58</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>3.83</td>
<td>3.52</td>
<td>4.15</td>
<td>3.88</td>
<td>3.69</td>
<td>3.58</td>
<td>3.63</td>
<td>3.57</td>
</tr>
<tr>
<td>2</td>
<td>2.86</td>
<td>2.60</td>
<td>3.34</td>
<td>2.85</td>
<td>2.88</td>
<td>2.75</td>
<td>2.80</td>
<td>2.73</td>
</tr>
<tr>
<td>3</td>
<td>2.45</td>
<td>2.20</td>
<td>2.80</td>
<td>2.40</td>
<td>2.55</td>
<td>2.45</td>
<td>2.48</td>
<td>2.34</td>
</tr>
<tr>
<td>4</td>
<td>2.21</td>
<td>1.98</td>
<td>2.50</td>
<td>2.15</td>
<td>2.38</td>
<td>2.21</td>
<td>2.38</td>
<td>2.16</td>
</tr>
<tr>
<td>5</td>
<td>2.07</td>
<td>1.83</td>
<td>2.34</td>
<td>1.98</td>
<td>2.26</td>
<td>2.09</td>
<td>2.22</td>
<td>2.00</td>
</tr>
<tr>
<td>7</td>
<td>1.89</td>
<td>1.64</td>
<td>2.16</td>
<td>1.89</td>
<td>2.13</td>
<td>1.94</td>
<td>2.00</td>
<td>1.81</td>
</tr>
<tr>
<td>10</td>
<td>1.73</td>
<td>1.48</td>
<td>2.01</td>
<td>1.60</td>
<td>2.02</td>
<td>1.82</td>
<td>1.88</td>
<td>1.70</td>
</tr>
<tr>
<td>15</td>
<td>1.66</td>
<td>1.35</td>
<td>1.88</td>
<td>1.46</td>
<td>1.93</td>
<td>1.72</td>
<td>1.77</td>
<td>1.58</td>
</tr>
<tr>
<td>20</td>
<td>1.59</td>
<td>1.28</td>
<td>1.82</td>
<td>1.38</td>
<td>1.88</td>
<td>1.67</td>
<td>1.71</td>
<td>1.47</td>
</tr>
<tr>
<td>30</td>
<td>1.46</td>
<td>1.20</td>
<td>1.75</td>
<td>1.29</td>
<td>1.82</td>
<td>1.61</td>
<td>1.68</td>
<td>1.44</td>
</tr>
<tr>
<td>40</td>
<td>1.42</td>
<td>1.16</td>
<td>1.72</td>
<td>1.24</td>
<td>1.79</td>
<td>1.59</td>
<td>1.65</td>
<td>1.42</td>
</tr>
<tr>
<td>50</td>
<td>1.40</td>
<td>1.13</td>
<td>1.69</td>
<td>1.20</td>
<td>1.73</td>
<td>1.57</td>
<td>1.62</td>
<td>1.40</td>
</tr>
<tr>
<td>∞</td>
<td>1.30</td>
<td>1.03</td>
<td>1.60</td>
<td>1.13</td>
<td>1.73</td>
<td>1.51</td>
<td>1.56</td>
<td>1.32</td>
</tr>
</tbody>
</table>

The following points should be noted:

1. Standard errors for the same school size are similar for each test listed. The similarity is indeed remarkable in view of (i) the differing overall zero errors (relative to M.H.T. 40) of these tests (see Chapter 8, Table 1, page 116); (ii) the variation of the time intervals between the construction of M.H.T. 40 and of the other tests listed - a range of from one to nine years; and (iii) the very wide range of schools involved (54 in Table XVI above). It seems reasonable to generalise - to say, for example, that a standard error of approximately 1.7 points is characteristic, for schools of size 15, of teachers' assessments rescaled on Moray House verbal reasoning tests; or, for schools of size 30, one of approximately 1.6.
(2) Equally striking is the initial rapid diminution, in every column, of standard error with increase in school size from 1 to 3, and the subsequent levelling out from size 4 onwards. The graph below, showing the relationship between school size and standard error of a rescaled assessment, is typical of all the nine columns in Table XVI. It is constructed from the average standard errors across all nine columns and, it is suggested, might be used to estimate the mean standard error, at any school size, of teachers' assessments rescaled on any Moray House verbal reasoning test.

In the previous chapter it was shown that, under the rather different conditions of the enquiry there reported, the standard error of a rescaled assessment was relatively insensitive to differences in school size. The results of the present enquiry confirm the earlier findings. These studies give further grounds for belief that the precision of individual rescaled assessments in small schools is higher than that expected by earlier writers.

* The graph is accurately drawn on the inserted page.
STANDARD ERRORS OF RESCALED ASSESSMENTS AND SCHOOL SIZE

\[ m \quad 1 \quad 2 \quad 3 \quad 4 \quad 5 \quad 7 \quad 10 \quad 15 \quad 20 \quad 30 \quad 40 \quad 50 \quad \infty \]

\[ s.e. \quad 3.747 \quad 2.855 \quad 2.480 \quad 2.264 \quad 2.123 \quad 1.960 \quad 1.810 \quad 1.699 \quad 1.633 \quad 1.568 \quad 1.534 \quad 1.510 \quad 1.434 \]

SCHOOL SIZE

S.E.


H.IV - DISCUSSION OF RESULTS

The main outcome of the two preceding studies, one based on Darlington data without a crossover design, the other on Edinburgh and Midlothian data using a crossover design, is in each case an array of standard errors of a rescaled assessment, with each member of an array appropriate to a particular size of school.

The interpretations of these two arrays of standard errors have certain similarities and differences reflecting those of the experimental designs used.

The basic similarities are as follows:

1. In both studies data from each of a number of schools are employed, so that the variability of school test means can be estimated if one test, rather than another, is used for rescaling. It was shown earlier that the contribution of this source of unreliability to the error variance (and hence to the standard error) of a rescaled assessment cannot be neglected.

2. In both studies the error variance of a rescaled assessment is derived without obtaining the rescaled estimate themselves. The derivation is based on the following facts:

   (i) After rescaling, assessment school means are the same as test school means.

   (ii) After rescaling, assessment within-school s.d.'s are the same as test within-school s.d.'s.

   (iii) Arrays of rescaled assessments derived from separate rescalings of the same original assessments within schools are necessarily perfectly correlated.

   These three facts lead to the new and important equation
\[
E(se^2_{\text{rescaled assessment}}) = \sigma^2_{st} + \frac{1}{m} \sigma^2_{ct.s} + \frac{m-1}{m} \sigma^2_{cas}
\]
(see page 195)

(3) In both studies the error variance of a rescaled assessment is found to be relatively stable over a wide range of school sizes. (See Table XI, page 202, and Table XVI, page 221). This result is contrary to opinions expressed by previous writers who have neglected to take into account the proportion \(\sigma^2_{st}\) of test school means error variance \(\left(\sigma^2_{st} + \sigma^2_{ct.s} / m\right)\) which may contribute substantially to that error variance and which is independent of school size.

The main differences* between the two studies are as follows:

(1) In the earlier study, the two tests used were similarly standardised on the year-group of which the data studied are a sample. Hence the expectation of mean test zero error is zero, and that of overall s.d. the same for each test. Any observed differences in these respects are the result of sampling error only. In the later study, the two tests were standardised at different times on different 'national' samples one of which is possibly superior in performance to the other. A test zero error is therefore to be expected, and possibly also a difference in overall s.d.

It follows that in the earlier study the expected difference between an estimate rescaled on test X and the same estimate rescaled on test Y is zero. In the later study the expected difference between the two estimates is the mean test zero error.

* In the Darlington study the test variate consisted of \((VR + E + A)\). In the Edinburgh-Midlothian study it was \(VR\) only. This difference, which has no theoretical implications, is of no importance in the present discussion.
(2) In the earlier study, no attempt was made to allow for practice effect. Standardisation of both tests on the same year-group, although it eliminates the overall mean practice effect, cannot remove differential school practice effect, which inflates the school \( \times \) test interaction. In the later study, the crossover design eliminates the effect of practice on each school test mean individually; hence inflation of the school \( \times \) test interaction by differential school practice effect is excluded. Other things being equal, therefore, the school \( \times \) test interaction will be larger with the first (non-crossover) design than with the second (crossover) one, and it is probable that the variance component which in the first design has been termed \( s^2_{st} \), which contributed to that interaction, is larger also in the first case than in the second. It follows that the error variance of a rescaled assessment, of which \( s^2_{st} \) is part, is likely to be larger without elimination of differential school practice effect than with it.

More concisely, starting from the model on page 213, which is appropriate to the crossover design, and using it in the first (non-crossover) design (that is, not taking advantage of the potentialities of the model to eliminate practice), we readily obtain, in terms of variance components, the following expectation:

\[
E (\text{School } \times \text{ test interaction mean square in the incomplete analysis}) = 2m^* (\sigma^2_{st} + \sigma^2_{so}) + \sigma^2_{ct,s}
\]

However, because the analysis is incomplete, separation of \( (\sigma^2_{st} + \sigma^2_{so}) \) into its two components is not possible, and their sum emerges as a single quantity. What we have called \( s^2_{st} \) in the first (non-crossover) design estimates this quantity, which includes \( \sigma^2_{so} \), a measure of school \( \times \) occasion interaction (i.e. differential school practice effect). We have noted already that this
\( s^2_{st} \) contributes in the non-crossover design, to the error variance of a rescaled assessment, which is therefore inflated to an unknown amount by differential practice effect not isolated in that design.

It must not be concluded, however, that the crossover design is always to be preferred to the earlier one in deriving the error variance of a rescaled assessment. Much depends on the circumstances of the particular case. If those responsible for Darlington’s testing programme wish to know what would happen to the rescaled assessments if test Y is used for rescaling instead of test X (administered 6 weeks earlier), the non-crossover design will provide the right answer. If, however, they had wished to know what would have happened to the rescaled assessments if, on a single occasion of testing, they had employed test Y instead of test X, the crossover design would have furnished a more nearly correct answer.

The latter answer would not however have been entirely correct. It is impossible to know exactly what would have happened had test Y been substituted for test X. Some statements can be made with a fair degree of precision after administering two tests closely together in time in a crossover design. For example, we can say that had test Y been used instead of test X, the year-group mean would have altered by the amount of the estimated zero error. We can also say that the school x test interaction relevant to the substitution of test Y for test X would have been more precisely estimated if the crossover design were used than if it were not. The crossover design enables us to eliminate the mean practice effect from the zero error and the school differential practice effect from the interaction.

What the design does not do—and no design can ever do—is to eliminate the differential practice effect from the individual children’s test scores and
rescaled assessments. To estimate the interaction variances $\sigma_{ct.s}^2$ and $\sigma_{ca.s}^2$, each of which contributes to the error variance of a rescaled estimate, at least two tests must be administered to each child; and for these individual children it is impossible to separate out, and hence to eliminate, the effect of differential practice from the first to the second test. For the hypothetical situation in which test Y is substituted for test X, the estimates of the variance components $\sigma_{ct.s}^2$ and $\sigma_{ca.s}^2$ obtained from the crossover design, are over-estimates, by some unknown – and unknowable – amount. Since both of these variance components are included in the equation leading to the error variance of a rescaled assessment, it follows that the latter is also over-estimated for the substitution situation.

The error variance of a rescaled assessment, obtained in the crossover experiment, is therefore an upper bound to the undiscoverable error variance in the substitution situation.
CHAPTER TWELVE - MAKING ASSESSMENTS AND CHOOSING A RESCALING TEST

A - GENERAL

A.1 - PURPOSE - 'FORWARD-LOOKING' AND 'BACKWARD-LOOKING'

The making of assessments and the choice of a test or examination on which to rescale them are matters of some complexity into which educational as well as statistical considerations enter.

The first consideration is the purpose served by the assessments. As Wiseman (95), page 143 points out, purpose is bound up with validity in the sense - or rather senses - in which psychometrists use the word.

Wiseman makes a distinction between the 'forward-looking' and 'backward-looking' purposes of tests or examinations. This distinction is useful in discussing teachers' assessments. Scores in 'forward-looking' examinations characteristically have more or less predictive validity in respect of a defined criterion as yet in the future; 'backward-looking' examinations characteristically have better or worse content validity in respect of past achievement. Predictive validity is measured by the correlation between scores on the 'forward-looking' examination and scores on the criterion. It is therefore a statistical concept. Content validity is judged by the extent to which the test 'reflect(s) adequately the content of the teaching' or (preferably) 'measure(s) the degree of achievement of aims and purposes' (ibid. page 144). It is thus 'mainly a qualitative rather than a quantitative method: logical analysis and subjective judgement are involved, not measuring and correlating'. (ibid. page 144)
An obvious example of a 'forward-looking' examination is the test, or group of tests, set to 11-year-old children at the end of the primary stage of their education. The primary purpose of this examination is prognostication of a definable type of secondary school success (there is, of course, more than one possible criterion). The examination achieves this purpose in so far as it succeeds in this prognostication. When the criterion is academic success the typical validity coefficient is .85 or even .90 (Emmett (21), Richardson (71), Bosomworth (3), and others); when the criterion is success of a less academic sort the predictive validity is only slightly lower (Emmett (22)).

Primary teachers' assessments are now taking their place alongside the 11+ tests, and almost certainly in the changing educational scene, their use will increase. It has already been pointed out (Chapter 9, page 126) that their predictive validity, after rescaling, appears to be as high as that of the tests, and, according to at least one study (Yates and Pidgeon (96)), even higher.

An example of an examination originally intended to be 'backward-looking' is the 0-level examination taken by some children four or five years after entry to secondary school. This examination achieves its 'backward-looking' purpose to the extent that it possesses content validity. To judge this sort of validity is, however, no easy task. Though Watts and Slater (93) did not mention content validity - an unfamiliar concept at the time they were writing (1950) - the idea of it was in their minds when they listed their objections to the use of an external examination, taken at 15+ or 16+, as a criterion of success subsequent to 11-plus. Cureton (45), page 651) points to the difficulties of establishing content validity, and Wiseman stresses the dangers of relying on it when inadequately demonstrated by means of an attempt to match the content of the
examination to that of the syllabus. The syllabus-content approach, he says, is 'a reactionary instrument helping to encapsulate method within the shell of tradition and accepted practice' (95) page 116. The aim should be to avoid this educationally undesirable outcome by compiling the examination with the aims and purposes of the syllabus in mind rather than the syllabus itself - the spirit rather than the letter.* 'The goal-orientated approach is exactly the opposite: it evaluates learning - and teaching - in terms of the aims of the curriculum, and so fosters critical awareness, good method, and functional content' (ibid. page 116).

Wiseman notes the difficulty of defining these aims in specific terms but insists on the attempt to do so. He would surely agree with Cureton who sums up the same issue as follows: 'Those educators who insist (and rightly, we believe) that other aims** are at least equally important, and in the aggregate probably much more important, could advance their cause most rapidly and effectively by setting about the task of specifying the materials, actions, situations and scoring criteria implied by the abstract terms which define these other aims. They will find the task difficult but in most cases possible. When they have accomplished it they will find that teachers will use the materials, set up appropriate school situations, and teach the desired acts' (43, page 652).

* 'I should like to see all science examinations scored on a rough basis of one-quarter of the marks for technique of presentation, one-quarter for power of selection and arrangement of the relevant material, one-quarter for mastery of the appropriate scientific method, and only the remaining quarter for the knowledge of the facts, the jargon, and the technical methods which now loom so large because though educationally unimportant they are so easy to teach and to assess.'

This was part of the Presidential address of Professor F. R. Winton to the Science Masters' Association in 1949. Evidently he is of the same mind as Professor Wiseman.

** Aims, that is, other than 'command of fundamental processes'.
It seems that those compiling the external examination meet most of the objections of Watts and Slater in so far as they succeed in achieving the aims indicated by Wiseman and Cureton.

Over these issues, as elsewhere, a sense of proportion must be preserved. It is one thing to have reservations about an external examination such as the O-level G.C.E.: it is another to reject it entirely. To obtain at least a modest crop of O-level passes is a necessary, though not a sufficient, condition of grammar school success. Says Wiseman: 'If a pupil fails to achieve passes in two or three subjects at the Ordinary level, then no matter how much he has profited in other directions he has nevertheless failed to justify his selection for a specialised education' (ibid. page 149).

The preceding discussion has relevance for the secondary teachers' assessments, which are made when the O-level examination is taken, and which are made available to the external examiners. In Great Britain these assessments play a deplorably ancillary role. They are used mainly to assist the external examiners to reach decisions about candidates with border-line scores on the examination. We shall return to these secondary assessments shortly. In the meantime, there must be considered a further aspect of the examination situation which is relevant to both primary and secondary teachers' assessments.

It has just been said that the G.C.E. O-level examination is 'backward-looking'. But this is an over-simplification. 'Backward-looking' in intention, it is also 'forward-looking' in use. It marks not only the end of the first phase of an academic course, but also the beginning of another. Wiseman points out that schools use it as a gateway to the sixth form. It is a fact also that though 'Matriculation' is a thing of the past, the 15- or 16-year-old school
leaver is still liable to find certain jobs open to him only if he possesses O-level passes in specified subjects. It is scarcely necessary to mention the present use of A-level success as an essential pre-requisite to University entry, a not unsuccessful 'forward-looking' use of another examination also intended to be 'backward-looking'.

In a publication as yet little known in this country, Bloom and Peters ((2), page 119) have this to say: 'In effect, the A.P.S.* operates on the thesis that if we know the past academic history of the student we can predict what he will do in the future.' Wiseman (ibid. page 151) remarks rather more specifically: 'Any achievement examination at the end of a more or less self-contained phase of education, and followed by different, selective, higher stages of education and training, must play its part in selection for these stages. Both experience and research tell us that, in selecting for specialised training or education, we ignore at our peril evidence of progress and achievement in previous courses.'

It is not the present intention, nor is this the place, to discuss at length the philosophical issues raised by these two statements. It may be said quite briefly, however, that prediction of this sort is of more general application. 'The human predictor interprets the signal by an act of recognition which puts it into some general category. We then assume that the future will have some general likeness with futures we have met before which followed this kind of signal, and this is the kind of future we prepare for.' (Bronowski (4) page 119). The writers quoted earlier would surely agree with Bronowski when he

* Academic Prediction Scale
says: 'The predictor ... does not assume the future already to exist, to be dredged or conjured up in advance at our bidding. It makes no larger claim than that the future can in general be predicted, within defined limits of uncertainty. And since there are uncertainties, the prediction will sometimes be wrong.' (ibid. page 115).

An intentionally 'backward-looking' examination, then, with good content validity in some area, will ipso facto give rise to marks possessing predictive validity in the same area. This is the reason for the general observation that subject examinations are on the whole the best predictors of later success in the same or related subjects. A reasonable extrapolation from this observation is that the better the content validity of the examination, the higher will be its predictive validity.

Is the converse true? Is the deliberately 'forward-looking' 11+ battery of tests indebted to its content validity for its undoubted success (when soberly assessed) in prediction?

There is no simple answer to the question. Although prediction is the aim of these tests, those constructing them do have in mind content validity in the sense defined above. This is perhaps seen at its best in the more imaginatively constructed English attainment tests. It is relevant too that test constructors are now beginning to compile tests of arithmetic attainment that will reflect, and in the best possible way 'feed back' and reinforce, the aims underlying the changes at present taking place in the teaching and learning of arithmetic in the primary schools. Although normally justified in terms of construct validity (high loading in the general ability factor) the 'intelligence' or reasoning test may be legitimately assessed also from the viewpoint of content validity.
The ability to perceive and handle relations, which it aims to test, is no longer to be thought of in terms of immutable innate capacity - if indeed it ever was by informed educationists and psychologists. It is rather to be thought of as an ability central to understanding in most areas of learning, an ability which it is the responsibility of every teacher to foster in his pupils within the framework of his subject.

In this sense, then, it may be claimed that predictive validity is achieved in these tests by virtue of their content validity. But the constructor of 'objective' tests has another aim which, though of less educational significance, is nevertheless important. For reasons too well known to need restatement here, the tests he constructs must be reliable in the several senses in which the psychometrist uses the word. It is indeed for these reasons that the test is 'objective'.

It is unfortunate, though perhaps inevitable, in the climate of anxiety engendered by selection for secondary schools, the chief instrument of which is 'the 11+', that the educational and non-educational requirements of the tests should conflict. The result has been that many teachers, often with the best of intentions, have subordinated the first to the second, and have unduly practiced or coached their pupils in the method of answering test questions, to the detriment of their pupils' liberal education. Unless test constructors can find some way of 'building into' their tests safeguards against this misuse and - unless there is a general lowering of the temperature over selection, these ills will continue.

There are signs, indeed, that the temperature will fall. A number of local authorities have 'abolished the 11+'. Often this means, in practice, that whatever modifications they make in their secondary schools arrangements, they still
use tests, but instead of relying entirely on test results, they now supplement them with teachers' assessments rescaled on the test scores. Though it is to be feared that the teachers may replace the tests as whipping-boy, the educational effects of the change ought to be beneficial. With the burden of sole responsibility for the outcome removed from the tests, one may hope that the pressure to expend school (and extra-school) time in preparing for them will decrease. Their 'built-in' content validity, no longer distorted by the effects of wrong-headed and time-consuming coaching and practice, should improve, and their predictive validity rise in consequence.*

One further advantage may be looked for. Already teachers' rescaled assessments are known to be highly valid predictors of later success. This degree of validity has been achieved even though the assessments have been rescaled on school test means and s.d.'s which, for the reasons given above, are probably less accurate than they might be as indicators both of the relative standing of the schools, and of their within-school dispersions. It is reasonable to hope that the predictive validity of assessments will rise if these school means and s.d.'s become more accurate as the distortions brought about by differential coaching and/or practice diminish.

B - PROCEDURES FOR MAKING ASSESSMENTS

B.I - MAKING 'll-PLUS' ASSESSMENTS

We return to teachers' assessments. Despite their increasing use in ll+

---

* This is not to say that coaching and/or practice should disappear entirely. There is a place for test preparation, properly controlled. Reference has been made elsewhere to Ortar's findings that controlled practice increases test validity (see page 210).
procedures there appears to be little uniformity in the methods by which they are made. Some authorities ask their teachers simply to assess for academic potential. 'Rank the pupils in order of potential academic success' is the gist of a fairly typical instruction to the teacher. Occasionally this is elaborated a little. 'In your ranking take into account personality factors you consider relevant'. Other authorities prefer specific 'subject' assessments, because they believe them to be more valid than 'general suitability' estimates, or because such assessments can be combined in different ways with alternative criteria in view. Since Vernon et al (91) page 137 summarised the position in 1957, little change has been reported.

The plan, sketched in bare outline below, is an attempt to help teachers prepare their original assessments in a more systematic fashion. It makes use of the 'forward-looking' and 'backward-looking' concepts in the context of assessments. *

We shall take it that whether these assessments are 'forward-looking' or 'backward-looking' they are numerically expressible in terms of marks or ranks.

Where the purpose of the assessments is 'forward-looking', the crucial evidence of their efficiency in achieving this purpose is their predictive validity in respect of some defined criterion. Obviously this crucial evidence is itself the criterion, and in practice the assessments must be made in the light of information available at present. It has been shown that the 'forward-looking' and 'backward-looking' aspects of examinations are closely related and we assume this to be true also of teachers' assessments. In the words of Bloom and Peters the teacher predicts what his pupils 'will do in the future' in the light of their

---

* See page 228.
'past academic history'. To paraphrase Bronowski*, he must assume that their academic futures will have some general similarity to academic futures which have previously followed these kinds of academic history. Since the similarity is not more than general, since no two children are exactly alike, and since a subsequent change in circumstances may alter the criterion, it is accepted that he must at times be wrong in his prediction.

To reduce these errors of prediction to a minimum, the teacher must first study the academic history of his pupils - the 'evidence of progress and achievement in previous courses', as Wiseman** puts it, in the best possible way. It is suggested that he will find it useful to adapt the procedure Wiseman and Cureton independently recommend for the compilation of an external examination of good content validity. The plan, in barest outline, is as follows.

The teacher should first define the aims and purposes of the curriculum his pupils have been following. Indeed, if he is a good teacher he will have done this already. It scarcely needs saying that these aims should include the general second-order aim of relating the specific purposes of the curriculum his pupils have been following under his tutelage to those of the curricula they are to study later.

He should then devise a number of questions which, in his view, have good content validity for his classroom situation. These questions he will both ask and answer himself. Of course, the questions need not be similar to those set in an external examination. For instance, one of the aims of the curriculum might be the development of his pupils' capacity to handle mathematical concepts, as distinct from arithmetic processes, a capacity highly relevant to success in later stages of

* See page 232.

** See page 232.
education. In this case, he might ask 'how individual children succeeded with the examples he devised after reading Sealey's book.' Obviously he will use records extensively, but he need not be tied to them. Many of his questions he will 'mark by impression', in whole or in part. In answering such questions as these with his pupils successively in mind, he will judge the degree to which he has helped them to achieve the aims and purposes of the curriculum.

Finally, he should express these judgements as numerical assessments. The only new aspect of this procedure is its insistence on systematically relating the assessments, via relevant questions, to underlying aims. All teachers do something like this in evaluating their pupils' progress, but where important decisions hinge on the result, some systematisation seems desirable.

This is as far as 'backward-looking' will help the teacher in making his assessments. But they are 'forward-looking' in intention, and, as was said earlier, the crucial test is their predictive validity in respect of the chosen criterion. The teacher needs therefore to be told in due course of the outcome for each of his pupils in the next stage of education. He should find this information useful in two ways. Firstly, it should enable him to identify patterns associated with future success or failure, and so to recognise these patterns as valid 'signals' in prediction for other pupils later on. Secondly, knowledge of the outcome should 'feed back' and so help him to review his aims and purposes, and hence to modify both his curriculum and the questions he asks in arriving at his assessments.

B.II - ASSESSMENTS ACCOMPANYING O-LEVEL AND A-LEVEL EXAMINATIONS

It is customary for secondary teachers to prepare assessments of their pupils' relative standings in various subjects when these pupils take O-level and A-level examinations. So far as we can discover, scant attention is paid to these assessments by Examining Boards. One hesitates therefore to suggest that teachers should take more trouble over them than they must. However the conviction is increasingly wide-spread that the opinions teachers have of their pupils' abilities and potentialities should carry more weight than they do. It is significant that the new C.S.E. Examining Boards give effect to this conviction in the heavy reliance they place on teachers' judgements of their pupils. Perhaps the older Examining Boards will be impelled to follow their example. Hence proposals for systematising the procedure to be followed in arriving at assessments at both of these stages may not be out of place.

This procedure is very similar to that proposed in the previous paragraphs. There is indeed little to add except in justification of that similarity. The purposes served by teachers' assessments, were they allowed to play their proper part in O- and A-level examinations, are similar to those of the external examinations that pupils take at these times. Despite the alleged terminal roles played by these examinations, they are de facto predictive instruments. Whatever the opinion of the S.S.E.C., O-level results are not only a record of past achievement, but (as said earlier) the gateway to the sixth form. Whatever the view of the Crowther Committee, the A-level examination is not inaccurately described as a university entrance examination.

These examinations, then, are both 'backward-looking' and 'forward-looking'. So too should be the teachers' assessments that accompany them. The procedure suggested previously is therefore appropriate in these cases also.
CHOOSING A RESCALING TEST

C.I - PREVIOUS WORK

The most extensive accounts yet published of British follow-up studies in which rescaled assessments are used are those of McClelland (54), and Yates and Pidgeon (96). McClelland's study, first published in 1945, is an acknowledged classic, and the enquiry of Yates and Pidgeon, which followed in 1957, is a not unworthy successor. From America comes the more recent study of Bloom and Peters (2), who are concerned with the selection of high school pupils for college. McClelland appears to have assumed that the best rescaling test available to him was the Qualifying Examination (54, Appendix A, page 245) in English and arithmetic, an 'old-style' examination. 'We use as our standard or comparison the teachers' marks scaled by the soundest possible method, that is, on the Qualifying examinations scores' (54, page 45). He offers no justification, however, for believing this to be the 'soundest possible method'. He concludes (ibid. page 47) that 'the only satisfactory procedure is to scale the estimates separately for English and arithmetic, and that the estimates should be scaled upon the results of a uniform examination'. He explicitly rejects the intelligence test because the school means on this test do not correlate perfectly with the Qualifying examination school means - \( r = 0.92 \) for the 12 school means in his Table XVIII (ibid. page 41). Yates and Pidgeon, however, prefer the verbal intelligence test alone, on the grounds that it is less affected by differences in standards of teaching (96, page 87). As a supplementary argument they adduce the similarity of two correlations: (i) that (within schools) between
teachers' judgements of the suitability of their pupils for grammar school education' (ibid. page 87) and verbal intelligence scores alone; and (ii) that (within schools) between the same judgements and the composite of intelligence, English and arithmetic. Vernon et al, on the other hand, believe the composite to be better than the intelligence test alone, 'because the battery is more valid than any single test' ( (91), page 145).* Bloom and Peters find higher predictive validities for college first-year performance after rescaling high school grades on the 'aptitude' tests than when the rescaling is based on 'achievement' tests ( (2), page 81), and prefer to rescale specific subject grades on the aptitude test.

The findings of these writers are thus inconclusive and contradictory. It is evident that there is room here for further research. The suggestions which follow are intended to indicate one line along which research might be directed.

C.II - THE 'IDEAL' RESCALING TEST

Of the writers mentioned above, Vernon et al come nearest to the basic requirement. 'The crucial factor in improving the accuracy of scaling is the validity of the external test or tests against which the scaling is done' ( (91), page 145). So far as it goes this is sound enough, and in the absence of more specific information, it is probably as safe a guide as any to the choice of a rescaling test. However, as will be shown below, the argument could be misleading.

(a) Validity of Test Scores

The overall predictive validity of test scores from a number of schools is

* It will be shown (page 243) that this argument is of doubtful validity.
expressed as a single correlation coefficient between the whole array of these scores and the whole array of criterion scores. Within these arrays, however, heterogeneity may occur which will affect the validity of assessments rescaled on the test scores. The discussion of this heterogeneity will be clearer if we think specifically of 1L+ allocation.

If the whole arrays of both test and criterion scores are broken down into subsets corresponding to the primary schools from which the pupils come, the following determiners of overall validity of test scores may be identified:

(i) Correlation between school test and criterion means.

(ii) Correlation between school test and criterion s.d.'s.

(iii) Correlation within schools between individual pupils' test and criterion scores.

Each of these correlations contributes to the overall validity of the test, though not independently, as an examination of the equation in the footnote will show.*

---

* If \( x \) and \( c \) are the test and criterion scores of a child in a school of size \( m \) with means \( \bar{x} \) and \( \bar{c} \), and s.d.'s \( s_x \) and \( s_c \); and if \( r_{xc} \) is the within-school correlation; then it is easy to show that the mean within-child variance of the discrepancies \( (x - c) \) for this school is:

\[
s_{(x-c)}^2 = \frac{1}{2m} \left[ \left( \frac{\bar{x} - \bar{c}}{m} \right)^2 + (m - 1) \left( s_x - s_c \right)^2 + 2(m - 1) s_x s_c (1 - r_{xc}) \right]
\]

The three terms in the square brackets show exactly how the defect from perfect overall correlation comes about. The first relates to discrepancies between a school's mean on \( x \) and \( c \); the second to discrepancies between the corresponding s.d.'s; and the third to discrepancies between pupils' individual scores within a school. These three sources of discrepancy relate respectively to (i), (ii) and (iii) on page 242 of the text. The effect should be tried of putting \( x = c \); \( s_x = s_c \); and \( r_{xc} = 1 \).
(b) **Validity of Rescaled Assessments**

If we now consider assessments rescaled, for example, by Method 3 (that is, using test school means and s.d.'s for the rescaled assessments), we see that only (i) and (ii) above are relevant to the validity of the rescaled assessments. The within-school correlation between test scores and criterion scores is replaced by that between assessments and criterion scores. That part of the overall test validity resulting from the correlations in (i) and (ii) is transferred unchanged to the overall assessment validity; that resulting from the correlation in (iii) is not. Therefore, a test with satisfactory correlations of types (i) and (ii) is always a good rescaling test, even though its overall validity may be low because of low type (iii) correlation. On the other hand, a test with poor correlations of types (i) and (ii) is always a bad rescaling test, even if it has high overall validity because of high type III correlation. It is clear that the argument of Vernon *et al* quoted earlier is incomplete and for that reason possibly misleading.

A more precise formulation of their statement that validity is the crucial factor is as follows:-

The best rescaling test to use is that which bestows on the assessments rescaled on it maximum validity in respect of the criterion to be predicted.

The statistical equations which follow on adopting this point of view are simple and will be worked out shortly. At this point we shall anticipate, stating the specification of the 'ideal' rescaling test $A$ as follows:-

(i) Over a set of $L$ schools, the school means $\bar{a}_k$ ($k = 1, 2, \ldots, L$) on the
'ideal' rescaling test are identical with the corresponding criterion means \( \bar{c}_k \); that is, \( \bar{a}_k = \bar{c}_k \).

(ii) The school s.d.'s \( s_k \) on the 'ideal' rescaling test are linear functions of the school s.d.'s \( s_k \) on the criterion. The constants of proportionality are the within-school correlation coefficients \( r_{cz} \) between original assessments \( z \) and criterion scores \( c \); that is, \( s_k = r_{ck} z_k s_{ck} \).

The rescaling equation embodying this specification is

\[
y_a = \bar{c} + r_{cz} \cdot \frac{s_c}{s_z} (z - \bar{z})
\]

in which \( y_a \) is the assessment after rescaling on the 'ideal' test \( A \); \( z \) is the corresponding original or 'raw' assessment; \( \bar{z} \) and \( s_z \) are the original school mean and s.d.; \( \bar{c} \) and \( s_c \) are the criterion school mean and s.d.; and \( r_{cz} \) is the within-school correlation between assessments (either before or after rescaling) and the criterion.

Assessments rescaled on this 'ideal' test will have maximum validity in respect of the criterion predicted and the error variance of prediction of \( c \) by \( y_a \) will be a minimum. A fuller discussion of the properties of the 'ideal' rescaling test and its implications follows in the 'statistical' section which follows.

C.III - IMPLICATIONS FOR RESEARCH

When the criterion scores become available, the specification of the 'ideal' rescaling test can help in the empirical search for tests which, when suitably standardised overall, conform as closely as possible to this specification, which may be briefly summarised:-
For school \( k (k = 1, 2, \ldots, \ell) \),

\[
\bar{a}_k = \bar{c}_k; \quad s_a = r_{ck} s_k s_{ck}
\]

The following lines of research are suggested by this concept of the 'ideal' test:

(i) A review of previous follow-up studies. The object would be to discover which tests or batteries giving school means and school s.d.'s which correlated best with the criterion means and s.d.'s for the same groups of testees.

(ii) Previous follow-up studies will not have been designed with the concept of the 'ideal test' in mind, and may not be entirely appropriate for this line of enquiry. In the design of further projected follow-up studies this defect should be remedied by employing a suitable design to highlight the correlations in (i).

(iii) In all these enquiries, the within-school correlations between original assessments and criterion should be studied.

For this study to be worth while, a large number of original records would need to be consulted, and a great deal of new work would have to be done. Almost certainly the co-operation of a sizeable group of investigators would be necessary. But the results might well repay the effort. At present, all that we have are the overall validities (encouraging enough) from a number of unco-ordinated studies. If we are to discern any patterns that may exist, we shall need to know, from each of many follow-up studies, the proportions of the covariance between rescaled assessments and criterion scores coming from: (a) correspondence between school and criterion means; (b) correspondence between school and criterion s.d.'s; and (c) within-school correlations for individual children.

It must be recognised that these three statistics determine the efficiency of the rescaling procedure, and we should aim at evaluating the relative contribution of each to that efficiency.
We do not know in advance whether any patterns of relationship will emerge between scores on particular tests competing for the role of rescaling test and scores on particular criteria; or whether the within-school correlations between the assessments concerned and these criteria are consistent enough to be useable. However, the fact that rescaled assessments have been shown in several enquiries (54), (96), (2) to have greater predictive validity than the original assessments provides reason to believe that such patterns and consistent correlations do exist and that they may be discoverable.

If so, it may be possible to reach the point when knowledge of the purpose assessments are to serve in a particular situation will enable us to recommend the use of a particular rescaling test - one that in a systematic study of the sort sketched above has been shown to conform most nearly to the specification of the 'ideal' rescaling test for that situation. It may even be possible to recommend a particular rescaling equation if reports of previous studies of similar situations have shown within-school assessment criterion correlations consistent from one study to another.

D - STATISTICAL DEVELOPMENT

In this section we examine the conditions under which the overall validity of rescaled assessments is maximised, and other related matters.

D.I - VALIDITY OF RESCALED ASSESSMENTS: THE GENERAL CASE

By methods similar to those used previously (see page 167, F.I. Chapter 11), it is easy to show that the overall validity of rescaled assessments is in general
\[
rcy(T) = \frac{\Sigma \bar{mc}y + \Sigma (m - 1) \bar{r}_{cz} s_c s_y}{(\Sigma m c^2 + \Sigma (m - 1) s_c^2) (\Sigma m y + \Sigma (m - 1) s_y^2)}
\]  

(1)

in which \(c\) and \(y\) are criterion and rescaled assessment means for a particular school with \(m\) children; \(s_c\) and \(s_y\) are the corresponding s.d.'s for that school; and \(r_{cz}\) is the correlation* between criterion scores and original assessments \(z\) in the same school. The summation signs denote summation over schools, and relate to every factor in the terms they precede.

**D.II - THE 'IDEAL' RESCALING TEST**

We obtain next the characteristics of the 'ideal' rescaling test. Rescaling on such a test will maximise the overall validity of the assessments and will also minimise the error variance (or error sum of squares) of prediction of the criterion scores by the rescaled assessments.

Suppose we have such a test \(A\), scores on it being \(a\). Then, after rescaling on \(A\)** by Method 3, an original assessment \(z\) becomes:-

\[
y_a = \bar{a} + \frac{s_a}{s_z} (z - \bar{z})
\]  

(2)

also, \(c = \bar{c} - (c - \bar{c})\)

The error of prediction of \(c\) by \(y_a\) is

\[
(c - y_a) = (\bar{c} - \bar{a}) + (c - \bar{c}) - \frac{s_a}{s_z} (z - \bar{z})
\]

Squaring and summing over the \(m\) errors in one school, we obtain

\[r_{cz} = r_{cy}\text{, since a linear change of scale does not affect the correlation.}\]

\[**Again we assume a general mean of zero for both test and criterion.\]
\[ \Sigma (c - y_a)^2 = m (\bar{c} - \bar{a})^2 + (m - 1) (s_c^2 + s_a^2 - 2s_c s_ar_c) \]  \hspace{2cm} \text{(3)}

Over all schools:

\[ \Sigma \Sigma (c - y_a)^2 = \Sigma m (\bar{c} - \bar{a})^2 + \Sigma (m - 1) (s_c^2 + s_a^2 - 2s_c s_ar_c) \]  \hspace{2cm} \text{(4)}

Using partial differentiation in equation (3) with respect to \( \bar{a} \) and \( s_a \), and equating the derivatives to zero, we obtain

\[ \frac{\Sigma (c - y_a)^2}{\bar{a}} = -2m (\bar{c} - \bar{a}) = 0 \]

\[ \frac{\Sigma (c - y_a)^2}{s_a} = 2(m - 1) (s_a - r_{cz} s_c) = 0 \]

whence \( \bar{a} = \bar{c} \), and \( s_a = r_{cz} s_c \) \hspace{2cm} \text{(5)}

A test \( A \) having these characteristics (\( \bar{a} = \bar{c}; \ s_a = r_{cz} s_c \)) is the 'ideal' rescaling test. School by school, the test mean (\( \bar{a} \)) and the criterion mean (\( \bar{c} \)) are the same, and school by school, the test s.d. (\( s_a \)) is the product of the 'raw' assessment validity (\( r_{cz} \)) and the criterion s.d. (\( s_c \)). The error variance of prediction of the criterion by assessments rescaled on a test with these specifications is an absolute minimum.

**D.III - MAXIMUM VALIDITY OF RESCALED ASSESSMENTS**

The overall validity of the rescaled assessments is maximised when the rescaling test is the 'ideal' test \( A \). Using equations (1) and (5), we obtain the absolute maximum overall validity:-
Equation (6) is useful in two ways. Firstly, the inclusion of \( r_{cz} \) highlights the importance of teachers' capacities to predict successfully, each for his own school, the positions on the criterion of the children in that school.

Secondly, for a given efficiency of within-school prediction, equation (6) provides an upper limit to the overall validity attainable in practice when, as perforce it must be, rescaling is based on a test whose specifications do not tally with those of the 'ideal' test.

D.IV - MINIMUM ERROR VARIANCE OF PREDICTION: 'UNAVOIDABLE ERROR'

Error variance of prediction is complementary to validity. The minimum error variance of prediction will next be derived.

Substituting the optimal values of \( \bar{a} \) and \( s_a \) from equation (5) in equation (3), and dividing by the degrees of freedom, we obtain the absolute minimum error variance for a single school:

\[
\text{One-school EV of prediction} = s_c^2 (1 - r_{cz}^2), \quad \text{. . . . . . (7)}
\]

an equation with a familiar form.

Summing over all schools and dividing by the 'total' degrees of freedom, we obtain the absolute minimum overall error variance of prediction of the criterion by 'ideally' rescaled assessments:

\[
* \quad \text{Upper limit (teachers all predict perfectly): } r_{cy(T)} = 1
\]

\[
\text{Lower limit (teachers predict at random): } r_{cy(T)} = \left( r_B / r_T \right)
\]
Total EV of prediction = $\sum (m - 1) s_c^2 (1 - r_{cz}^2) / (\Sigma m - 1)$ ... (8) (minimum)

Equation (7) gives the 'unavoidable' error variance for a particular school and equation (8) that for the whole range of schools.

D.V. - 'AVOIDABLE' ERROR VARIANCE

'Avoidable' error variance is the error variance that would have been avoided if the 'ideal' rescaling test had been substituted for that used. 'Avoidable' error variance is thus the difference between the observed and the minimum error variances.

If the rescaling test is X, and $y_X$ is an assessment rescaled on X, then

$$y_X = \bar{X} + \frac{s_X}{s_Z} (z - \bar{z})$$

By a development parallel to that giving equation (3), page 248, we obtain

$$\sum (c - y_X)^2 = m (\bar{c} - \bar{x})^2 + (m - 1) (s_c^2 + s_x^2 - 2r_{cz} s_c s_x)$$ ... (9)

whence

One-school EV of prediction = $\frac{m}{m - 1} (\bar{c} - \bar{x})^2 + (s_c^2 + s_x^2 - 2r_{cz} s_c s_x)$ ... (10)

(observed)

Subtracting equation (7) from equation (10), we obtain

'Avoidable' Error Variance of prediction (one school)

$$= \frac{m}{m - 1} (\bar{c} - \bar{x})^2 + (s_x^2 - r_{cz} s_c^2)^2$$ ...

* 'Unavoidable' in the sense that it cannot be decreased further for this degree of within-school efficiency of prediction by the teachers ($r_{cz}$).
which shows exactly how 'avoidable' error variance is the result of the actual rescaling test's departures from the specification of the 'ideal' test - the extent to which \( \bar{x} \) differs from \( \bar{c} \) and \( s_x \) from \( r_{cz} s_c \).

The overall 'avoidable' error variance of prediction is therefore:

\[
EV_{\text{overall}} = \sum (\bar{c} - \bar{x})^2 + \sum (m - 1) (s_x - r_{cz} s_c)^2 / (2m - 1) . \cdot \cdot (12)
\]

D.VI - ERROR VARIANCE OF PREDICTION BY THE RESCALING TEST

Discussion has been concentrated until now on the predictive validity of the rescaled assessments. We turn now to that of the rescaling test itself, as reflected in the error variance of prediction associated with its use.

The equation for the test \( X \) may be written

\[
x = \bar{x} + (x - \bar{x})
\]

From this starting point we proceed as before and obtain:

'Avoidable' error variance of prediction by the rescaling test (one school)

\[
= \frac{m}{m - 1} (\bar{c} - \bar{x})^2 + (s_x - r_{cz} s_c)^2 + 2s_c s_x (r_{cz} - r_{cx}) \cdot \cdot \cdot (13)
\]

In this case the 'avoidable' error variance consists of three parts. The first two taken together constitute the 'avoidable' error variance of prediction by the assessments rescaled on \( X \) (see equation (11)); the third is a function of the difference between the within-school validities of assessments and test scores.

Subtracting equation (11) from equation (13), we obtain:

Difference between 'avoidable' error variances of prediction (1) by assessments
rescaled on X and (ii) by the scores on X themselves (one school)

\[ 2s_c s_x (r_{cz} - r_{cx}) \]  \hspace{1cm} (14)

Equation (14) enables us to compare, for each school, the validities of assessments rescaled on a test and the test scores themselves. The difference in error variance of prediction is positive in the direction \((EV_{asst} - EV_{test})\) if \(r_{cz} > r_{cx}\); in this case the validity of the assessments is the greater. If \(r_{cz} < r_{cx}\), the difference is negative; in this case the validity of the test scores is the greater. The corresponding overall difference between the error variance of prediction by assessments and by test scores

\[ 2\Sigma (m - 1) s_c s_x (r_{cz} - r_{cx}) / (\Sigma m - 1) \]  \hspace{1cm} (15)

D.VII - REDUCTION OF ERROR VARIANCE OF PREDICTION AS A RESULT OF RESCALING

Hitherto it has been tacitly assumed that the rescaling procedure is worthwhile. It is essential that this assumption be examined, and appropriate that it be examined at this stage, since in doing so methods will be used closely similar to those in the preceding paragraph.

In terms of predictive validity in respect of a criterion, rescaling is profitable if, and only if, the overall validity of assessments after rescaling is higher than that of the original assessments. In terms of error variance of prediction of the criterion, rescaling is profitable if, and only if, the overall error variance of prediction is less, after rescaling, than before rescaling. Here we shall work with error variance.
An original assessment may be written as
\[ z = \overline{z} + (z - \overline{z}) \]
The same assessment after resealing on X becomes
\[ y_x = \overline{x} + \frac{sx}{sz} (z - \overline{z}) \]
The criterion score corresponding to the assessment is
\[ c = \overline{c} + (c - \overline{c}) \]
From these equations we derive, for a single school:
\[ \Sigma (c - z)^2 = m (\overline{c} - \overline{z})^2 + (m - l) \left( s_c^2 + s_z^2 - 2r_{cz} s_c s_z \right) \] \[ \Sigma (c - y_x)^2 = m (\overline{c} - \overline{x})^2 + (m - l) \left( s_c^2 + s_x^2 - 2r_{cx} s_c s_x \right) \]
Rewriting the last two equations in a slightly different form and dividing by the degrees of freedom we obtain:
\[ \overline{EV} (c - z) = \frac{m}{m - l} (\overline{c} - \overline{z})^2 + \left\{ (s_z - r_{cz} s_c)^2 + s_c^2 (1 - r_{cz}^2) \right\} \]
and
\[ \overline{EV} (c - y_x) = \frac{m}{m - l} (\overline{c} - \overline{x})^2 + \left\{ (s_x - r_{cx} s_c)^2 + s_c^2 (1 - r_{cx}^2) \right\} \]
The corresponding overall error variances are:
\[ \overline{EV} (c - z) \text{ (total)} = \]
\[ \left[ \Sigma m (\overline{c} - \overline{z})^2 + \Sigma (m - l) \left\{ (s_z^2 - r_{cz} s_c)^2 + s_c^2 (1 - r_{cz}^2) \right\} \right] / (\Sigma m - l) \]
\[ EV(c - y_x) \text{(total)} = \left[ \Sigma_m (\bar{c} - \bar{x})^2 + \Sigma (m - 1) \left\{ (s_x - r_{cz} s_c)^2 + s_c^2 \left( 1 - r_{cz}^2 \right) \right\} \right] / (\Sigma m - 1) \] (21)

Our conclusions from these illuminating equations are as follows:

(i) In so far as the error sum of squares is the result of discrepancies between assessment and criterion means, comparison of the first terms on the right of equations (18) and (19) shows that rescaling is profitable if

\[ |\bar{c} - \bar{z}| > |\bar{c} - \bar{x}| \]

(a)

(ii) In so far as the error sum of squares is the result of discrepancies between assessment and criterion s.d.'s within schools, comparison of the first terms in the 'curly' brackets on the right of these equations shows that rescaling is profitable if

\[ |s_z - r_{cz} s_c| > |s_x - r_{cz} s_c| \]

(b)

It will be seen that discrepancies are at issue between (i) obtained means (before and after rescaling), and 'ideal' means, and between (ii) obtained s.d.'s (before and after rescaling), and 'ideal' s.d.'s. A smaller absolute difference between \(\bar{c}\) and \(\bar{x}\) than between \(\bar{c}\) and \(\bar{z}\) will favour rescaling, as will also a smaller absolute difference between \(s_x\) and \(r_{cz} s_c\) than between \(s_z\) and \(r_{cz} s_c\).

The differences for means are independent of the differences for s.d.'s. Hence it can happen that the first may work in one direction and the second in the other. It seems a priori probable that the differences will be in the directions shown at (a) and (b). The examiners' criterion scale is presumably

* Absolute values, regardless of sign.
common to all schools, while that of the rescaled assessments has been made common. For the original assessments, however, there is no common scale.

Nevertheless, there is no certainty that the error variance of prediction will be reduced by rescaling. The previous theoretical development makes clear the conditions under which reduction will occur; whether or not these conditions obtain so that reduction does occur must be settled by empirical enquiry.

E - NUMERICAL EXAMPLES

We use again the Darlington data employed previously to illustrate different methods of rescaling (see Chapter 10, page 149).

School by school computations based on equations (18)-(21) would be laborious. The work can be simplified, however, as follows:-

Starting from equation (16), summing over all schools, and rearranging terms, we obtain, for the criterion and the original estimates:-

$$\Sigma (c - z)^2 =$$

$$\Sigma \bar{c}^2 + m \bar{z}^2 - 2m \bar{c} \bar{z} + \Sigma \left[ \Sigma (c - \bar{c})^2 + \Sigma (z - \bar{z})^2 - 2 \Sigma (c - \bar{c})(z - \bar{z}) \right]$$

Abridging the notation as in 3(a) - 10(a) (Chapter ), we have

$$(C - Z)_T = \left[ C_B + Z_B - 2 (CZ)_B \right] + \left[ C_W + Z_W - 2 (CZ)_W \right]$$

$$= C_T + Z_T - 2 (CZ)_T \quad \ldots \quad \ldots \quad \ldots \quad \ldots$$  (22)

Similarly, for the criterion and the rescaled assessments, we can now write down, or obtain from equation (17)

$$(C - Z)_T = C_T + Z_T - 2 (CZ)_T \quad \ldots \quad \ldots \quad \ldots \quad \ldots$$  (23)
It will be recalled (see pages 131 to 134), that $C_T$ and $X_T$ are the total sums of squares for criterion and rescaling test, and that $(C_Y)_T$ is the total sum of products for the criterion and rescaled assessments. These have been computed and reported previously (see page 154). We still have $Z_T$ and $(C_Z)_T$ to compute for the original estimates.

Before doing so, however, it should be noted that the rescaling test, and hence the rescaled assessments, have been standardised on the same basis as the criterion, so that the total sums of squares are very nearly the same for all three. In order to make a proper comparison with the original assessments the latter (expressed as percentages in the first place) must also be put on the same overall scale so as to have the same total sum of squares. This procedure (which amounts to multiplying all assessments by a constant) in no way changes the relative magnitudes of the school means, or of the school s.d.'s, for the original assessments; it merely alters all the teachers' initial arbitrary scales in exactly the same proportion so that for all schools combined the overall mean and overall s.d. of the original assessments are the same as for the rescaled assessments. In particular it should be noted that this overall change of scale for the original assessments does nothing towards the alignment of the individual school means and s.d.'s with those of the rescaled assessments obtained by the normal piece-meal school-by-school rescaling procedure. It does however, impart meaning to comparisons between original and rescaled assessment school means and s.d.'s.

The labour is shortened and the rescaling procedure entirely avoided by making use of the relations:-
\[ z_1 = k z_2; \quad \Sigma \Sigma z_1^2 = k^2 \Sigma \Sigma z_2^2; \quad \Sigma \Sigma z_1 c = k \Sigma \Sigma z_2 c, \]
in which \( z_1 \) and \( z_2 \) are original assessments on the first and second scales respectively, and \( k \) is the constant of proportionality common to all schools.

An analysis of variance between and within schools is therefore carried out for the original assessments (without rescaling); and an analysis of covariance between and within schools for the original assessments (without rescaling) and the criterion scores. We next obtain \( k^2 \) as the ratio between the total sum of squares for the rescaling test and that for the original assessments. Finally the analysis of variance just completed is adjusted by multiplying each of its three sums of squares by \( k^2 \); and the corresponding analysis of covariance by multiplying each of its three sums of products by \( k \). The outcomes, with much less labour, are the analyses of variance and covariance that would have been obtained if we had first adjusted all the original assessments by multiplying them by \( k \).

The results of these computations are reported below, together with the other analyses required for the comparisons to be made.

**TABLE I: SUMS OF SQUARES AND SUMS OF PRODUCTS**

For \( y, z \) and \( c \).

18 schools; 935 children.

<table>
<thead>
<tr>
<th>Source</th>
<th>SS (y)</th>
<th>SS (z)</th>
<th>SP (yc)</th>
<th>SP (zc)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Schools</td>
<td>20,334</td>
<td>1,397</td>
<td>20,223</td>
<td>-2,003</td>
</tr>
<tr>
<td>Within Schools</td>
<td>150,587</td>
<td>169,524</td>
<td>135,681</td>
<td>143,089</td>
</tr>
<tr>
<td>Total</td>
<td>170,921</td>
<td>170,921</td>
<td>155,904</td>
<td>141,086</td>
</tr>
</tbody>
</table>
In the analyses of variance of y and z, the total sums of squares are the same because they have been made so. Between schools, however, the sum of squares for z is much smaller than that for y. That for z is indeed so small that the corresponding mean square is below expectation ($MS_B/MS_W = .4$). Clearly the school original assessment means are all similar in magnitude, in contrast to the differences in magnitude among the rescaling test school means; there is no correlation between original assessment means and rescaling test means.

The correspondence between original assessment means and criterion means is no better. In the analysis of covariance for z and c, the between-school sum of products is low and negative, indicating, if anything, a negative correlation between original assessment means and criterion means.

Table II below reports the correlations derived from Table I.

**Table II: Correlations Between Y and C, and Z and C**

<table>
<thead>
<tr>
<th>Source</th>
<th>$r_{yc}$</th>
<th>$r_{zc}$</th>
</tr>
</thead>
<tbody>
<tr>
<td>Between Schools</td>
<td>.9683</td>
<td>-.3659</td>
</tr>
<tr>
<td>Within Schools</td>
<td>.9003</td>
<td>.8949</td>
</tr>
<tr>
<td>Total</td>
<td>.9086</td>
<td>.8222</td>
</tr>
</tbody>
</table>

The between-school correlations differ greatly; $r_{zc}$ is not significant at the 5 per cent level. Within schools, the correlations are similar, as might be expected, since they are the averages of the same within-school correlations differently weighted. Of the overall correlations, $r_{yc}$ is numerically the higher,
and, though no exact test appears available in this situation, the statistical significance of the difference is certainly very large.* It seems quite safe to conclude that rescaling has improved prediction in this case.

To complete this empirical enquiry, equations (22) and (23) are applied to the data of Table I.

(i) The error sum of squares of prediction by the rescaled assessments $y$ is $$E(V(C - Y)) = 31,373, \text{ df } = 934,$$ whence the error variance $$EV(C - Y) \text{ (overall)} = 33.6,$$ and the corresponding standard error is 5.8.

(ii) The error sum of squares of prediction by the original assessments $z$ is $$E(V(C - Z)) = 61,009, \text{ df } = 934,$$ whence the error variance $$EV(C - Z) \text{ (overall)} = 65.3,$$ and the corresponding standard error is 8.1.

Thus the effect of rescaling for these data has been to reduce the average standard error of prediction from 8.1 to 5.8.

To sum up: before rescaling, the assessment school means were all similar, indicating the usual tendency of teachers to pitch their standard similarly regardless of real inter-school differences. The correlation between the original means and the criterion means was, if anything, negative (-.3659). By rescaling, the correlation between the assessment school means and the criterion means was forced to the same magnitude as the correlation between the rescaling test school means and the criterion school means, which was high (.9683). The overall correlation

---

* Assuming the r's to be independent, the application of Fisher's z test to the difference between them produced a value of t corresponding to a probability 'off the scale' at the lower end. However, the r's are not independent, since they relate to the same children. The true probability level is therefore even lower than that demonstrated on the conservative assumption of independence.
between individual assessments and individual criterion scores rose from .8222 before the assessments were rescaled, to .9086 afterwards. The standard error of prediction of criterion scores by the assessments fell correspondingly from 8.1 before rescaling, to 5.8 afterwards. Rescaling has therefore been well worth while.
CHAPTER THIRTEEN

CURRENT PROCEDURES FOR RESCALING ASSESSMENTS

The wide use of assessments rescaled by labour-saving variants of Method 3 (equating means and s.d.'s), and the ready availability of accounts of these variants (see page 126), makes a detailed description of them unnecessary here. What follows, therefore, is a brief technical summary; an outline of a new procedure for adjusting rescaled assessments for age differences; and an illustrative example.

The most usual procedures can be classified as follows:

A.I - LINEAR TRANSFORMATION

Each teacher's set of marks is adjusted for mean and standard deviation only; the shape of the distribution is unchanged.

The standard equation for effecting a linear transformation is

\[ y = k \bar{x} + \frac{s_x}{s_z} (z - \bar{z}), \] with \( k = 1 \) (see Method 3, page 140).

The reader is reminded that \( z \) is a child's mark originally assigned by the teacher in a particular school; \( \bar{z} \) is the mean of all the marks for that school, and \( s_z \) their standard deviation; \( x \) and \( s_x \) are respectively the test score mean and standard deviation for the same school; and \( y \) is the child's rescaled mark or estimate.

If a calculating machine is available, the array of rescaled marks for each school can be readily obtained using the equation in the form
\[ y = (\bar{x} - k'\bar{z}) + k'z \]

where \[ k' = s_x/s_z = \sqrt{\frac{\Sigma (z - \bar{z})^2}{\Sigma (x - \bar{x})^2}} \]

Alternatively, the equation may be used for each school separately to construct a straight line graph from which the \( y \)'s can be read off directly from the \( z \)'s.

The advantage of using this equation is that the rescaled marks differ from the original marks in level and scale only; any relative spacings among the children's marks, which the teacher might wish to indicate, are retained. Its disadvantage is the labour involved in deriving a separate equation for each of a number of schools.

Some L.E.A.'s employ a simplified procedure which leads to approximately the same results. For each school separately the quartiles and the ends of the ranges for the \( z \)'s and the \( x \)'s are equated and a graph drawn by joining up these four points ((54), page 41). This method is not entirely successful in retaining the teachers' spacings, since the four points joined do not necessarily fall on a straight line, in which case use of the graph will result in some distortion of the distribution. Furthermore, the method may operate unfairly in some instances; for example, the whole of school 2 in McClelland's Table XII (ibid. page 35), in which the mean of the teachers' rescaled estimates is 4 - 5 points below that of the rescaling test.

A better approximation is obtained by increasing the number of percentile points equated, and then drawing a line through them by eye. This method is advocated by Vernon et al (91), page 195). If, however, the intention is to retain the teachers' spacings, the writers are mistaken in suggesting that 'a smoothed
line or curve' should then be drawn through the points. Distortion of the original distribution of marks is avoided only if a straight line is drawn through the points.

A.II - NON-LINEAR TRANSFORMATION

Rescaling methods based on non-linear transformations give, for each school, a set of marks adjusted to have the same mean and the same standard deviation as scores on the rescaling test, but with a distribution different from that of the teacher's original marks.

Most non-linear transformations depend on the equating of percentiles or ranks. The n'th percentile score in the array of test scores is assigned to the child whose teacher's mark is at the n'th percentile in the array of teacher's marks. After rescaling by this method, the mean, the standard deviation, and the shape of the distribution for each school set of teacher's marks are the same as for the school set of corresponding test scores.

Since in using the isopercentile method, the spacings of the original marks are lost, it is pointless to require teachers' marks in the first place if this method is to be used; rank orders will serve equally well.

The simplest and possibly the most accurate non-linear transformation is the rescaling method proposed by the National Foundation for Educational Research (96). The teacher simply ranks the children in his school. Next, scores on the rescaling test for the same children are themselves ranked. Finally the highest test score is assigned to the child ranked first by the teacher, the next highest score to the child ranked second; and so on, regardless of the identities of the children who obtained these test scores. The National Foundation method is thus seen to be the method of equating percentiles carried to its logical extreme.
Obviously, this simple reshuffling of test scores must give a set of teachers' rescaled marks for the school having the same mean, the same standard deviation and the same distribution as the school's test scores.

On balance the greater simplicity of the National Foundation method probably outweighs the slight loss of information which results when it is used instead of the more accurate rescaling equation.

*E.I - Age Allowances with Teachers' Assessments*

Objective test norms are so constructed that, in transforming raw scores to standardised scores or 'quotients', due allowance is made for age differences among the children tested. Rescaled estimates require a similar correction for age differences.

Attempts made by teachers to allow for age differences in allotting their original marks or ranks to the children in their schools are necessarily arbitrary and therefore likely to be inconsistent. It is better that teachers should ignore age difference when allotting their original marks or ranks, leaving it to a suitable statistical procedure to make subsequent adjustments for age differences which are consistent for all teachers.

If this course is adopted, it is logical that the rescaled marks which are initially allotted to the children and which are to be adjusted subsequently to allow for age differences should themselves be uncorrected in this respect. It follows that the scores on the test used for rescaling should be uncorrected for age differences also.

Vernon et al clearly state ([91], page 194) the difficulties arising in administering age allowances with rescaled assessments and propose a neat
solution as follows: '... the estimates do not normally make any allowance for age; hence they should preferably be scaled against test scores, not quotients. Once scaled, they can themselves be turned into quotients or other scores with an age allowance' (ibid. page 145).

No exception can be taken to this suggested procedure. Its one disadvantage is that it entails an extra standardisation in order to obtain the appropriate conversion table.

This labour can be avoided if instead of raw scores, on the rescaling test, standardised scores ('quotients') on the test, uncorrected for age are employed. After deciding on some convenient single age, the original test raw scores should be transformed into quotients using the line in the conversion table for the chosen age, and no other. These unadjusted test quotients should then be used, together with the teachers' original unadjusted 'raw' marks or ranks, to give unadjusted rescaled marks, to which finally appropriate corrections for age differences can be consistently applied. The size of the required corrections can be determined accurately by the appropriate regression method. Alternatively, the age allowance determined for the test quotients, and incorporated in the test conversion table, may be employed. In the latter case, the results will be rather less accurate, since the assumption is unwarrantable that the teachers' assessments should have exactly the same age allowance as the test quotients.

The suggested procedure is not only logical; it is also simple to work. To illustrate: suppose that the rescaling method to be used is that advocated by the National Foundation. The procedure is then as follows:-

(i) Each teacher ranks the children in his school group, taking care that his judgements are not influenced by differences in age among the children.
(ii) The rescaling test is administered to all the children in all school groups. The raw scores obtained are converted into quotients uncorrected for age using only the line in the conversion table appropriate to one particular month of age. It is convenient to choose the age of the oldest month-group in the complete year-group. All corrections for age are then positive, so that errors in calculation are less likely to occur.

(iii) Within each school group the uncorrected quotients from (ii) are ranked. The highest quotient is assigned to the child ranked first by the teachers; the next highest to the child ranked second; and so on.

(iv) A ready reckoner is prepared relating total age allowance to age at test date. For example, if the oldest month group is 11 y. 11 m. at test date, and the age allowance per month is 0.7 points of quotient, the ready reckoner (giving the total allowance to the nearest whole number) is:

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>ADD</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>6</td>
<td>5</td>
<td>4</td>
<td>4</td>
<td>3</td>
<td>2</td>
<td>1</td>
<td>1</td>
<td>0</td>
</tr>
</tbody>
</table>

(v) Each child's hitherto uncorrected rescaled estimate or mark is now corrected by adding the proper age allowance derived from the ready reckoner, having regard to the child's age at test date.

Before the ready reckoner can be constructed, the age allowance must be estimated. This can be done accurately by determining the regression of uncorrected quotient on age for the complete year group. Usually a sufficiently close approximation to the age allowance is that incorporated in the original conversion table, from which it may be deduced as follows:-
For a fixed raw score chosen from anywhere in the table, find the quotients $Q_1$ and $Q_2$ corresponding respectively to the lowest and highest ages listed. If $n$ is the interval in months between the lowest and highest ages,

$$\text{Age allowance per month} = \frac{(Q_1 - Q_2)}{n}$$

**B.II - Numerical Example**

A numerical example will illustrate the procedure (see the table on page 269). It relates to a school group of 20 children, part of a complete year-group with ages ranging from 11 y. 0 m. to 11 y. 11 m. The complete year-group has previously taken the rescaling test, for which the age-allowance is 0.7 points of quotient per month. This age-allowance is to be used also in adjusting the uncorrected rescaled estimates. Column 1 gives the ranks assigned by the teachers to the children and Column 2 their ages, the mean of which is 11 y. 6 m. for these 20 children. Column 3 gives the uncorrected quotients (with no age allowance) obtained by the children on the rescaling test; the quotients have been derived by entering the 11 y. 11 m. line of the conversion table for the rescaling test with the children's raw scores on the test. In Column 4 the uncorrected quotients of Column 3 have been placed in the teacher's rank order, and await adjustment for age. Column 5 gives the age-allowances, obtained by entering the ready reckoner previously prepared (see page 266). Column 6, obtained by adding the corresponding entries in Columns 4 and 5, gives the teacher's estimates or marks, adjusted for age, and now comparable with marks similarly constructed for other schools.

It will be noted that the adjustment for age has raised the school mean from 101.4 for the uncorrected quotients to 104.0 for the corrected quotients, an average increase of 3.6 points. This increase corresponds closely to that expected,
since the school mean age 11 y. 6 m. is 5 months below 11 y. 11 m., the age on which the uncorrected quotients were based; and $5 \times 0.7 = 3.5$.

Column 7, obtained by summing corresponding entries in Columns 3 and 5, shows the corrected quotients obtained by the same 20 children on the rescaling test itself. Thus both the teachers' rescaled marks and also the quotients on the rescaling test are readily obtained with this procedure by referring to one line only in the test conversion table and to the ready reckoner; a considerable saving in time and labour.
SCHOOL X: n = 20

<table>
<thead>
<tr>
<th>Teacher's Ranks</th>
<th>Ages (from 11:11 line)</th>
<th>Test Quotients (uncorrected)</th>
<th>Teacher's Marks (uncorrected)</th>
<th>Age allowance (from ready reckoner)</th>
<th>Teacher's Marks (corrected)</th>
<th>Test Quotients (corrected)</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>11:10</td>
<td>114</td>
<td>134</td>
<td>1</td>
<td>135</td>
<td>115</td>
</tr>
<tr>
<td>2</td>
<td>11:3</td>
<td>134</td>
<td>122</td>
<td>6</td>
<td>128</td>
<td>140</td>
</tr>
<tr>
<td>3</td>
<td>11:8</td>
<td>106</td>
<td>119</td>
<td>2</td>
<td>121</td>
<td>108</td>
</tr>
<tr>
<td>4</td>
<td>11:4</td>
<td>122</td>
<td>114</td>
<td>5</td>
<td>119</td>
<td>127</td>
</tr>
<tr>
<td>5</td>
<td>11:5</td>
<td>119</td>
<td>112</td>
<td>4</td>
<td>116</td>
<td>123</td>
</tr>
<tr>
<td>6</td>
<td>11:11</td>
<td>105</td>
<td>110</td>
<td>0</td>
<td>110</td>
<td>105</td>
</tr>
<tr>
<td>7</td>
<td>11:6</td>
<td>112</td>
<td>106</td>
<td>4</td>
<td>110</td>
<td>116</td>
</tr>
<tr>
<td>8</td>
<td>11:0</td>
<td>110</td>
<td>105</td>
<td>8</td>
<td>113</td>
<td>118</td>
</tr>
<tr>
<td>9</td>
<td>11:9</td>
<td>101</td>
<td>105</td>
<td>1</td>
<td>106</td>
<td>102</td>
</tr>
<tr>
<td>10</td>
<td>11:7</td>
<td>100</td>
<td>101</td>
<td>3</td>
<td>104</td>
<td>103</td>
</tr>
<tr>
<td>11</td>
<td>11:1</td>
<td>99</td>
<td>100</td>
<td>7</td>
<td>107</td>
<td>106</td>
</tr>
<tr>
<td>12</td>
<td>11:7</td>
<td>105</td>
<td>99</td>
<td>3</td>
<td>102</td>
<td>108</td>
</tr>
<tr>
<td>13</td>
<td>11:3</td>
<td>89</td>
<td>99</td>
<td>6</td>
<td>105</td>
<td>95</td>
</tr>
<tr>
<td>14</td>
<td>11:9</td>
<td>93</td>
<td>97</td>
<td>1</td>
<td>98</td>
<td>94</td>
</tr>
<tr>
<td>15</td>
<td>11:8</td>
<td>97</td>
<td>93</td>
<td>2</td>
<td>95</td>
<td>99</td>
</tr>
<tr>
<td>16</td>
<td>11:10</td>
<td>99</td>
<td>90</td>
<td>1</td>
<td>91</td>
<td>100</td>
</tr>
<tr>
<td>17</td>
<td>11:2</td>
<td>82</td>
<td>89</td>
<td>6</td>
<td>95</td>
<td>88</td>
</tr>
<tr>
<td>18</td>
<td>11:10</td>
<td>90</td>
<td>82</td>
<td>1</td>
<td>83</td>
<td>91</td>
</tr>
<tr>
<td>19</td>
<td>11:6</td>
<td>71</td>
<td>80</td>
<td>4</td>
<td>84</td>
<td>75</td>
</tr>
<tr>
<td>20</td>
<td>11:1</td>
<td>80</td>
<td>71</td>
<td>7</td>
<td>78</td>
<td>87</td>
</tr>
</tbody>
</table>

Means: 11:6 101.4 101.4 3.6 105.0 105.0
S.D.'s: 15.10 15.10 14.92 15.30

(Age allowance used for both test quotients and rescaled marks = 0.7 points of quotient per month.)
CHAPTER FOURTEEN

FINAL REMARKS AND CONCLUSIONS

The intention of this study has been to fill a gap in our present knowledge about rescaling procedures, the theory on which they are based, and the outcome of their use. One of the most interesting developments in examinations over recent years has been the increasing regard paid to the opinions expressed by teachers of their pupils' capacities and potential. This increasing regard is well shown by the increase in the number of local education authorities now supplementing external test procedures by teachers' numerical assessments. As one example, chosen out of many, the Birmingham Education Authority, in its pre-secondary testing procedure, now gives exactly the same weight to teachers' assessments as to the total over the three examinations taken by all the children. It is gratifying that the initial (and very natural) reluctance of teachers to accept this responsibility has now been overcome.

All this is as it should be. Vernon et al. ([91], page 137) have already admirably summarised the 'numerous excellent reasons' for making use of teachers' judgements, and there is no need to repeat these reasons here. If, however, rescaled assessments are to be used extensively, it is most desirable that those concerned with examination procedures should know more than they do at present about the statistics of these assessments. A vast amount of research has been done in the field of test construction, test reliability and test validity. There is a need for parallel work in the field of assessment-making, rescaling procedures, choice of rescaling tests, and of course, assessment reliability, in addition to the
validity studies we have at present. Teachers' assessments have been used to an increasing extent following the publication in 1942 of McClelland's massive Dundee research, probably the first in which these assessments were extensively used. Nevertheless research into their characteristics, other than their validity, has been scanty. McClelland's study, in which he made use of the method of rescaling which has since become standard procedure, rightly enjoys great prestige, despite the absence of any real discussion of the rationale of the method. Possibly that prestige has itself lent his study an authority which has inhibited further research of a more fundamental nature.

It is believed that the work reported here may be of value, firstly, in drawing attention to the pressing need for such a study, secondly, in indicating the directions in which it might usefully be pursued; and thirdly, in pointing to the use of an analysis of variance as a particularly appropriate method of enquiry especially suited to these purposes. Rescaling procedures make use of inter-school differences in means and scatters, and the suitability of the method of analysis of variance in such a situation is obvious.

This work, however, is only a beginning. Among important issues receiving no mention are the use made of teachers' assessments in the Valentine 'Quota' scheme now being used in Walsall, and the Peaker scheme with which the West Riding is experimenting. An 'ideal' rescaling equation has been developed, but the suggestion has not been followed up that it should be used as a tool in the search for the most appropriate rescaling tests in particular circumstances. Moreover, the present study is concerned only with the reliability of the rescaling procedure. The 'teacher-teacher' reliability of an original or raw assessment, that is, the degree of agreement between assessments made independently by two teachers of the same children's qualities is an entirely different matter still awaiting study.
Moreover, for much of the work reported here, repetition is desirable. After a theoretical discussion of the conditions affecting the reliability of the rescaling procedure, Darlington data are used to estimate its magnitude. Further empirical enquiries, should be undertaken, using the same methods, but with other data. Assuming the theoretical discussion of the inter-relations of assessments, test scores and criterion to be acceptable, the nature of the criterion used still imposes limitations on the validity of the conclusions reached as to the best rescaling procedure. Further work should be done along the same lines in which a more ultimate criterion is employed. Again, the results of the empirical study of the incidence of 'flop' scores disagree with the reported experience of McIntosh et al. Although discussion with one of the authors suggests that their conclusions were based mainly on impression, further evidence is clearly required.

The chief emphasis in the present study has been on theoretical aspects rather than on their applications. The aim of the study will have been achieved if others are stimulated to endorse or criticise the theoretical developments and to use them in wider empirical investigation.
PART THREE

THE RELIABILITY OF MARKING IN A GENERAL EXAMINATION
CHAPTER FIFTEEN

THE 'BAGRUT' EXAMINATION IN ISRAEL

A - INTRODUCTION

In the enquiry which follows, use is made of the procedures worked out in the preceding chapters for estimating the reliability of test results and teachers' assessments.

There is a serious lack of published information in this country about the reliability of external examinations such as the G.C.E. at both its levels. The early enquiries of Hartog and Rhodes (35) do not appear to have influenced the Examining Boards to publish the results of their own studies of examiner reliability. To quote Professor Oliver (95): 'The problems of secondary school examining are no doubt different, yet some of the basic questions that Dr. Wiseman asks about tests ... could equally well be asked about the examinations. Such questions might prove embarrassing. A fundamental concept in measurement - and examining is a form of measurement - is the 'standard error of score'. What is known about the standard error of a mark in any G.C.E. subject examined by any examining body? Dr. Wiseman surely makes out the case for a rapprochement between the test statistician and the examiner. The importance of 18 - plus examining would warrant the examining bodies in seeking the co-operation of those who have successfully encountered problems essentially similar to theirs'.(page 178). Professor Wiseman (95) himself says, in the course of two pages of criticism of the same nature: 'But what is disturbing to the teacher is the absence of any published evidence .... Boards seem to have strong objections to revealing their mysteries to 'outsiders', even on matters where publication would reassure teachers and raise their level of confidence in the examination'.(page 153).
Professor Oliver's own account of the Joint Matriculation Board's experimental English test of which he was the architect is a model for others to follow (57). It is all the more regrettable, therefore, that a recent publication from the same Board, concerned with the marking of A-level History scripts, should report a study so ill-designed as to enhance, rather than dispel, misgivings (58) and (68).

Evidence recently reported from India is no more reassuring. The levels of marking of scripts, thoroughly randomised before distribution to the markers, showed startling variations from one marker to another (83). Moreover, foreknowledge of the pass-mark by the markers played havoc with the frequency distributions of the marks they assigned.

In the absence of adequate information about public examinations in Britain the present study is based on the results of a similar examination conducted in Israel.

B - THE 'BAGRUT' EXAMINATION *

The 'Bagrut' is an examination of 'Matriculation' type administered to Israeli pupils towards the end of their secondary school careers when they are 17 or 18 years old. It is similar to corresponding examinations in Britain in that it is a 'subject' examination. The examinees must possess high academic qualifications and they are 'briefed' and controlled by Chief Examiners.

However, the 'Bagrut' differs from the corresponding British examinations in two important particulars:

* During a Unesco mission to Israel, I was asked to study and report on the efficiency of this examination. My acknowledgements are due to the Ministry of Education and Culture in Israel for permission to make use of the material.
(1) Every script is assessed independently by at least two markers, and the average of their marks is employed. In Britain (so far as one can discover) only one marker sees each script, except for sample checking.

(2) In its 'Bagrut', Israel makes full use of teachers' quantified assessments of candidates' achievements in the various subjects as a routine procedure. These assessments carry approximately the same weight as the external marks, to which they are added. The final score for every candidate depends on the aggregate of the two. By contrast, in Britain, teachers' assessments are employed at this educational level only to assist adjudication in borderline cases.

C - PURPOSES OF THIS STUDY

The purposes of this study are as follows:-

(1) To show how the method of analysis of variance can be used to study the examiner reliability of an external examination. For such an enquiry, replication of marking is essential, and this requirement is met in the 'Bagrut' examination, unlike its British counterparts.

(2) To examine the conditions determining the reliability of the assessments made by the candidates' own teachers after these assessments, originally arbitrary as to scale, have been rescaled on the external examination. As will be shown, this aspect of the study has important implications for the manner in which the scripts should be batched before distribution to the markers.

(3) To discuss the implications of the statistical enquiry for the external part of the 'Bagrut' examination as now conducted, and to make recommendations based on the results of the enquiry for improving its efficiency.
D - LIMITATIONS OF THE STUDY

A full inquiry into the efficiency of the Matriculation examination would involve follow-up studies to establish validity, as well as an investigation of the examination's reliability.

The only follow-up study of the Israel Matriculation Examination appears to be that of Guttman and Ortar relating to drop-out in the University. They find an inverse correspondence between high Matriculation grades in certain subjects and early drop-out at the University. This finding is one indication of the validity of the Matriculation Examination, but further evidence, based perhaps on a study of the correspondence between Matriculation and University grades, would be welcome.* Meanwhile, we must be content with the light thrown indirectly on the validity of the examination by a study of its reliability.

A full reliability investigation would involve replication of the whole examination, that is, the administration of at least two similar papers in each subject to the candidates, with each of these two papers assessed independently by at least two markers. The replication of the paper would provide information about the variability of the candidates from one occasion to another; the replication of the marking would throw light on the additional and independent variation caused by discrepancies between markers. Since two examinations could not be arranged, all that was possible was the study on a single examination, so that only the second of the two sources of variation, that ascribable to markers, could be investigated. It must be borne in mind therefore that the present study is incomplete in that it does not take into account variability caused by differences in performance among the candidates that would certainly have occurred had the papers been different or had the examination been held on a different day.

*See, for example: Parkyn, G. 'Success and Failure at the University' Wellington N Z C R
CHAPTER SIXTEEN

EXPERIMENTAL ENQUIRY

A - MATERIAL STUDIED

Discrepancies between the marks awarded to the same script by different markers are the result of differences in standard, in spread, and in rank order from one marker to another,* and these sources of undesirable variation were investigated with a sample of Hebrew Composition scripts written by candidates in the 1960 Matriculation Examination.

A.1 - SAMPLE USED IN THE ENQUIRY

The original data sheets showing the allocation of markers to schools contained the following data:

(i) The identities of the two markers allotted to each school.

(ii) For each school, the two arrays of candidates' marks, one array for each marker.

From these sheets, that pair of markers was selected which had been allotted the largest number of schools in common. These markers and schools constituted a good random sample of all markers and all schools. The statistical work reported below was carried out on this sample.

A.11 - RESULTS OF THE ENQUIRY

The results reported below related to this pair of markers, X and Y, each of whom marked the Hebrew Composition scripts of 373 candidates unequally divided among 11 schools. This pair of markers and these schools were typical.

* See page 242, footnote.
of them all.

Table I presents for each school:

(i) The averages of the two arrays of marks awarded to the scripts by the two markers (\( \bar{x} \) and \( \bar{y} \)).

(ii) The differences between these two averages (\( \bar{x} - \bar{y} \)).

(iii) The standard deviations (spreads) of the two arrays of marks (\( s_x \) and \( s_y \)).

(iv) The product-moment correlation (a measure of degree of correspondence) between these arrays of marks (\( r_{xy} \)).

(v) The standard error of single marking, which is a measure of the spread of marks that would be awarded to a single script if it were marked singly by each of a number of markers (\( s_{e1} \)).

(vi) The standard error of double marking, which is a measure of the spread of the averages of two marks that would be awarded to a single script if it were marked by each of a number of pairs of markers (\( s_{e2} \)).

<table>
<thead>
<tr>
<th>School</th>
<th>n</th>
<th>( \bar{x} )</th>
<th>( \bar{y} )</th>
<th>(( \bar{x} - \bar{y} ))</th>
<th>( s_x )</th>
<th>( s_y )</th>
<th>( r_{xy} )</th>
<th>( s_{e1} )</th>
<th>( s_{e2} )</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>77</td>
<td>40.4</td>
<td>42.1</td>
<td>-1.7</td>
<td>4.8</td>
<td>4.4</td>
<td>0.27</td>
<td>4.2</td>
<td>3.0</td>
</tr>
<tr>
<td>2</td>
<td>63</td>
<td>39.9</td>
<td>39.1</td>
<td>+0.8</td>
<td>6.5</td>
<td>6.1</td>
<td>0.44</td>
<td>4.8</td>
<td>3.4</td>
</tr>
<tr>
<td>3</td>
<td>14</td>
<td>41.5</td>
<td>37.3</td>
<td>+4.2</td>
<td>4.3</td>
<td>4.3</td>
<td>0.24*</td>
<td>4.8</td>
<td>3.4</td>
</tr>
<tr>
<td>4</td>
<td>22</td>
<td>34.5</td>
<td>35.5</td>
<td>-1.0</td>
<td>5.6</td>
<td>6.7</td>
<td>0.60</td>
<td>4.0</td>
<td>2.8</td>
</tr>
<tr>
<td>5</td>
<td>47</td>
<td>37.7</td>
<td>40.1</td>
<td>-2.4</td>
<td>6.5</td>
<td>7.1</td>
<td>0.37</td>
<td>5.7</td>
<td>4.0</td>
</tr>
<tr>
<td>6</td>
<td>31</td>
<td>38.9</td>
<td>36.5</td>
<td>+2.4</td>
<td>5.0</td>
<td>7.2</td>
<td>0.39</td>
<td>5.2</td>
<td>3.7</td>
</tr>
<tr>
<td>7</td>
<td>14</td>
<td>37.9</td>
<td>39.4</td>
<td>-1.5</td>
<td>4.7</td>
<td>5.7</td>
<td>0.83</td>
<td>2.5</td>
<td>1.8</td>
</tr>
<tr>
<td>8</td>
<td>23</td>
<td>36.4</td>
<td>38.9</td>
<td>-2.5</td>
<td>3.3</td>
<td>4.7</td>
<td>0.26</td>
<td>4.0</td>
<td>2.8</td>
</tr>
<tr>
<td>9</td>
<td>18</td>
<td>31.3</td>
<td>34.6</td>
<td>-3.3</td>
<td>3.7</td>
<td>5.0</td>
<td>-0.05*</td>
<td>5.1</td>
<td>3.6</td>
</tr>
<tr>
<td>10</td>
<td>11</td>
<td>40.3</td>
<td>40.1</td>
<td>+0.2</td>
<td>5.1</td>
<td>4.3</td>
<td>0.36*</td>
<td>3.8</td>
<td>2.7</td>
</tr>
<tr>
<td>11</td>
<td>53</td>
<td>39.9</td>
<td>39.4</td>
<td>+0.5</td>
<td>8.1</td>
<td>6.6</td>
<td>0.56</td>
<td>5.0</td>
<td>3.5</td>
</tr>
</tbody>
</table>

All 373 38.7 39.2 -0.5 6.2 6.2 0.39 4.7 3.3
A.III - INFERENCES FROM TABLE I

The fact that the school averages differ from school to school (columns $\bar{x}$ and $\bar{y}$) is to be expected, since it is well known that schools differ in average achievement. Of greater interest are the discrepancies between corresponding entries in these two columns, which point to differences in average standards of marking by the two markers $X$ and $Y$, and which are shown in the column $(\bar{x} - \bar{y})$. The fact that they range from +4.2 to -3.3 indicates not only differences, but also variations, in markers' standards from one school to another. (The weighted correlation between these school average marks is 0.66).

The next two columns show the extent to which each of the two markers has spread his marks for the same scripts within each school. All the standard deviations are low, indicating a reluctance on the part of the markers to use the full extent of the 0-65 scale (a random sample of 109 scripts showed the highest mark awarded to be 56 and the lowest 27, with the marks of 103 scripts, or 94 per cent, concentrated between 32 and 50). Moreover, the standard deviations differ, within each school, indicating that the markers differ in the spread of the marks they award to the same set of scripts.

The column headed $r_{xy}$ shows, for each school, the correlation between the two arrays of marks awarded, one array by each markers, to the same set of scripts. Ranging from .83 to -0.05, with an average value of 0.39, they are generally low, and reflect the lack of correspondence between the two marks on each script, which results from differences in rank order. Those marked with an asterisk do not differ significantly from zero at the 5 per cent level.
The last two columns, headed $s_1$ and $s_2$, show the standard errors of single and double marking respectively. These standard errors indicate the accuracy of the marking procedure's outcome.

The 'All' row at the foot of the table shows the weighted averages of the figures in the body of the table.

Let us now sum up these statistics. Firstly, looking at the whole set of 11 schools, we see that overall the two markers do not differ in standard by more than half a mark. This overall agreement, however, offers no comfort to the candidates from individual schools, since sometimes X's school average is higher than Y's, and sometimes the opposite. Were all the discrepancies in one direction, and reasonably constant in size, it would have been possible to treat the error as systematic and to remove it by adding a constant to all of X's, or all of Y's marks. Since, however, the average discrepancy varies from school to school, as shown in the $(x - y)$ column, this possibility is denied us. These discrepancies in standard between markers for the same schools have implications for the method of batching scripts before dispatch to the markers which will be discussed later.

Secondly, the notorious reluctance of the majority of examiners to employ the whole of the scale available to them is well exemplified here in the $s_x$ and $s_y$ columns. This reluctance arises from a pre-occupation with absolute standards which again will be the subject of later discussion. It is relevant, however, to note at this point the desirability on purely technical grounds of widely dispersed marks. By refusing to use the range of marks more extensively, markers deny themselves the first requisite for reliable marking.

Thirdly, the highly variable discrepancies between the two examiners' evaluations of the same individual scripts, probably the most serious defect
revealed, in this analysis, by the generally low though variable correlations $r_{xy}$ and by the entries in the $s_{e_1}$ and $s_{e_2}$ columns, clearly result from different subjective evaluations on the part of the examiners. Marking essays equitably is a notoriously difficult task, but probably the situation here is worse than it need be. Positive recommendations for improving marker reliability will follow. Meanwhile, it may be noted, as a fair inference from the reported standard errors of marking, that these two markers had marked all the 3,780 scripts in Hebrew Composition presented by candidates in 1960, then:

(i) 50 per cent of the scripts, or about 1,900, would have their two marks differing by $\pm 5$ or $\pm 5$ or more;

(ii) 20 per cent of the scripts, or about 760, would have their two marks differing by $\pm 9$ or $\pm 9$ or more;

(iii) 10 per cent of the scripts, or about 380, would have their two marks differing by $\pm 11$ or $\pm 11$ or more;

(iv) 5 per cent of the scripts, or about 190, would have their two marks differing by $\pm 13$ or $\pm 13$ or more.

To see these figures in perspective, it should be remembered that the maximum possible mark in this examination was 65.

To provide empirical verification of the above statements, which are based on the application of statistical theory to the data of Table I, a random sample of 203 scripts was drawn and for each script the difference between the two marks awarded was found. These differences, tabulated in the form of a frequency distribution, are presented in Table II.
TABLE II - HEBREW COMPOSITION: DIFFERENCES BETWEEN MARKS
FOR EACH OF 203 SCRIPTS

<table>
<thead>
<tr>
<th>Difference:</th>
<th>19 or 20</th>
<th>17 or 18</th>
<th>15 or 16</th>
<th>13 or 14</th>
<th>11 or 12</th>
<th>9 or 10</th>
<th>7 or 8</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency:</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>10</td>
<td>6</td>
<td>12</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Difference:</th>
<th>5 or 6</th>
<th>3 or 4</th>
<th>1 or 2</th>
<th>-1 or 0</th>
<th>-3 or -2</th>
<th>-5 or -4</th>
<th>-7 or -6</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency:</td>
<td>11</td>
<td>22</td>
<td>29</td>
<td>14</td>
<td>21</td>
<td>24</td>
<td>18</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Difference:</th>
<th>-9 or -8</th>
<th>-11 or -10</th>
<th>-13 or -12</th>
<th>-15 or -14</th>
<th>-17 or -16</th>
<th>-19 or -18</th>
<th>Total</th>
</tr>
</thead>
<tbody>
<tr>
<td>Frequency:</td>
<td>11</td>
<td>13</td>
<td>3</td>
<td>2</td>
<td>2</td>
<td>3</td>
<td>N=203</td>
</tr>
</tbody>
</table>

Mean difference = -0.85; standard deviation = 6.8

In the random sample referred to in Table II, the pair of marks differed:

(i) For 105 (or 51 per cent), by +5 or -5 or more;
(ii) for 45 (or 22 per cent), by +9 or -9 or more;
(iii) for 27 (or 13 per cent), by +11 or -11 or more;
(iv) for 10 (or 5 per cent), by +13 or -13 or more.

The figures in this sample are, if anything, slightly worse than expectation.

The discussion above relates to the differences between the same two markers' assessments of each of a number of different scripts. We now ask: What are the expected differences among the marks that would be awarded by a number of different markers to the same script? The situation is hypothetical, since in fact usually not more than two markers are concerned with the script. Nevertheless, the answer to the question is important. Various other pairs of markers might
have been allotted to the script, and we need to know the extent of the
differences in marks to be expected as a result. The answer to the question
raised for the hypothetical situation furnishes the required information.
The average standard error ($s_{e1}$) of a mark awarded to a script by either marker
X or marker Y is 4.7 marks. It is a fair inference that if a large number of
markers, each as variable as X or Y, were each to mark just one script independently,
the marks that any two markers would award to the one script would differ:

(i) in 50 per cent of the awards by 6 marks or more;
(ii) in 20 per cent of the awards by 12 marks or more;
(iii) in 10 per cent of the awards by 16 marks or more;
(iv) in 5 per cent of the awards by 18 marks or more.

However, in the 'Bagrut' Examination, the final assessment of any script
is normally based on the average of the marks awarded by two markers. Now if
the standard error of marking for a single marker is 4.7 marks, it follows that
the standard error of the average of two is 3.3 marks. It is then a fair
inference that if a large number of pairs of markers, each as variable as X and Y,
were each to mark just one script independently, the average marks that any two
pairs of markers would award to the one script would differ:

(i) in 50 per cent of the awards by 4 marks or more;
(ii) in 20 per cent of the awards by 8 marks or more;
(iii) in 10 per cent of the awards by 11 marks or more;
(iv) in 5 per cent of the awards by 13 marks or more.

These ranges for averages of two marks are narrower than the corresponding
ranges for single marks, but they are still uncomfortably wide. To illustrate
what might happen: the pair of markers actually concerned might award an average
mark of 44 out of 65 to a script (67 per cent - a good pass); but there is one chance in 20 that another pair, who might have been allotted to the same script, would have awarded it an average mark of 31 (47 per cent - a clear fail).

It should be borne in mind that the two markers, the 11 schools, and the 373 scripts which are the subject of the present study are typical of the whole set of markers, schools and scripts. A study of any other pair, marking for a different set of scripts from a different set of schools, would have yielded similar results.

B - PRELIMINARY ANALYSIS OF VARIANCE

The application of more sophisticated methods of analysis furnishes more detailed information. It is assumed that the candidates, schools and examiners are random samples.

To begin with, the division of candidates by schools is ignored. There are then simply 373 pairs of scores (an X- and a Y-score) for each candidate.

On the assumption stated above, the correlation between the two arrays of scores estimates the overall examiner reliability of the examination.

Table III shows the results of this preliminary analysis of variance.

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidates</td>
<td>372</td>
<td>20,533,050</td>
<td>55.196</td>
<td>2.477 (372,372)</td>
</tr>
<tr>
<td>Examiners</td>
<td>1</td>
<td>50,454</td>
<td>50,454</td>
<td>2.264 (1,372)</td>
</tr>
<tr>
<td>C x E</td>
<td>372</td>
<td>8,288,556</td>
<td>22.281</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>745</td>
<td>28,872,060</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
B.I - INFERENCES FROM THIS ANALYSIS

(i) There are significant differences among the candidates ($P < .01$) but not between the examiners. These results show that the totals or means over two examiners discriminate among the candidates, and that the overall standards of marking are similar.

(ii) Although the discrimination among the candidates is statistically significant, it is nevertheless unsatisfactory. The mean standard error of measurement of a candidate's mark from either examiner is

\[ \sqrt{22.281} = 4.722 \]

Also, the mean s.d. of a single examiner's marks is

\[ \sqrt{\frac{1}{2} (55.196 + 22.281)} = 6.224 \]

The s.e. of measurement is therefore about $\frac{3}{4}$ of the s.d. of marks, or about $\frac{1}{8}$ of the probable range of marks. This degree of precision is low in comparison with that of test scores (see page 184).

(iii) The correlation between the two examiner's arrays of marks is estimated by

\[ r_{XY} = \frac{F_c - 1}{F_c + 1} \quad (F_c = \frac{6}{OE}) \]

\[ = .426. \]

This is the single-examiner reliability.

(iv) In Israel, however, the aggregate or mean of the two marks assigned to the scripts is used. The s.e. of measurement of the mean of two examiner's marks is estimated as

\[ \sqrt{\frac{1}{2} (22.281)} = 3.338, \]

and the s.d. of the means of two examiner's marks as

\[ \sqrt{\frac{1}{2} (55.196)} = 5.253. \]
Using the means of two marks thus increases the precision of marking. The s.e. of measurement of a candidate's mean is now about 60 per cent of the overall s.d. of the candidate means, or about $\frac{1}{10}$ of the probable overall range of means.

(v) The inferred correlation between the totals (or means) for two different pairs of examiners is

$$r(x + y)(x' + y') = \frac{F_c - 1}{F_c}$$

$$= .598.$$  

This is the inferred reliability of a pair of markers.

B.II - IMPLICATIONS OF THESE RESULTS

The 'briefing' of the 'Bagrut' examiners seems reasonably efficient, though probably less so than their British counterparts. Also, Hebrew Composition is regarded in Israel as difficult to assess. Nevertheless, the fact remains that the reliability of single marking in a public examination was as low as .426; and this in a country which values academic excellence no less than we do in Britain and which examines along similar lines. The result is disturbing, and it would be reassuring to have firm evidence, based on the results of a properly designed enquiry, that our own reliability of marking is substantially higher.

C - A MORE DETAILED ANALYSIS

C.I - MODEL

In the preliminary study, the division by schools was ignored. It will now be taken into account. The model used is of the now familiar form:-
\[ x_{ijk} = M + s_k + a_j + s_{akj} + c_{ik} + ca_{ijk} \]  \hspace{1cm} (1)

in which

\( x_{ijk} \) is the mark awarded by examiner \( j \) to candidate \( i \) in school \( k \);
\( M \) is a component common to all marks;
\( s_k \) is a component common to all marks in school \( k \);
\( a_j \) is a component specific to marker \( j \);
\( s_{akj} \) is a component arising from interaction between school \( k \) and examiner \( j \);
\( c_{ik} \) is a component specific to candidate \( i \) in school \( k \);
\( ca_{ijk} \) is a component expressing the interaction between examiner \( j \) and candidate \( i \) in school \( k \).

The usual assumptions are made about the independence of these components and their zero expectations; and \( i = 1, 2, \ldots, m; \ j = 1, 2, \ldots, n; \ k = 1, 2, \ldots, \ell \).

**C.II - RELIABILITIES**

(i) Sum over \( j \) to obtain the total \( C_{ik} \) for candidate \( ik \) over \( n \) examiners:

\[ C_{ik} = \sum_j x_{ijk} = n s_k + \sum_j a_j + \sum_j s_{akj} + n c_{ik} + \sum_j ca_{ijk} \]  \hspace{1cm} (2)

The expectation of the variance of candidate totals within schools is therefore

\[ E(\text{V. of candidate totals within schools}) = n^2 \sigma_{c.s}^2 + n \sigma_{ca.s}^2 \]  \hspace{1cm} (3)

In equation (3), the first term on the right is 'true' variance, and the second is 'error' variance.

The expectation of the within-school reliability of candidate totals (or means) over \( n \) examiners is therefore

\[ \rho = \frac{n \sigma_{c.s}^2}{n \sigma_{c.s}^2 + \sigma_{ca.s}^2} \]  \hspace{1cm} (4)
(ii) Also from (2), we obtain the expected variance of candidate totals over all schools, the same n examiners marking all the scripts:

\[ E(V \text{. of candidates}) = n^2 (\sigma_{c,s}^2 + \sigma_s^2) + n(\sigma_{ca,s}^2 + \sigma_{sa}^2) \]  

whence the expectation of the overall reliability of the totals (n means) over n examiners is

\[ \rho = \frac{n (\sigma_{c,s}^2 + \sigma_s^2)}{n (\sigma_{c,s}^2 + \sigma_s^2) + (\sigma_{ca,s}^2 + \sigma_{sa}^2)} \]  

(iii) Summing again in equation (2) over i, we obtain the school k total \( S_k \) over m candidates and n examiners:

\[ S_k = \sum_{ij} \sum_{ijk} = mn s_k + m \sum_{j} c_{ik} + m \sum_{j} a_{ij} + n \sum_{i} c_{j} + \sum_{ij} c_{ijk} \]  

whence the expectation of the variance of school totals, the same n examiners marking all the scripts, is

\[ E(V \text{. of school totals}) = (m n^2 \sigma_{c,s}^2 + m^2 n^2 \sigma_s^2) + (m n \sigma_{ca,s}^2 + m^2 n \sigma_{sa}^2) \]  

On repeating the experiment with a different sample of n examiners, all of whom mark all the scripts, the terms in equation (7) including 'a' will change, while those without it will not. Therefore in equation (8), the first bracket includes 'true' variance, and the second 'error' variance. Hence the expectation of the reliability of school totals over n examiners is

\[ \rho_{\text{schools}} = \frac{n (\sigma_{c,s}^2 + m \sigma_s^2)}{n (\sigma_{c,s}^2 + m \sigma_s^2) + (\sigma_{ca,s}^2 + m \sigma_{sa}^2)} \]  

D - NUMERICAL RESULTS

D.I - ANALYSIS OF VARIANCE

The analysis of variance gave the results shown in Table IV.
### TABLE IV - FULL ANALYSIS OF VARIANCE

(373 candidates; 11 schools; 2 examiners)

<table>
<thead>
<tr>
<th>Source</th>
<th>d.f.</th>
<th>SS</th>
<th>MS</th>
<th>F</th>
</tr>
</thead>
<tbody>
<tr>
<td>(1) Schools</td>
<td>10</td>
<td>3,133.000</td>
<td>313.300</td>
<td>4.837 (10,10)**</td>
</tr>
<tr>
<td>(2) Examiners</td>
<td>1</td>
<td>50.454</td>
<td>50.454</td>
<td>2.390 (1,162)</td>
</tr>
<tr>
<td>(3) S x E</td>
<td>10</td>
<td>647.706</td>
<td>64.771</td>
<td>3.069 (10,162)**</td>
</tr>
<tr>
<td>(4) Candidates within schools</td>
<td>362</td>
<td>17,400.050</td>
<td>48.066</td>
<td>2.277 (362,362)**</td>
</tr>
<tr>
<td>(5) C x E within schools</td>
<td>362</td>
<td>7,640.850</td>
<td>21.107</td>
<td></td>
</tr>
<tr>
<td>Total</td>
<td>745</td>
<td>28,872.060</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

** indicates P < .01.

Note that pooling lines (1) and (4), and (3) and (5), gives the analysis of Table III in which the division by schools was ignored. 'Collapsing' the analysis in this way is analogous to the 'collapsing' in each of the previous Sections and also in the Introductory Chapter (see pages 31, 63, and 183). The result is the same in each case - loss of information about possible stratification within the sample.

### D.II - ESTIMATION OF VARIANCE COMPONENTS

(i) The 'average' number of candidates per school is

\[
m_o = \frac{\sum m_k}{K} - \frac{(\sum m_k^2)}{K} = 32.547 \quad ((44), \text{ page 105}), \quad (76), \text{ page 234})
\]
(ii) The expectations of the mean-squares in the above analysis are:

\[ E(\text{MS - Schools}) = \sigma_{ca.s}^2 + n \sigma_{c.s}^2 + m \sigma_{sa}^2 + m n \sigma_s^2 \]

\[ E(\text{MS - Examiners}) = \sigma_{ca.s}^2 + m \sigma_{sa}^2 + l m \sigma_a^2 \]

\[ E(\text{MS - S x E}) = \sigma_{ca.s}^2 + m \sigma_{sa}^2 \]

\[ E(\text{MS - C w S}) = \sigma_{ca.s}^2 + n \sigma_{c.s}^2 \]

\[ E(\text{MS - E x C w S}) = \sigma_{ca.s}^2 \]

Writing the mean-squares for these expectations, and putting \( n = 2 \), \( m = 32.547 \), we obtain the following estimates of the variance components:

\[ s_{ca.s}^2 = 21.107; \quad s_{c.s}^2 = 13.479; \quad s_{sa}^2 = 1.3416; \quad s_a^2 = 0; \quad s_s^2 = 3.4038. \]

**D. III - ESTIMATION OF VARIANCES, ERROR VARIANCES AND RELIABILITIES: INDIVIDUAL MARKS AND SCHOOL MEANS**

The estimates of the variance components derived above are now used, together with equations (3), (4), (5), (6), (8) and (9), to obtain the following estimates of variances, error variances, and reliabilities.
TABLE V - ESTIMATED VARIANCES, S.D.'S AND RELIABILITIES

\( n = 1 \) and \( 2; \ m = 16, 32 \) and \( 65 \).

(i) Candidate Marks

<table>
<thead>
<tr>
<th>Equ.</th>
<th>Variance of Candidate Marks</th>
<th>s.d. of Candidate Marks</th>
<th>Error Variance of a Candidate's Mark</th>
<th>s.e. of a Candidate's Mark</th>
<th>Examiner Reliability</th>
</tr>
</thead>
<tbody>
<tr>
<td>(3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Within Schools</td>
<td>Overall</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 Marker</td>
<td>2 Markers</td>
<td>1 Marker</td>
<td>2 Markers</td>
<td>Equ.</td>
</tr>
<tr>
<td></td>
<td>34.586</td>
<td>24.032</td>
<td>39.331</td>
<td>28.107</td>
<td>(5)</td>
</tr>
<tr>
<td></td>
<td>5.882</td>
<td>4.902</td>
<td>6.271</td>
<td>5.302</td>
<td></td>
</tr>
<tr>
<td></td>
<td>(3)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>School Means</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>1 Marker</td>
<td>2 Markers</td>
<td>1 Marker</td>
<td>2 Markers</td>
<td>Equ.</td>
</tr>
<tr>
<td></td>
<td>6.907</td>
<td>5.808</td>
<td>5.278</td>
<td>5.576</td>
<td>4.813</td>
</tr>
<tr>
<td></td>
<td>2.628</td>
<td>2.410</td>
<td>2.297</td>
<td>2.361</td>
<td>2.194</td>
</tr>
<tr>
<td></td>
<td>(8)</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>Examiner Reliability</td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td></td>
<td>0.390</td>
<td>0.561</td>
<td>0.429</td>
<td>0.601</td>
<td>(6)</td>
</tr>
</tbody>
</table>

(ii) School Means

<table>
<thead>
<tr>
<th>Equ.</th>
<th>m</th>
<th>16</th>
<th>32</th>
<th>65</th>
<th>16</th>
<th>32</th>
<th>65</th>
</tr>
</thead>
<tbody>
<tr>
<td>(8)</td>
<td></td>
<td>Variance of School Means</td>
<td>6.907</td>
<td>5.808</td>
<td>5.278</td>
<td>5.576</td>
<td>4.813</td>
</tr>
<tr>
<td></td>
<td></td>
<td>s.d. of School Means</td>
<td>2.628</td>
<td>2.410</td>
<td>2.297</td>
<td>2.361</td>
<td>2.194</td>
</tr>
<tr>
<td>(8)</td>
<td></td>
<td>Error Variance of a School Mean</td>
<td>2.661</td>
<td>1.990</td>
<td>1.666</td>
<td>1.330</td>
<td>.995</td>
</tr>
<tr>
<td></td>
<td></td>
<td>s.e. of a School Mean</td>
<td>1.631</td>
<td>1.410</td>
<td>1.291</td>
<td>1.153</td>
<td>.998</td>
</tr>
<tr>
<td>(9)</td>
<td></td>
<td>Examiner Reliability</td>
<td>.615</td>
<td>.657</td>
<td>.684</td>
<td>.761</td>
<td>.793</td>
</tr>
</tbody>
</table>
D.IV - DISCUSSION OF RESULTS

In Table V (i), the overall figures are all slightly higher than the corresponding 'within-school' figures - the effect of the significant inter-school differences brought out in Table IV, page 289. In all cases, whether there is one examiner or a pair, the ratio of s.e. of a mark to s.d. of marks is rather high. The reliabilities are correspondingly low, though they are higher for pairs than for single markers.

In Table V (ii), corresponding figures for school means are reported for three different sizes of school. Note the decrease in error variance of a school mean as the size increases. From equation (8), we may write:

$$E(e.v. \text{ of a school mean}) = \left(\frac{\sigma_{sa}^2}{n}\right) + \left(\frac{\sigma_{ca.s}^2}{mn}\right)$$

which decreases as m increases, but which will never become zero unless $\sigma_{sa}^2 = 0$. In fact, with these data the derived s.e. of a school mean is relatively insensitive (particularly with two examiners) to increase in the number of candidates per school (c.f. pages 198 to 202, where the relation between the standard errors of rescaled assessments and school size was discussed).

Moreover, compared with the error variances of a candidate's mark, those of school means are quite small, a fact which is reflected by the considerably higher inter-school examination reliabilities.

To sum up: the agreement between examiners for individual candidates is relatively poor, while that for school means is better.

The relatively high examiner reliability of school means augurs well for the use of the 'Bagrut' examination as a rescaling instrument for the teachers' assessments which play an important part in deciding the outcomes for individual
candidates. The efficiency of the examination as a rescaling instrument will be returned to later.

D.V - FURTHER USES OF ESTIMATES OF VARIANCE COMPONENTS

(i) The expected variance of a random selection of candidates' marks can be written down from equation (1):-

\[ \sigma^2_{\text{candidates}} = \sigma^2_{c.s} + \sigma^2_{a} + \sigma^2_{ca.s} + \sigma^2_{s} + \sigma^2_{sa} \]

Using the estimate of these variance components reported on page we obtain

\[ s^2 = 13.479 + 0 + 21.107 + 3.404 + 1.340 \]
\[ \text{candidates} = 39.330 \]

Thus the proportion of variance specific to individual candidates (corresponding to \( \sigma^2_{c.s} \)) is approximately one-third of the total, and the major source of error (corresponding to \( \sigma^2_{ca.s} \)) is approximately 54 per cent of the total. This error is the result of the poor correspondence between the examiners' marks within the schools (see Table I, page 278).

A break-down of this sort points the direction in which improvement should be sought. In this case, greater precision would be obtained if there were closer agreement between the marks assigned by the two examiners to the same script. The way to achieve this in practice is through more efficient 'briefing' of examiners (see page 304).

The alternative way of increasing the proportion of the variance specific to the candidate is to increase the number of examiners who mark his script. For the 'Bagrut' examination there are two examiners per script. The expected
variance of the means of two examiners' marks for a random selection of scripts is therefore

$$\sigma^2 = \sigma^2_{c.s} + \sigma^2_a / 2 + \sigma^2_{ca.s} / 2 + \sigma^2_s + \sigma^2_{sa} / 2,$$

(candidates' means)

which is estimated by

$$s^2 = 13.479 + 10.553 + 3.404 + 0.670$$

(candidates' means)

$$= 28.106.$$

The proportion of variance specific to the candidate has now risen to approximately 48 per cent, and that arising from poor agreement between the examiners has fallen to approximately 38 per cent. Increasing the number of markings per script from one to two has been well worth while.

(ii) Once the variance components have been estimated, the enquiry can be extended to include making inferences about the outcome if the examination conditions were changed. This line of enquiry will be developed next.
ALTNERNATIVE PROCEDURES FOR DISTRIBUTING SCRIPTS TO MARKERS

A - THREE DISTRIBUTION SCHEMES

In the present investigation, the same pair of examiners marked the scripts from all the schools. Other pairs of examiners marked other scripts from other schools. We shall designate by Scheme I a procedure in which two examiners are employed in the whole examination; and by Scheme II a procedure in which one pair marks scripts from one school, while other pairs similarly mark other scripts from other schools. In both Schemes I and II the scripts are batched so that the school groups are preserved intact.

A third possible procedure, designated Scheme III, is that in which the scripts are randomised before they are batched and sent to the examiners.

A.I - DESCRIPTION OF THREE MARKING SCHEMES

The difference between the three schemes is shown in the following diagram:

<table>
<thead>
<tr>
<th>School</th>
<th>1</th>
<th>2</th>
<th>3</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidates</td>
<td>C₁ C₂ C₃ ...</td>
<td>C₁ C₂ C₃ ...</td>
<td>C₁ C₂ C₃ ...</td>
</tr>
<tr>
<td>Scheme I</td>
<td>X₁ X₁ X₁ ...</td>
<td>X₁ X₁ X₁ ...</td>
<td>X₁ X₁ X₁ ...</td>
</tr>
<tr>
<td>Markers</td>
<td>Y₁ Y₁ Y₁ ...</td>
<td>Y₁ Y₁ Y₁ ...</td>
<td>Y₁ Y₁ Y₁ ...</td>
</tr>
<tr>
<td>Scheme II</td>
<td>X₁ X₁ X₁ ...</td>
<td>X₂ X₂ X₂ ...</td>
<td>X₃ X₃ X₃ ...</td>
</tr>
<tr>
<td>Markers</td>
<td>Y₁ Y₁ Y₁ ...</td>
<td>Y₂ Y₂ Y₂ ...</td>
<td>Y₃ Y₃ Y₃ ...</td>
</tr>
<tr>
<td>Scheme III</td>
<td>X₁ X₂ X₃ ...</td>
<td>X₄ X₅ X₆ ...</td>
<td>X₇ X₈ X₉ ...</td>
</tr>
<tr>
<td>Markers</td>
<td>Y₁ Y₂ Y₃ ...</td>
<td>Y₄ Y₅ Y₆ ...</td>
<td>Y₇ Y₈ Y₉ ...</td>
</tr>
</tbody>
</table>
In Scheme I, the same pair of examiners $X_1$ and $Y_1$, chosen at random from the pool of examiners, mark all the scripts from all the schools.

In Scheme II, different pairs of examiners, $X_1$ and $Y_1$, $X_2$ and $Y_2$, ..., are assigned at random, one pair to each school.

In Scheme III, the assignment of pairs of examiners to individual candidates is completely random.

Scheme I

The preceding discussion has related throughout to Scheme I. By making the reasonable assumption that the estimates of the variance components reported on page 290 will be constant in the face of procedural changes, we can infer the outcomes of these changes.

Scheme II

From equation (2), page 287, remembering that in Scheme II the examiners are different from one school to another, we obtain the expectation of the variance of candidate totals over all schools:

$$\text{E} (V - \text{candidates}) = (n^2 \sigma_s^2 + n^2 \sigma_{c.s}^2) + (n \sigma_a^2 + n \sigma_{sa}^2 + n \sigma_{ca.s}^2) \quad \cdots (10)$$

On repeating the marking with other sets of examiners in accordance with Scheme II, only the terms in the first bracket in equation (10) relate to mark components common to both markings. The terms in the first bracket are therefore 'true' variance, and those in the second are 'error' variance. The expectation of the overall reliability of candidates' mean marks over $n$ examiners is therefore:

$$\rho = \frac{n (\sigma_s^2 + \sigma_{c.s}^2)}{n (\sigma_s^2 + \sigma_{c.s}^2) + (\sigma_a^2 + \sigma_{sa}^2 + \sigma_{ca.s}^2)} \quad \cdots (11)$$

(Candidates overall Scheme II)
In the same way, starting from equation (7), it is easy to show that the expected reliability of school means under Scheme II conditions is

\[
\rho = \frac{n \left( m \sigma_s^2 + \sigma_c^2 \right)}{n \left( m \sigma_s^2 + \sigma_c^2 \right) + m \left( \sigma_a^2 + \sigma_{sa}^2 \right) + \sigma_{ca}^2}
\]  

(Schools)  

Scheme II

Scheme III

In Scheme III the scripts are allotted to the examiners at random, or, which amounts to the same thing, the examiners are allotted to the scripts at random.

Starting from equation (2), and taking account of this randomisation, we find that the expectation of the overall reliability of candidate totals over n examiners is the same as for Scheme II, and is given by equation (11).

However, the random allocation affects the expectation of the variance of school totals, which becomes

\[
E(V - \text{School Totals}) = m n^2 \left( m \sigma_s^2 + \sigma_c^2 \right) + m n \left( \sigma_a^2 + \sigma_{sa}^2 + \sigma_{ca}^2 \right).
\]

The expectation of the reliability of a school total is therefore

\[
\rho = \frac{n \left( m \sigma_s^2 + \sigma_c^2 \right)}{n \left( m \sigma_s^2 + \sigma_c^2 \right) + \sigma_a^2 + \sigma_{sa}^2 + \sigma_{ca}^2}
\]

(Schools)  

Scheme III

A.II - RESULTS WITH DIFFERENT SCHEMES

The several reliabilities are brought together for purposes of comparison in Table VI.
### TABLE VI RELIABILITIES WITH SCHEMES I, II AND III

Number of Examiners = n.

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Candidate Totals (or Means)</th>
<th>School Totals (or Means)</th>
</tr>
</thead>
<tbody>
<tr>
<td>I</td>
<td>$n \left( \sigma_s^2 + \sigma_{c.s}^2 \right) \over n \left( \sigma_s^2 + \sigma_{c.s}^2 \right) + (\sigma_a^2 + \sigma_{ca.s}^2)$</td>
<td>$n \left( m \sigma_s^2 + \sigma_{c.s}^2 \right) \over n \left( m \sigma_s^2 + \sigma_{c.s}^2 \right) + (m \sigma_{sa}^2 + \sigma_{ca.s}^2)$</td>
</tr>
<tr>
<td>II</td>
<td>$n (\sigma_s^2 + \sigma_{c.s}^2) \over n (\sigma_s^2 + \sigma_{c.s}^2) + (\sigma_a^2 + \sigma_{sa}^2 + \sigma_{ca.s}^2)$</td>
<td>$n (m \sigma_s^2 + \sigma_{c.s}^2) \over n (m \sigma_s^2 + \sigma_{c.s}^2) + (m \sigma_{sa}^2 + \sigma_{ca.s}^2)$</td>
</tr>
<tr>
<td>III</td>
<td>$n (\sigma_s^2 + \sigma_{c.s}^2) \over n (\sigma_s^2 + \sigma_{c.s}^2) + (\sigma_a^2 + \sigma_{sa}^2 + \sigma_{ca.s}^2)$</td>
<td>$n (m \sigma_s^2 + \sigma_{c.s}^2) \over n (m \sigma_s^2 + \sigma_{c.s}^2) + (m \sigma_{sa}^2 + \sigma_{ca.s}^2)$</td>
</tr>
</tbody>
</table>

The following may be inferred from Table VI.

(a) **Reliability of Candidate Totals (or Means) over n Examiners**

The expected examiner reliabilities for Schemes II and III are the same, but both are smaller than that for Scheme I, unless $\sigma_a^2 = 0$. The reliability is depressed in Schemes II and III if 'zero' error is present from one examiner to another, that is, if the examiners' marking standards differ.

(b) **Reliability of School Totals (or Means) over n Examiners**

(i) The expected examiner reliability for Scheme II is smaller than that for Scheme I unless $\sigma_a^2 = 0$. The depression is again caused by 'zero' error among the examiners.
(ii) The expected examiner reliability for Scheme III is larger than that for Scheme II, and, although it cannot be said with certainty, probably larger than that for Scheme I in practice, unless $\sigma_a^2$ is large.

A.III - APPLICATION TO NUMERICAL DATA

Table VII below shows the standard errors of marking and the corresponding reliabilities derived from Table VI using the Hebrew Composition data. For candidates, they relate to the means over two examiners. For schools, they relate to the means of 32 candidates (approximately the average school size) over two examiners.

**TABLE VII - ESTIMATED STANDARD ERRORS AND RELIABILITIES WITH DIFFERENT MARKING SCHEMES**

(n = 2 Examiners)

<table>
<thead>
<tr>
<th>Scheme</th>
<th>Candidate Means (Overall)</th>
<th>School Means (m = 32)</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>s.e.</td>
<td>R</td>
</tr>
<tr>
<td>I</td>
<td>3.350</td>
<td>.601</td>
</tr>
<tr>
<td>II</td>
<td>3.350</td>
<td>.601</td>
</tr>
<tr>
<td>III</td>
<td>3.350</td>
<td>.601</td>
</tr>
</tbody>
</table>
A. IV — DISCUSSION OF NUMERICAL RESULTS

Since \( \sigma^2_a = 0 \) for these data (see page 290), the standard errors and 2-examiner reliabilities of candidate means are the same no matter which scheme is used. This would not necessarily be the case with other data.

For the same reason, there is no difference between the standard errors and the 2-examiner reliabilities of school means with Schemes I and II. With Scheme III, however, the standard error is lower, and the reliability correspondingly higher.

Scheme III, it will be recalled, corresponds to the situation in which the batch of scripts each examiner receives is a random sample from the complete set. As Table VI shows, the main reason for the improvement in reliability of school means is the reduction of the effect of the Schools \( \times \) Examiners interaction. This is a measure of the bias occurring whenever a particular examiner encounters scripts from particular schools. If an examiner marks all the scripts from one school, the bias accumulates. If, however, because of previous randomisation, the scripts for a single school are divided among different examiners, then several biases will tend to cancel out.

To sum up the results with these data: for individual scripts, the method of distributing the scripts among the examiners makes no difference to the outcome. For the means of schools of similar size, the examiner reliability is highest with Scheme III, in which each examiner's batch of scripts is a random sample.

There are other reasons for preferring Scheme III. If each examiner receives a random sample of scripts, the expectation of his mean mark and the dispersion of his marks is the same as for every other examiner. Using the
quality control techniques which are commonplace in industry (62) it would be easy to set control limits for both examiner mean and standard deviation. An objective check on marking standards would thus be possible. It is remarkable that this obvious method of control has not yet been adopted.

There are other, non-statistical, reasons for preferring the randomisation procedure. One does not need to be a psychologist to imagine the effect on an examiner's standard of encountering a good school batch of scripts immediately after a poor one. Furthermore, shifts in standard are almost certain to occur through interruptions and fatigue (97). With the system of distribution normally used, these shifts may be expected to affect numbers of candidates from one school. With the Scheme III procedure, individual candidates from a school would still be affected, but over the whole school these errors would tend to cancel out. The precision of measurement of the school means would therefore be greater, a matter of some importance if the use of teachers' assessments rescaled on the examination is contemplated.
A. I - CONCLUSIONS: (1) REGARDING RESCALING OF ASSESSMENTS

In its 'Bagrut' examination, Israel makes use of teachers' assessments. It is known that these assessments, in their original form, are not comparable from one school to another. The question has therefore been asked whether the external examination itself is sufficiently precise in its measurements to warrant its use as an instrument for rescaling. The results of this study make possible at least a partial answer to this question.

The efficiency of a rescaling instrument, whether it be a test in an examination, depends on the reliability of school means and school s.d.'s. The results of this study suggest that provided two examiners mark each script the precision of the school means is fairly high. The results reported in Table VII, page 299, indicate that the inferred standard error of a school mean with double marking is approximately one mark with Schemes I and II, dropping to half this value with Scheme III. Having regard to the observed range of school mean marks (which Table I, page 278, shows to be approximately 6 marks) this degree of precision, so far as school means are concerned, is surely enough to warrant the use of the examination for rescaling purposes.

Little has yet been said in this study about school standard deviations, the precision of which is also an important factor in a rescaling instrument. Using the method of differences within pairs on the standard deviations in Table I, the standard error of a school standard deviation was inferred to be
.54 of a mark, the variate being \((x + y) / 2\). This again is a sufficiently high degree of precision to warrant the use of the examination as a rescaling instrument.

**A.II - CONCLUSIONS: (2) REGARDING THE EXTERNAL EXAMINATION**

The general heterogeneity of the results of this enquiry is disturbing. The low correlations between the marks awarded in Hebrew Composition reflect the extent of the present disagreement over rank order between the markers, who, as previously pointed out, are typical of all the markers concerned. Superimposed on this lack of agreement about individual scripts are further disagreements over the average quality of whole school sets of scripts, sometimes in one direction, sometimes the other; and over the dispersion of the marks they award to the same sets of scripts. Each of these sources of discrepancy contributes to the unreliability of the marks awarded to a single script; so much so that the average of these awards, on which the final assessment of the quality of a script depends in the majority of cases, must also be an unreliable assessment, though in general less so than that of each individual marker. The overall picture is one of standards shifting from script to script and from school to school, and despite the intervention of a third marker when discrepancies are large, the possibility cannot be ignored that injustice results for a number of candidates.

**B - GENERAL DISCUSSION OF MARKING DIFFICULTIES**

Marking difficulties are inevitable in examinations such as this, designed to provide searching tests for a large group of candidates, many of them very intelligent, at the upper end of secondary schools in a State which
has a high regard for academic excellence and which examines accordingly. Evaluative judgement on the part of markers is essential if the examination is to be appropriate to the situation just defined. No solution to the problems arising in marking which involves the elimination of evaluative judgement on their part is admissible. Nevertheless, the unreliability revealed by the statistical study made earlier is too great to be tolerated and some means of reducing it must be sought.

It is a truism that in principle no complete solution is possible to the problem of bringing together those two notorious incompatibilities, the necessity of evaluative judgement and reliable assessment. Nevertheless, the attempt must be made, and justice - perhaps not least among the virtues - is most nearly served when the highest possible measure of agreement exists among informed judges about the judgements each must make. The aim should therefore be, not to eliminate evaluative judgement, but to seek common agreement among the markers in regard to the judgements they are called upon to make, through discussions among themselves before they embark on the task of marking scripts in the mass.

C - CRITICISMS OF CURRENT PROCEDURES

C.I - INADEQUATE BRIEFING

The requirements stated above call for a considerable extension of the current system of briefing markers. At present each receives a set of general instructions and a key or marking scheme specific to the examination paper which concerns him. Usually there is also some discussion of marking procedures among those members of the marking panel (often by no means all) who do come together for this purpose.
The initial discussions appear to be inadequate; the general instructions issued to all markers open to objection on several grounds; and the marking schemes not sufficiently specific in the guidance they give.

The discussions preceding the marking operation appear inadequate for two main reasons: attendance at these discussions is not obligatory; and the marking scheme is prepared beforehand. High marker reliability is achieved when the mark awarded to a script is largely independent of the identity of the markers who happen to be allotted to it. Unless all markers attend the preliminary meetings, the marks awarded are probably less reliable than they need to be. Markers are more likely to agree among themselves over a marking scheme in the framing of which they have taken an active part, than over one of which the details are communicated to them through the post.

In order to reach the closest possible agreement on the marking scheme, and consistency in its interpretation, all markers should therefore be required to attend the initial briefing meetings, at which they analyse and discuss the questions in the paper under the chairmanship of a specially appointed Chief Examiner. At this meeting they are not presented with a fait accompli, but instead collaborate in the drawing up of a provisional marking scheme. The Chief Examiner will have selected previously a number of scripts representing - in his view - various grades of performance from excellent to its opposite. Photostat copies of these scripts are supplied to all markers who independently assess them in accordance with the analytic marking scheme provisionally agreed. Problems will of course arise which must be settled in a subsequent meeting or meetings at which the analytic scheme is further discussed and refined.
The aim should be an agreed marking scheme the intensive detail of which goes far beyond that at present employed. Matters of general policy are first discussed: the intention of each question, and hence, in broad terms, what constitutes a good answer; the relative importance of each question, and hence the different weights to be assigned to each. Next, for each question individually, methods of break-down are discussed and specifically defined with the aim of allocating maximum marks to each aspect or sub-section of it. For this purpose the answers in the photostat copies of the sample of scripts are helpful, and they assist also in reaching agreed decisions as to the proportions of these maximum marks to be awarded in specific examples drawn from the scripts.

Once agreed by all, the marking policy must be adhered to by all without deviation; each marker becomes part of the marking machine. At intervals, the marked scripts are sampled and checked by the Chief Examiner, who can call in any marker for discussion of deviations from the agreed scheme.

The general aim is thus to produce a scheme which can be applied with a fair measure of objectivity in the majority of cases, the 'run-of-the-mill' scripts. There will always be some, of course, which defy assessment by set rules, but the number of these is reduced to manageable proportions and can be handled individually in discussion with the Chief Examiner.

C.II - THE UNDESIRABILITY OF GRADE-FIXING BY MARKERS

Markers are at present requested to total their own marks and to convert them into percentages. They are further requested to pay special attention to those scripts whose total marks place them near one of the border-lines between the grades. After reaching a decision about such a border-line script,
the marker is to modify the marks he has previously awarded so that border-line marks such as 54 are avoided.

Underlying this instruction is the assumption of a pre-determined relationship between marks awarded and grades finally assigned. Reasons for doubting the necessity for this assumption will be presented shortly. We shall first note, however, some of its consequences in practice.

The statistical evidence presented points to unreliability of marking. If markers are not consistent in the marks they award, the consistency of the grades they assign, which, on the above assumption, are directly related to the marks, must also be doubtful. At the border-lines, in particular, difficulties must arise. The thought of a number of markers, or pairs of markers, adjudicating in crucial cases is therefore disturbing.

A further consequence of the fixed relationship between marks and grades is the reluctance of markers to make full use of the range of marks available to them. That they do in fact restrict their range was established earlier. (In fairness it should be said that the same is true of markers in many comparable examinations elsewhere.) Probably the reason is as follows. It is a matter of common experience that in many human activities, including the sitting of examinations, average performance is the most frequent, and performances deviating considerably from this average in either direction relatively few. A legitimate interpretation of this distribution of performance

* Grades are based on percentages:

<table>
<thead>
<tr>
<th>Per cent:</th>
<th>95+</th>
<th>85-94</th>
<th>75-84</th>
<th>65-74</th>
<th>55-64</th>
<th>45-54</th>
<th>35-44</th>
</tr>
</thead>
<tbody>
<tr>
<td>Grade:</td>
<td>10</td>
<td>9</td>
<td>8</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>4</td>
</tr>
</tbody>
</table>

In practice, no grade lower than 4 appears to be awarded.
in terms of grades is placement of the majority of individuals in, or somewhere near, the middle grades, and few in those at the extremes. If grade-mark relationship is fixed, and known to the marker in advance, his 'built-in' concept of relative frequencies within grades must influence the marks he awards. The result is parsimonious award of high and low marks, and a clustering over the middle range of the mark scale. It has been already pointed out that such a restriction cannot but reduce reliability.

It is of course recognised that the concept of an absolute scale of quality must underlie the grades finally awarded to the scripts. It is open to question, however, whether this consideration should dominate the marking procedure. No logical connection exists between a mark of 95 per cent or above and a grade of 10; between a mark in the 85-94 per cent range and a grade of 9; and so on. The distinction of a grade of 10 is rightly reserved for scripts of the highest merit, but by what process of logic can it be concluded that for every paper in every year the mark range corresponding to this distinction should be 95 per cent or above? Although open to criticism on the ground that it is insufficiently detailed, an analytic scheme of marking does exist now. The resulting scale of marks must of necessity be arbitrary. How can it possibly be tied to grades at this stage?

The confusion becomes worse when we take into account the inevitable fluctuation in difficulty of examination papers from one year to another. Since the total mark awarded to a script is the aggregate of a number of 'submarks', each out of a sub-total determined on grounds that must be arbitrary; and since there is no satisfactory means of knowing in advance the difficulty of the current paper relative to that of papers set in previous years; the present assumption that the mark scale - quality relationship is constant over
the years appears unwarrantable. If empirical evidence is needed to substantiate this rather obvious train of reasoning, it is furnished by the fluctuations over the years in the pass-rate, now tied to a fixed pass-mark, which are in fact observed for a number of subjects to a disturbing extent. It is unlikely that these fluctuations reflect real differences in the quality of candidates presented in successive years. More probably they are the result of the interaction between arbitrary marking procedures, papers of differing difficulties, and a rigid pass-mark. If this is so, the serious situation exists that final outcome for a candidate is influenced not only by the markers allotted to his scripts, but also by the year in which he happens to sit the examination.

Marking procedures must be arbitrary; differences in difficulty of papers are unavoidable. There remains the pass-mark, which can, and should, be altered from year to year; and so, too, the marks corresponding to other inter-grade border-lines.

D - SUGGESTIONS FOR REFORM OF THE MARKING PROCEDURE

It must be concluded from the several arguments above that the responsibility of passing or failing a candidate, or of assigning him to any particular grade, should not devolve upon those carrying out the general marking procedure. The task of a marker is to award marks in accordance with an analytic scheme. This is a heavy responsibility, but it is legitimately his. It is not legitimate, however, to require him also to place the scripts he marks in a number of grade categories, and, in particular, to adjudicate in crucial cases.
A safer policy would be to remove from the main body of markers the responsibility, and even the possibility, of deciding, during the marking process which of the scripts before them merit pass or fail, or any other 'grade. At this stage the marker's sole task should be that of assigning marks, or rather sub-marks, in accordance with the previously agreed scheme; not to the script before him as a whole; not even to the complete answers to separate questions in the script; but only to the elements within these answers agreed in the marking scheme. Various suggestions may be made in order to assist him in this relatively humble but important endeavour. Firstly, the task of totalling the sub-marks he awards to the several elements, and hence arriving at the global 'raw' totals for the scripts, should be performed by someone else. He is not then deflected from his main task, that of awarding sub-marks, nor is he given opportunity to reflect that perhaps he is awarding too many high, or too many low grades, a dangerous reflection leading to unconscious shift of standards. Secondly, in order to prise the marker loose from his customary habit of thinking in terms of percentages, the maximum global 'raw' total for any examination should be some unfamiliar number such as 137, which is difficult to transform mentally into anything. Thirdly, the scripts sent to any marker should be a random sample, in random order, of the whole set of scripts. This is borne out by the fact that the differences between markers' standards are not constant from school to school (see Table I). All these suggestions are directed at one objective: to assist the marker in his legitimate task of assigning sub-marks as nearly automatically as possible, and to deflect him from attempting a task not rightfully his, that of grading the scripts for overall merit while marking is going on.
The task of final grading remains. Cutting-lines must be fixed to
demarcate the grades in terms of the total marks awarded to the scripts.
On the reasonable assumption that average achievement in a subject does not
fluctuate unduly from one year to the next, the percentages of scripts in the
several grades last year provide a useful starting-point. A random sample of
this year's scripts is carefully drawn and divided provisionally into grades
in the same proportions as last year's grades. In this way, tentative
cutting-lines, which will probably differ from last year's cutting-lines, are
arrived at. The Chief Examiner and a small committee of experienced helpers
then read through the sample of scripts, paying particular attention to those
with marks placing them near the several tentative cutting-lines between the
grades. In consequence of this the committee may decide to alter some or
all of the tentative cutting-lines. For example, in the crucial case of
the cutting-line between 'pass' and 'fail' they may decide that some scripts
which have marks placing them below the tentative cutting-line and which therefore
fall into the provisional 'fail' category are in fact of 'pass' standard.
Once decided upon, these cutting-lines are employed in assigning the whole body
of scripts to grades.

At this point the scale may be changed from that of the original global
'raw' scores to one with the virtue of familiarity. This may be accomplished
quite simply by equating the original 'raw' score cutting-lines to the cutting-
lines on the familiar scale and arriving at intermediate scores within the
grades by simple interpolation. The construction of a point-by-point conversion
table shortens the work.

By the use of procedures such as this, marks are relegated to their proper
instrumental role.
REFERENCES


<table>
<thead>
<tr>
<th></th>
<th>Author(s)</th>
<th>Title and Publication Details</th>
</tr>
</thead>
</table>


New York: Wiley.

34. Guttman, L. (1945) A Basis for Analysing Test-Retest Reliability
Psychometrika, X, 255-282.

London: Macmillan

London: Methuen


38. Hoyt, C. (1941) Test Reliability estimated by Analysis of Variance
Psychometrika, VI, 155-60.


40. Jackson, R. W. B. (1939) Reliability of Mental Tests

41. Jackson, R. W. B. (1940) Application of the Analysis of Variance and Covariance Methods to Educational Problems
Toronto: University of Toronto.

Bulletin 12, Department of Educational Research, Toronto: University of Toronto.

43. Johnson, P. O and Neyman, J. (1936) Tests of Certain Linear Hypotheses and their Application to some Educational Problems
Statistical Research Memoirs I, 57-93.

44. Kempthorne, O. (1952) The Design and Analysis of Experiments
New York: Wiley.

Psychometrika, II, 151-160.

46. Lawley, D. N. (1950) A Method of Standardising Group Tests
<table>
<thead>
<tr>
<th>No.</th>
<th>Author(s)</th>
<th>Year</th>
<th>Title</th>
<th>Publisher/Location</th>
</tr>
</thead>
<tbody>
<tr>
<td>52.</td>
<td>Lord, F. M.</td>
<td>1964</td>
<td>Nominally and Rigorously Parallel Test Forms</td>
<td>Psychometrika, XXIX, 335-346.</td>
</tr>
<tr>
<td>58.</td>
<td>OP 17 (1964)</td>
<td></td>
<td>The Marking of Scripts in Advanced Level History</td>
<td>Joint Matriculation Board.</td>
</tr>
</tbody>
</table>
63. Pearson, K. (1901)  
On Lines and Planes of Closest Fit to Systems of Points in Space  
Phil. Mag. II, 559-72.

64. Pilliner, A. E. G. (1952)  
The Application of Analysis of Variance to Problems of Correlation  

65. Pilliner, A. E. G. (1958)  
The Rescaling of Teachers Estimates  

Zero Error in Moray House Verbal Reasoning Tests  

The Measurement of Educational Achievement in Israel  
Report to UNESCO.

68. Pilliner, A. E. G. (1965)  
Review of 'The Marking of Scripts in Advanced Level History'  

69. Rajaratnam, N. (1960)  
Reliability Formulas for Independent Decision Data when Reliability Data are Matched  
Psychometrika, XXV, 261-271.

Generalizability of Stratified-Parallel Tests  
Psychometrika, XXX, 39-55.

71. Richardson, S. C. (1956)  
Some Evidence Relating to the Validity of Selection for Grammar Schools  

72. Rulon, P. J. (1939)  
A Simplified Procedure for Determining the Reliability of a Test by Split Halves  

73. Sandon, F. (1956)  
A Regression Control Chart for Use in Personnel Selection  

74. Schultz, E. F. (1955)  
Rules of Thumb for Determining Expectations of Mean Squares  
Biometrics XI, 123-135.

75. Sealey, L. G. W. (1960)  
The Creative Use of Mathematics in the Primary School  
London: Blackwell.

76. Snedecor, G. W. (1946)  
Statistical Methods  
Iowa: State College Press.

77. Spearman, C. (1904)  
The Proof and Measurement of Association between Two Things  
Amer. J. Psychol. XV, 72-101.
78. Spearman, C. (1910)  
Correlation Calculated with Faulty Data  

79. Spearman, C. (1936)  
Note 'On the Reliability and Validity of Examinations'  
in Essays and Examinations  

Selection for Secondary Education without a Written Examination  
London: Murray.

81. Tanner, J. M. (1953)  
Times Educational Supplement, 21 November 1953.

82. Tanner, J. M. (1961)  
Education and Physical Growth  
London: University of London Press.

83. Taylor, H. J. (1964)  
Operation Passmark: An Account of the Method Used in the Matriculation Examination of 1963  
Gauhati: University of Gauhati.

84. Technical Recommendations for Psychological Tests and Diagnostic Techniques (1954)  

85. Thomson, G. H. (1928)  
The Mental Age Concept and Standardisation of Group Tests  
Psychol. Review XXXV, 398-413.

86. Thomson, G. H. (1932)  
The Standardisation of Group Tests and the Scatter of Intelligence Quotients  

87. Thomson, G. H. (1945)  
The Distribution of Intelligence Among University and College Students  

88. Thomson, G. H. (1946)  
The Factorial Analysis of Human Ability  
London: University of London Press.

89. Tippett, L. H. C. (1941)  
The Methods of Statistics  
London: Williams and Nozgati.

90. Vernon, P. E. (1954)  
Symposium on the Effects of Coaching and Practice in Intelligence Tests  

91. Vernon, P. E. (Ed.) (1957)  
Secondary School Selection: A British Psychological Society Inquiry  
London: Methuen.