Individual experimental work in physics in the junior forms of grammar schools

Joyce, L. S.

INDIVIDUAL EXPERIMENTAL WORK IN PHYSICS IN THE JUNIOR FORMS OF GRAMMAR SCHOOLS.

By

L. S. JOYCE

Being a thesis submitted for the degree of M. Ed. in the University of Durham. March 1949.

INDIVIDUAL EXPERIMENTAL WORK IN PHYSICS IN THE JUNIOR FORMS OF GRAMMAR SCHOOLS

CONTENTS

CHAPTER 1

BRIEF HISTORICAL INTRODUCTION

1. The Heuristic Method.

The introduction of science into the curriculum of secondary grammar schools was initiated in the middle of the nineteenth century and during the past hundred years there have been many changes, or developments, in the methods of instruction. Previous to the advocacy of the heuristic method by Professor H.E. Armstrong in 1888 practically no individual experimental work was done by the pupils. Instruction in Physics and Chemistry was in most cases limited to the imparting of "facts" with, in some cases, a few demonstrations by the teacher. Armstrong's method[1] was largely a reaction against the preceeding didactic methods. Science was something one "did" and not a matter of "chalk and talk". The basic idea was that the pupils should be guided by the teacher to discover facts and relations for themselves by means of actual experiment. The value of group discussions was also recognised and stressed. Most teachers were stimulated by the new method and were willing to appreciate the value of a more experimental approach to science. Many however attempted to carry the method to extremes and three major objections were soon apparent.

(a) Progress was extremely slow since very little ground could be covered in a term,

(b) the method was very difficult from the point of view of teacher control when large classes were involved,

(c) with the continual growth of scientific knowledge the possibility and wisdom of attempting to "discover" all facts by individual experiment became more and more absurd.

As a consequence of these objections the method was slowly subjected to modifications during the early years of the present century, but the accent on at least some form of individual experimental work by the students remained.

2. Practical Work by the Pupils.

The arguments in favour of methods which entail the performance of experiments by the pupils themselves have shown gradual changes, and the following are a representative selection.

(a) Practical work can assist the pupil to appreciate and memorise the factual material of the subject.

(b) The pupil learns to make accurate observations and measurements, and to prepare clear descriptions of his work.

(c) Since many new facts in science can only be obtained by actual manipulation of apparatus and the designing of experiments to test new hypotheses, the pupil must have personal experience of the necessary technique.

(d) The teaching of scientific method is more effective when individual practical work is undertaken.

(e) In many cases the pupil derives much emotional satisfaction from his experimental work.

Although their individual reasons may differ, most teachers now believe that practical work in some form is essential for the effective teaching of science. Unfortunately however, towards the end of the nineteenth century and during the early years of the present century, the faculty school of psychology was still prevalent and influenced the form of the practical work undertaken.

> "The Laboratory method originally (1880-1910) laid emphasis upon a psychology which has now been abandoned. Laboratory exercises were planned to provide for the training of the faculties. It was held that such work gave opportunities for the cultivation of accuracy in observing changes in phenomenon, for developing systematic habits of work, and for training in the power of reasoning from a particular set of observations to a general law. The keeping of a Laboratory manual was held to give valuable training in habits of neatness and precision of expression."[2]

Period after period was devoted to the almost aimless repetition of stereotyped experiments. Pupils of eleven and twelve for example often spent lesson after lesson measuring the density of solids by various methods. Again an over insistence on the need for inculcating so called habits of neatness as reflected in the condition of the pupils' laboratory note books, and accounts of their experiments, led in many cases to the slavish following of typewritten instructions. The observations and results tended to be as stereotyped as the actual experiments. The dangers of such a system were not too well appreciated and men like F.W. Westaway, continued to criticize the system.

> "A laboratory is not a place either for the mechanical repetitions of a cloistered cell or for the dusty ritual of an antiquary's den."
>
> "Is there not some confusion between instruction in science and instruction in scientific technique."
>
> "The ritual of the laboratory must not be confused with the spirit of science."[3]

It is probably that these retrograde developments were in some measure the natural outcome of large classes and the need to prepare the pupils for external examinations. When classes were small, and the external examination system was more flexible, or non existent, methods of a freer nature, allowing more scope to the individual pupils, were developed. A good example was the "Project" method which became very popular in the United States. The basic principle was the assignment of useful tasks to small groups of pupils. For example one group might be required to arrange the lighting of the school stage, and another required to instal a small telephone system. The co-ordination of the work was difficult and trouble was often encountered in devising suitable projects for young pupils. It has never become popular with science teachers in England, for use as a method complete in itself. Most teachers feel that such a method fails to provide the child with a background of coherent knowledge of the fundamentals of the subject. Many teachers however do find that the assignment of an occasional "project" to older pupils has a stimulating effect.

3. General Science.

During the past twenty years a growing doubt of the value of much of the experimental work being done in schools has arisen. This movement though slow at first is gaining momentum and is closely allied with the growth in popularity of General Science in the schools. The Science Masters' Association have published two reports[4,5] giving, a brief history of the General Science Movement, and suggested syllabi. Their publication has aroused considerable controversy on the subject but there is no doubt with regard to the growing popularity of the subject.[6] In 1924 only 1,266 candidates offered General Science as a subject in the School Certificate Examinations and by 1938 this number had grown to 8,752. In 1942 the number had risen to 17,617 and candidates were presented from approximately 680 schools. In the same year the total number of candidates taking Physics as a separate subject was 23,686. The Science Masters' Association reports seem to have created an impression that in General Science the accent should be on demonstration work by the teacher rather than on individual work by the pupils, although no actual suggestion to that effect was made in the reports. The "General Science" approach to the study of science is still in the experimental stage but certain factors are emerging. The field of study is wider and shallower and the "Topic" method, whereby the teacher suggests to the class topics of general interest, which are then studied scientifically, provides a useful approach to the subject. Individual experimental work must to some extent be replaced by demonstration lessons by the teacher, and some quantitative and mathematical aspects of the various single subjects will have to be sacrificed. The immediate reason for the former change is governed by the time factor since the range is wider and in most schools the time allotted to science has not been increased. Many critics of General Science have attacked the subject on the grounds of the lack of opportunity for the pupils to do so much individual experimental work, but many of its supporters contend that much of the practical work done by the pupils in the past was comparatively valueless. Whatever may be the ultimate result of the arguments one interesting result of the "General Science" controversy is that interest in the value of Demonstration Experiments by the teacher has increased.

4. Demonstrations by the Teacher.

In the early days of science as a subject in the school curriculum the practical work was in the main confined to demonstration experiments by the teacher. Such methods are again becoming prominent and their advocates have several strong arguments.

(a) More ground can be covered in the same time.

(b) Many experiments involve apparatus that is too expensive or complicated for use by the pupils alone.

(c) The teacher being in full control the actual manipulation of the apparatus is less likely to distract the pupil's attention from the real object of the experiment.

(d) The method ensures better control of the class and this factor is important with large classes and teachers of weak disciplinary powers.

(e) The pupils <u>see</u> the measurements made in the correct manner and may even be allowed to participate in the actual measurements.

(f) Large scale experiments are possible and tend to be more impressive.

The series of Christmas Lectures for juvenile audiences, given each year at the Royal Institution, are splendid examples of the use and value of demonstrations in the teaching of science. It must however be rememebered that most of these have been delivered by first class lecturers and have involved an immense amount of preparation, and costly apparatus. The average school teacher could not devote an enormous amount of time to preparation, but it is at least arguable that the preparation of demonstration lessons is less exacting than the preparation of lessons involving individual practical work by the pupils, since for school work, demonstration experiments, when once designed and constructed, can often be stored for future use.

5. <u>The Relative Merits of Individual Experimental Work and Demonstration Lessons.</u>

When considering the relative merits of the two methods there is a rather dangerous tendency to concentrate on the time factor alone and attach too much significance to the general statement that the Demonstration method enables more ground to be covered in the same time. This statement is doubtless true so far as more advanced students are concerned and is probably true with younger students, if the alternative is implied that they should perform <u>all</u> the experimental work in the course themselves. In most cases however the conditions are not so rigid and the decision rests between a method where all the experimental work is done through the medium of demonstration experiments and a method using a combination of demonstation experiments with a good proportion of suitable experiments performed by the pupils themselves. With this restriction the time factor of course ceases to be so important but still exists. F.W. Westaway summarises the position when he states:-

> "If it can be shown that the lecture room method is as good as the laboratory method both as to training and as to knowledge imparted a great saving of time might be effected in our science teaching."[3]

Many attempts have been made to assess the relative values of the two methods but the majority of the researches have been American in origin. The experiments have not been limited to one particular science but have dealt with Chemistry, Physics and Biology.[7.8.9.] The general technique has been to start with two parallel classes, efforts being made to "match" the classes with regard to initial ability. The classes are then taught by the same teacher, but one of the classes is taught purely by demonstration methods while the other is concentrated on individual experimental work. Care is taken to see that the two classes cover the same ground. At the end of a set period both classes are given a written test and in some cases a further test after an extra interval of six months. The tests are generally intended to measure the extent to which the pupils have learned new facts or laws, or have mastered the principles involved. The results have never been highly significant although in some cases there are indications that the "demonstration" pupils do better on the immediate test and worse on the delayed test. All the experiments of this type have involved only small samples and the concentration appears to have been on the effect of the two methods of instruction on the achievement of the pupils as assessed by written, or pen and paper, tests.

This is rather unfortunate since there is no doubt that many of the significant objectives of science teaching can not be tested efficiently by such tests. Any method of instruction applied to a pupil will have intellectual, physical and emotional effects and in comparing the values of two methods all three effects should be considered. Even if the "Demonstration" method and "Individual Experiment" method produce no significant difference in the attainments of the pupils as measured by educational tests there may be significant differences in other directions. One method may improve the practical skill and ability of the pupil, the other may have a more beneficial emotional effect, or produce important developments in the interests, attitudes or personality of the pupil. The problem is still further complicated by the fact that, even if we could evaluate all the changes of pupil behaviour produced by the two methods, the question of the relative importance of the various changes would still be a matter of subjective opinion. One investigator might for example consider an increase in practical ability or skill to outweigh the disadvantages of less progress in the acquisition of scientific information and knowledge, while a second investigator might place a higher premium on the development of character and personality. The weight attached to any particular outcome will clearly depend upon the objectives desired, and hence in any comparison of two methods of instruction it is essential to have a clear statement of what changes in pupil behaviour are desired, or expected, and if possible, some objective measures of these changes. In the teaching of science many of the changes desired can be objectively measured by the conventional written tests, but this is not so with outcomes such as the development of practical powers or skills. Attempts must be continued to devise objective tests of these outcomes, rather than to depend upon subjective assessments or judgements.

6. References.

1.	H.E. Armstrong.	The Teaching of Scientific Method. MacMillan, 1910.
2.	Report.	Science in General Education. Appleton Century, 1937.
3.	F.W. Westaway.	Science Teaching. Blackie, 1929.
4.	Science Masters' Association.	The Teaching of General Science. (Interim Report). Murray, 1936.
5.	Science Masters' Association.	The Teaching of General Science, Part II. Murray, 1938.
6.	D.H.J. Marchant.	An Inquiry Into the Present Position of the Teaching of General Science in Secondary Schools. School Science Review Vol.XXV No.96, 1944.
7.	E.R. Downing.	School Review. Vol.XXXIII pp.688-697, 1925.
8.	P.O. Johnson.	Journal of Educational Research. Vol.XVIII. pp. 103-111, 1928.
9.	F.A. Riedal.	School Science and Mathematics. Vol.XXVII pp. 512-519, 1927.

CHAPTER 2.

INTRODUCTION TO EXPERIMENTAL INVESTIGATIONS.

1. Introduction.

The writer decided to initiate some experimental research for the purpose of investigating the value of individual experimental work in Physics by pupils during their first two years in a secondary grammar school. The experiments were conducted with pupils at Chesterfield School, in North Derbyshire during the years 1947 and 1948. During those two years the normal entry to the school was approximately ninety boys. As in most schools the pupils on entry were allotted at random to three or four 1st Forms and for their first year in the school followed parallel courses of instruction. At the end of their first year the pupils were given examinations in all subjects and on the basis of their scores were graded into three 2nd Forms. This again is a customary procedure and from the research worker's point of view creates some difficulty. The reasons for the grading vary from school to school and it may even be that in their second year the pupils cease to follow the same curriculum. To the research worker one of the first needs is selected groups or classes since by careful selection he can usually increase the precision of his experiments. Random samples are satisfactory for most purposes but "matched" samples are even better. The technique of preparing "matched" samples or classes has been used with good effect in many cases but often at the expense of considerable administrative inconvenience. The problem of "matching" the 1st Forms in a secondary school is particularly difficult since on entry the only criterions available are, as a rule, age, scores on Intelligence tests, and entrance examination results. Since in the present case it was impossible to arrange "matched" samples the writer decided to design his experiments in relation to the classes as already organised in the school. During their first year in the school the pupils' instruction in Physics was confined to very elementary work on Heat, Density, Archimedes' Principle and Flotation. In the second year the instruction was mainly concerned with more advanced work on Heat.

2. The Objectives of Instruction in Physics.

Since one intention was to apply two methods of instruction, namely a "Demonstration" method and an "Individual Experiment" method to various classes it was necessary to consider what changes in pupil behaviour were desired or expected and how these changes could be assessed. The objectives or changes in pupil behaviour are of course a matter of opinion and closely related to the age of the pupils. The writer finally decided that the following, while by no means comprehensive, were representative of the major objectives with young pupils,

(a) the acquisition of a knowledge of the empirical facts, and principles of the course,

(b) the ability, to solve problems by the application of scientific principles and facts and, to apply their scientific knowledge to explain facts of everyday life,

(c) the ability, to manipulate simple apparatus and, make simple measurements and observations with a reasonable degree of speed and accuracy,

(d) the ability to make simple deductions from their measurements and observations,

(e) the ability to solve small problems of a practical nature,

(f) to provide the pupil with some sense of accomplishment, and pleasure in his work, and to increase his interest in the subject.

Progress towards achievement of objectives (a) and (b) can be assessed by means of conventional pen and paper achievement tests and the writer decided to construct what in future will be referred to as "Theoretical Tests" for this purpose. The objectives (c), (d) and (e) are distinctly practical in nature and it was decided to design special "Practical" or "Experimental" tests to measure progress towards attainment of these objectives. Assessment of progress towards objective (f) is particularly difficult and in this case it was decided to rely on subjective opinions. In designing the tests considerable reference was made to three American publications.[1.2.3.]

3. <u>The Theoretical Tests.</u>

In constructing a "Theoretical" Test the choice lies between an objective or New-Type test and the more conventional Essay-Type tests. The relative advantages and disadvantages of the two types have been discussed by several authors[4.] and there is no doubt that both types are valuable, but the latter type are probably more valuable when applied to older pupils, since they tend to put a high premium on the verbal factor and powers of self expression. Even with older pupils however there is some evidence to show that good correlation may exist between scores on new type tests in Physics and the more conventional essay type tests.[5.] It was finally decided, that new type tests were most suitable for the writer's purpose, the main reasons being as follows:-

(a) a large number of questions could be set, and thus the whole field of knowledge under test could be more adequately sampled,

(b) the marking could be made more objective,

(c) the tests do not put a high premium on verbal facility and literary skill.

Many forms of new-type question have been devised such as, the open or simple recall, true-false, and multiple choice, but there is no conclusive evidence to show that any particular type has undoubted superiority over the others, although it may be possible that some questions are expressed better in one form than another. The writer felt more confident in his ability to devise items of the open or recall type and thus most of the items used in the "Theoretical Tests" were of this type. In the very few cases where multiple choice items were used the guessing correction was not applied. In marking the tests one mark was awarded to each item, or question.

4. The Practical Tests.

The testing of laboratory technique and ability at Practical Physics is very difficult and the present position if far from satisfactory. In the School Certificate Examinations a Practical Physics test is only compulsory in one or two cases but is compulsory for all Higher School Certificate Candidates. The Practical Physics paper usually consists of four questions, the candidate being required to answer two, and in only a few cases do the examining boards send their own external examiners to invigilate the test. Several serious objections are apparent. The sampling of the syllabus is obviously small and if no external examiner is present the candidate is assessed not on what he does but on what he writes. The marking can hardly be anything but subjective, and the main value of the procedure seems to lie in the fact that it does ensure that the candidates have followed a course in Practical Physics in preparation for the examination. Very little work has been done on the design of new-type objective Practical Physics tests. J.W. Cox[6.] has carried out extensive research into the problems of measuring mechanical aptitudes and skills and has designed reliable tests of manual dexterity and mechanical aptitude. W.P. Alexander[7] has designed a Performance Scale to measure practical ability using the Passalong, Koh's Block Design and the Cube Construction tests. The Bennet-Fry[8] Mechanical Comprehension Test is a very good example of efforts to use pen and paper tests for testing mechanical aptitude. The writer had no intention of attempting to measure the Practical Ability or Mechanical Aptitude of the pupils "per se". The need was for a "Practical Test" of the abilities (c), (d) and (e) as given in Paragraph 2 page 11. Three methods of constructing such a test were considered.

(A) The pupil could be given a series of practical tests or problems involving measurements, observations and manipulations of apparatus. The pupil would then be observed at work and an attempt made to evaluate each step of the work as he proceeded. Such an individual testing method requires tact and sympathy since the continual close proximity of the examiner may have an adverse effect on the pupil's behaviour. One great objection to the technique is of course the time and labour required to administer such a test to large groups of pupils. A further objection is that the test cannot be applied in such a way that the attitude of the examiner is a constant factor.

(B) The examiner could set up apparatus and carry out simple experiments in front of the whole class. Any measuring devices employed could have scales large enough to be read by all members of the class. The class could then be asked to make certain measurements, observations and deductions. The method is attractive and very convenient for administration to large classes but suffers from the serious defect that the pupils are not active physical participators in the experiments and manipulations involved.

(C) The teacher or tester can give the pupils a series of individual experiments or operations to perform and base his assessment of the pupils' achievements on an evaluation of the product of their work. One advantage of this technique lies in the fact that the pupils are faced with real concrete situations. The examiner tends to be less obtrusive and evaluation of the end products makes objective marking reasonably easy. The pupil's speed and skill at manipulation can, to a limited extent, be evaluated by imposing a time limit for each experiment or operation. One objection to the technique is that failure to observe the pupil actually at work implies a great loss of valuable information. The emotional re-actions of the pupil and details of his technique are not observed.

The three methods each possess peculiar advantages and it may well be that a really satisfactory testing technique would be a combination of all three. It was however finally decided to concentrate on method (C) the deciding factors being:-

(a) The technique does present the pupils with concrete situations which are the essence of a "Practical Test".

(b) The technique is reasonably suitable for application to large classes.

5. Factors influencing Design of Practical Tests.

In designing the Practical Tests the following points were considered.

(a) The number of experiments, measurements, or problems should be as large as possible in order to obtain adequate sampling of the course of study and yet the time taken for the test must be kept within reasonable limits.

(b) The tests should mainly involve measurements and apparatus with which the pupils were already familiar.

(c) Where quantitative results were required the limits of permissible error must be carefully considered, taking into account the apparatus used and the age of the pupils.

(d) Efforts should be made to make the scoring as objective as possible.

At first it was hoped that it would be possible to include qualitative as well as quantitative experiments. It was soon found that the design of the latter was easier than the former particularly in view of the syllabi followed by the pupils and the desire for objective scoring. It was finally decided to employ quantitative experiments only and even these presented difficulties in certain branches of Physics. The design of suitable <u>short</u> experiments in such branches as Specific Heat and Latent Heat were particularly difficult.

6. <u>Preliminary Experiments on Practical Test Design.</u>

The Practical Tests were required for application to first and second year pupils, and early in 1947 it was considered advisable to carry out some minor preliminary experiments with third year pupils. Two third year classes were available and at intervals small experiments, based on the practical work that the pupils were supposed to have done in the previous two years, were designed and applied to the classes. By this means valuable experience was gained in three directions.

(a) Knowledge of the degree of accuracy that could reasonably be expected when the pupils were using certain measuring devices was obtained. At times for example a whole class would be asked to measure out 80 ccs of water using a measuring cylinder, or to measure the weight of an object with a spring balance.

(b) Knowledge of the time taken for the performance of certain experiments under examination conditions was obtained.

(c) Experience of various methods of administering small tests to large classes of thirty or more pupils was obtained.

The information gained in this manner was found to be very useful indeed and finally the writer had a collection of short practical test items which were considered suitable for first and second year pupils, and all of them had been tried out under examination conditions with <u>third</u> year pupils. It was of course realised that more accuracy and greater speed might reasonably be expected from these pupils than from first and second year pupils.

7. Statistical Analysis of Experimental Data.

Very few researches in the field of Education are now possible without at least some reference to statistical analysis. When the research is concerned with a methods experiment where the effects of two different methods of instruction are to be compared and criterion tests are to be devised, the statistical analysis becomes extremely important. However statistical theory must be the servant not the master. In the physical sciences we usually have very clear ideas and knowledge of what is being measured and the degree of accuracy with which it is being measured. In education matters are not so simple since in nearly all cases the human traits, abilities or things we are presuming to measure are essentially more complex than physical quantities like mass and length. The question of even the actual existence of the trait or ability which we are "measuring" is usually a highly controversial matter. In most cases we are really measuring changes in behaviour which we consider are a measurable tangible sign or indication of the particular ability or trait under consideration. For reasons of this kind it is important to avoid placing too much faith on the statistical results without first considering the reliability, and validity of the data on which it is based. In analysing the results of the experiments described in the following chapters the writer found the works of three authors particularly valuable.

(a) P.E. VERNON. The Measurement of Abilities. U.L.P. 1946.

(b) E.F. LINDQUIST. Statistical Analysis in Educational Research Houghton Mifflin Co. 1940.

(c) O.L. DAVIES. Statistical Methods in Research and Production. Oliver and Boyd. 1947.

To save the necessity for long discussions on statistical technique and the inclusion of long mathematical calculations in the succeeding chapters, the following device has been employed. Whenever a certain statistical technique has been used - for the first time - a reference is given to a page or section of one or more of the above works where fuller details of the technique can be found, the references being in such form as, P.E. Vernon page 16, or E.F. Lindquist, Chapter IV etc. In the majority of cases the raw scores from which each statistic has been derived are given in full. Constant use was made of the methods of analysis known as the analysis of variance and the analysis of covariance. For convenience the symbols used in the following chapters are collected below with a brief explanation of their significance.

X = an individual raw score or measure,

N = Number of measures in each group,

ΣX = Total sum of all the X measures for a stated group.

M = Mean score for a group.

$\sum X^2 =$ Total sum of all the values of X^2 for a given group,

x = Deviation of each score from the mean of the group,

$\sum x^2 =$ Total sum of all the values of x^2 for a given group.

d.f= Degrees of Freedom.

S.D.= Standard Deviation of a group.

The main methods of Computational Procedure were as follows.

$$M = \frac{\sum X}{N} \quad ; \quad \sum x^2 = \left\{ \sum X^2 - \frac{(\sum X)^2}{N} \right\}$$

$$S.D. = \sqrt{\frac{\sum x^2}{(N-1)}}$$

When calculating Correlations between two measures of different traits it was necessary to calculate the value of $\sum xy$ and $\sum XY$ where these terms have the following meanings.

X = Score on one test } by the same pupil
Y = Score on other test }

$\sum XY$ = Sum of products of all the X and Y scores.

x = deviation of an X score from the mean of all the X's for the group.

y = deviation of the Y score of the same pupil from the mean of all the Y's for the group.

$\sum xy$ = The sum of products of x and y for all members of the group.

$$\sum xy = \sum XY - \frac{(\sum X)(\sum Y)}{N}$$

r = Product-moment Correlation Coefficient.

$$r = \frac{\sum xy}{\sqrt{(\sum x^2)(\sum y^2)}}$$

8. <u>References and Bibliography.</u>

1. R.W. Tyler. Constructing Achievement Tests. Ohio State University, 1934.

2. Report. Science in General Education. Appleton Century, 1937.

3. H.E. Hawkes and others. The Construction and Use of Achievement Examinations. Harrap.

4. P.E. Vernon. The Measurement of Abilities. U.L.P. 1946.

5. H.P. Wood. Objective Test Forms for School Certificate Physics. British Journal of Educational Psychology. Vol.XIII Nov.1943.

6. J.W. Cox. Manual Skill. C.U.P. 1934. Mechanical Aptitude. Methuen. 1928.

7. W.P. Alexander. Intelligence, Concrete and Abstract. C.U.P.

8. Bennet-Fry. Mechanical Comprehension Test. Physical Science Aptitude Test. Psychological Corporation, New York.

CHAPTER 3.

THE FIRST EXPERIMENT. PART I. CRITERION TESTS.

1. Details of Groups used in Experiment.

If practical work either in the form of demonstrations by the teacher or individual experimental work by the pupil is expected to have an effect on the pupil's ability as measured by a Theoretical Test then one might expect a reasonable positive correlation between scores on a Practical Test and a Theoretical Test. In July 1947 an experiment was made to investigate this and other matters. Four first year forms were available for the experiment. The pupils in these four forms IA, IB, IC, and ID had all entered the school in September 1946 and had then been allotted at random to the four forms. All four forms had followed the same curriculum and their instruction in science had consisted of:-

(a) Autumn Term, 1946. Physics.
(b) Spring Term, 1947. Chemistry.
(c) Summer Term, 1947. Biology.

The four forms had each received four thirty five minute periods of science per week, two of these periods being combined to form a double period when they received their instruction in one of the laboratories. The other two periods were taken in ordinary class rooms. In all, three teachers were responsible for their instruction in science. Teacher "ab" was responsible for Forms IA and IB; Teacher "c" for Form IC; Teacher "d" for Form ID. Early in July 1947, the four Forms had been given some revision work in all the three branches of Science and had then taken their normal, end of the school year, examination in Science.

2. The Design of the Experiment.

/It

It was decided to design and apply to all four forms, two criterion tests.

(a) A Theoretical Physics Test.
(b) A Practical Physics Test.

These tests were applied at the end of July 1947 and no warning was given to either the pupils or their teachers so that there was no possibility of special preparation or revision by the boys or the teachers. The main objects of the experiment were, to obtain some information about the reliability and validity of the type of Practical Test that was to be used, and to investigate the correlation between the scores on the two tests.

3. The Theoretical Test.

A copy of the Theoretical Test applied is given below and it consisted of 28 items mainly of the open or simple recall type. Items 3, 5, 18 and 24 are multiple choice items to which there are three alternative responses. The test was constructed after a careful study of the syllabus, a representative sample of the pupils notebooks, and a copy of a Physics Test which had been applied to the pupils in December 1946. The number of items may appear rather small but the test was designed to take one hour and all pupils finished the test in an hour and none finished in less than fifty minutes. The pupils were warned that they would be penalised for guessing and in scoring one mark was awarded for each correct item and the guessing correction was not applied for the multiple choice items.[1] This procedure is justified in view of the small number of multiple choice items in the whole test.[2] The morality of the technique is perhaps open to criticism.

1. P.E. Vernon p.248.

2. P.E. Vernon p.275.

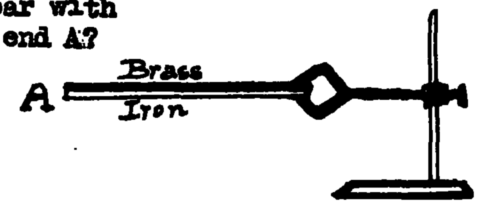
THE THEORETICAL TEST.

1. What is the normal temperature of a healthy person?

2. What is the name of the thermometer used by a doctor?

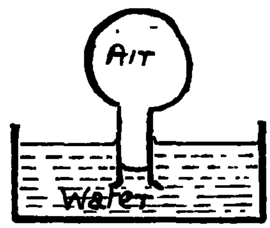
3. Do Telegraph wires sag more in summer than in winter?

4. A compound bar is made of brass and iron and clamped as shown in the diagram. On heating the bar with a bunsen burner, what happens to end A?



5. Does an iron ball weigh more when hot than when cold?

6. The sketch shows a flask containing air with its mouth under water at room temperature. What is observed if two warm hands are placed on the flask?



7. What is observed when the hands are removed?

8. What do you mean by the density of a substance?

9. One cubic centimetre of metal weighs 8 grams. What is the weight of 7 cubic centimetres of the metal?

10. A piece of glass weighs 24 grams and its density is 3 grams per c.c. What is the volume of the glass?

11. A piece of metal weighs 49 grams and has a volume of 7 ccs. What is the density of the metal?

12. The density of some wood is 42 lbs per cubic foot. What is the weight of a piece of furniture containing 8 cubic feet of the wood?

13. A measuring cylinder contains 92 ccs of water. Some metal is dropped into the cylinder and the reading of the water level is 126 ccs. What is the volume of the metal?

14. What is the density of pure water?

15. State the Principle of Archimedes.

16. A piece of wood weighs 82 grams and floats in pure water. What is the weight of the water displaced?

17. What volume of the wood is under water?

/18.

18. Which is heavier, a pint of milk or a pint of cream?

19. A piece of copper weighs 81 grams in air and apparently only 72 grams when completely immersed in water. What is the upthrust of the water on the copper?

20. What is the volume of the copper?

21. What is the density of the copper?

22. An empty beaker weighs 62 grams. 50 ccs of liquid are poured into the beaker and it is then found to weigh 122 grams. What is the weight of 50 ccs of the liquid?

23. What is the density of the liquid?

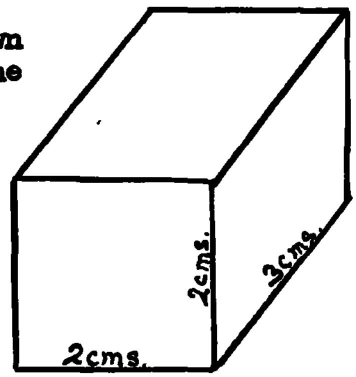
24. A piece of metal floats on mercury. Which has the greater density?

25. What is the name of the safety loading line marked on the side of all large ships?

26. Why does a man find it easier to float in the sea than in a river?

27. A block of metal has the dimensions shown and weighs 168 grams. What is the volume of the metal?

28. What is the density of the metal?



4. <u>The Practical Test.</u>

The practical test consisted of six major questions or problems and is given below. With the exception of Question 1 all the Experiments were closely related to actual experiments that the pupils were supposed to have either done or seen. The experiments were all divided into sub sections with the intention of making efforts at objective marking easier.

The Practical Test.

1. Name.................

Measure the length and breadth of the piece of cardboard correct to the nearest centimetre and then calculate its area.

/(a)

(a) Length of cardboard = cms.
(b) Breadth of cardboard = cms.
(c) Area of cardboard = sq. cms.

---

2. Name....................

The flask contains a boiling liquid and the beaker contains cold water. Measure the following temperatures:-

(a) Temperature of cold water = ^{o}C
(b) Temperature of boiling liquid = ^{o}C
(c) Temperature of vapour above boiling liquid = ^{o}C.

---

3. Name....................

You have a piece of metal, a piece of wood, thread and a measuring cylinder. Determine the following quantities:-

(a) Volume of metal = ccs.
(b) Volume of metal and wood together = ccs.
(c) Volume of wood. = ccs.

---

4. Name....................

The test-tube is loaded at the bottom so that it will float upright in water. You may assume that one cubic centimetre of water weighs one gram. Float the test-tube in water in the measuring cylinder and so determine:-

(a) Volume of test-tube under water = ccs.
(b) Weight of test-tube. = grams.

---

5. Name....................

You have a spring balance, beaker of water, thread and a piece of metal. You may assume that one cubic centimetre of water weighs one gram. Find:-

(a) Weight of metal in air = grams.
(b) Weight of metal when immersed in water= grams.
(c) Upthrust by the water on the metal = grams.
(d) Volume of metal = ccs.
(e) Density of the metal = grams per cc.

Name....................

You have a pipette, spring balance, a beaker of liquid and a small bottle. Find:-

(a) Weight of the bottle when empty = grams.
(b) Weight of bottle + 25 ccs of liquid = grams.
(c) Weight of 25 ccs of liquid = grams.
(d) Density of liquid = grams per cc.

---

5. Preparation of apparatus and Laboratory.

None of the Forms contained more than twenty pupils. Five complete and practically identical sets of the apparatus needed for the carrying out of each experiment were prepared and then distributed around the laboratory, in such a manner than no two sets of the same apparatus were adjacent. A large white card inscribed with the appropriate number of the experiment was placed beside each set of apparatus. For convenience these cards were mounted on wood blocks so that the cards were vertical and the number was printed on both sides of the card. The details of the apparatus for the individual experiments were as follows:-

Experiment 1.

The pieces of thick cardboard were cut on a guillotine and were rectangular in shape the dimensions being 18 cms x 12 cms. Half metre rulers graduated in inches and centimetres were provided.

Experiment 2.

The beakers of cold water were of 1000 ccs. capacity to minimize fluctuations in temperature during the course of the examination. The flasks contained saturated salt solution and the thermometers provided were graduated in degrees from -10°C to 110°C and all five were tested and chosen so that they gave the same readings correct to the nearest degree at all the temperatures involved in the experiment.

Experiment 3.

The pieces of metal were copper cylinders and were of the same volume - 8 ccs. The pieces of wood (oak) were also cylinders and had a volume of 6 ccs. The measuring cylinders were of 50 ccs capacity and were graduated in single ccs.

/Experiment 4.

Experiment 4.

The five test-tubes were carefully adjusted with wax and lead shot until they had the same weight (23 grams). The measuring cylinders had a capacity of 100 ccs. and were graduated in single ccs.

Experiment 5.

The spring balances provided were of the usual type with hooks at the bottom and were graduated in single grams from 0 to 100 grams. The pieces of metal were copper cylinders filed until they all had practically the same weight of 70 grams.

Experiment 6.

The spring balances were similar to those used in experiment 5, except that they were fitted with scale pans instead of hooks. The bottles were 50 cc density bottles and were adjusted to have the same weight (21 grams) by tying fine copper wire round the necks. The liquid provided was salt solution with a density of 1.10 grams per cc.

6. The Administration of the Practical Test.

A satisfactory technique for administering the test had been devised as the result of some experience in preliminary ventures with third year pupils (Chapter 2 - 6. page 16). Six VIth Form students volunteered to assist in the conduct of the examinations and were each given the task of controlling one particular experiment. Before entering the laboratory the pupils were given some verbal instructions as follows:-

(1) You are going to have a practical examination in Physics and will have to attempt six experiments.

(2) You will be allowed twelve minutes for each experiment and if you do not finish an experiment in that time you will have to leave it and go on to the next one.

(3) You will be provided with slips of paper for each experiment and they contain spaces in which you must enter your name, and the results of your experiment. Rough work can be done on the back of the slips.

(4) When you have finished one experiment, if you have time, reset the apparatus in its original condition ready for the next boy, and hold up your hand. Someone will then collect your slip of paper and tell you which experiment you have to do next.

/(5)

(5) Once you have entered the laboratory you may neither talk nor ask any questions.

(6) You may not ask any questions now.

The reason for the last instruction was to ensure that all classes received the same initial information before entering the laboratory. The pupils were then admitted to the laboratory, allotted at random to their initial experiment, and given the corresponding slip of paper, giving details of the experiment. It was considered inadvisable to give the pupils a sheet containing the instructions for all six experiments in case it caused distraction and increased any temptation to copy. Moreover if a pupil finished one experiment in less than twelve minutes he would have been able to devote some time to a preliminary consideration of the details of the other experiments. When a pupil had finished Experiment 1, the VIth Former in charge of that experiment collected the pupil's slip and made sure that the apparatus was re-arranged in its original state. The VIth Former then placed a fresh Experiment 1 slip beside the apparatus ready for the next candidate. At the end of twelve minutes the slips of all the pupils doing Experiment 1 were collected, even if they had not finished, and they were moved on to Experiment 2. The pupils who had finished Experiment 2, were moved on to Experiment 3, etc. The VIth Former in charge of Experiment 2, actually measured the temperature of the cold water in the beakers and made a pencil note of the result on the back of each pupil's slip <u>after</u> it was collected. This was to avoid any errors due to fluctuations in temperature during the test. The same VIth Former also made sure that none of the flasks were boiled dry and had spares available. The VIth Formers in charge of Experiments 3& 5, replaced the old pieces of thread by new pieces as each pupil finished. In no case did a pupil fail to complete the purely experimental part of each experiment before the

/end

end of the twelve minutes. At no time were there more than four pupils doing each experiment at the same time. Since five sets of apparatus were prepared for each experiment emergencies due to breakages etc. caused no serious trouble.

7. Comments on Technique of Administration.

Some of the experiments required less time than the others for completion by the average student. A pupil who finished a particular experiment quickly was at liberty to gaze around the laboratory and perhaps gain valuable information about the other experiments before commencing them. Moreover the pupils were not given identical treatment in so far as they did not all perform the six experiments in the same order. One solution to these problems was considered. Six separate rooms could be used, one for each experiment, and arrangements made for all pupils to start with Experiment 1 and then after twelve minutes pass to Experiment 2 in another room and so eventually to Experiment 6. This solution is very attractive but for a single class the time required to complete the test is increased by sixty minutes and it needs more space than is usually available in the average school. In preliminary experiments in administration, it was found that the Practical Tests could be applied without the assistance of the VIth Formers but for really efficient technique at least one spare administrator was needed. It is interesting to note that the VIth Formers took a keen interest in the work and their criticisms, which were always constructive, were very valuable indeed.

8. The Scoring of the Practical Test.

All the experiments were quantitative in nature and as a consequence

/the

the marking could be made at least objective in character, since the "correct answers to each item were known. The major problems in preparing a marking scheme were to decide what were reasonable permissible limits of error for each item, and how many marks should be awarded to each item. In new type pen and paper examinations there is considerable experimental evidence to show that weighting the marks according to the estimated difficulty of individual items is usually pointless since there tends to be a very high correlation between weighted and unweighted total scores.[1] This will probably not be so true for examinations containing a small number of items and twenty items which is the total for the practical test is comparatively small. In the Practical Test another justification for weighting the marks can be advanced. All measurements of physical quantities are subject to experimental errors and we could give more marks for a more "accurate" answer to each item. Considerable thought was given to this aspect of the problem. In general a very high degree of accuracy in experimental work is neither obtained, nor even desired from young pupils. The pupils had actually been accustomed and trained to read thermometers to the nearest degree, spring balances to the nearest gram, and volumes of liquids in measuring cylinders to the nearest cubic centimetre. It was finally decided to award only one mark to each item, and in those items involving actual measurements a departure of one gram, one cubic centimetre, or one degree from the "correct" value was marked as "correct". An exception was made in the case of Experiment 1 where only the responses 18cms; 12cms; and 216 square centimetres were accepted as correct. Items 5e and 6d were marked

/correct

1. P.E. Vernon P. 275.

correct if the results expressed either as vulgar fractions or decimal fractions were correctly deduced from the experimental results. No pupil was awarded a mark for an item involving a deduction from preceding experimental data if the preceding data had not been marked as "correct". A marking scheme on the above lines was prepared and the scripts were marked three times in all, twice by the writer and once by a colleague. No serious discrepancies were detected. It is important however to point out that the scoring was not entirely objective since the allocation of a mark to each item depended ultimately on a subjective opinion of the reasonable permissible limits of error. The total score of each pupil for all twenty items is referred to in future as his Practical Test Score.

9. <u>The Experimental Test.</u>

The Practical Test contained a number of items involving pure measurement, and manipulation of apparatus, by the pupils. These items were 1a, 1b, 2a, 2b, 2c, 3a, 3b, 4a, 5a, 5b, 6a, and 6b. In this list of items there is some doubt about the wisdom of including 3a, 3b and 4a since they are really deductions from two measurements and are therefore to this extent similar in character to items 3c, 5c, and 6c. Since however the actual measurements from which 3a, 3b, and 4a were deduced were not recorded and they were the primary recorded data for Experiments 3 and 4 it was finally decided to include these items in the list of experimental type items. The total scores on the twelve items listed above were calculated and in future are referred to as the pupil's Experimental Test, Score. This score was regarded as a measure of the pupil's ability to carry out very simple measurements and manipulations of apparatus. The setting of a time limit to each experiment was intended to penalize a pupil, to

/some

some extent, for lack of manipulative skill. In any practical task two factors are distinguishable on the basis of which the pupil's ability can be assessed. These two factors are the accuracy of the result and the rate at which it is attained. In this particular test no pupil failed to complete the items included in the Experimental Test and thus his score on this test is not influenced to any great extent by his rate of working. The Practical Test Score is however influenced to some extent by the pupil's rate of working. For example in Experiment 6 the pupil had to use a pipette and if it took him a long time to measure out 25 ccs of liquid then he had less time in which to complete items 6c and 6d. It should be noted that the skill with which, for example, the pupil used the pipette in Experiment 6 could to some extent be checked by his answer to items 6b, provided that he could use the spring balance correctly.

10. <u>The Raw Scores on the Criterion Tests.</u>

The raw scores obtained by the pupils on the three criterion tests:-

(a) Theoretical Test,
(b) Practical Test,
(c) Experimental Test,

are given below in tabular form. To economize on space the names of the pupils have not been given but each individual pupil can be identified by means of his form and a letter. For example Pupil "e" Form 1B scored, 14 on the Theoretical Test, 14 on the Practical Test, and 10 on the Experimental Test. A list of the individual responses of each pupil to each item in all the tests was prepared but has not been included below. Summaries of the results for each test with some of the more important statistics used in the later analysis of the results have also been given below.

Theoretical Test Scores.

Pupil	Form IA	Form IB	Form IC	Form ID
a	11	20	16	14
b	20	19	14	18
c	18	2	18	13
d	16	9	9	15
e	15	14	11	16
f	15	17	13	10
g	16	8	14	20
h	24	16	21	24
i	9	16	17	15
j	18	21	25	18
k	7	24	9	16
l	6	15	19	15
m	17	15	11	8
n	17	20	14	20
o	23	10	12	18
p	11	11	6	20
q	9	10	20	25
r	17	7	14	
s	18	19	13	
t	16	10	10	

(a) **Theoretical Test Summary.**

	Group.				
Statistic.	IA	IB	IC	ID	All Forms
ΣX	303	283	286	285	1157
N	20	20	20	17	77
M	15.150	14.150	14.300	16.765	15.026
ΣX^2	5051	4605	4502	5089	19247
Σx^2	460.560	600.560	412.200	311.065	1861.863
S.D.	4.923	5.622	4.658	4.410	4.950
Range.	6 - 24	2 - 24	6 - 25	8 - 25	2 - 25

/Practical

Practical Test Scores.

Pupil	Form IA	Form IB	Form IC	Form ID
a	15	13	10	10
b	19	10	6	14
c	11	10	15	16
d	12	11	8	15
e	14	14	10	12
f	15	11	5	16
g	7	10	13	16
h	13	6	11	13
i	7	10	14	13
j	16	13	16	15
k	15	15	9	11
l	8	13	15	9
m	10	10	12	14
n	7	8	7	20
o	16	11	3	7
p	12	11	17	16
q	10	8	14	20
r	14	10	15	
s	12	13	12	
t	8	6	11	

(b) Practical Test Summary.

	Group.				
Statistic.	IA	IB	IC	ID	All Forms
ΣX	241	213	223	237	914
N	20	20	20	17	77
M	12.050	10.650	11.150	13.941	11.870
ΣX^2	3137	2381	2775	3499	11792
Σx^2	232.960	112.560	288.560	194.931	942.777
S.D.	3.501	2.434	3.897	3.490	3.522
Range.	7 - 19	6 - 15	3 - 17	7 - 20	3 - 20

/Experimental

Experimental Test Scores.

Pupil	Form IA	Form IB	Form IC	Form ID
a	11	10	7	8
b	11	7	4	11
c	8	7	9	11
d	8	8	6	11
e	9	10	7	9
f	11	7	4	12
g	3	7	8	11
h	8	4	7	8
i	6	6	11	9
j	10	10	10	10
k	10	10	7	8
l	7	7	12	6
m	8	7	8	10
n	6	8	4	12
o	10	9	3	4
p	9	8	11	11
q	7	6	11	12
r	9	8	9	
s	9	9	11	
t	5	6	9	

(c) Experimental Test Summary.

	Group.				
Statistic.	IA	IB	IC	ID	All Forms
ΣX	165	154	158	163	640
N	20	20	20	17	77
M	8.250	7.700	7.900	9.588	8.312
ΣX^2	1447	1236	1388	1643	5714
Σx^2	85.760	50.200	139.800	80.122	394.532
S.D.	2.122	1.626	2.712	2.238	2.278
Range	3 - 11	4 - 10	3 - 12	4 - 12	3 - 12

/11.

11. The Reliability of the Theoretical Test.[1]

In order to obtain an estimate of the Reliability of the Theoretical Test the scores on odd and on even numbered questions were totalled separately for all pupils, and then inter-correlated. This gave a correlation coefficient of,

$$r = 0.733 \pm 0.036$$

where 0.036 is the Probable Error.

When corrected by the "Spearman-Brown" Prophecy formula[2] this gave a Reliability Coefficient R of

$$R = \frac{2r}{1 + r} = 0.846$$

Since the reliability of a test is almost synonymous with its thoroughness it can be increased by the addition of more items, these items being of course homogeneous with the original items. As a matter of interest it was decided to determine how long the test should be to obtain a reliability coefficient of 0.90. Using the "Spearman Brown" formula again it was found that a test 1.7 times as long would be required. This would involve a test of about 50 questions instead of 28 questions and would require approximately one hundred minutes for completion by the pupils. The increase in reliability was, in future tests, not considered more important than, the dangers of fatiguing the pupils with longer tests and, the administrative inconvenience of longer tests, and as a consequence all the Theoretical tests in this work were restricted to approximately thirty questions or items.

The reliability coefficient as calculated by the "Split Half" method is really a "Consistency Coefficient" or a measure of the self-

/consistency

1. P.E. Vernon p.145; 2 P.E. Vernon p.147.

consistency of the test. As Cattell[1] points out it might be advisable to retain the term "Reliability Coefficient" for correlations obtained on re-applying the same, or an equivalent, test after a reasonable lapse of time. In this experiment no opportunity for a re-application of the same test occurred. During the following year however the same pupils were given a similar type of Theoretical Physics Test based on their second year work in Physics and the inter-correlation of the test scores with the above scores was 0.671 (Chapter 8 - 9 p. 127.). In all cases the test papers were marked by two independent teachers and the writer, and no serious discrepancies in the final scores were discovered.

12. The Reliability of the Practical Test.

An estimate of the reliability of the Practical Test is very difficult. ~~The one~~ One reasonable method would be to re-apply the test after a reasonable lapse of time and correlate the two scores. This method suffers from the defect that the pupils might remember and be influenced by their previous responses and the alternative of setting a similar form of test instead is rather difficult. The "Split-Half" method of calculating the Reliability Coefficient or Consistency Coefficient tacitly assumes two things:-

(a) The test should contain a large number of questions or items.

(b) The items should be graded in difficulty. In actual fact it would probably be sufficient if the items were grouped in pairs of approximately equal difficulty and character, since when we correlate the total scores on alternate items we are more or less assuming that it is possible to split the test into two parts of equal length and difficulty.

The Practical Test consisted of a total of twenty items and, of these, twelve items were similar in so far as they involved actual

/measurements

1. R.B. Cattell. A Guide to Mental Testing U.L.P. p XV.

measurements and observations. These twelve items were grouped together to form what has been called, the Experimental Test. The remaining eight items involved simple deductions from the measurements recorded. The author decided to attempt to obtain some estimate of the reliability of the Practical Test by a modification of the "Split Half" method. The first step was to split the whole test into 10 pairs of items, the members of each pair to be as far as possible similar in difficulty and character. In some cases the pairing was obvious but in others the pairing is doubtful and is based solely on the subjective opinions of the writer and a colleague. The final pairing decided upon is shown below and for clearness the items involving measurement and observation only are shown in red.

1a	2a	3a	5a	6a	2c	5c	5e	4b	3c
1b	2b	3b	5b	6b	4a	6c	6d	5d	1c

Reference to the actual Practical Test will quickly show that the pairing of 1a with 1b for example is reasonable but that the pairing of 2c with 4a for example is very doubtful indeed.

The total score for each pupil on the items in the first row was correlated with the total score for each pupil on the items in the second row. This gave a correlation coefficient of

$r = 0.653 \pm 0.044$ where 0.044 is the probable error.

When corrected by the "Spearman Brown" formula this produced a Reliability Coefficient

$$R = \frac{2r}{1 + r} = 0.790$$

It is realised that the validity of this coefficient is not high and that it might be altered and even reduced by a re-arrangement of the pairs. The pairing of the items was performed before the construction of a table showing the total number of correct responses to

/each

each item. It is interesting to consult this table since it gives an objective measure of the difficulty of each item from the pupil's point of view (Chap. 3 - 15 p. 45.).

As a matter of interest the author decided to correlate the scores on alternate items in the Practical Test, without previous re-arrangement into similar pairs. For convenience the two halves of the test are given below and as before the experimental items are marked in red. ✓

1a	1c	2b	3a	3c	4b	5b	5d	6a	6c
1b	2a	2c	3b	4a	5a	5c	5e	6b	6d

It will be noted that the first row contains five experimental items whereas the second row contains seven experimental items. The result of correlating the scores on the items in the first row with the scores on the items in the second row gave

$$r = 0.789 \pm 0.029$$

When corrected by the "Spearman Brown" formula

$$R = \frac{2r}{1 + r} = 0.883.$$

The writer believes that the previous value of R = 0.790 is more valid but feels that even it can only be considered as a very approximate estimate.

13. Normality of Distribution of Test Scores.

The majority of good reliable well standardised objective educational achievement tests are well known to give a close approximation to a normal distribution when applied to a large number of pupils. In tests which are mainly diagnostic in character, the questions are intentionally designed with the purpose of discovering what parts of the subject have been mastered by the pupils, or vice versa. Such tests tend to give a

/pronounced

pronounced negatively shewed distribution since there is usually a deficiency of the more difficult type of question. Most examinations in schools are a combination of diagnostic and achievement tests and the writer's tests were intended to be of this character. A considerable number of statistical methods of analysing results assume that the scores being examined are normally distributed. It was decided to examine the distribution of the scores on all three tests and since it is generally accepted that small samples are only able to detect large divergences from normality the tests for normality were applied to the whole sample of 77 pupils. One fallacy in interpreting such tests for normality must be emphasized. A good reliable achievement test tends to give a normal distribution but the fact that an achievement test gives a normal distribution is not by any means certain evidence of its reliability. We are however justified in regarding it as a piece of corroborative evidence. The results of applying the χ^2 test for normality of distribution to the three criterion tests are shown below and reference to fuller details of the statistical technique involved are quoted.

<u>Theoretical Test.</u>

χ^2 Test for Normality of Distribution[1.2.]

Scores	fo	fe	$\frac{(fo - fe)^2}{fe}$
25 and over 22 - 24	2 4	1.69 4.24	0.0008
19 - 21	12	10.24	0.3025
16 - 18	20	16.32	0.8298
13 - 15	16	18.17	0.2591
10 - 12	11	14.25	0.7412
7 - 9 4 - 6 3 and below	9 2 1	8.01 3.08 1.00	0.0007
			χ^2 = 2.1341

1. Lindquist Chapter II. 2. Vernon p.102.

$\left(\begin{array}{ll} \text{Mean} & = 15.026 \\ \text{S.D.} & = \ \ 4.950 \\ \text{N} & = 77 \end{array}\right)$ $\left(\begin{array}{l} \text{fo = Frequency observed in each class} \\ \text{fe = Frequency expected in each class} \end{array}\right)$

Degrees of Freedom = 6 - 1 - 2 = 3.

For three degrees of freedom tables[1] show that χ^2 exceeds 2.13 more than 50% of the time. In more than fifty cases out of a hundred similar samples we might expect as great or greater deviations of the distribution from normality.

We can therefore have a high degree of confidence in the hypothesis that the Theoretical Test tends to give a normal distribution.

Practical Test.

χ^2 Test for Normality of Distribution.

Scores	fo	fe	$\frac{(fo - fe)^2}{fe}$
18 and over 16 - 17	3 8	3.16 6.08	0.3352
14 - 15	16	11.86	0.8313
12 - 13	14	16.09	0.2715
10 - 11	19	16.86	0.2716
8 - 9	7	12.55	2.4544
6 - 7 4 - 5 3 and below	8 1 1	6.78 2.62 1.00	0.0154
			χ^2 = 4.1794

$\left(\begin{array}{ll} \text{Mean} & = 11.870 \\ \text{S.D.} & = \ \ 3.522 \\ \text{N} & = 77 \end{array}\right)$ Degrees of Freedom = 6 - 1 - 2 = 3

For three degrees of freedom tables show that χ^2 exceeds 4.18 almost 24% of the time. On the basis of this result we have no justification /for

1. O.L. Davies p.268.

for rejecting the hypothesis of normality of distribution but our degree of confidence is not extremely high. An examination of the fo column shows a tendency towards a negatively skewed distribution and reference to Chapter 3 - 10 p 32 , shows that two pupils obtained the maximum possible score of 20.

Experimental Test.

χ^2 Test for Normality of Distribution.

Scores	fo	fe	$\frac{(fo - fe)^2}{fe}$
12 and over.	4	4.08	5.2528
11	12	5.01	
10	10	8.47	0.2764
9	11	11.70	0.0410
8	13	13.48	0.0171
7	12	12.94	0.0755
6	7	9.24	0.5430
5	1	6.30	4.4587
4	5	3.70	0.2575
3 and below	2	2.08	
			χ^2 = 10.9920

(Mean = 8.312)
(S.D. = 2.278)
(N = 77)

Degrees of Freedom
= 8 - 1 - 2 = 5

For five degrees of freedom tables show that χ^2 exceeds 10.99 only slightly more than 5% of the time. A divergence from normality such as exists here would occur by chance approximately only once in twenty similar samples of 77 pupils. Our confidence in the hypothesis of normality of distribution for the experimental test scores is as a consequence very low indeed. An examination of the actual frequencies observed in each class indicates a negative skew or tendency for the

/pupils

pupils to score high marks. It will be noticed that although no pupils scored no marks there are four pupils with the maximum possible score of 12. There seems little doubt that the experimental test has a very pronounced diagnostic character.

14. <u>The practical Test: Examination of Responses to Individual Items.</u>

Some evidence as to the relative difficulty of the various items can be obtained by tabulating the number of correct responses to the various items. This has been done in the table shown below and for convenience the items of the Experimental Test are shown in red. The results of the analysis are interesting and it should be remembered that Form IA and Form IB had been instructed by the same teacher. It is at once obvious that many of the experimental items were answered correctly by a very large percentage of the pupils and are therefore mainly of diagnostic value. For example very few pupils failed to answer items 5a, 5b and 6a correctly. These three items all involved a simple measurement of weight using a spring balance. Again item 2a involving a simple measurement of temperature was answered correctly by almost ninety percent of the pupils. One interesting test of the reliability of certain items is possible since in cases where two items involve similar measurements or manipulations we would expect the number of correct responses to be similar. A good example of this is provided by 1a and 1b which both involve measurements of length and the total number of correct responses is almost the same for the two items. A more rigid investigation would involve an examination of the individual pupil's scores to see if a pupil who got 1a correct also got 1b correct. This was done and it was found that 54 pupils got 1a and 1b correct.

/The

The items 2a, 2b, and 2c all involved the reading of temperatures yet the total correct responses are 68, 41 and 30 respectively. An examination of the individual responses seemed to indicate that many errors were due to unskilled manipulation in so far as care was not taken to ensure that in 2b the bulb of the thermometer was in the liquid and in 2c that the bulb was in the vapour. Items 3a, 3b and 4a involve similar measurements and processes and in these cases an examination of the individual scores showed that 24 pupils got all three items correct and 16 pupils got two out of the three items correct and 15 pupils got only one of these items correct. Items 5a, 5b and 6a were answered correctly by almost all the pupils but 6b is not quite identical with these items since in this case the result depends upon the pupil's ability to measure out 25 ccs of the liquid with the pipette. For 6b the fall in the number of correct responses from 6a is pronounced. This is very reasonable evidence that this item did measure the skill with which the pupils could use the pipette, since the responses to items 5a and 6a show that very few pupils had difficulty with the actual weighing. This analysis of the responses to the individual items does tend to give some added confidence in the reliability of the Practical Test and even in the validity of some of the items.

/Practical

Practical Test: Responses to Individual Items.

Item Number	Number of Correct Responses.				
	IA	IB	IC	ID	All Forms.
1a	14	13	14	15	56
1b	13	15	15	16	59
1c	12	11	10	12	45
2a	16	19	17	16	68
2b	14	6	9	12	41
2c	4	5	14	7	30
3a	15	13	6	11	45
3b	9	9	5	10	33
3c	7	8	1	5	21
4a	11	10	8	12	41
4b	9	9	6	8	32
5a	19	20	20	17	75
5b	18	20	20	17	75
5c	15	16	17	16	64
5d	10	7	10	12	39
5e	7	4	4	6	21
6a	19	19	17	16	71
6b	13	5	13	15	46
6c	11	4	13	14	42
6d	5	0	3	2	10
Total No. of Pupils.	20	20	20	17	77

15. The Practical Test: Discriminative Value of Individual Items.

The total number of correct responses to any item is no valid indication of the true difficulty or discriminative value of the item from the educational point of view. A reliable item for example should be answered correctly by more pupils whose total score on the test exceeds the median score than by those whose total score lies below the median value. This effect of course may not be so pronounced in items which are mainly of diagnostic value. In order to investigate the validity

/of

of each item in the Practical Test, from this point of view, the following procedure was adopted. The total sample of 77 pupils was divided into four groups the dividing points of the groups being approximately the lower quartile, the median, and the upper quartile, for the whole sample. The first group or "1st Quarter" consisted of the nineteen pupils with the lowest total scores. The second group or "2nd Quarter" consisted of the nineteen pupils with the next lowest total scores etc. The last group or "4th Quarter" included the twenty pupils with the highest total scores. The total number of correct responses to each item by the members of each group were then determined and the results of this analysis are given below.

Discriminative Value of Individual Items.

Question or Item No.	Number of Correct Responses by			
	4th Quarter	3rd Quarter	2nd Quarter	1st Quarter.
1a	19	16	14	7
1b	19	17	15	8
1c	15	13	11	6
2a	19	18	16	15
2b	19	11	6	5
2c	14	5	8	3
3a	17	14	9	5
3b	14	11	4	4
3c	11	6	2	2
4a	16	11	7	7
4b	12	12	4	4
5a	20	19	17	19
5b	20	19	18	18
5c	19	16	16	13
5d	18	8	8	5
5e	10	8	2	1
6a	20	19	17	15
6b	18	15	8	5
6c	18	13	7	4
6d	6	1	1	2
No. of pupils in each group.	20	19	19	19

For an item to be valid the number of correct responses to each item should decrease regularly as we pass from the "4th Quarter" to the "1st Quarter". This is true for the majority of the items but it is at once obvious that for some of the items the decrease is so slight that it can have little significance. Such items are mainly diagnostic in value. It is however essential to remember that such items are not without value and an educational test can and usually does combine the advantages of diagnostic and achievement tests[1]. In general the items involving deductions from the actual measurements naturally show the most discriminative value. In item 2c there is an increase in the number of responses as we pass from the "2nd Quarter" to the "3rd Quarter". The whole 77 responses to this item were re-examined in the hope of finding some explanation, but without success. In items 5a and 6d the slight increases are obviously not significant. Regarded as an achievement test there is a lack of sufficiently difficult items. In every case except that of item 6d the correct responses were made by 50% or more of the "4th Quarter" pupils. One obvious method of improving the test suggests itself. The accuracy demanded could be increased for all the experimental items. An alternative would be to have two limits of permissible error for each measurement and award 2 marks per item for the more accurate and 1 mark per item for the less accurate response. The application and even extension of this principle seems on initial consideration valid and easy. The writer however had carried out some initial research of an exploratory nature and as a result of this research

/considered

1. H.E. Hawkes. Achievement Examinations p.26.

considered the above principle invalid or at least impracticable with the young pupils involved. As an example the case of weighing with a spring balance can be considered. The balances that the pupils were using could only be expected to weigh correctly to the nearest gram and they had only been instructed to weigh correct to the nearest gram. If for example the pupils had been instructed to read thermometers correct to the nearest half degree instead of to the nearest degree then it might have been valid to award more marks for higher accuracy in such a case. An examination of the individual experiment slips was interesting in this connection. It will be remembered that errors of one degree, one gram or one cubic centimetre were accepted as correct. In the vast majority of cases the errors of those pupils who failed to score on the measurement tests were very large indeed. With older pupils, using more sensitive measuring devices the principle of giving more credit for greater accuracy would be easier to carry out and more valid and this point will be discussed in more detail later.

16. <u>The Validity of the Practical Test.</u>

The validity of the Practical Test is very difficult to assess with any degree of confidence. The reliability of the test has been discussed from several points of view but the validity of a test can only be assessed with confidence if it correlates effectively with other reliable measures of the skills abilities or processes it is supposed to test. The writer discussed with several teachers the possibility of the teacher being capable of making a subjective estimate of the pupils ability in Practical Physics, and the majority felt that with first year pupils such an estimate would be extremely difficult and unreliable. An estimate based on their

/practical

practical notebooks and accounts of experiments is of course of little value and certain to be heavily biased by verbal ability. We have no objective criterion of a pupil's ability in Practical Physics - at present - and when objective criterions are lacking we must fall back on subjective opinions. A copy of the Practical Test was shown to six experienced Teachers of Physics and all considered that in general it was a reasonable and valid test of ability in Practical Physics in relation to the age of the pupils and their syllabus, the major objections being that question 1 was too simple and that the pupils would never complete the test in 75 minutes.

17. <u>The Theoretical Test: Examination of Responses to Individual Items.</u>

The four forms to which the tests were applied were originally random samples. The writer had not been responsible in any way for the instruction of the forms and it was possible that the Theoretical Test was not a fair sampling of the work covered by all the forms. For example some of the items might have been heavily biassed in favour of one form. The total correct responses to the items by the pupils of each form are tabulated below. In no case does a particular item appear to be very heavily biassed in favour of some forms. If such cases had been detected the question of discarding the results for such items from the total test scores would have been considered. The table does give some indication of the relative difficulty of the items and is partial evidence that the four forms although instructed by different teachers had followed the same syllabus and covered the same ground.

/Theoretical

Theoretical Test: Responses to Individual Items.

Question Number.	Number of Correct Responses. 1A	1B	1C	1D	All Forms
1	8	7	6	4	25
2	19	17	16	17	69
3	15	16	16	15	62
4	8	6	5	3	22
5	11	13	10	12	46
6	10	11	10	10	41
7	7	8	7	7	29
8	14	12	11	9	46
9	17	15	15	15	62
10	8	11	9	6	34
11	8	7	9	8	32
12	14	11	12	11	48
13	14	16	15	13	58
14	14	13	14	12	53
15	10	11	4	11	36
16	10	12	10	9	41
17	9	7	8	8	32
18	3	4	5	5	17
19	13	10	13	9	45
20	12	9	7	9	37
21	10	8	7	10	35
22	16	14	15	17	62
23	3	1	4	5	13
24	16	14	19	17	66
25	18	15	19	16	68
26	3	6	7	10	26
27	7	4	8	9	28
28	6	5	5	8	24
Number of Pupils	20	20	20	17	77

18. The Theoretical Test: Discriminative Value of Individual Items.

In order to examine the validity of the various items in the Theoretical Test the same procedure as in Chapter 3 - 15 page 45, was adopted. The results of this analysis are shown below. For the majority of the items the number of responses decreases as we pass from the 4th Quarter to the 1st Quarter. Two outstanding exceptions are items 18 and 24, both of /which

which are multiple choice items. There is some experimental evidence to show that multiple choice items may be less reliable than simple recall items. Both of these items, or variants of them have been used in some well standardised reliable objective Physics Tests and it was finally decided that there was not sufficient justification for removing these two items from the test. It is interesting to note that fourteen of the items, namely Items, 9, 10,11,12, 13, 16, 17, 19, 20, 21, 22, 23, 27 and 28 were all of a mathematical nature involving some simple calculations. The discriminative value of this group of items is quite good but not obviously better than the discriminative value of the non-mathematical items. More elaborate determination of the discriminative value of the items was not considered necessary, nor justified.

Theoretical Test: Discriminative Value of items.

Question Number.	Number of Correct Responses by 4th Quarter	3rd Quarter	2nd Quarter	1st Quarter.
1	11	8	3	3
2	18	18	18	15
3	20	17	15	10
4	8	5	5	4
5	13	12	12	9
6	17	11	7	6
7	16	4	5	4
8	16	13	10	7
9	19	16	16	11
10	14	13	4	3
11	14	8	9	1
12	17	14	13	4
13	20	16	13	9
14	19	17	13	4
15	14	8	10	4
16	15	12	8	6
17	15	11	5	1
18	6	1	8	2
19	18	14	10	3
20	17	13	4	3

Cont'd.......

Table Cont'd....

Question Number.	Number of Correct Responses by			
	4th Quarter	3rd Quarter	2nd Quarter	1st Quarter.
21	17	10	5	3
22	18	14	16	14
23	7	4	1	1
24	15	18	19	14
25	20	16	17	15
26	13	8	5	2
27	14	7	5	2
28	11	10	2	1

19. Reaction of Pupils to the Tests.

Some subjective opinions about the attitude of the pupils are of interest. The pupils displayed little enthusiasm for the Theoretical Test but their re-action to the Practical Test was rather impressive. The discipline during the conduct of the Practical Test was extremely strict but the pupils were obviously absorbed and interested in their tasks. It was the subjective opinion of several teachers who saw the tests in progress, and the VIth Formers who were assisting, that it was a long time since they had been pupils so obviously enjoying an examination. When the tests were finished and discipline was relaxed the pupils were very anxious to know when they could have another similar examination. The writer allowed a week to elapse and then asked how many of the pupils would like to stay after school some night for a similar examination. Eighty per cent of the pupils were keen to do so because "It was real fun doing something," and "It isn't like a real examination."

CHAPTER 4

THE FIRST EXPERIMENT PART II. CORRELATION OF TEST SCORES

1. Introduction.

It is well known that correlation coefficients based on small samples are unstable and unreliable, and most research workers tend to consider such coefficients as almost worthless when derived from samples of less than fifty pupils. A further complication is created by the fact that undue homogeneity or heterogeneity in the sample may be responsible for an unduly high value of the correlation coefficient.[1] In this particular experiment there were four groups or forms with a grand total of 77 pupils. Two methods of calculating the correlation coefficients in such a manner as to utilize all the results are possible.[2]

(1) The product-moment correlation coefficient can be calculated for the whole sample considered as a single intact group of 77 pupils. The value so obtained is usually referred to as r total. If this method is adopted it is wise to examine the samples to discover whether undue homogeneity or heterogeneity is present.

(2) The correlation "within classes" or forms can be calculated, by applying the methods of analysis of covariance. This value is usually referred to as $r_{\text{within forms}}$ and is essentially the average of the correlations between the two test scores that would be obtained for the separate forms or classes if all the forms had received the same instruction in the two subjects involved. Such coefficients are not affected by differences in the mean scores for each form and can be regarded, to a certain extent as the coeffient that would be obtained from a single total class or form of 74 pupils. This method can only be adopted if we can assume homogeneity of correlation from form to form, or in other words can assume that the correlation coefficients for the separate forms are the same except for chance differences.

---

1. P.E. Vernon, pp. 140-142. 2. Lindquist, pp. 219-228.

The writer decided to calculate the correlation coefficients for the criterion tests by both the above methods. As a preliminary it is important to examine the results for the four forms to see if undue homogeneity or heterogeneity is present. On entry to the school in September 1946 the pupils had been allotted at random to the four forms and it is interesting to note that in July 1947 after one year in the school the pupils were graded into three second year forms on the results of their year's work in all subjects, and approximately one third of each first year form was promoted to each of the three graded second year forms. This is evidence in support of the efficiency of the initial random sampling.

2. The Age of the Pupils.

There was a possibility that the whole group might contain some abnormally older or younger pupils or that one particular form might contain an undue proportion of the older or younger pupils. The table below gives an analysis of the ages of the pupils. The ages given are those on 31st December 1946 and the means have only been given correct to the nearest month since greater precision was considered unnecessary. All ages are given in months.

Group	N	Mean	Median	Range	S.D
Form 1A	20	137	139	127 - 143	5.8
Form 1B	20	136	135	128 - 148	7.3
Form 1C	20	136	136	125 - 143	5.2
Form 1D	17	137	136	123 - 147	6.5
All Forms	77	137	139	123 - 148	6.1

An examination of the anlysis shows that the deviations from the mean for the whole sample never exceed 2.5 times the standard deviation. It should be pointed out here that in samples of 20 pupils we would expect the measures to range from approximately Mean + 2.0 x SD to Mean - 2.0 x SD and in samples of 100 pupils to range from Mean + 2.5 SD to Mean - 2.5 SD. The pupil whose age was 123 months was actually pupil "p" in Form 1D and reference to the Raw Scores on the various tests shows that his scores were not exceptional. The analysis shows that from the points of view of mean age and dispersion we have no valid reasons for suspecting either undue heterogeneity or homogeneity in the four forms.

3. <u>Examination of Individual Scores on all Tests for Abnormally High or Low Scores.</u>

In random samples of known size we can state with some confidence between what limits we expect the individual scores on a test to lie if we have some justification for assuming the distribution of scores to be normal for the universe from which the samples are drawn. In the three tests used in this experiment we have no valid reasons for rejecting the hypothesis of normality of distribution with the exception of the Experimental Test. If therefore on examination of the individual scores we find a pupil with a score deviating by considerably more than 2.5 x S.D from the mean of the total sample we would have some justification for believing that a highly improbable event had occurred and might reject that pupil's score from our analysis of the results. An examination of the data (Ch.3 - 10, p.32) shows that no highly improbable scores are present. Pupil "C" in Form 1B Theoretical Test has a score of 2 which deviates by 2.63 x S.D from the mean of the whole sample of 77 pupils but the deviation is an isolated case and is not so large as to justify his exclusion. It is interesting to note that this same pupil had scores of 10 in the Practical Test, 7 on the Experimental test and an age of 143 months. In actual fact the odds against a score falling outside the limits Mean + 2.5 x S.D., and Means - 2.5 x S.D. are practically 80 to 1.

4. <u>Influence of Different Teachers on Homogeneity of Total Sample.</u>

The four forms had not been instructed by the same teacher and this might have produced undue heterogeneity in the total sample. An examination of the results for the Theoretical Test (Chapter 3-10, page 32) shows that the mean total scores for the four forms are different, that for Form 1D being the highest. From the point of view of testing for homogeneity it is important to discover whether these differences in means for the various forms are signficant of real differences possibly caused by the teacher variable, or whether they may be explained away in terms of chance fluctuations in random sampling. The most convenient statistical technique for examining this problem is R.A. Fisher's technique of analysis of variance.[1,2] This technique assumes that whatever factors may have caused a significant difference in the means of the four forms then these same factors will not have caused significant differences in the variances of the four forms. In the following sections the technique of analysis of variance has been applied to determine the significance of the difference in means for the four forms with respect to all three Criterion Tests. Only the final variance tables are given together

1. E.F. Lindquist, Ch.V. 2. O.L. Davies, Ch.V.

with the tests for homogeneity of variance but all the data from which they are derived are given in Chapter 3-10 p. 32 .

5. **The Theoretical Test: Significance of the Difference in Means.**

The Null Hypothesis:- The difference in means for the four forms on the Theoretical Test may be explained away in terms of chance fluctuations in random sampling.

ANALYSIS OF VARIANCE

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
Form 1A	460.560	19	24.240
Form 1B	600.560	19	31.608
Form 1C	412.200	19	21.694
Form 1D	311.065	16	19.442
Between Forms	77.478	3	25.826
Within Forms	1784.385	73	24.444
Total	1861.863	76	

$$F = \frac{\text{Between Forms (error) Variance}}{\text{Within Forms Variance}} = \frac{25.826}{24.444} = 1.057$$

For $df_1 = 3$ and $df_2 = 73$ suitable tables [1,2] show that F must exceed 1.65 to be significant at even the 20% level.

We can as a result of this analysis have high confidence in the null hypothesis provided that the assumption of homogeneity of variance is justified. The most satisfactory test for homogeneity of variance is the Bartlett Test[3] and on applying this test to the above results it was found that $\chi^2_0 = 1.189$ with three degrees of freedom. This is not even significant at the 70% level.

We can therefore feel confident that the difference in means for the

1. O.L. Davies p.272.
2. Fisher & Yates - Statistical Tables for Biological, Agricultural and Medical Research. Oliver and Boyd.
3. O.L. Davies p.113.

four forms on the Theoretical Test is not significant. The standard error of the mean for Forms 1A, 1B and C is 1.106 and for Form 1D is 1.199 these values being calculated by dividing the Within Forms variance by the number of pupils in each form and extracting the square root.[1] For convenience the various means are given below again.

Statistic	1A	1B	1C	1D	All Forms
Mean	15.150	14.150	14.300	16.765	15.026
Standard Error	1.106	1.106	1.106	1.199	0.563

6. <u>The Practical Test: Significance of the Difference in Means.</u>

<u>The Null Hypothesis:-</u> The difference in means for the four forms on the Practical Test may be explained away in terms of chance fluctuations in random sampling.

<u>ANALYSIS OF VARIANCE</u>

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
Form 1A	232.960	19	12.261
Form 1B	112.560	19	5.924
Form 1C	288.560	19	15.187
Form 1D	194.931	16	12.183
Between Forms	113.766	3	37.922
Within Forms	829.011	73	11.356
Total	942.777	76	

1. E.F. Lindquist P.102.

Applying the Bartlett Test for homogeneity of variance gave $\chi_0^2 = 4.173$ with three degrees of freedom and reference to tables shows that this is not significant at the 20% level. We can therefore have some confidence in the hypothesis of homogeneity of variance.

$$F = \frac{\text{Between Forms Variance}}{\text{Within Forms Variance}} = \frac{37.922}{11.356} = 3.339$$

For $df_1 = 3$ and $df_2 = 73$ F need only exceed 2.74 to be significant at the 5% level of confidence and 4.08 to be significant at the 1% level. Interpolating by means of a nomogram[1] the value of $F = 3.34$ is approximately significant at the 2.5% level.

As a result of the above analysis we can have little confidence in the null hypothesis, since such large differences in means would only be obtained once in approximately every 40 cases due to chance variations alone.

The most likely cause of the difference in means is the teacher variable, and to some extent different teachers are synonymous with different methods of instruction. The result is interesting in view of the fact that there was considerable subjective evidence to show that Teacher "d" favoured a method of instruction involving a considerable amount of individual experimental work by the pupils. The above analysis does not indicate that the performance of 1D is superior to that of all the other three forms. The significance of the individual differences between the four forms can however be evaluated by means of Students "t" test.[2] The standard error of the mean for any single form can be calculated by dividing the Within Forms Variance by the number of pupils in the form and then extracting the square root. The result of this analysis is given below.

---

1. O.L. Davies p.285.

2. E.F. Lindquist p.56, p.102.

Statistic	1A	1B	1C	1D
Mean Score	12.050	10.650	11.150	13.941
Standard Error	0.758	0.758	0.758	0.817

The standard error for a difference in means between any two of the forms 1A, 1B, and 1C is 1.072. The standard error for a difference in means between Form 1D and any one of the others is 1.112.

To be significant at the 5% level the difference in means between any two of the Forms 1A, 1B, and 1C, must exceed 2.187 and none of them satisfy this condition.

To be significant at the 5% level the difference in means between Form 1D and any one of the other three forms must exceed 2.269. Thus the differences in means between Form 1D and 1C and also between Form 1D and 1B are the only differences which are significant at the 5% level of confidence. The difference in means between Form 1D and 1B is actually significant at the 1% level. It must be emphasized that the above analysis does not prove that these differences in means are due to the teacher variable unless we can feel confident that all other extraneous factors which might have influenced the performance of the pupils had been completely equalised. It is well to remember that Forms 1A and Form 1B were both instructed by teacher "ab". The writer after consideration of these points decided that the whole sample of 77 pupils could be regarded as neither unduly homogeneous nor heterogeneous.

7. <u>The experimental Test: Significance of Difference in Means.</u>

<u>The Null Hypothesis:</u> The difference in means for the four forms

on the Experimental Test may be explained away in terms of chance fluctuations in random sampling.

ANALYSIS OF VARIANCE

Source of Variation	Sum of Squares	Degrees of Freedom	Variance
Form 1A	85.760	19	4.514
Form 1B	50.200	19	2.642
Form 1C	139.800	19	7.358
Form 1D	80.122	16	5.008
Between Forms	38.650	3	12.883
Within Forms	355.882	73	4.875
Total	394.532	76	

Applying the Bartlett Test for homogeneity of variance gave $\chi^2_0 = 4.725$ with three degrees of freedom and reference to tables shows that this is not signficant at the 18% level. We can as a consequence have some confidence in the hypothesis of homogenity of variance.

$$F = \frac{\text{Between Forms Variance}}{\text{Within Forms Variance}} = \frac{12.883}{4.875} = 2.64$$

For $df_1 = 3$ and $df_2 = 73$ F must exceed 2.74 to be significant at the 5% level. Interpolating by means of a nomogram the value of $F = 2.64$ is significant at the 6% level.

We are as a consequence of the above analysis not justified in rejecting the null hypothesis. For convenience the means and standard errors are given below for the four forms.

Statistic	1A	1B	1C	1D
Mean	8.250	7.700	7.900	9.588
Standard Error	0.494	0.494	0.494	0.536

It should be pointed out here that our confidence in the normality of distribution for the scores on the Experimental Test is not high (Ch.3-13 p.39.) but departure from normality has to be very significant indeed before tests like the above become invalid. The analysis gives no valid evidence for undue homogeneity or heterogeneity in the group of 77 pupils.

8. Correlation of Theoretical and Practical Test Scores.

Let $\left\{\begin{array}{l} x \text{ refer to the Theoretical Test Scores} \\ y \text{ refer to the Practical Test Scores} \end{array}\right\}$

ANALYSIS OF COVARIANCE

Group	$\sum x^2$	$\sum y^2$	$\sum xy$	$r = \frac{\sum xy}{\sqrt{\sum x^2 \sum y^2}}$
Form 1A	460.560	232.960	+107.850	+0.329
Form 1B	600.560	112.560	+100.050	+0.385
Form 1C	412.200	288.560	+113.100	+0.328
Form 1D	311.065	194.931	+69.765	+0.283
Within Forms	1784.385	829.011	+390.765	+0.321
Total	1861.863	942.777	+474.260	+0.358

Tables have been prepared to give the minimum value of r that will be significant at any given level. None of the values of r for the individual forms are significant at the 5% level but there appears to be little justification for rejecting a hypothesis of homogeneity of correlation for the four forms.

For samples of 77 pupils the minimum values of r required for significance at various levels are as follows:-

Level of Significance	10%	5%	2%	1%	0.1%
Minimum value of r	0.189	0.225	0.265	0.293	0.368

As a consequence of this both r within forms and r total are highly significant. Some measure of the reliability of these two coefficients can be obtained by applying R.A. Fisher's "z" technique.[1]

At the 5% level we can be confident that the true r within forms lies somewhere within the limits +0.104 to +0.509.

At the 5% level we can be confident that the true r total lies somewhere within the limits +0.145 to +0.537.

Correlations are often lowered or attenuated as the result of errors of measurement. In other words the correlations are lowered because the scores being correlated are only in part a true measure of what they <u>purport</u> to measure and in part the result of chance errors of measurement. If however the Reliability Coefficients of the two sets of scores are known then Spearman[2] has shown that a correction for attentuation may be applied. His formula is

$$r_{corrected} = \frac{r_{observed}}{\sqrt{(R_1 \times R_2)}}$$

where R_1 and R_2 are the Reliability Coefficients of the two tests.

In this case we have Consistency Coefficients rather than Reliability Coefficients available for both tests as follows:-

$$\begin{cases} R_1 = \text{Reliability Coefficient for Theoretical Test} = 0.846 \text{ (Ch.3.11.p3}\ldots \\ R_2 = \text{Reliability Coefficient for Practical Test} = 0.790 \text{ (Ch.3.12.p 38)} \end{cases}$$

1. E.F. Lindquist pp.211-214. 2. E.F. Lindquist p.233.

$$\therefore r_{\text{total corrected}} = \frac{0.358}{\sqrt{(0.846 \times 0.790)}} = 0.438$$

This value is very highly significant but it must be remembered that both R_1 and R_2 are really Consistency Coefficients and R_2 is very doubtful in both origin and value. As Lindquist has pointed out:-

> "The mistake has frequently been made of interpreting a correlation coefficient corrected for attentuation as the "true" correlation between the traits which the tests are supposed to measure, rather than as the estimated correlation between perfectly reliable measures of whatever the tests actually do measure".

The corrected value is in fact of little practical value.

The product -moment correlation coefficients are based on the assumption of linearity of regression and a scattergram for the two tests is given below. An examination of the scattergram, which is rather coarse in its grouping does give some justification for the assumption of linearity of regression. Since there was no pronounced indication of curvilinear regression more accurate tests for linearitywere not applied.[1]

<u>Scattergram: Theoretical and Practical Test Scores.</u>

---

1. E.F. Lindquist p.235.

Practical Test Scores

Theoretical Test Scores	3.4	5.6	7.8	9.10	11.12	13.14	15.16	17.18	19.20	n	mean
23-25						2	3		1	6	15.50
20-22			1		1	3	2		2	9	14.44
17-19			2	2	3	4	4			13	12.27
14-16		2	3	5	3	5	3			21	10.95
11-13	1	1		1	4		2			9	10.67
8-10		1	3	3	3	1	1			12	10.08
5-7			1	1			1	1		4	12.50
2-4				1						1	10.00
n	1	4	10	13	14	15	16	1	3	77	
mean	12.00	13.25	13.70	12.15	14.07	17.40	16.88	6.00	21.67		

The correlation between the Theoretical and Practical Test Scores is low. Now certain items in the Theoretical Test are similar in nature to some of those in the Practical Test. For example Item 13 in the Theoretical Test is very similar to items 3a, 3b, and 3c in the Practical Test. A fundamental difference of course lies in the fact that in the latter the pupil is presented with a concrete situation and it is probable that the capacity of the pupil to deal with such situations is the very essence of practical ability in Physics. An examination of the two tests (Chapter 3 - 3; Chapter 3 - 4;) shows that items 13, 16, 17, 19, 20, 21, 22 and 23 on the Theoretical Test are similar to the items of questions 3, 4, 5 and 6 of the Practical Test. The scores for each pupil on these restricted portions of the tests were computed and the correlation between the scores was found. The result was as follows:-

$r_{total} = 0.435$ and $r_{within\ forms} = 0.418$

The increased value of r was of course expected.

9. <u>Correlation of Theoretical and Experimental Test Scores.</u>

let $\left\{\begin{array}{l}\text{x refer to Theoretical Test Scores.}\\ \text{y refer to Experimental Test Scores.}\end{array}\right\}$

<u>ANALYSIS OF COVARIANCE</u>

Group	Σx^2	Σy^2	Σxy	$r = \frac{xy}{\sqrt{(\Sigma x^2 . \Sigma y^2)}}$
Form 1A	460.560	85.760	+24.250	+ 0.122
Form 1B	600.560	50.200	+59.900	+ 0.345
Form 1C	412.200	139.800	+112.800	+ 0.470
Form 1D	311.065	80.122	+6.353	+ 0.040
Within Forms	1784.385	355.882	+203.303	+ 0.255
Total.	1861.863	394.532	+210.377	+ 0.245

None of the values of r for the individual forms are significant at the 5% level with the exception of that for Form 1C. The value of r for Form 1D appears to be very low but tests[1] showed that none of the <u>differences in r</u> for the individual forms were significant at the 5% level.

Both $r_{within\ Forms}$ and r_{total} are significant at the 5% level, but not at the 2% level.

Applying Fisher's "Z" technique to examine the reliability of the coefficients shows that at the 5% level we can be confident that the true r_{total} lies somewhere within the limits +0.022 to +0.445.

A scattergram for the two tests is given below, and further tests for linearity of regression were not applied.

---

1. E.F. Lindquist p.214.

Scattergram: Theoretical and Experimental Test Scores.

EXPERIMENTAL TEST SCORES

THEORETICAL TEST SCORES	3	4	5	6	7	8	9	10	11	12	n	Mean.
23-25						2		3		1	6	9.67
20-22					1	1		2	4	1	9	10.11
17-19		1		1	2	2	4	2	2	1	15	8.67
14-16	1	3	1	2	3	4	4	1	2		21	7.29
11-13	1	1			1	2	1		3		9	8.00
8-10				4	3	1	2	1		1	12	7.75
5-7					1	1		1	1		4	9.00
2-4					1						1	7.00
n	2	5	1	7	12	13	11	10	12	4	77	
mean	14.00	15.00	16.00	12.29	12.33	15.46	14.82	17.80	15.67	18.50		

10. Discussion on Correlations and Validity of Tests.

If the correlation between two tests is known then there is a possibility that a pupil's probable score on the first test may be forecasted from a knowledge of his score on the second test and vice versa. The accuracy or extent to which such forecasting will be more accurate than pure chance guessing is usually expressed in terms of the "forecasting efficiency" which is equal to $100(1 - \sqrt{1 - r^2})$ %[1]. The forecasting efficiency for the various correlation coefficients, (within forms) obtained in the previous paragraphs are tabulated below

1. P.E. Vernon p.127

Inter Test Correlations	r within forms	Forecasting Efficiency
Theoretical and Practical. Restricted Items.	+0.418	9.2%
Theoretical and Practical. Complete Test Scores.	+0.321	5.3%
Theoretical and Experimental Test Scores.	+0.255	3.3%

This table makes it quite clear that any attempt to forecast a pupil's score on the Practical or Experimental Test from a knowledge of his score on the Theoretical Test is practically valueless being little better than a pure chance guess.

The low values of the correlations obtained are not unexpected and they are in some degree corroborative evidence of the validity of the tests. High correlations would have indicated that the tests were to a large extent measures of the same abilities or traits. The tests were designed to measure different abilities or outcomes and hence we would expect low correlations. When correlating the scores on any two tests we can regard one of the tests as a mixture of three components, so far as its efficiency as a measuring device is concerned.

(a) A component consisting of that group of factors which is measured to some extent by both tests. The magnitude of this "Communality" is indicated to some extent by the correlation of the two tests.

(b) A component consisting of that group of factors or abilities which is peculiarly measured by the test under consideration. This property of the test is often referred to as its "Specificity" or uniqueness.

(c) A third component which is due to the lack of perfect reliability and validity of the test. The test is subject to errors in measurement and may in addition be measuring abilities that were not anticipated when the test was designed. This residual component is often combined with (b) under the general term of "Specificity".

An examination of the correlation coefficients brings forward several important points. The communality of the Experimental and Theoretical Tests is very low. There seems no doubt that these two tests are to a large degree measuring different abilities. When the Theoretical Test is correlated with the Practical Test the communality appears to increase and increases still further when we restrict the correlation to certain portions of the two tests. This apparent increase in communality might be expected from a subjective examination of the material of the various tests.

If the performance of practical work by the pupil is likely to influence the pupil's success and progress as measured by a Theoretical Test then the communality of the various tests is of great interest. The problem reduces, or more correctly increases, to that of discovering what factor or factors are responsible for the communality. It may be that both tests are measuring some ability or abilities that are specific to Physics alone or it may be that they are measuring in common some factors of a more general nature. The solution to this problem is naturally very complex, would involve the application of the methods of multiple factor analysis,[1,2] and was outside the scope of the writer's present enquiry. This question however will be discussed at more length in a later chapter.

There is at least some indication that the communality of the tests may, in part, be due to a general group factor such as the Numerical (N) factor of L.L. Thurstone.[3] This "N" factor is of course concerned with facility with numbers rather than general mathematical

1. L.L. Thurstone. Multiple Factor Analysis. C.U.P.
2. G.H. Thomson. The Factorial Analysis of Human Ability. U.L.P. 1939.
3. L.L. Thurstone. Primary Mental Abilities. Un. Chicago Press. 1938.

or arithmetical ability. The Theoretical Test can be divided into two sections, the "numerical" section including items 9, 10, 11, 12, 13, 16, 17, 19, 20, 21, 22, 23, 27 and 28, and the "non numerical" section consisting of the remainder. The Experimental Test was composed entirely of quantitative items and might therefore be influenced by the numerical ability of the pupils. The inter correlations between the two sections of the Theoretical test and the Experimental Test were calculated by the methods of analysis of covariance and the final results for the complete group of 77 pupils are given below, with their probable errors.

Tests Inter Correlated	r total	$r_{within\ forms}$
Theoretical (non numerical) - Experimental.	0.197 ± 0.073	0.161 ± 0.075
Theoretical (numerical) - Experimental.	0.411 ± 0.064	0.389 ± 0.065
Theoretical (numerical) - Theoretical (non numerical)	0.561 ± 0.053	0.560 ± 0.053

The results do give some indication that the numerical factor may be responsible for <u>some</u> of the communality between the Experimental and Theoretical Test scores. The comparatively high correlation of 0.560 between the two sections of the Theoretical Test suggests that these two sections are to some extent measuring similar abilities. The abilities in common may be the general intelligence or "g" factor, the verbal or "v" factor, some factor connected with memory or retentivity ability, and some factor, or factors, which are specific to Physics.

The above discussion on the sources of the communality is mainly speculative in nature, since with the numbers involved in the experimental group the reliabilities of the correlation coefficents which are quoted are very low. For example at the 5% level the true r total for the Theoretical "numerical" and Theoretical "non numerical" scores lies within the limits 0.384 and 0.698. The discussion does however suggest possibilities for further research.

CHAPTER 5.

The Second Experiment. 1st Year Pupils. Criterion Tests and Methods Experiment.

1. The Groups used in the Experiment.

In September 1947 the new entries to the school were assigned at random to three forms and these forms followed the same curriculum. In Physics the forms began a course in Elementary Physics lasting for a term of thirteen weeks, the syllabus being confined to elementary work on Heat, Density, Specific Gravity and Flotation. In all the pupils received four periods per week devoted to Physics, two single period lessons of thirty five minutes duration being taken in ordinary class rooms and one double period of seventy minutes being taken in the laboratory. So far as general instruction was concerned the major difference in the treatment of the forms was that the classes were not taught the same subjects at the same time or even on the same day. This difficulty might have been avoided or reduced by a cyclic interchange of the time tables of the three forms every week but such an arrangement was not possible in the present case for administrative reasons. Two of the forms 1D and 1P were both taught Physics by the same teacher, who was recognised as a good disciplinarian, and these two forms were used as the experimental groups in the present experiment.

2. The Age of the Pupils.

Although the pupils were assigned to the classes at random it was possible that one class or form might contain an undue proportion of older or younger pupils. The table below gives an analysis of the ages of the pupils. The ages being in months as on 31st December 1947.

AGE ANALYSIS.

Group.	N	Mean.	Median.	Range.	S.D.
Form 1P	28	141.2	141	136-147	3.2
Form 1D.	21	141.2	141	137-147	2.9
Both Forms	49	141.2	141	136-147	3.1

Now assuming that age is normally distributed, then in samples of this size we would expect to find the ages ranging from Mean - 2.33 x S.D. to Mean + 2.33 x S.D.(for samples of 50 pupils). In samples of 25 pupils the range might be less, from say Mean - 2.05 S.D. to Mean + 2.05 S.D. An examination of the above age analysis shows no reason for doubting the hypothesis that the two classes are random samples. It will be noticed by reference to Chapter 4 - 2 p. 54. that the pupils used in this experiment were slightly older than those used in the previous experiment.

3. The Design of the Experiment.

Both the forms, 1P and 1D, were taught by the same teacher and it was decided to carry out a simple methods experiment as follows.

(a) Form 1P was taught Physics with the accent on individual experimental work by the pupils.

(b) Form 1D was taught with the accent on demonstration work by the teacher.

(c) Care was taken to ensure that the same amount of factual knowledge was taught or presented to both forms and the homework and examples given in class were the same for both forms. It is of course appreciated that the conditions under which the pupils did their homework were not identical but there is at least some evidence that for boys in their third year at a Grammar School this factor may have little effect on their achievement.[1]

---

1. Sutcliffe and Canham. Experiments in Homework and Physical Education John Murray, 1937.

(d) At the end of the course of instruction in Physics both forms were given a Theoretical Test and a Practical Test of the type described in the previous chapters.

No effort was made to devise special experiments the principle being that whenever it was considered necessary for an experiment to be performed then in the case of Form 1P the pupils themselves, often working in pairs, performed the experiment, whereas in the case of Form 1D the Teacher performed the experiment with the active assistance of the pupils. In all cases experiments which were unsuitable for individual experimental work were demonstrated by the teacher. The obvious result of this technique was that the pupils in Form 1P received more experience in the actual handling and manipulation of apparatus. It is interesting to note that never once during the course of the experiment did the teacher find it necessary to slow down the speed of the work with Form 1D in order to keep the two forms parallel. The apparatus was always set out before the commencement of each lesson.

The main aim of the experiment was to compare the effects of the two methods of instruction on the achievement of the pupils as measured by both the Theoretical and Practical Tests. It was also a secondary aim to endeavour to obtain more information about the reliability and validity of the type of Practical Test that was used. In addition to the two criterion tests a group intelligence test was also applied to the pupils.

4. The Criterion Tests.

The Theoretical test was very similar to the one used in the previous experiment. It consisted of thirty items mainly of the "open" or recall type, and a copy of the complete test is given below. Two multiple choice items 12, and 29 were included. The pupils were warned that they would be penalized for guessing and in scoring one mark was awarded for each correct item and the guessing correction was not applied. The test was administered to the classes at the end of December, 1947, and a copy of the test is given below.

THE THEORETICAL TEST

1. On the Fahrenheit Scale pure water boils at________________

2. On the Centigrade Scale pure water freezes at________________

3. The normal temperature of a healthy person is________________

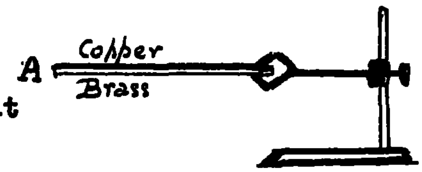
4. What is the name of the thermometer used by a doctor?

5. A doctor's thermometer is marked from__________to__________

6. A man appears to be asleep and his temperature is found to be 70° F. What would you conclude from this?

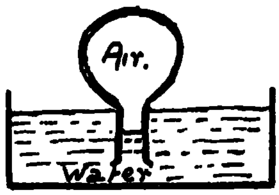
7. Explain why Telegraph wires sag more in summer than in winter.

8. A compound bar is made of copper and iron and clamped as shown in the diagram. The end A is heated with a bunsen. Explain what happens.



9. Explain why it is unwise to put a thick glass vessel into hot water.

10. The sketch shows a flask containing air with the neck under water, at room temperature. Explain what happens if two warm hands are placed on the flask.



11. Explain what happens when the hands are removed.

12. Does an iron ball weigh more when hot than when cold?

13. What do you mean by the density of a substance?

14. 1 cc. of iron weighs 8 grams. What is the weight of 7 ccs. of iron?

15. A piece of metal weighs 490 grams and has a volume of 70 ccs. What is the density of the metal?

16. A piece of glass weighs 24 grams and its density is 3 grams per cc. What is the volume of the glass?

17. The density of some wood is 40 lbs per cubic ft. What is the weight of a piece of furniture containing 7 cubic feet of wood?

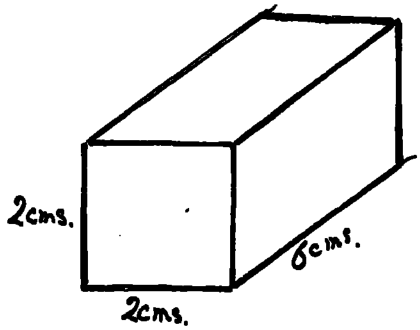
18. A measuring jar contains 73 ccs of water. Some metal is dropped into the jar and the reading of the water level is 96 ccs. What is the volume of the metal?

19. An empty beaker weighs 70 grams. 50 ccs. of liquid are poured in and the weight is then 130 grams. What is the weight of 50 ccs. of the liquid?

20. What is the density of the liquid?

21. A piece of metal has the dimensions shown and weighs 180 grams. What is the volume of the metal?

22. What is the density of the metal?



23. State the Principle of Archimedes.

24. A piece of copper weighs 80 grams in air and apparently only 72 grams when completely immersed in water. What is the upthrust of the water on the metal?

25. What is the volume of the copper?

26. What is the density of the copper?

27. A piece of wood weighs 80 grams and floats in pure water. What is the weight of the water displaced?

28. What is the volume of the water displaced?

29. Which is heavier, a pint of milk or a pint of cream?

30. Why does a man float more easily in sea water than in river water?

The Practical and Experimental Tests.

The Practical Test was identical with the one used in the previous experiment (Chapter 3 - 4.p.24.) and was applied to the pupils early in 1948. The test was administered and scored in the same manner as that previously adopted and again the total score on the items 1a, 1b, 2a, 2b, 2c, 3a, 3b, 4a, 5a, 5b, 6a, and 6b was obtained for each pupil and given the title of Experimental Test.

The Group Intelligence Test.

In October 1947 the Northumberland Standardised Tests (1925 Series) III General Intelligence (C. Burt) was applied to the pupils and the results were made available to the writer. This group test is well standardised and reliable and consists of nine highly valid sub-tests as follows:-

Test 1. Understanding Instructions;

Test 2. Opposites; Test 3. Similarities; Test 4, Mixed Sentences; Test 5, Completing Sentences; Test 6, Selecting Reasons; Test 7, Simple Reasoning; Test 8, Following an Argument; Test 9, Detecting Absurdities.

It should be noted that this test contains no sub-tests devoted to the completion of number series and it is very probable that the test is to a large extent a measure of the general intelligence ability or "g" factor of Spearman and the "v" group factor of verbal ability. The writer has no reliable evidence on this point but there is some evidence to show that similar verbal intelligence tests are loaded with the "g" and "v" factors.[1]

5. The Raw Scores on the Criterion Tests.

The raw scores obtained by the pupils on all four tests,

(a) Theoretical Test,

(b) Practical Test,

(c) Experimental Test,

(d) Intelligence Test,

are given below in tabular form. As before the names of the individual pupils have not been quoted but each individual pupil can be identified by means of his Form and a letter. In addition to the raw scores

1. W.P. Alexander. Intelligence Concrete and Abstract p.96 C.U.P. 1935.

summaries of the results for each test with some of the more important statistics used in the later analysis of the results have also been given below. All the tests were scored by two independent examiners and no serious discrepancies or differences in opinion were discovered.

Raw Test Scores 1P.

Pupil.	Theoretical Test.	Practical Test.	Experimental Test.	Intelligence Test.
a	11	11	7	273
b	14	13	9	297
c	19	17	11	278
d	22	18	10	307
e	17	6	5	290
f	21	11	8	310
g	19	5	4	269
h	19	11	9	276
i	15	16	10	260
j	18	12	8	258
k	18	12	9	281
l	18	14	9	253
m	22	14	9	282
n	15	14	9	268
o	21	11	8	301
p	19	8	7	290
q	20	13	9	295
r	18	9	6	280
s	16	8	6	284
t	22	12	7	247
u	23	16	9	282
v	22	11	9	320
w	18	10	8	265
x	14	9	7	281
y	13	12	8	260
z	21	9	7	284
a'	16	7	5	291
b'	20	15	10	254

Raw Test Scores, 1D.

Pupil.	Theoretical Test	Practical Test	Experimental Test	Intelligence. Test
a	8	6	5	277
b	18	7	6	273
c	14	14	9	275
d	23	17	12	271
e	20	10	7	267
f	21	15	10	307
g	25	9	8	301
h	26	18	11	276
i	15	7	5	294
j	21	5	4	221
k	18	7	7	260
l	24	8	5	290
m	19	13	10	256
n	22	13	8	270
o	19	13	9	289
p	19	12	10	313
q	17	10	7	270
r	21	8	5	305
s	14	11	7	264
t	20	12	8	279
u	9	10	8	305

Summaries of Test Scores.

(a) Theoretical Test.

Statistic	Form 1P	Form 1D	Both Forms.
ΣX	511	393	904
N	28	21	49
$M = \frac{\Sigma X}{N}$	18.250	18.714	18.449
ΣX^2	9585	7795	17380
Σx^2	259.236	440.275	702.066
$S.D = \sqrt{\frac{\Sigma x^2}{N-1}}$	3.098	4.691	3.825
Range.	11→23	8→26	8→26

(b) Practical Test.

Statistic	Form 1P	Form 1D	Both Forms.
ΣX	324	225	549
N	28	21	49
$M = \frac{\Sigma X}{N}$	11.571	10.714	11.204
ΣX^2	4038	2667	6705
Σx^2	288.884	256.284	553.932
$S.D = \sqrt{\frac{\Sigma x^2}{(N-1)}}$	3.271	3.579	3.394
Range.	$5 \rightarrow 18$	$5 \rightarrow 18$	$5 \rightarrow 18$

(c) Experimental Test.

Statistic	Form 1P	Form 1D	Both Forms.
ΣX	223	161	384
N	28	21	49
$M = \frac{\Sigma X}{N}$	7.964	7.667	7.837
ΣX^2	1853	1331	3184
Σx^2	76.957	96.668	174.719
$S.D = \sqrt{\frac{\Sigma x^2}{n-1}}$	1.688	2.199	1.889
Range.	$4 \rightarrow 11$	$4 \rightarrow 12$	$4 \rightarrow 12$

(d) Northumberland Test - Intelligence.

Statistic	Form 1P	Form 1D.	Both Forms.
ΣX	7,836	5,863	13,699
N	28	21	49
$M = \frac{\Sigma X}{N}$	279.857	279.190	279.571
ΣX^2	2,201,844	1,646,029	3,847,873
Σx^2	8,885.68	9,140.82	18,035.74
$S.D = \sqrt{\frac{\Sigma x^2}{N-1}}$	18.14	21.38	19.38
Range.	$247 \rightarrow 320$	$221 \rightarrow 313$	$221 \rightarrow 320$

6. The Reliability of the Tests.

(a) The Theoretical Test. Using the results for both forms the correlation of the scores on even and odd items gave a value of

$r = 0.472 \pm 0.075$ where 0.075 is the Probable Error.

Corrected by the Spearman Brown formula this gave a "Reliability" or "Consistency" Coefficient of $R = 0.642$.

This "Consistency" Coefficient is rather low and compares unfavourably with the value $R = 0.846$ obtained for a similar test used in the previous experiment but it must be remembered that this time the number of cases involved was only 49 instead of 77 and the two tests were not identical.

(b) The Practical Test. The estimation of the Reliability of this test was very difficult but as a matter of interest the correlation of the scores on two halves of the test was calculated. As before (Chapter 3 - 12 p 37) an effort was made to split the test into pairs of items of equal difficulty the final pairing decided upon being as follows, the experimental items being given in red.

1a	2a	3a	5a	6a	2c	5c	5e	4b	3c
1b	2b	3b	5b	6b	4a	6c	6d	5d	1c

The total score for each pupil on the items in the first row was correlated with the total score for each pupil on the items in the second row and this gave a correlation coefficient of

$r = 0.589 \pm 0.063$

Corrected by the Spearman Brown formula this gave a consistency coefficient of $R = 0.741$.

It must be again emphasized that this value can only be regarded as a very approximate estimate of the Reliability of the Practical Test.

7. The Practical Test: Discriminative Value of Individual Items.

Since the reliability and validity of the Practical Test was a matter of crucial importance it was considered advisable to again make some examination of the validity of the various items. The whole group of 49 pupils was divided into four groups the dividing points of the groups being approximately the upper quartile the median and the lower quartile. The group with the highest total scores contained 13 pupils, and is referred to as the "4th Quarter", while the other three groups each contained 12 pupils. The totalnumber of correct responses to each item made by the members of each group are tabulated below. The result of the analysis is very similar to that carried out in the previous experiment (Chapter 3 - 15 p. 45) In general the number of correct responses to each item does decrease as we pass from the "4th Quarter" to the "1st Quarter" the major exceptions being item 2c and item 6b. The individual scripts were re-examined but no errors in the scoring of these items were detected. Reference will show that item 2c in a previous analysis (Chapter 3 - 15 p. 45) showed a similar discrepancy. The writer decided that much of the trouble might be due to failure on the part of the pupils to clean the thermometer after answering item 1b. The general tendency for the experimental itemsto lack discriminative value is not quite so pronounced as in the previous experiment. In the "4th Quarter" three items were answered correctly by less than 50% of the pupils and in the "1st Quarter" thirteen of the items were answered correctly by less than 50% of the pupils. A more mathematical method of evaluating the discriminative value or validity of the individual items is available. The biserial correlation coefficient [1]

1. E.F. Lindquist. p 241-243.

r_{bis} can be calculated for the scores on the test and the responses to the individual items. This was done but the standard errors were not calculated since there is considerable doubt about the value of such calculations[1]. It is at once obvious that in general the discriminative value of the purely experimental items is less than that for the others. This is of course to be expected since a correct response to the latter was dependent upon correct responses to the former. The comparatively high values of r_{bis} for items 3a, 3b and 4a are interesting in view of the fact that these are the very items whose inclusion in the Experimental Test was of doubtful validity.

Practical Test: Discriminative Value of Individual Items.

Question or Item No.	Number of Correct Responses by 4th Quarter.	3rd Quarter.	2nd Quarter.	1st Quarter	r_{bis}
1a	13	11	11	8	+0.068
1b	13	12	11	8	+0.185
1c	11	8	9	5	+0.207
2a	13	9	11	9	+0.168
2b	10	8	8	4	+0.208
2c	10	5	7	4	+0.227
3a	8	6	4	0	+0.554
3b	4	4	2	0	+0.470
3c	4	2	2	0	+0.538
4a	10	7	3	2	+0.445
4b	9	5	3	2	+0.390
5a	13	11	11	10	+0.135
5b	13	10	9	8	+0.210
5c	12	8	7	7	+0.253
5d	11	7	3	2	+0.536
5e	11	5	6	1	+0.466
6a	12	12	10	7	+0.179
6b	10	8	1	4	+0.378
6c	10	8	1	1	+0.542
6d	5	0	0	0	+0.845
No of Pupils in each Group	13	12	12	12	

1. E.F. Lindquist P. 241-243

8. <u>Preliminary Examination of Random Groups.</u>

The two forms were initially random sample and so far as age was concerned there was no reason to doubt the efficiency of the random sampling. The results of the Northumberland Test however provided another method of checking the validity of the initial random sampling. This test was of course originally devised for administration to children in the age range of 10 to 12 years and was applied to Elementary or Primary School children. The mean score obtained by the pupils in the present experiment was approximately 280 with a standard Deviation of 19 whereas in the table of norms quoted by Burt we have:- Age last birthday 11 years, Average Score 205, standard deviation 39; Age last birthday 12 years, Average score 232, standard deviation 42. From the results of the test it would have been possible to convert the raw scores into Intelligence Quotients for each pupil. There is considerable evidence that the I.Q. remains reasonably constant for a given individual during both childhood and adult life[1]. For the writer's purpose however it was considered more reliable to compare the performance of the two forms as measured by the raw scores on the Intelligence Test. If the two forms were random samples we would expect no significant difference in their mean performance as measured by the Intelligence Test. The most convenient method of testing this was by the use of the Analysis of Variance[2] and only the final variance table is given below although all the data required for constructing the table is presented on page 80 .

<u>The Null Hypothesis:-</u> The difference in means for the two forms on the Intelligence Test may be explained away in terms of chance fluctuations in random sampling

---

1. C.S. Slocombe. The Constancy of "g". British Journal of Psychology XXVI p. 17, 1926.

2. E.F. Lindquist, Chapter V.

Form IP Mean = 279.857
Form ID Mean = 279.190
Both Forms Mean = 279.571

Analysis of Variance.

Source of Variation.	Sum of Squares.	Degrees of Freedom.	Variance.
Form IP	8885.68	27	329.099
Form ID	9140.82	20	457.041
Between Forms	9.24	1	9.24
Within Forms.	18026.50	47	383.54
Total.	18035.74	48	

$$F = \frac{\text{Between Forms Variance.}}{\text{Within Forms Variance.}} = \frac{9.24}{383.54}$$

and since this is less than unity we have no valid reason for rejecting the null hypothesis.

The analysis of Variance as applied above involves the assumption of homogeneity of variance. This assumption can be tested by means of the "F" test[1] applied to the two individual forms as follows.

$$F = \frac{\text{Form ID Variance}}{\text{Form IP Variance}} = \frac{457.041}{329.099} = 1.39$$

Tables show that with $df_1 = 20$ and $df_2 = 27$ then F must exceed 1.42 to be significant at the 20% level. We can as a consequence have high confidence in the assumption of homogeneity of variance for the two forms.

The normality of distribution of the scores also needs consideration and the result of applying the χ^2 test for normality of distribution to the scores is given below.

1. E.F. Lindquist Page 60.

Intelligence Test.

χ^2 Test for Normality of Distribution.

Scores	f_o = Frequency observed	f_e = Frequency expected	$\frac{(f_o - f_e)^2}{f_e}$
320 and over 310-319 300-309	1 2 6	0.91 1.95 4.30	0.4729
290	7	7.06	0.0006
280	8	9.87	0.3540
270	11	9.70	0.1742
260	8	7.57	0.0244
250 240 230 229 and under	4 1 0 1	4.54 2.10 0.98 0.02	0.3521
			χ^2 = 1.3782

Mean = 279.571
N = 49
S.D. = 19.384

Degrees of Freedom = 6 - 1 - 2 = 3.

Now for 3 degrees of freedom χ^2 exceeds 1.378 slightly more than 70% of the time. We can therefore accept the hypothesis of normality of distribution with confidence.

The method was applied to the whole group since small samples can only detect very large divergences from normality. The analysis shows that we can have a very high degree of confidence in the hypothesis of normality of distribution. An examination of the individual scores however was necessary. In samples of 50 pupils we would expect the scores to lie within the range of Mean + 2.33 x S.D. to Mean - 2.33 x S.D. There was one outstanding case in the results. Pupil "j" in Form ID had a raw score of 221 which was slightly more than 3 x S.D. below the mean of the whole group. There were as a consequence some grounds for rejecting this pupil's scores in the future analysis. Further consideration however was needed since the majority of the scores were within the limits, Mean $\pm$ 2.33 x S.D. The odds against a single score lying outside these limits are approximately 50 to 1. An examination of the same pupil's scores on the other tests showed that his score on the Theoretical Test was quite high at 21 whilst his scores on the Practical and Experimental Tests were low, but not unduly low. It was finally decided that there was no valid reason for rejecting the scores of pupil "j" from the future analysis.

In general terms we can with reasonable confidence assume that Forms ID and IP were random samples from the first year entries to the school.

9. Preliminary Examination of Criterion Test Scores.

On each criterion test the mean scores for the two forms are of course different and the important point is to discover if the differences are significant of real differences due to the effect of the two different methods of instruction or whether they may be due to chance fluctuations in random sampling. Before investigating this matter it is important to see if there are any scores of an improbably high or low value. As already pointed out, in a normal distribution, for a sample of 25 pupils, we would expect the scores to lie within the range Mean $\pm$ 2.05 S.D and for a sample of 50 pupils, within the range, Mean $\pm$ 2.33 S.D. Small samples can only detect large divergences from normality of distribution and evidence was produced in the previous chapters to show that criterion tests similar to those used here probably tend to give a normal distribution.

An examination of the raw scores for the theoretical Test, (Chapter 5 - page 77.), showed that there might be scores which were rather lower than might be reasonably expected. In the case of Form IP the score of 11 by Pupil "a" was rather low but not excessively so when it is noticed that the next higher score is 13. Moreover the Intelligence test score by pupil "a" in Form IP was almost equal to the mean score on the Intelligence Test. In Form ID there were two doubtful scores namely Pupil "a" with 8 and Pupil "u" with 9. The next highest score in Form ID was however 14. For Form ID a score of 8 is approximately 2.28 x S.D. below the mean of the Form and the odds against a score lying outside the limits Mean $\pm$ 2.28 S.xD. are approximately 43 to 1. The scores on the Intelligence Test for pupils "a" and "u" in Form ID are not unduly low or high. Considering the whole group of 49 pupils pupil "a" of Form ID with a score of 8 is approximately 2.73 x S.D. below the mean of the whole group. It was finally decided that there was not sufficient justification for rejecting Pupil "a" of Form ID from the future analysis, particularly since his divergence from the mean of his own group was not extremely large.

An examination of the raw scores for both the Practical and Experimental Tests discloses the presence of no improbably high or low scores. In order to examine the significance of the differences in means for the two forms on the various tests the technique of analysis of variance was employed and only the final Analysis of Variance tables are given in the following sections.

10. The Theoretical Test. Significance of Difference in Means.

The Null Hypothesis:- The difference in means for the two forms on the Theoretical Test may be explained away in terms of chance fluctuations in random sampling and is not significant of a real difference due to the different treatments of the two forms.

$$\left[\begin{array}{l}\text{Form IP} \quad \text{Mean} = 18.250 \\ \text{Form ID} \quad \text{Mean} = 18.714 \\ \text{Both Forms Mean} = 18.449\end{array}\right]$$

Analysis of Variance.

Source of Variation.	Sum of Squares.	Degrees of Freedom.	Variance.
Form IP	259.236	27	9.601
Form ID	440.275	20	22.014
Between Forms	2.555	1	2.555
Within Forms	699.511	47	14.883
Total	702.066	48	

$$F = \frac{\text{Between Forms Variance}}{\text{Within Forms Variance}} = \frac{2.555}{14.833}$$

and since this F is less than 1 we have no valid reason for rejecting the null hypothesis.

The above analysis assumes homogeneity of variance for the two groups or forms. Applying the "F" test to the individual variances for the two forms we have

$$F = \frac{22.014}{9.601} = 2.29$$

and with $df_1 = 20$ and $df_2 = 27$ we find that F must exceed 1.97 to be significant at the 5% level and 2.63 to be significant at the 1% level. Use of a suitable Nomogram[1] showed that F = 2.29 was approximately significant at the 2.5% level. Our confidence in the hypothesis of homogenity of variance was therefore very low indeed. The individual scores have already been considered and there appeared to be little justification for rejecting the low scores which are responsible for the greater variance of Form ID. As a matter of interest the analysis of variance

---

1. O.L. Davies P. 285.

was repeated with the score of pupil "a" for Form ID rejected and the result was still not significant. It must be pointed out here that the tests of significance involved in the analysis of variance are flexible and can be applied with reasonable accuracy to cases where the homogeneity of variance is not very marked[1].

The χ^2 test for normality of distribution of all the scores was applied and the final analysis is given below. It was not considered worth while to apply the χ^2 test to the individual forms.

The Theoretical Test.

χ^2 Test for Normality of Distribution.

Scores.	f_o = Frequency observed.	f_e = Frequency expected.	$\frac{(f_o - f_e)^2}{f_e}$
26 and over	1	1.19	
24 - 25	2	2.41	0.2042
22 - 23	7	5.06	
20 - 21	10	8.13	0.4301
18 - 19	14	9.99	1.6096
16 - 17	4	9.42	3.1185
14 - 15	7	6.80	0.0059
12 - 13	1	3.75	
10 - 11	1	1.58	0.6667
9 and under	2	0.67	
		χ^2	= 6.0350

Mean = 18.449
N = 49
S.D. = 3.825

Degrees of Freedom = 6 - 1 - 2 = 3.

Now for 3 degrees of freedom χ^2 exceeds 6.0350 in approximately 12% of random samples.

We can therefore hardly reject the hypothesis of normality of distribution, even though the results show a distinct tendency towards a negatively skewed distribution.

1. O.L. Davies p. 113.

We can as a result of the above analysis claim with reasonable confidence that the difference in means on the Theoretical Test is not significant and that the two methods of instruction have probably had no significant effect, on the progress of the forms as measured by the Test, since there were no important uncontrolled factors that might have affected the result.

11. The Practical Test: Significance of Difference in Means.

The Null Hypothesis. The difference in means for the two forms in the Practical Test may be explained in terms of chance fluctuations in random sampling.

$$\begin{bmatrix} \text{Form IP} & \text{Mean} = 11.571 \\ \text{Form ID} & \text{Mean} = 10.714 \\ \text{Both Forms} & \text{Mean} = 11.204 \end{bmatrix}$$

Analysis of Variance.

Source of Variation.	Sum of Squares.	Degrees of Freedom.	Variance.
Form IP	288.884	27	10.699
Form ID	256.284	20	12.814
Between Forms	8.764	1	8.764
Within Forms	545.168	47	11.599
Total	553.932	48	

$$F = \frac{\text{Between Forms Variance}}{\text{Within Forms Variance}} = \frac{8.764}{11.599}$$

and since this F is less than one we can have no valid reason for rejecting the null hypothesis.

Applying the "F" test for homogeneity of variance to the separate forms we get

$$F = \frac{12.814}{10.699} = 1.20$$

and with $df_1 = 20$ and $df_2 = 27$ tables show that this value of F is not even significant at the 20% level. We can as a consequence have high confidence in the hypothesis of homogeneity of variance.

The χ^2 test for normality of distribution of the scores for the whole sample is given below.

<u>Practical Test.</u>

<u>χ^2 Test for Normality of Distribution.</u>

Scores	f_o = frequency observed.	f_e = frequency expected	$\frac{(f_o - f_e)^2}{f_e}$
18 and over	2	1.11	
16 - 17	4	2.75	0.3827
14 - 15	6	6.18	
12 - 13	11	9.90	0.1222
10 - 11	10	11.34	0.1584
8 - 9	8	9.25	0.1689
6 - 7	6	5.39	0.0261
5 and under.	2	3.08	
		χ^2 =	0.8583

Mean = 11.204
N = 49
S.D. = 3.394

Degrees of Freedom = 5 - 1 - 2 = 2

Now for 2 degrees of freedom χ^2 exceeds 0.8583 slightly more than 75% of the time.

We can therefore accept the hypothesis of normality of distribution with confidence.

As a result of the above analysis we can have high confidence in the hypothesis that the two methods of instruction produced no significant difference in the two forms so far as those abilities measured by the Practical Test were concerned.

12. <u>The Experimental Test: Significance of Difference in Means.</u>

<u>The Null Hypothesis:</u> The difference in means for the two forms on the Experimental Test may be explained in terms of chance fluctuations in random sampling.

$$\begin{bmatrix} \text{Form IP} & \text{Mean} = 7.964 \\ \text{Form ID} & \text{Mean} = 7.667 \\ \text{Both Forms} & \text{Mean} = 7.837 \end{bmatrix}$$

<u>Analysis of Variance.</u>

Source of Variation.	Sum of Squares.	Degrees of Freedom.	Variance.
Form IP	76.957	27	2.758
Form ID	96.668	20	4.833
Between Forms	1.094	1	1.094
Within Forms	173.625	47	3.694
Total.	174.719	48	

$$F = \frac{\text{Between Forms Variance.}}{\text{Within Forms Variance.}} = \frac{1.094}{3.694}$$

And since F is less than one there is no significant difference and we can as a consequence have high confidence in the null hypothesis. Applying the "F" test for homogeneity of variance to the two forms gives

$$F = \frac{\text{Form ID Variance}}{\text{Form IP Variance}} = \frac{4.833}{2.758} = 1.75$$

and with $df_1 = 20$ and $df_2 = 27$ we find that F must exceed 1.97 to be significant at the 5% level and 1.70 to be significant at the 10% level. We can therefore have confidence in the hypothesis of homogeneity of variance.

The result of applying the χ^2 test for normality of distribution to the scores of the whole group of 49 pupils is given below.

Experimental Test.

χ^2 Test for Normality of Distribution.

Scores.	f_o = Frequency observed.	f_e = Frequency expected	$\frac{(f_o - f_e)^2}{f_e}$
12	1	0.67	
11	2	1.63	1.2962
10	6	3.87	
9	11	7.01	2.2710
8	9	9.64	0.0425
7	9	10.06	0.0112
6	3	8.02	3.1421
5	6	4.85	0.0012
4 and under	2	3.25	
		χ^2	= 6.7642

Mean = 7.837
N =49
S.D. = 1.888

Degrees of Freedom
= 6 - 1 - 2 = 3.

With three degrees of freedom χ^2 exceeds 6.764 in almost 7% of random samples.

We cannot completely reject the hypothesis of normality of distribution but our confidence in such a hypothesis is rather low and there is considerable indication of a negatively skewed distribution.

As a consequence of the above analysis we can have high confidence in the hypothesis that the two methods of instruction produced no significant difference in the two forms so far as their abilities as measured by the Experimental Test was concerned.

13. The Individual Items in the Theoretical Test.

The previous analysis showed that the two methods of instruction had produced no significant difference in the mean ability of the two forms as measured by the criterion tests. An examination of the Theoretical and Practical Tests (p 74) shows that there is a close correspondence between some items on both tests. For example Item 18 in the Theoretical Test was the counterpart to some extent of Items 3a and 3b in the Practical Test. Items 19 and 20 in the Theoretical Test correspond to Items 6a, 6b, 6c and 6d in the Practical Test. In many cases there were of course items in the Theoretical Test to which there were no corresponding items in the Practical Test. We might therefore expect that the two methods of instruction were only likely to cause statistically significant differences for those items in the Theoretical Test that had corresponding or similar items in the Practical Test. It must be emphasized here that, apart from question 1, the Practical Test was designed to sample the practical work actually performed by the pupils of Form IP and demonstrated to the pupils of Form ID by the Teacher. An analysis of the number of correct responses to each item is given below and for ease in comparison the number of correct responses by each form has also been expressed as a percentage, of the number of pupils in each form (correct to the nearest whole number).

The Theoretical Test.

Table of Number of Correct Responses, To each Question or Item.

Question No.	No. of correct Responses.		% of correct Responses.		Question No.	No. of correct Responses.		% of correct Responses.	
	IP	ID	IP	ID		IP	ID	IP	ID
1	25	16	89	76	16	23	19	82	90
2	24	17	86	81	17	22	17	79	81
3	12	11	43	52	18	24	21	86	100
4	26	20	93	95	19	22	15	79	71
5	3	17	11	81	20	11	8	39	38
6	18	17	64	81	21	17	18	61	86
7	23	18	82	86	22	10	10	36	48
8	11	5	39	24	23	15	18	54	86
9	11	11	39	52	24	27	21	96	100
10	26	18	93	86	25	23	13	82	62
11	6	7	21	32	26	13	8	46	38
12	27	10	96	48	27	24	18	86	86
13	20	16	71	76	28	19	13	68	62
14	27	20	96	95	29	3	2	11	10
15	19	17	68	81	30	6	2	21	10

Total number of pupils in $\left\{\begin{array}{l}\text{Form IP} = 28\\ \text{Form ID} = 21\end{array}\right\}$

The analysis shows that probably the test contained too many easy questions and was as a consequence rather diagnostic in character although this characteristic was also obvious from the test for normality,of distribution which indicated a definite tendency towards a positively skewed distribution. However the present important point is to discover whether the difference in the percentage of correct responses by each form is statistically significant. For example 54% of Form IP answered item 23 correctly whereas Form ID produced 86% of correct responses to the same item. There is the possibility that such a difference might be attributed to chance and not be indicative of a real difference caused by the different treatments of the two forms.

One statistical method of investigating the problem is to apply a 2 x 2 Contingency Table and the details of its application to item 23 are given below:-[1.2.]

	No. of Correct Responses.	No. of Wrong Responses.	Total.
Form IP	15 (18.8571)	13 (9.1429)	28
Form ID	18 (14.1429)	3 (6.8571)	21
Total.	33	16	49

The null hypothesis is that there is no significant difference between the performance of the two forms. Now $\frac{33}{49}$ of the <u>total</u> sample gave the correct response. We might therefore have expected

$\frac{33}{49} \times \frac{28}{1} = 18.8571$ of Form IP to make the correct response and

$\frac{33}{49} \times \frac{21}{1} = 14.1429$ of Form ID to make the correct response. The expected frequencies on this basis are shown in parenthesis in the above table. For each cell the difference or deviation which equals (Frequency observed) - (frequency expected) = $\pm$ 3.8571.

Applying the correction for continuity[3] the value of χ^2 is given by

1. E.F. Lindquist, p 41. 2. O.L. Davies, p.190. 3. O.L. Davies, p. 190

$$\chi^2 = (3.8571 - 0.5000)^2\left\{\frac{1}{14.1429} + \frac{1}{6.8571} + \frac{1}{18.8571} + \frac{1}{9.1429}\right\}$$

or $\chi^2 = 4.2705$ and has one degree of freedom.

The significance of this value of χ^2 can be obtained with some degree of confidence from the normal tables giving the percentage points of the χ^2 distribution provided that no expected frequency is less than five.

The above process and analysis was applied to all the items of the Theoretical Test with the final results shown below. The items for which one or more of the expected frequencies were less than five, have been marked with an asterisk, although in many of these cases the frequencies expected were only very slightly less than five.

Values of χ^2 for 2 x 2 Contingency Table Applied to each item of Theoretical Test.

Question No.	χ^2 from 2 x 2 Contingency Table.	Form giving Greater % of Correct Responses.	Question No.	χ^2 from 2 x 2 Contingency Table.	Form giving Greater % of Correct Responses.
1 *	0.7002	IP	16 *	0.3403	ID
2 *	0.0031	IP	17 *	0.0236	ID
3	0.1383	ID	18 *	1.6421	ID
4 *	0.0667	ID	19	0.0575	IP
5	21.6870	ID	20	0.0448	IP
6	0.9188	ID	21	2.9503	ID
7 *	0.0031	ID	22	0.2975	ID
8	0.6980	IP	23	4.2705	ID
9	0.3864	ID	24 *	0.0213	ID
10*	0.1159	IP	25	1.5896	IP
11	0.3687	ID	26	0.1701	IP
12	12.9288	IP	27 *	0.0000	-
13	0.0022	ID	28	0.1690	IP
14 *	0.2714	IP	29	0.0186	IP
15	0.4904	ID	30	0.5260	IP

For 1 degree of freedom we have

$$\chi^2 = 2.706 \text{ at the } 10\% \text{ level.}$$
$$= 3.841 \text{ at the } 5\% \text{ level.}$$
$$= 6.635 \text{ at the } 1\% \text{ level.}$$

An examination of the results shows that for items 5 and 12 the differences are significant at the 1% level, and for item 23 the difference is significant at the 5% level, while for item 21 the difference is significant at the 10% level. These are the only cases where significant differences are observed.

Now item 5 - "A Doctor's thermometer is marked form - to - " is not the type of question which we would expect to be considerably influenced by the two methods of instruction. It is much more probable that the teacher did not stress the matter equally with the two forms.

Item 12 - "Does an iron ball weigh more when hot" is again a question which is not likely to be influenced by the two methods of instruction. In fact this question or item is a multiple choice item and there are grave doubts as to the validity of even applying the 2 x 2 Contingency Table to this item.

Item 23 - "State the Principle of Archimedes - is however in a different category. Here we might reasonably expect some connection with the method of instruction. Form ID had seen experiments designed to verify the principle whereas Form IP had performed indentical experiments. It is conceivable that the teacher in the case of Form ID was better able to direct the pupil's attention to the principle whereas in the case of Form IP their attention was more concerned with manipulation of the apparatus. Other items, involving applications of the Principle show no significant differences.

Item 21 - Here the difference is not highly significant and we would not expect a great deal of effect due to the different methods.

To summarize we can be fairly confident that the two methods have produced practically no significant differences in the responses of the pupils to the individual items of the Theoretical Test.

14. The Individual Items in the Practical Test.

The same analysis as in the previous section was applied to the individual items in the Practical Test, and the results of the analysis are given below. Those items for which one or more of the expected frequencies were less than five have been marked with an asterisk.

The differences in responses for items 4b and 5b were apparently significant at the 5% level and for 5b the lowest expected frequency was almost four. For item 5c the difference in the number of correct responses was significant at the 1% level. The items 5b and 5c both dealt with the Principle of Archimedes and it will be remembered that the difference in responses to item 23 on the Theoretical Test was also significant at the 5% level. To advance an argument that Form ID would make better responses to questions on the Principle of Archimedes on the strength of such flimsy evidence would however be very dangerous. As already pointed out, significant differences restricted to a few isolated items might be due to more stress having been laid on certain parts of the syllabus by the teacher when dealing with one of the forms. In the case of Item 4b which was concerned with flotation it was Form IP that made the greater percentage of correct responses.

The general conclusion must be that there is no valid evidence to show that the two methods of instruction have produced significant differences in the responses of the two forms to the individual items.

The Practical Test.

Table of Number of Correct Responses, to each Question of Item.

Question of Item No.	No of correct Responses.		% of Correct Responses.		χ^2 from 2 x 2 contingency Table.
	IP	ID	IP	ID	
* Ia	26	17	93	81	0.6688
* b	27	17	96	81	1.6724
c	22	11	79	52	2.6477
* 2a	23	19	82	90	0.3403
b	19	11	68	52	0.6456
c	12	14	43	67	1.8589
3a	11	7	39	33	1.8295
* b	8	2	29	10	1.6360
* c	6	2	21	10	0.5260
4a	16	6	57	29	2.8900
b	15	4	54	19	4.6577
* 5a	24	21	86	100	1.6119
* b	19	21	68	100	6.2636
c	15	19	54	90	10.0900
d	13	10	46	48	0.0428
e	15	8	54	38	0.6166
* 6a	25	10	89	76	0.6998
b	13	10	46	48	0.0426
c	13	7	46	33	0.3963
* d	2	3	7	14	0.1159

Total number of pupils { Form IP = 28, Form ID = 21.

15. General Conclusions from Methods Experiment.

The two methods of instruction appear to have produced no significant differences in the performance of the two forms as measured by the three criterion tests. This result is of course restricted to the particular forms, and teacher involved in the experiment. The two methods may have produced differences of a significant nature, so far as abilities or outcomes not measured by the criterion tests are concerned. One method might have had a more beneficial effect on the interest of the pupils in Physics or have given them more emotional satisfaction. This aspect of the problem is of course extremely difficult. The subjective opinion of the Teacher was to the effect that no noticeable difference in the interest displayed, or enthusiasm shown by the pupils, was detected. So far as the examinations were concerned the pupils were enthusiastic about the Practical Test and obviously enjoyed the whole process.

The Theoretical Test had a reasonably high Consistency Coefficient and by subjective standards was typical of many achievement tests in Physics applied to first year pupils in Grammar Schools. There was as a consequence reasonable justification for considering that the two methods had produced no significant difference in the two forms as assessed by normal methods.

The Practical Test was however a more difficult problem since it is not customary to give such tests to pupils of this age in Grammar Schools. Some efforts have been made in the previous chapters to assess the reliability and validity of the Practical Test. Both the reliability and validity are difficult to assess and the validity in particular is difficult since no objective standard of ability in Practical Physics is available, at present, with which to compare the results of the test.

In the absence of such an objective standard the validity of the test must finally depend upon subjective opinions. The most that can be claimed with certainty is that the Practical and Experimental Tests were essentially measuring different outcomes to those measured by the Theoretical Test and the two methods of instruction produced no significant differences for the two forms with respect to those abilities which were measured by these tests.

In the following chapter an account of further efforts to investigate the validity of the Practical Test is given.

CHAPTER 6

THE SECOND EXPERIMENT: CORRELATION OF TEST SCORES

1. Introduction.

In this experiment four test scores were available for each pupil and the calculation of the various correlation coefficients should produce some useful information.

The calculation of the correlation between the Theoretical and Practical Test Scores should give at least some indication of the degree to which the tests were measuring different abilities, and hence might give some added confidence in the validity of the Practical Test. The calculation of the correlation between the Intelligence Test, and the Practical Test or Experimental Test might give some indication of the degree to which these latter tests were measuring the general intelligence "g" factor, the verbal "v" factor, or more specific factors.

As pointed out in Chapter 4 correlation coefficients based on small samples are usually unstable and unreliable so it was decided to calculate the coefficients for the whole sample of 49 pupils. In the previous chapter some evidence was produced to show that the two forms did not show either undue homogeneity or heterogeneity within the total sample. The two groups however had received different treatments so it was considered advisable to calculate both r_{total} and $r_{within\ forms}$ in each case. The original data from which the correlations were calculated has already been given in the previous chapter and in the following sections only the final analysis of covariance tables are given.

2. Correlation of Theoretical and Practical Test Scores.

Let $\begin{cases} x \text{ refer to the Theoretical Test Scores.} \\ y \text{ refer to the Practical Test Scores.} \end{cases}$

ANALYSIS OF COVARIANCE

Group	$\sum x^2$	$\sum y^2$	$\sum xy$	$r = \dfrac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}}$
Form 1P	259.236	288.884	58.000	+0.212
Form 1D	440.275	256.284	116.286	+0.346
Within Forms	699.511	545.168	174.286	+0.282
Total	702.066	553.932	169.510	+0.272

For samples of 49 pupils the minimum value of r required for significance at various levels is given below.

Level of Significance	10%	5%	2%	1%
Minimum value of r	0.238	0.281	0.332	0.365

As a consequence of this r within forms is just significant at the 5% level. In fact, applying R.A. Fisher's "Z" technique the reliability of the coefficient is low since at the 5% level we can only be confident that r within forms lies somewhere within the limits 0.000 to 0.522. There is no reason for rejecting the hypothesis of homogeneity of correlation for the two individual forms, although neither of the values of r for the separate forms are significant at the 5% level. The above analysis of course is based on the assumption of linear regression. A scattergram is given below.

Scattergram: Theoretical and Practical Test Scores

Theoretical Test Scores

PRACTICAL TEST SCORES

	8+	10+	12+	14+	16+	18+	20+	22+	24+	26+	n	Mean
18+								1		1	2	24.00
16+				1		1		2			4	19.00
14+				2		1	2	1			6	18.00
12+			1	1		5	2	2			11	18.18
10+	1	1		1	1	2	3	1			10	16.60
8+				1	1	2	2		2		8	19.25
6+	1			1	2	2					6	15.00
4+						1	1				2	20.00
n	2	1	1	7	4	14	10	7	2	1	49	
Mean	8.00	11.00	12.00	12.00	7.75	10.71	10.90	14.43	8.50	18.00		

Considering the smallness of the total sample there appears to be little justification for rejecting the hypothesis of linear regression or applying more exact tests.

Taking the reliability of the Theoretical Test as 0.642 and the reliability of the Practical Test as 0.741, $r_{\text{within forms}}$ corrected for attenuation by Spearman's formula was 0.409. This value of r is significant at the 1% level but is of doubtful validity since the reliability of the Practical Test is open to doubt.

3. Correlation of Theoretical and Experimental Test Scores.

Let $\left\{\begin{array}{l}\text{x refer to Theoretical Test Scores}\\ \text{y refer to Experimental Test Scores}\end{array}\right\}$

ANALYSIS OF COVARIANCE

Group	$\sum x^2$	$\sum y^2$	$\sum xy$	$r = \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}}$
Form 1P	259.236	76.957	30.250	+0.214
Form 1D	440.275	96.668	59.000	+0.286
Within Forms	699.511	173.625	89.250	+0.256
Total	702.066	174.719	87.592	+0.250

Neither r total nor r within forms is significant at the 5% level. In fact in 5% of cases we might even get a negative value of r.

A scattergram is given below and suggests no reason for rejecting a hypothesis of linearity of regression, particularly in view of the smallness of the total sample.

Scattergram: Theoretical and Experimental Test Scores

THEORETICAL TEST SCORES

EXPERIMENTAL TEST SCORES

	8+	10+	12+	14+	16+	18+	20+	22+	24+	26+	n	Mean
12								1			1	23.00
11						1				1	2	22.50
10				1		2	2	1			6	19.33
9				3		4	1	3			11	18.55
8	1		1			2	3	1	1		9	18.56
7		1		2	1	2	2	1			9	17.33
6					1	2					3	17.33
5	1			1	2		1		1		6	16.83
4						1	1				2	20.00
n	2	1	1	7	4	14	10	7	2	1	49	
Mean	6.50	7.00	8.00	8.00	5.75	8.07	7.60	9.14	6.50	11.00		

4. Correlation of Theoretical and Intelligence Test Scores.

let $\left\{\begin{array}{l} x \text{ refer to Theoretical Test Scores} \\ y \text{ refer to Intelligence Test Scores} \end{array}\right\}$

ANALYSIS OF COVARIANCE

Group	Σx^2	Σy^2	Σxy	$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}}$
Form 1P	259.236	8885.68	+424.000	+0.279
Form 1D	440.275	9140.82	-71.857	-0.036
Within Forms	699.511	18026.50	+352.143	+0.099
Total	702.066	18035.74	+348.429	+0.098

The result is very interesting since both r total and r within forms are very low indeed and neither of them are significant at even the 10% level. For form 1D alone r would have been negative implying that large scores on the Intelligence Test correspond on the average with low scores on the Theoretical Test. No significance however can be attached to this result since it is derived from a sample of only 21 cases. Moreover there is considerable evidence to indicate that all tests of mental abilities tend to give positive inter correlations, and moreover tests of manual, physical and other "non intellectual" abilities usually correlate positively with each other and with tests of mental abilities. A scattergram gave no obvious indication of curvilinear regression.

5. Correlation of Intelligence and Practical Test Scores.

let $\left\{\begin{array}{l} x \text{ refer to Practical Test Scores} \\ y \text{ refer to Intelligence Test Scores} \end{array}\right\}$

ANALYSIS OF COVARIANCE

Group	$\sum x^2$	$\sum y^2$	$\sum xy$	$r = \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}}$
Form 1P	288.884	8885.68	-227.714	-0.142
Form 1D	256.284	9140.82	+260.143	+0.170
Within Forms	545.168	18026.50	+32.429	+0.010
Total	553.932	18035.74	+66.286	+0.021

In this case both r total and r within forms are extremely small and have practically no significance. The values of r for the individual forms are also without significance since for samples of 28 pupils r must exceed 0.374 to be significant at the 5% level and for samples of 21 pupils r must exceed 0.433 to be significant at the 5% level. A scattergram suggested no reason for rejecting a hypothesis of linear regression

6. Correlation of Intelligence and Experimental Test Scores.

let $\begin{Bmatrix} x \text{ refer to Experimental Test Scores} \\ y \text{ refer to Intelligence Test Scores} \end{Bmatrix}$

ANALYSIS OF COVARIANCE

Group	$\sum x^2$	$\sum y^2$	$\sum xy$	$r = \frac{\sum xy}{\sqrt{\sum x^2 . \sum y^2}}$
Form 1P	76.957	8885.68	-11.143	-0.013
Form 1D	96.668	9140.82	+193.333	+0.206
Within Forms	173.625	18026.50	+182.190	+0.103
Total	174.719	18035.74	+184.571	+0.104

Once again r total and r within forms have very little significance and neither of the values of r for the individual forms are significant. Form 1D had not so much experience in the actual manipulation of the apparatus as Form 1D and the fact that the value of r for Form 1D is small but positive is interesting. In view of the smallness of the samples however this result is not significant and it would be absurd to claim that the correlation between the Experimental Test Scores and the Intelligence Test Scores would be greater for forms instructed by the Demonstration Method.

7. Discussion of Results of Correlation Analysis.

For convenience the correlation coefficients obtained in the previous paragraphs are collected below into a matrix and the coefficients given in the table are in all cases the values of r total.

Test	Theoretical	Practical	Experimental	Intelligence
Theoretical	-	0.272	0.250	0.098
Practical	0.272	-	-	0.021
Experimental	0.250	-	-	0.104
Intelligence	0.098	0.021	0.104	-

For samples of ~~45~~ 49 pupils r must exceed 0.281 to be significant at the 5% level and 0.238 to be significant at the 10% level.

Considering first the correlations between the Theoretical and Practical and Experimental Tests there was considerable similarity between these results and those obtained in the first experiment. The forecasting efficiency is about 3% and any attempt to forecast a pupil's ability as measured by the Practical Test from a knowledge of his score on the Theoretical Test would be practically useless. Again, the correlation between Theoretical and Practical was slightly larger than that between Theoretical and Experimental and this could reasonably be expected. The low values of the correlations are at least corroborative evidence in support of the claim that the tests were measuring different abilities but this evidence would be enhanced of course if more definite evidence of the Reliability of the Practical Test was available.

Considering the correlations between the Intelligence Test and the other criterion tests it is at once obvious that any real correlations which may exist are probably extremely small. Unfortunately very low correlation coefficients must be derived from very large samples in order to be significant and highly reliable. For example a correlation coefficient of 0.098 must be derived from a sample of 400 pupils in order to be significant at the 5% level and from a sample of almost 1000 pupils in order to be significant at the 1% level.[1] There is considerable evidence to show that the majority of the verbal group intelligence tests are heavily loaded with the general intelligence or "g" factor and the verbal or "v" factor. For example W.P. Alexander[2] produced some evidence to show that this was true for such tests as the Otis Group Test, the Terman Group Test and the Simplex Test, particularly when those sub-tests which were dependent on number were omitted. The Northumberland Test, Intelligence III, contains no sub-tests devoted to number and as a consequence there is probably a considerable loading of "g" and "v" in the complete test. A considera-

1. E.F. Lindquist p.212. 2. W.P. Alexander. Intelligence Concrete and Abstract.

tion of the correlations obtained in the present experiment even though obtained from a very small group and therefore of low reliability do suggest that neither the Theoretical nor Practical Tests are heavily loaded with the "g" and "v" factors. Since both these tests were in the nature of attainment tests, a low correlation with a measure of "g" is by no means unexpected.

It may be that the Practical and Experimental Tests are loaded with W.P. Alexander's "F" or Practical Ability factor and that their communality with the Theoretical Test is due to a numerical ability and a memory or retentivity ability. Only an application of the methods of factor analysis to a large experimental group could provide an adequate solution to these problems and it is very probable that such an analysis should, in the first place be conducted with older age groups. It is at least interesting to note that W.P. Alexander working with pupils of mean age seventeen years, from a technical high school in Chicago found the following factors present in the Science tests applied to them: "g" factor 12%, "v" factor 31%, "X" factor 55%.[1] The Science tests used were apparently normal school achievement tests although no details were given. The low percentage of "g" is noteworthy but the significance of the X factor is doubtful although Alexander did suggest it might be a psychological factor connected with the "interests" or "long-term persistence" of the pupil. A possibility is that this "X" factor may account to a large extent for the medium correlation which exists between the numerical and non-numerical sections of the Theoretical Test, rather than the memory and retentivity abilities which were suggested in the first experiment.

1. W.P. Alexander. Intelligence Concrete and Abstract Ch. VI. C.U.P. 1935

CHAPTER 7.

The Third Experiment: 2nd Year pupils: The Criterion Tests and Scores.

1. The Groups used in the Experiment.

In September 1947 the pupils who had been in Forms 1A, 1B, 1C and 1D, the previous year were regrouped into three graded Forms, - classified as 2A, 2B and 2C, - on the basis of their progress in all subjects of the curriculum. All three forms then commenced a course in Elementary Heat lasting for thirteen weeks, and consisting of four thirty-five minute periods per week. Forms 2B and 2C were taken by the same teacher referred to in future as "Teacher bc" and Form 2A was taken by "Teacher a". The three groups were not random samples. Initial measures of the abilities of each pupil as measured by the Theoretical and Practical Tests used in the first experiment were available. All three groups followed the same curriculum and each group received one double period per week in the Physics Laboratory.

2. The Age of the Pupils.

Some details of the ages of the pupils are given below, the ages being in months, as on 31st December, 1947.

Group	N	Mean	Median	Range	S.D.
2A	22	147.0	146	135 - 160	6.7
2B	20	147.5	147	137 - 159	6.6
2C	23	151.0	153	140 - 157	5.4
All Pupils	65	148.6	149	135 - 160	6.3

Considering the whole group of 65 pupils the ages lie within the range Mean $\pm$ 2.18 x S.D. approximately and there are no abnormally old or young pupils. Form 2C consisted of a group of pupils having less dispersion than the other two groups and a rather higher mean age.

3. The Initial Status of the Groups.

Initial criterion measures of the abilities of all the pupils in Physics as measured by the Theoretical Test and Practical Test described in the First Experiment, (Chapters 3; 4) were available. The Mean scores and some relevant data for these criterion tests are given below and the individual scores are given in a later section (p 123)

INITIAL THEORETICAL TEST SCORES

Group	n	Mean	Variance	S.D.
2A	22	17.773	23.557	4.85
2B	20	14.950	16.681	4.08
2C	23	12.261	22.564	4.75

INITIAL PRACTICAL TEST SCORES.

Group	n	Mean	Variance	S.D.
2A	22	12.409	12.541	3.54
2B	20	11.850	12.765	3.57
2C	23	11.217	10.087	3.18

The fact that the three forms were not random samples is reflected in the above summary. Since initial criterion measures were available, the effect of different methods of instruction on the progress of the forms might have been tested by application of the methods of analysis of covariance[1]. Since such a method of analysis tends to increase the precision of a methods experiment its use is highly desirable. This method of analysis is however generally limited to cases where the

1. E.F. Lindquist, Chapter VI

experimental groups are true random samples. The three groups involved were not random samples, but even if they had been true random samples there would have been chance differences in initial mean scores for the three groups. By means of the methods of analysis of variance the hypothesis that the actual differences in means was no greater than might have been obtained with true random samples was tested. The differences in initial means for the Theoretical Test were found to be highly significant. In similar situations experimenters, starting with non random samples have often discarded from their final analysis the results for such pupils as were necessary to make the initial differences in means and variances for the groups no greater than might be reasonably expected in true random groups. This procedure is dangerous since the ability of a class to benefit from a certain method of instruction is affected by the status of <u>all</u> the pupils in the class. The procedure might be justified if it involved for example the rejection of only one pupil's score in each group but even then would be dangerous.

Application of the analysis of variance showed that the differences in means for the initial Practical Test were not significant at the 5% level. This might appear to give some statistical justification for regarding the three groups as equivalent to possible random samples, so far as their ability as measured by the Practical Test was concerned. However even this possibility was considered improper in view of the facts that, the Practical Test was of rather uncertain reliability, and the original grading of the three groups was based on reasonably valid measures of the pupils average ability in all subjects of the curriculum.

A simple methods experiment might also have been applied to the forms and at the close of the experiment the results for certain pupils in each form discarded in such a manner as to make the means and standard deviations of the initial scores alike for all three forms. This technique of using "matched" groups would have made the final true experimental groups very small indeed, and the initial disparity in the groups as a whole was so large that this factor along would have had a very grave effect in the precision of such an experiment.

For the above reasons it was considered unwise to compare the effect of different teaching methods with the three second year forms since the precision of such an experiment was certain to be very low indeed.

4. <u>The Design of the Experiment.</u>

It was finally decided to conduct an investigation similar in nature to that of the First Experiment (Chapter 3). The masters concerned were to teach their forms by the method which appeared to them most suitable for the pupils.

"Master a" taught Form 2A with the accent on Individual Experimental work by the pupils. "Master bc" taught Form 2B with the accent on Individual Experimental work and Form 2C with the accent on Demonstration work by the teacher. Precautions were taken to see that all three forms were taught as far as possible the same amount of factual knowledge. They received the same homework. Teacher "bc" was convinced that with Form 2C he would never have been able to cover the same ground if he had allowed the pupils to do the experiments individually. His attitude was that with the weaker form demonstration of the experiments enabled him to accentuate the important features to better advantage. At the end of the thirteen weeks the three forms were all given a Theoretical and a Practical Test.

As a result of the experiment it was hoped that some information about the reliability and validity of the Practical Test, and its correlation with the Theoretical Test would be obtained. Since initial criterion scores were also available it was hoped that the correlations between the initial and final criterion tests would give some further evidence as to the validity and reliability of the Practical Tests used in this and the previous experiments.

5. <u>The Final Theoretical Test.</u>

A copy of the Theoretical Test applied to the pupils in December, 1947 is given below. It consisted of thirty four items and the majority were of the open or recall type. Question or item 32 was really a multiple choice question with six possible responses and item 33 is open to objection on the score of guessing but, since it was the only item of this type, it was decided to apply no corrections for guessing. The pupils were not given previous warning that they were to be given the test and it was designed after a careful study of their syllabus, classwork and homework. One mark was awarded for each correct response.

THE FINAL THEORETICAL TEST.

1. The Boiling point of Mercury is________ ^{o}C.
2. The Freezing point of Mercury is________ ^{o}C.
3. The Boiling point of Alcohol is_______ ^{o}C.
4. The Freezing point of Alcohol is_______ ^{o}C.
5. Convert 50^{o}C into ^{o}F.
6. Convert -10^{o}C into ^{o}F.
7. Convert 77^{o}F into ^{o}C.
8. Convert -13^{o}F into ^{o}C.

9. What is a Calorie?

10. What is a British Thermal Unit?

11. Heat which when supplied to a body produces a change in state but no change in temperature is called ______________________

12. What is the Specific Heat of a substance?

13. How much heat is required to raise the temperature of 60 grams of water from $10^{o}C$ to $50^{o}C$.?

14. How much heat is required to raise the temperature of 12 lbs of water from $40^{o}F$ to $50^{o}F$.?

15. A piece of metal weighs 80 grams and is at $30^{o}C$. When 160 calories are given to the metal its temperature rises to $38^{o}C$. What is the Specific Heat of the Metal?

16. A piece of metal of Specific Heat 0.1 is given 200 calories and its temperature rises from $20^{o}C$ to $70^{o}C$. Find the mass of the metal.

17. 360 grams of water at $100^{o}C$ are poured into a copper calorimeter of mass 600 grams and temperature $20^{o}C$. The temperature of the mixture is $80^{o}C$. How much heat is gained by the calorimeter?

18. What is the specific heat of the calorimeter?
The Latent Heat of Fusion of Ice is 80 calories per gram. The Latent Heat of Vaporization of water is 540 calories per gram.

19. How much heat is required to convert 8 grams of ice at $0^{o}C$ into water at $0^{o}C$.?

20. How much heat is given out when 12 grams of steam at $100^{o}C$ change to water at $100^{o}C$.

21. How much heat is required to convert 20 grams of Ice at $0^{o}C$ into water at $60^{o}C$?

22. How much heat is required to convert 10 grams of water at $40^{o}C$ into steam at $100^{o}C$?

23. A bunsen is placed under a beaker containing 100 grams of water at $20^{o}C$. Two minutes later the temperature of the water is $40^{o}C$. How much heat is supplied per <u>minute</u> to the water by the bunsen?

24. How long from the start will it be before the water boils?

25. If after boiling for 30 minutes, 60 grams of water have been converted into steam calculate a value for the Latent Heat of steam.

26. Why does a gas tap feel colder than the bench?

27. Why is it usually warmer upstairs than downstairs in a cinema?

28. What would you do if you wanted to keep a block of Ice in the house for a long time if you had no refrigerator?

29. What is proved by the experiment showing water boiling in the top of a test tube and ice at the bottom?

30. Who invented the Miner's Safety Lamp?

31. Why is a wooden wash-tub better than a metal one?

32. Which of the following is the best conductor of heat? Asbestos, Iron, Air, Wood, Copper, Water.

33. At night near the coast the wind usually blows from________to________

34. Explain briefly the connection between convection and density.

6. The Final Practical Test.

The design of a suitable Practical Test was more difficult than in the previous experiments. Suitable short experiments dealing with Specific Heat and Latent Heat were particularly difficult to design. In this connection it is interesting to note that many teachers consider quantitative work on these branches of Physics not suitable for second year pupils[1] The pupils involved here however had determined Specific Heats using thick calorimeters and had performed several experiments dealing with the rate at which a bunsen supplied heat to calorimeters containing water. A copy of the Practical Test applied to the pupils is given below. Experiments 1, 2, 4, and 5 were very closely related to experiments either performed by the pupils or demonstrated to them by the teacher. Experiment 3 was in the nature of a problem and was new to the pupils although it was fundamentally based on work done by the pupils during the previous year.

---

1. The Teaching of General Science pt. II (Section 12) S.M.A. John Murray 1938.

THE PRACTICAL TEST.

Name.................................

1. Measure the temperature of the water in the beaker with the Centigrade thermometer and then calculate its temperature in degrees Fahrenheit.

(a) Temperature of water = ^{o}C

(b) Temperature of water = ^{o}F

Name.................................

2. Measure the temperature of the boiling liquid with the Fahrenheit thermometer and then calculate its temperature in degrees Centigrade.

(a) Temperature of boiling liquid = ^{o}F

(b) Temperature of boiling liquid = ^{o}C

Name.................................

3. Weigh the copper cube and then the copper cylinder on the spring balance.
Find the volume of the copper cube in cubic centimetres.
Calculate the volume of the cylinder from these measurements.

(a) Weight of cube = grams.

(b) Weight of cylinder = grams.

(c) Volume of cube = ccs.

(d) Volume of cylinder. = ccs.

Name.................................

4. The thick calorimeter weighs 1,000 grams. Fill it with tap water, read its temperature and then empty the water into the sink.

Next pour in the boiling water and note the steady temperature of the "mixture" after it has been stirred.

Finally measure the volume of the water in the calorimeter.

(a) Temperature of cold water = ^{o}C.

(b) Temperature of mixture = ^{o}C.

(c) Volume of water in calorimeter = ccs.

(d) Rise in temperature of calorimeter = C^{o}.

Cont'd from No.4.

(e) Fall in temperature of boiling water = 0^o

(f) Heat lost by hot water = calories

(g) Specific heat of calorimeter =

Name..............................

5. <u>Don't touch the bunsen, tripod, or gauze.</u>

There are 60 ccs of tap water in the calorimeter. Note its temperature and when told to do so place it on the centre of the gauze. Take its temperature half a minute later and again after a further half minute.

(a) Temperature of water at start = ^oC.

(b) Temperature of water half a minute later = ^oC.

(c) Temperature of water after another half minute = ^oC.

(d) Heat supplied to water in first half minute = calories.

(e) How long from the start would it take for the water to boil? Answer = minutes.

---

7. <u>The Preparation and administration of the Practical Test.</u>

The test was administered with the assistance of VIth Formers and the technique was similar to that described in the previous experiments. There were five experiments but Experiments 1 and 2 were short so a period of fifteen minutes was allowed for Experiments 1 and 2 combined, and a further fifteen minutes was allowed for each of the Experiments 3, 4 and 5. No pupil failed to complete the purely experimental portions in the time allowed. Some details of the various experiments are given below.

<u>Experiment 1.</u> The water was contained in 1000 ccs beakers and taken direct from the tap. The VIth Former in charge noted the temperature of the water in each beaker when each pupil had finished and made a note of the result on the back of the pupil's answer paper. In marking an answer correct to $1C^o$ was marked as correct.

Experiment 2. The liquid was saturated salt solution contained in a flask and in marking an error of 1F° was accepted as correct.

Experiment 3. The copper cylinders and cubes were as near identical as possible for each pupil. Spring balances graduated in single grams and a half metre rule graduated in inches and centimetres was provided. For items 3a and 3b an error of one gram was accepted as correct but for 3c only 8ccs was accepted as correct since the cubes were of 2 cm edge.

Experiment 4. Large thick brass calorimeters filed to have a weight of 1,000 grams and covered in felt were used and each pupil was provided with a small beaker of boiling water, tripod, gauze, and bunsen. The measuring cylinders provided were graduated in intervals of 2ccs. The VIth Former in charge of the experiment made a note of the temperature of the tap water and checked the volume of water in the calorimeter for each pupil making a pencil note of the result on the back of the pupil's answer paper. In item 4a an error of 1C° was allowed and for item 4c an error of 2ccs was allowed. The temperature of the mixture was checked from a knowledge of the volume of boiling water, the initial temperature of the calorimeter and its Specific Heat. An error of 5 C° was allowed. This may appear rather generous but the writer gave the same experiment to ten VIth Formers and the mean value of the difference between the observed and calculated value for the temperature of the mixture was approximately 3C°.

Experiment 5. This was rather difficult from the point of view of ensuring objective marking. The bunsens were shielded to avoid draughts and during the whole test no other bunsens in the laboratory were turned off or on since such alterations might have affected the pressure of the gas supply. Three VIth Formers performed the complete experiment with each set of apparatus immediately before and after the whole test was completed. The average of the six values for items 5b and 5c were taken as the "correct" value for each set and it should be noted that none of the six values deviated by more than 2C° from the mean. In marking an "error" of 1C° was allowed for Item 5a and "errors" of 3 C° were accepted for items 5b and 5c. For 5e an answer correct to the nearest minute was accepted.

As in the previous Practical Tests only one mark was awarded to each item and it will be noted that a very generous permissible error was allowed for some of the items in Experiment 4 and 5.

8. The Raw Scores on the Criterion Tests.

The raw scores obtained by the pupils on the various tests together with a brief summary of the major statistics for each test are given below.

As in the previous experiments the Practical Test was subdivided to produce a new score under the heading of Experimental Test. This was the total score obtained by each pupil on the items of the Practical Test involving pure measurement and manipulation and included items 1a, 2a, 3a, 3b, 3c, 4a, 4b, 4c, 5a, 5b and 5c. The item 3c is to some extent of doubtful right to be included in this list. For convenience in comparison the scores of the pupils on the Initial Tests are also tabulated below.

TEST SCORES FORM 2A.

Pupil	Theoretical Test		Practical Test		Experimental Test	
	Final	Initial	Final	Initial	Final	Initial
a	22	19	11	10	7	7
b	18	18	13	15	10	9
c	14	18	14	14	11	11
d	16	17	15	11	9	7
e	13	13	11	5	9	4
f	8	8	8	10	8	7
g	21	21	11	11	9	7
h	14	16	11	10	9	6
i	21	24	14	13	10	8
j	20	24	8	13	8	8
k	22	25	10	16	8	10
l	16	21	10	13	8	10
m	13	15	9	13	7	9
n	11	17	12	10	11	8
o	11	20	10	8	9	8
p	8	17	9	7	9	6
q	16	20	11	20	8	12
r	23	23	12	16	9	10
s	8	6	11	17	10	11
t	28	21	16	16	11	11

Cont'd........

Cont'd........

Pupil	Theoretical Test		Practical Test		Experimental Test	
	Final	Initial	Final	Initial	Final	Initial
u	13	17	10	14	9	9
v	8	11	12	11	9	8

TEST SCORES FORM 2B.

Pupil	Theoretical Test		Practical Test		Experimental Test	
	Final	Initial	Final	Initial	Final	Initial
a	13	16	8	10	7	7
b	16	20	10	13	8	10
c	15	13	11	16	9	11
d	18	11	9	10	8	7
e	13	14	11	14	9	10
f	15	14	9	13	7	8
g	9	15	11	15	9	11
h	13	16	9	6	9	4
i	18	20	13	16	9	11
j	12	16	6	7	6	3
k	22	24	7	15	7	10
l	26	19	15	15	9	12
m	15	16	10	11	8	8
n	12	7	10	15	9	10
o	11	15	12	10	9	7
p	21	12	14	3	9	3
q	12	9	10	10	8	7
r	19	14	11	15	9	9
s	16	18	11	12	9	9
t	9	10	8	11	8	9

TEST SCORES FORM 2C.

Pupil	Theoretical Test		Practical Test		Experimental Test	
	Final	Initial	Final	Initial	Final	Initial
a	6	11	10	15	10	11
b	8	14	9	6	8	4
c	9	18	11	11	9	8
d	4	2	9	10	8	7
e	10	7	8	12	8	8
f	7	9	9	8	8	6
g	13	15	9	15	9	11
h	15	19	12	13	9	9
i	14	10	8	6	6	6
j	9	10	11	16	9	12
k	11	9	8	7	7	6
l	10	18	9	16	8	10
m	9	15	11	13	10	7
n	19	18	11	15	10	10
o	8	11	8	12	8	8
p	5	6	10	8	8	7
q	7	10	8	11	7	9
r	6	11	10	12	8	9
s	10	20	10	14	9	11
t	13	10	10	8	9	6
u	10	.7	10	10	9	8
v	10	16	10	8	8	5
w	8	16	10	12	10	9

SUMMARY OF TEST RESULTS

(a) Final Theoretical Test.

/Table over.

Statistic	Form 2A	Form 2B	Form 2C	All Forms.
ΣX	344	305	221	870
N	22	20	23	65
$M = \frac{\Sigma X}{N}$	15.636	15.250	9.609	13.385
ΣX^2	6052	5019	2387	13458
Σx^2	673.066	367.740	263.477	1813.380
S.D.	5.661	4.400	3.461	5.323
Range.	8 - 28	9 - 26	4 - 19	4.-28

(b) Final Practical Test.

Statistic	Form 2A	Form 2B	Form 2C	All Forms.
ΣX	248	205	221	674
N	22	20	23	65
$M = \frac{\Sigma X}{N}$	11.273	10.250	9.609	10.369
ΣX^2	2890	2195	2153	7238
Σx^2	94.372	93.740	29.477	249.200
S.D.	2.120	2.221	1.158	1.973
Range.	8 - 16	6 - 15	8 - 12	6 - 16

(c) Final Experimental Test.

Statistic	Form 2A	Form 2B	Form 2C	All Forms.
ΣX	198	166	195	559
N	22	20	23	65
$M = \frac{\Sigma X}{N}$	9.000	8.300	8.478	8.600
ΣX^2	1810	1394	1677	4881
Σx^2	28.000	16.200	23.739	73.600
S.D.	1.154	0.923	1.038	1.064
Range.	7 - 11	6 - 9	6 - 10	6 - 11

CHAPTER 8.

THE THIRD EXPERIMENT : ANALYSIS OF RESULTS.

1. The Reliability of the Final Theoretical Test.

Using the scores for the total group of 65 pupils the correlation between the total scores on the odd and even items was determined giving a value of,

$r = 0.703 \pm 0.042$ where 0.042 is the probable error.

Corrected by the "Spearman Brown Formula" this gave a Reliability or Consistency Coefficient,

$R = 0.826.$

This is a reasonably high value for such a test.

Now the Initial Theoretical Test which the pupils had received when in the First Year Forms was rather similar in type to this test. In fact both tests were designed to measure the same abilities. The treatments of the three forms in the period between taking the Initial and Final Theoretical Tests were not identical but one would still naturally expect a reasonably high correlation between the scores on the two tests if they were measuring the same abilities. The correlation between the scores on the two tests was accordingly worked out in order to obtain some estimate of the degree to which the two Theoretical Tests were measuring the same abilities.

Correlation of Initial and Final Theoretical Test Scores.

Let $\left\{\begin{array}{l}x \text{ refer to Initial Theoretical Test Scores} \\ y \text{ refer to Final Theoretical Test Scores.}\end{array}\right\}$

ANALYSIS OF COVARIANCE.

Group	Σx^2	Σy^2	Σxy	$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$
2A	515.904	673.066	+452.182	+0.767
2B	316.940	367.740	+163.250	+0.478
2C	496.410	263.477	+179.348	+0.496
Within Forms	1329.254	1304.283	+794.780	+0.604
Total	1670.960	1813.380	+1168.154	+0.671

For samples of 65 pupils the values of r required for significance at various levels are as follows:-

Level of Significance	10%	5%	1%	0.1%
Correlation coefficient	0.206	0.245	0.318	0.400

Both r total and r within forms are comparatively large and highly significant so there seems to be little reasonfor doubting that to a large extent the Initial and Final Theoretical Tests were measures of the ~~same~~ same ability or abilities. Application of Fisher's "z" technique shows that at the 5% level we can be confident that the true r total lies within the limits 0.510 to 0.786 and r within forms lies within the limits 0.422 to 0.735. All of the values of r for the individual forms are highly significant at the 5% level but the differences between them are not significant at the 5% level and thus we have no reason for rejecting the hypothesis of homogeneity of correlation. Taking the r within forms value as being the more stable and reliable value, the forecasting efficiency is 20%.

The reasonably high value of the correlation between the two tests does give some added confidence in the reliability and perhaps even the validity of the two tests. It may of course be that the high correlation is due to factors that are common to the two tests and were not included in the list of objectives which the tests were designed to measure. For example the high correlation might be due to the fact that both tests are to a large degree measuring the general intelligence "g" and verbal "v" factor. This is however unlikely in view of the extremely low correlation between a test which was almost the same as the Initial Theoretical Test and the Northumberland Intelligence Test (Chapter 6 - 4 page 108). The responses to each item were examined and no items of very doubtful discriminative value were detected. Both teachers responsible for the instruction of the forms agreed that the test was a fair sampling of the work done by the forms. Since a good achievement test tends to give a normal distribution it was decided to apply the χ^2 test for normality of distribution of the scores on the Theoretical Test using the whole sample of 65 pupils and the final analysis is given below.

FINAL THEORETICAL TEST.

χ^2 TEST FOR NORMALITY OF DISTRIBUTION.

Scores	f_o	f_e	$\frac{(f_o - f_e)^2}{f_e}$
28 and over	1	0.19	
25 - 27	1	0.75	0.6759
22 - 24	4	2.47	
19 - 21	6	6.06	
16 - 18	8	10.81	0.7304
13 - 15	15	14.08	0.0682
10 - 12	12	13.57	0.1816
7 - 9	14	9.59	2.0280
4 - 6	4	4.95	1.6190
3 and under.	0	2.53	
			$\chi^2 =$ 5.3031

$$\left[\begin{array}{l}\text{Mean} = 13.385\\ N = 65\\ \text{S.D.} = 5.323\end{array}\right] \qquad \left[\begin{array}{l} f_o = \text{frequency observed}\\ f_e = \text{frequency expected.}\\ \text{Degrees of Freedom} = 6 - 1 - 2 = 3.\end{array}\right]$$

For three degrees of freedom χ^2 exceeds 5.30 in almost 15% of random samples of this size and our confidence in the hypothesis of normality of distribution is as a consequence rather low but we have no justification for rejecting the hypothesis.

It should be noted that items involving some facility with numbers constituted 50% of the Initial Test and 70% of the Final Test. There is however some evidence that the correlation between scores on the numerical and non numerical items is by no means low. As a general result of the analysis described in the present section we can have a reasonable degree of confidence in the Reliability of the Final Theoretical Test.

2. The Reliability of the Final Practical Test.

An attempt was made to obtain some estimate of the reliability of the Practical Test by a modification of the split-half method. The items were grouped into pairs of items of estimated equivalent difficulty. With this particular test the pairing was rather difficult since there were eleven items involving what we might term pure measurement and observation. The pairing finally decided upon was as follows the experimental items being marked in red.

$$\begin{Bmatrix}1a\\2a\end{Bmatrix}\begin{Bmatrix}1b\\2b\end{Bmatrix}\begin{Bmatrix}3a\\3b\end{Bmatrix}\begin{Bmatrix}4a\\4b\end{Bmatrix}\begin{Bmatrix}5a\\5b\end{Bmatrix}\begin{Bmatrix}3c\\5e\end{Bmatrix}\begin{Bmatrix}4d\\4e\end{Bmatrix}\begin{Bmatrix}4f\\5d\end{Bmatrix}\begin{Bmatrix}5c\\4c\end{Bmatrix}\begin{Bmatrix}3d\\4g\end{Bmatrix}$$

The total score for each pupil on the items in the first row was correlated with the total score for each pupil on the items in the second row. Using the results for all 65 scripts this gave a correlation coefficient of

$r = 0.357 \pm 0.073$ where 0.073 is the Probable Error.

Corrected by the "Spearman Brown" formula this gave a Reliability coefficient of,

$$R = 0.526$$

This "Consistency" or Reliability Coefficient is very low but a brief examination of the pairs and reference to the actual test paper (p/20.) will show that in many cases the pairing is of necessity far from satisfactory. Correlating the scores on alternate items as set out in the original test gave a value of R = 0.555.

Both of these Reliability or Consistency Coefficients are very low but it must be noted that the test only contained twenty items and moreover there is considerable doubt as to whether the test has been split into two equivalent halves.

Since good achievement tests tend to give a normal distribution the χ^2 test for normality of distribution of the scores on both the Practical Test and the Experimental Test was applied to the whole sample of 65 pupils and the final analysis is given below.

FINAL PRACTICAL TEST.

χ^2 TEST FOR NORMALITY OF DISTRIBUTION.

Scores	f_o	f_e	$\frac{(f_o - f_e)^2}{f_e}$
16 and over	1	0.14	
15	2	0.47	0.7308
14	3	1.52	
13	2	3.79	
12	5	7.36	0.7567
11	15	11.05	1.4120
10	16	12.99	0.6975
9	10	11.76	0.2634
8	9	8.44	0.0345
7	1	4.62	4.0148
5 and under	1	2.86	
			χ^2 = 7.9097

$$\left[\begin{array}{lll} \text{Mean} & = & 10.369 \\ \text{N} & = & 65 \\ \text{S.D} & = & 1.973 \end{array}\right] \qquad \left[\begin{array}{l} fo = \text{frequency observed} \\ fe = \text{frequency expected} \\ \text{Degrees of Freedom} = 7 - 1 - 2 = 4. \end{array}\right]$$

For four degrees of freedom χ^2 exceeds 7.91 in almost 9% of random samples of this size, and as a consequence our confidence in the hypothesis of normality of distribution is rather low and moreover there is some evidence of a negatively skewed distribution.

FINAL EXPERIMENTAL TEST.

χ^2 TEST FOR NORMALITY OF DISTRIBUTION.

Scores	fo	fe	$\frac{(f_o - f_e)^2}{f_e}$
11 and over.	3 }	0.78 }	2.4599
10	7 }	5.34 }	
9	27	16.89	6.0516
8	19	23.39	0.8239
7	7	14.26	3.6962
6	2 }	3.87 }	1.2616
5 and under.	0 }	0.47 }	
			χ^2 = 14.2932

$$\left[\begin{array}{lll} \text{Mean} & = & 8.600 \\ \text{N} & = & 65 \\ \text{S.D.} & = & 1.064 \end{array}\right] \qquad \text{Degrees of Freedom} = 5 - 1 - 2 = 2.$$

For 2 degrees of Freedom χ^2 exceeds 13.80 in 0.1% of random samples and as a consequence the hypothesis of normality of distribution must be rejected. The distribution shows a marked negative skew.

An examination of the frequency distribution shows that both the Practical Test and Experimental Test tended to give a negatively skewed distribution. It is evident that the Experimental Test in particular is very diagnostic in character and although this characteristic was also observed in the Experimental Test applied to the First Year Pupils, it is much more pronounced in this case. One remedy as already pointed out would be to award a total of two marks to each item of the Experimental Test and have two limits of permissible error, two marks being awarded for the more correct and one mark being awarded for the less correct response. This solution is very attractive and can be applied with some accuracy when older pupils such as VIth Formers are being examined. With the type of experiments employed here, the apparatus used, and the pupils concerned such a technique would be very difficult to apply. The mere allocation of two marks to some items and one mark to other items, i.e. weighting the marks might be justified if there were certain proof that some of the items were more difficult that the others, since the whole test involves only a small number of items. A certain amount of weighting is already present in sofar as certain fundamental measurements such as, for example, the measurement of temperature are present in several of the experiments.

A certain amount of evidence as to the reliability of the individual items of the Experimental Test is provided by the fact that when groups of ten or twelve VIth Formers were given identical experiments and apparatus their readings for each item showed very little dispersion.

3. The Practical Test: Discriminative Value of Individual Items and Validity.

Since there was some indication that the Reliability of the Practical Test might be low it was considered advisable to examine the Discriminative Value of the individual items. The whole group of 65 pupils was divided into four sections the dividing points of the sections being approximately the upper quartile, the median, and the lower quartile. The group or section with the highest total scores contained 17 pupils and is referred to as the "4th Quarter" whilst the other three sections each contained

16 pupils. The total number of correct responses made by the members of each section, were calculated for each item and are tabulated below. For convenience the items of the Experimental Test are underlined in red.

PRACTICAL TEST: DISCRIMINATIVE VALUE OF INDIVIDUAL ITEMS.

Question or Item No.	Number of Correct Responses by				
	4th Quarter	3rd Quarter	2nd Quarter	1st Quarter	All Pupil[illegible]
1a	16	15	15	16	62
1b	8	2	2	2	14
2a	15	13	11	7	46
2b	7	1	0	1	9
3a	17	15	16	15	63
3b	16	13	13	12	5[illegible]
3c	9	6	6	4	25
3d	0	0	0	0	0
4a	17	16	16	15	64
4b	17	16	16	12	61
4c	15	16	15	11	58
4d	15	13	10	6	44
4e	14	11	6	1	32
4f	4	0	0	0	4
4g	1	0	0	0	1
5a	17	16	16	16	6[illegible]
5b	13	13	10	9	45
5c	8	5	3	1	17
5d	7	0	0	1	7
5e	2	0	0	1	3
No of Pupils in each group	17	16	16	16	65

In general the number of correct responses to each item does decrease as we pass from the 4th Quarter to the 1st Quarter. The lack of discriminative value of the Experimental items is very marked and it is evident that most of them are mainly diagnostic in character. Items 1a, 4a and 5a all involved the same process of measuring the temperature of some cold water with a Centigrade Thermometer and it is interesting to note that the totals of correct responses to these items were 62, 64 and 65 respectively. These items were obviously very easy so far as the pupils were concerned but it should be noted that they were consistently easy. Items 3a and 3b both involved the measurement of weight with a Spring Balance and the totals of correct responses to these items were 63 and 54 respectively. The items involving deductions from the results of the observations and measurements naturally show greater discriminative value than the items of the Experimental Test.

The validity of the Practical and Experimental Tests must finally depend to a large extent upon subjective opinions. However the Initial Practical Test administered to the pupils when in their first year at the school and this Final Practical Test were designed to measure similar objectives or abilities and as a consequence a reasonably high correlation between the scores on these two tests might be expected. The correlations for both the Practical and Experimental Tests were calculated and the results are given below.

Correlation of Initial and Final Practical Test Scores.

Let $\left\{\begin{array}{l} x \text{ refer to Initial Practical Test Scores} \\ y \text{ refer to Final Practical Test Scores.} \end{array}\right\}$

ANALYSIS OF COVARIANCE.

Group	Σx^2	Σy^2	Σxy	$r = \frac{\Sigma xy}{\sqrt{(\Sigma x^2 \cdot \Sigma y^2)}}$
2A	263.374	94.372	+37.545	0.238
2B	242.540	93.740	+23.750	0.158
2C	221.910	29.477	+30.957	0.383
Within Forms	707.824	217.589	+92.252	0.235
Total.	743.740	249.200	+114.431	0.266

Correlation of Initial and Final Experimental Test Scores.

Let $\begin{Bmatrix} x \text{ refer to Initial Experimental Test Scores} \\ y \text{ refer to Final Experimental Test Scores.} \end{Bmatrix}$

ANALYSIS OF COVARIANCE.

Group	Σx^2	Σy^2	Σxy	$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 \cdot \Sigma y^2}}$
2A	81.455	28.000	+9.000	0.188
2B	130.200	16.200	+16.200	0.353
2C	98.609	23.739	+22.565	0.466
Within Forms	310.264	67.939	+47.765	0.329
Total.	311.446	73.600	+49.600	0.328

The correlations obtained are very low. So far as the individual forms are concerned the only coefficient which is significant at the 5% level is that for Form 2C with the Experimental Tests. The differences between the values of r for the separate forms are not significant at the 5% level. For a total sample of 65 pupils r must exceed 0.245 to be significant at the 5% level and 0.318 to be significant at the 1% level.

Using Fisher's "z" technique we can feel confident at the 5 per cent level that for the Practical Tests the true value of r total lies within the limits +0.023 and +0.480. For the Experimental Tests, at the 5% level, r total lies within the limits + 0.091 and +0.530

The low correlations tend to cast serious doubt on the reliability and validity of at least one of the tests. It may be for example, that the Initial Practical Test had a high validity while that of the Final Practical Test was low, or vice versa. It should be noted that the Consistency or Reliability Coefficient of the Initial Practical Test was estimated at 0.790 (Chapter 3 - 12 page 38) while that for the Final Practical Test was estimated at 0.526. Using these figures, which are of course only very approximate estimates then r total for the two Practical Tests when corrected for attenuation is still only 0.412 instead of 0.266. A further point of note is that we can have little confidence in the hypothesis that the Final Practical Test gives a normal distribution and no confidence at all in such a hypothesis for the Final Experimental Test. In conjunction with the fact that the actual design of the Final Practical Test was very difficult, the above factors seem to indicate that the Initial Tests were probably of higher validity and reliability than the Final Tests.

4. Correlation of Final Theoretical and Practical Test Scores.

Even though the reliability and validity of the Final Practical Test were doubtful the correlation between the Final Theoretical and Practical Test Scores were calculated for the whole sample of 65 pupils. The three forms had been given different treatments but is is probably correct to say that there is a widespread tendency in the teaching of Physics to adopt demonstration methods with weaker classes. The reasons for this tendency are generally the subjective opinions that, discipline needs to be stricter with weaker pupils and that by the adoption of such a method a weaker class can cover the same amount of ground, in the same time as a better class where the pupils are allowed to do some individual experimental work. The advisability of such a technique is doubtful but it is probably true that the methods of teaching the three classes involved in this experiment were to a large extent in conformity with popular practice, and as a consequence we have no justification for considering the whole group as either unduly homogeneous or heterogeneous.

Before actually calculating the correlation coefficient a scattergram was constructed.

<u>Scattergram: Final Theoretical and Practical Test Scores.</u>

<u>Practical Test Scores.</u>

Theoretical Test Scores	6	7	8	9	10	11	12	13	14	15	16	n	Mean
28+											1	1	16.00
25+										1		1	15.00
22+		1			1	1	1					4	10.00
19+			1			3			2			6	11.50
16+				1	2	2		2		1		8	11.50
13+			2	4	3	4	1		1			15	10.13
10+	1		2	1	6		2					12	9.58
7+			4	3	1	5	1					14	9.71
4+				1	3							4	9.75
n	1	1	9	10	16	15	5	2	3	2	1	65	
Mean	12.00	22.00	11.11	10.90	11.56	14.13	13.60	18.00	18.67	21.00	28.00		

The scattergram gives no pronounced indication of curvilinear regression and in view of the smallness of the total sample there is little justification for rejecting a hypothesis of linear regression or applying more exact tests of linearity of regression.

<u>Correlation of Final Theoretical and Practical Test Scores.</u>

Let $\left\{\begin{array}{l}x \text{ refer to Theoretical Test Scores} \\ y \text{ refer to Practical Test Scores.}\end{array}\right\}$

ANALYSIS OF COVARIANCE.

Group	$\sum x^2$	$\sum y^2$	$\sum xy$	$r = \frac{\sum xy}{\sqrt{(\sum x^2 \cdot \sum y^2)}}$
2A	673.066	94.372	+105.182	0.417
2B	367.740	93.740	+81.750	0.440
2C	263.477	29.477	+20.478	0.232
Within Forms	1304.283	217.589	+207.410	0.389
Total	1813.380	249.200	+313.769	0.467

The values of r for Forms 2B and 2C are almost significant at the 5% level and there is no reason for rejecting the hypothesis of homogeneity of correlation for the three forms. Both r total and r within forms are highly significant at the 1% level. The correlation is higher than that obtained with the first year forms, but is still quite low. A subjective opinion is that both the Final Tests gave more weight to numerical calculations than was the case with the tests applied to the first year pupils and hence larger correlation coefficients might reasonably be expected. At the 5% level of confidence r total lies within the limits 0.250 to 0.639 and r within forms lies within the limits 0.150 and 0.573.

5. Correlation of Final Theoretical and Experimental Test Scores.

The Final Experimental Test was very diagnostic in character, and did not give a normal distribution. Moreover the range of the scores was very small indeed. Despite this it was decided to construct a scattergram and calculate the correlation coefficient for the whole sample of 65 pupils.

Scattergram.

Final Theoretical Test Scores.

Experimental Test Scores

	4+	7+	10+	13+	16+	19+	22+	25+	28+	n	Mean
11			1	1					1	3	17.67
10	1	3			1	2				7	12.71
9		5	5	9	3	3	1	1		27	14.04
8	3	5	4	1	4	1	1			19	11.58
7		1	1	3			2			7	14.71
6			1	1						2	13.00
n	4	14	12	15	8	6	4	1	1		
Mean	8.50	8.71	8.42	9.13	8.63	9.17	7.75	9.00	11.00		

A scattergram such as the above is not very satisfactory since there are only six rows. Our confidence in the hypothesis of linearity of regression is not high and yet there is no very pronounced indication of curvilinear regression. The correlation between the two test scores is obviously low. If the regression were actually curvilinear then the product moment correlation coefficient would of course underestimate the degree of relationship between the two variables.

Correlation of Final Theoretical and Experimental Test Scores.

Let $\left\{\begin{array}{l}x \text{ refer to Theoretical Test Scores.} \\ y \text{ refer to Experimental Test Scores.}\end{array}\right\}$

ANALYSIS OF COVARIANCE.

Group	Σx^2	Σy^2	Σxy	$r = \frac{\Sigma xy}{\sqrt{\Sigma x^2 . \Sigma y^2}}$
2A	673.066	28.000	+1.000	0.007
2B	367.740	16.200	+7.500	0.097
2C	263.477	23.739	+11.304	0.143
Within Forms	1304.283	67.939	+19.804	0.067
Total.	1813.380	73.600	+39.000	0.107

The correlation coefficients are all extremely small. For samples of 65 pupils r should exceed 0.206 in order to be significant at at the 10% level. As pointed out in the experiments with the first year pupils we would naturally expect the correlation between the Experimental and Theoretical Tests to be low since the tests were intended to be measures of different abilities. However it is important to realise that correlations between the scores on two tests are systematically lowered or attenuated as the result of errors of measurement. The reliability of the Theoretical Test is probably quite high since its Consistency Coefficient by the "split half" method was 0.826 and it correlated quite highly with the Initial Theoretical Test. The reliability of the Experimental Test is however probably low and as a consequence considerable attenuation is probably present, so far as the correlations of the two tests are concerned.

6. Discussion on Results of Correlation Analysis.

The relatively high consistency coefficient (0.826) for the Final

Theoretical Test and its relatively high correlation of 0.671 with the Initial Theoretical Test (Consistency Coefficient 0.846) does tend to give added confidence in the reliability and even validity of the Theoretical Tests used in all the experiments. It must be admitted of course that the high correlation between the tests might have been due to the two tests measuring some common factors not included in the original list of objectives which the tests were designed to measure. This possibility has of course been considerably reduced by virtue of the fact that a test almost identical with the Initial Theoretical Test had a very low correlation with a reliable verbal group Intelligence Test. (Chapter 6 - 4 page 108)

The correlations between the Initial and Final, Practical and Experimental Tests were very low and as already pointed out appear to cast doubt on the validity and reliability of the Final rather than the Initial Tests. The low correlations are however not entirely unexpected since it is well known that many performance tests, and tests of occupational abilities are often of poor reliability. W.P. Alexander[1] working with a group of 100 elementary school children with an age range of 124 to 166 months obtained a correlation of 0.335 between the scores on the Pass along Test and KOHS Block Design Test. For the same group the correlation between two of COX'S tests of Mechanical Aptitude, viz Test E_3 and Test D, was 0.283. The majority of these tests were of course designed for application to older students. The fact that the correlation between the Initial and Final Practical and Experimental Tests were in this experiment only 0.266 and 0.328 is certainly not in itself sufficient evidence that practical tests of

1. W.P. Alexander. Intelligence Concrete and Abstract - C.U.P. 1935.

the type used are in general unsatisfactory. All the available evidence however does appear to indicate that the Final Practical Test was of less validity than the Initial Practical Test.

With regard to the correlations between the scores on the Final Theoretical Test and the Final Practical and Experimental Tests the fact that r total for the Theoretical and Practical Tests is 0.467 while r total for the Theoretical and Experimental Tests is only 0.107 is interesting. The general trend of these coefficients is similar to the results for similar correlations calculated in the first and second experiments. Low correlations are to be expected if the two tests being correlated are valid measures of different abilities. The extremely low value of 0.107 may be considerably attenuated by a low reliability for the Experimental Test. In October 1947 an initial exploratory experiment similar to the present one was applied to two third year forms. After one month's revision work they were given a Theoretical and Practical Test based on their first and second year work. These tests contained many items that were later used in the tests which have already been quoted. The Theoretical test of 40 items had a self Consistency Coefficient of 0.823 based on 66 scripts. The Practical Test gave a distribution which was reasonably normal and the correlation between the Practical and Theoretical Test Scores was 0.101.

Even allowing for the possibility of low reliability for the Practical Tests used in all the three experiments which have been described in full it is almost certain that they were to a large extent measuring abilities not measured by the Theoretical Tests. The difficulties encounter in designing the Practical Tests for application to young pupils, and assessing their validity have been stressed because it may be that tests of this nature are more suitable and useful when applied to more advanced pupils.

CHAPTER 9.

Conclusions: Suggestions for Further Research.

In the research discussed in the previous chapters two major but complementary problems were considered. The primary problem, of course, was to determine whether the two different methods of instruction had produced any measurable differences in the progress of the two forms. The other problem was the design of reliable and valid objective tests of ability in Practical Physics. The two problems are by no means distinct and it is almost certain that the first problem can never be completely solved until a satisfactory solution of the second problem has been attained and accepted as valid by a representative body of teachers and physicists.

1. The Methods Experiment. (vide Chapters 5 and 6.)

The two methods were applied to two random groups of first year pupils and the effect of the two methods on the mean abilities of the pupils in each group as measured by the Theoretical and Practical Tests was determined. The Theoretical Test employed was of reasonably high reliability and validity and conventional in type. The results showed that no significant differences in those abilities which were measured by this test were produced by the two methods. The Practical and Experimental Tests were designed to measure different abilities or outcomes to those measured by the Theoretical Test and the reliability and validity of these tests were rather uncertain, depending to a large extent on subjective opinions. Again the two methods produced no significant difference so far as those abilities which were actually measured by these two tests were concerned. It is realised that the above results are restricted to the particular forms, school and teacher involved in this experiment. As a result of the experiment it might be argued that the two methods had produced no significant differences for the two forms. This may not be true since other factors are involved.

A method of instruction will have intellectual, physical and emotional reactions on the pupil and it may be that the emotional effects on the child are the ones of greatest importance. The young pupil entering a physics laboratory for the first time is usually intrigued by the sight of the apparatus; his curiosity is obvious and his interest in the subject is aroused and stimulated. If in the course of the year he is not allowed to use at least some of the apparatus his interest may wane and be replaced by a sense of frustration. The objective measurement of the interest of a pupil in a given subject is of course extremely difficult. Recently some research into the interests of pupils in a single grammar school has been undertaken and the results though restricted in their nature scope and validity are interesting.[1] It is claimed for example that pupils of 11+ were very interested indeed in learning how to weigh and not so interested in finding the density of a solid, even though the latter involved the use of a balance. In general with pupils of 11+ and 12+ the interest in those branches of science involving weighing and measurement was high but tended to decrease with age. As already emphasized the assessment of the interest of a pupil in a given subject is extremely difficult. A new type achievement test in Physics will to some extent be a measure of the pupil's interest in Physics if it is not overloaded with items involving difficult numerical calculations. The assumption is of course that a person is well informed in those subjects in which he is most interested. If the Theoretical Test employed in the

1. N.L. Houslop and E.J. Weeks. The Interests of Schoolchildren. 1947 The School Science Review Nos. 109; 110 Vol XXX.

methods experiment is taken as a measure - and probably a very poor measure - of the pupil's interest in Physics then once again we could claim that the two methods produced no statistically significant difference for the two forms. This was still true when the total scores on the non-mathematical items of the Theoretical Test were examined by means of the analysis of variance.

Since the two methods produced no measureable differences in those abilities or outcomes that were measured by the criterion tests it might appear that both methods are equally effective. There are however strong subjective opinions by experienced teachers that the interest of the pupil in Science is stimulated by the performance of experimental work, and since there is at present no really valid objective measure of the pupil's interest there is some justification for teaching Physics to young pupils by a method which does include a reasonable amount of individual experimental work.

This simple solution to the problem is however complicated by the fact that the personality and ability of the actual teacher is an important factor. Some teachers have a genius for stimulating interest by means of demonstration experiments, some find disciplinary difficulties in controlling classes performing individual experiments etc. Since it is possible that the influence of the two methods is not very pronounced in any direction it then appears justifiable at present that each teacher should use the method which his experience has shown is best suited to his own personal interests and abilities. The prevalent position in grammar schools is summarized by the following quotation from a recent work, which was compiled by a panel of about thirty experienced Science Teachers, with the assistance of about 200 corresponding members of the Science Masters Association and the Incorporated Association of Assistant

Masters.

> "In most schools Science is taught by a combination of classroom and laboratory methods. The more theoretical parts of the subject are dealt with by lecturing, discussion and oral and written questioning familiar in other subjects. Where possible, the teacher's descriptions are amplified and enlivened by demonstration experiments, and where suitable the problems involved are investigated practically in the laboratory by the pupils themselves".[1]

The "methods" experiment described in Chapters 5 and 6 has produced no evidence that might cast doubt on the wisdom of the above procedure so far as first year students of Physics are concerned.

2. Objective Practical Physics Tests: Young Pupils.

Considerable research is needed to produce really satisfactory objective tests of ability in Practical Physics. The two tests used in the present research were based solely on one method of approach to the problem and were fundamentally efforts to measure outcomes which are not generally assessed when dealing with young pupils.

The fact that the test administered to the second year pupils was more difficult to design than that given to the first year pupils is not necessarily discouraging. It must be noted that the syllabus followed by the second year pupils was probably not particularly appropriate for pupils of their age and even with pen and paper objective Physics Tests difficulty is often experienced in designing good items to test certain branches or aspects of the subject. The reliability of the tests may not be extremely high but this is also true of many performance tests and tests of occupational abilities. Unfortunately the writer had no opportunity to re-administer the Practical Tests to the first and second year pupils after a reasonable interval of time, such as three months and so obtain more valid estimates of their reliability.

---

1. The Teaching of Science in Secondary Schools. Joint Committee of I.A.A.M. and S.M.A. John Murray, 1947.

The validity of practical physics tests is particularly difficult to assess. The normal procedure of estimating a test's validity by determining how it correlates with known valid objective measures of those abilities which the test is designed to measure is almost impossible. No valid and generally accepted criterion measure of a young student's ability at Practical Physics exists at present. The validity must depend to a large extent on subjective opinions. Even here difficulty is encountered and few teachers of Physics are prepared to classify first year pupils with regard to their practical ability even by a coarse method involving the use of a five point scale.

If however the reliability of tests of this type can be established then a low correlation with a reliable new-type Theoretical Test in Physics will at least indicate that the Practical Test measures outcomes not measured by the pen and paper test. It may be that the ability of a pupil in Practical Physics is closely related to the Practical Ability or "F" factor of W.P. Alexander and the Mechanical Aptitude or "m" factor of J.W. Cox. If this is so then scores on the Practical tests should correlate well with measures of these factors. Tests of these factors are however well known to be more reliable when applied to older pupils. It is true that W.P. Alexander has recently published norms for his Performance Scale - used as a measure of F - for a range of 7 to 16 years but he still points out that the scale is more effective between the ages of eleven and sixteen.[1]

The general tendency of the Experimental Tests to be diagnostic in character and give a positively skewed distribution has already been discussed. In the writer's opinion it is probable that more progress will be made, in

---

1. A Performance Scale for the Measurement of Practical Ability. W.P. Alexander. Nelson and Sons. 1946.

the initial stages of future research, by concentrating on the measurement of ability in Practical Physics with older pupils such as School Certificate and Higher School Certificate candidates. With such pupils tests that are less diagnostic in character can be designed and longer tests involving many items can be used.

3. Practical Physics and Older Pupils.

So far as older pupils are concerned the general tendency is to place more emphasis on the importance of individual experimental work by the students. The importance attached to such work and the time allotted to it, is of course governed to a large extent by the careers for which the students are preparing. By the age of 17 or 18 however physics teaching has become more specialised and less general in so far as its objectives are concerned. Candidates who take the Higher School Certificate or Intermediate B.Sc. examination in Physics are usually intending to follow some branch of Pure or Applied Science as a career. For such pupils ability at Practical Physics, "per se", and not as an aid to a fuller appreciation of the principles of the subject becomes important.

Even here caution is necessary. The need for accurate measurement, manipulation of apparatus, and the design and construction of new techniques is essential in all the experimental sciences. Lord Kelvin's dictum "We never know much about anything until we have contrived to measure it" is very pertinent. Despite this the actual accurate measurement of quantities, observation of experimental phenomenon and manipulation of apparatus may, to a large extent, be divorced from ability in Physics. Examples of eminent physicists with great practical ability are numerous. Examples at the other end of the scale are also well know. The following quotation is from a letter to "The Times" on September 4th, 1940 by Dr. F.W. Aston, F.R.S., concerning Sir J.J. Thomson.

> "Among great experimental physicists his lack of manipulative skill must have been well nigh unique, yet the simplicity and beauty of the methods of analysis and measurement which he originated - make them ideal for the actual operator". .....
>
> "This intuitive ability to comprehend the inner working of intricate apparatus without the trouble of handling it appeared to me then, and still appears to me now, as something verging on the miraculous, the hall mark of the great genius."

In the above quotations it is important to stress the words "unique" and "miraculous" but it must also be remembered that today we have two almost distinct types of physicist, the mathematical physicist and the experimental physicist. In rare cases of course we may have a first class combination of both in one individual. It is by no means a "sine qua non" that a good physicist today must be capable of making accurate measurements, and manipulating apparatus himself, but if he cannot, then he must have available the work of those who can, and should be in a position to appreciate their difficulties and limitations. In training a Physicist then, it is important that he should have at least some experience of accurate measurement and manipulation of apparatus. The need of course is widely appreciated and all the Examination Boards demand that the candidates for H.S.C. must have pursued a course of practical work and be examined in Practical Physics. It is the effectiveness and validity of the customary tests that cause most concern and they have shown comparatively little development during the past twenty five years.

4. Practical Tests. Present Position with Older Pupils.

Candidates for Higher School Certificate and Intermediate B.Sc., Physics are usually given a three-hour practical examination. The customary form of these examinations is rather unsatisfactory for several reasons. Four problems or questions are set and the candidate has to attempt two. The first points of importance are that the sampling of the course is limited and all the candidates do not attempt the same questions. The assessment is mainly based on the written account finally handed in by the candidate and tends to place a high premium on verbal ability. In scoring the scripts a certain amount of objectivity can be obtained by awarding marks for the intermediate and end products of the candidates work but the conditions of administration are such that marks cannot be awarded with great confidence for high accuracy. There is always a danger of attaching too much importance to the students written description of his work.

In fairness it must be admitted that, although the sampling is limited two well chosen practical problems can involve a reasonable variety of fundamental measurements and techniques. For example a question on the temperature coefficient of resistance involves measurement of temperature, length, and resistance; manipulation of a metre bridge or Post Office Box; and may even involve the application of graphical methods to complete the solution of the problem.

It is easy to criticize the above type of examination but as yet no satisfactory alternative has been produced, and any change in form must ultimately be accepted by, and imposed by the Examination Boards and the Universities. The validity and reliability of the customary practical tests are doubtful but apparently no figures have been published. The Examination Boards do not as a rule publish or make available data from which reliability and correlation coefficients can be calculated.

In 1929 the Joint Matriculation Board did conduct some research in connection with the correlation between Higher School Certificate practical examination results and written paper results. The conditions in the Higher School Certificate Examinations have changed very considerably since then and no information is available with regard to the sizes of the samples involved, and the probable errors of the coefficients. The following coefficients were supplied by and are quoted by kind permission of the Board.

Subject.	r.
Biology	0.65
Botany	0.40
Zoology	0.30
Physics	0.38
Chemistry	0.14

r = product moment correlation coefficient between written examination and practical examination test scores.

It would be dangerous to place much stress on the above figures but it is interesting to note that for Physics and Chemistry the coefficients were low thus indicating, at the worst, poor reliability and validity, or at the best, that the two tests were in general measuring different outcomes. The writer has been unable to obtain corresponding data for more recent years.

Another important point in connection with the practical examinations is the question of what weight should be attached to the scores obtained when they are combined with the pupils' scores on the written tests. In most casesthe weight given to the practical test is comparatively small and only ten to fifteen percent of the pupil's final total score is allotted to the practical work. In addition to this there is a general tendency for practical tests to be set and scored in such a manner that the results show little dispersion. This of course only aggravates the position since when several sets of marks are combined their relative weights or influence upon the final rank or order of the candidates depends upon the standard deviations rather than the means of the individual test. The low weight attached to the practical tests is probably due to the opinion that they are of low validity and reliability, and not due to a belief that Practical Physics is relatively unimportant.

Even if perfectly reliable and valid tests of Practical Ability were available considerable research would still be needed to settle the question of what weight should be given to the practical test scores. The "correct" weight would no doubt be influenced by the age of the pupils and the ultimate objectives of their course of instruction in physics. If it can be shown that the correlation between ability in experimental and theoretical physics is generally low then it would probably be wiser to refrain from combining the scores. An average, even when weighted, of virtually uncorrelated scores can have little significance from an educational and prognostic point of view.

5. <u>Objective Practical Tests. Older Pupils.</u>

Practical tests similar in general character to those used in the present research can be designed for application to more advanced students such as VIth Formers and Intermediate B.Sc. candidates. A three-hour test may consist of ten, or even more, distinct short problems or experiments, and the range and type of suitable problems is not seriously limited by the reduction of the time allowance from ninety minutes to twenty or less minutes per question. Problems which are in many cases fundamentally similar to those at present included in Intermediate B.Sc. tests can be completed in fifteen minutes by average students if they are not expected to produce time consuming verbal accounts of their work. The candidates must of course work fast but it is worth while to note that there is considerable evidence to show that the faster worker is usually more efficient. In fifteen minutes a student can for example; determine the focal length of a spherical mirror; measure the resistance of a coil using a metre bridge; measure the specific heat of a solid; determine the velocity of sound using a resonance tube and a tuning fork of known frequency; obtain a value for the acceleration due to gravity using a simple pendulum, etc. etc. Some of the advantages of practical tests of this nature are as follows:-

(a) The larger number of questions ensures better sampling of the course of study
(b) All the candidates attempt the same experiments.
(c) No premium is placed on verabl ability.
(d) More objective marking can be obtained.

Certain objections to practical tests of this type are obvious. The division of each experiment into well defined sections, each of which requires an answer, does assist objective marking but probably gives the candidate valuable hints on how to carry out the experiment. In some experiments such hints may be justified anddesirable but if it is desired to give more scope to the initiative of the candidate it is usually possible to reword the questions in such a manner that the hints are reduced to a minimum.

One rather more important objection to the tests is that certain techniques and manipulations cannot be adequately tested in fifteen or twenty minutes. In fifteen minutes there is not sufficient time for the recording of multiple check readings to any large extent and graphical solutions of problems are not possible. A further point is raised by the fact that the tests fail to measure or test the pupil's ability to make reasonable written reports of his observations and deductions, and give no indication of the candidates' method of approach to difficulties that may arise, in the course of his experiment. Many of the aspects of practical work not adequately tested by the new type tests are however of such a nature that they might be more efficiently tested through the medium of a written examination. It is possible that with older pupils a more satisfactory testing programme would be obtained by dividing the practical examination into two parts. The first part might consist of eight new type problems taking a total of about two hours for completion. During the remaining hour the examiner might demonstrate an experiment to the candidates using large scale instruments and perhaps tabulating certain readings on a blackboard. The candidates might then be asked to write an account of the experiment and make deductions from the readings. This latter technique has been used to some extent in America and might lead to useful results.

6. A design for Future Research.

The important point with regard to new-type tests of Practical Physics is of course the question of their validity and reliability. If they are reliable and valid then they should have a low correlation with the Theoretical or written tests which are in general measuring different outcomes. It is perhaps wise to point out here that some of the written papers set in Intermediate and H.S.C. Examinations do apparently attempt to measure or test some aspects of practical work since they include questions on the description and design of experiments.

Any attempt at serious research into the question of the validity and reliability of new-type practical tests with older pupils suffers from the drawback that large samples are needed for reliable experiments. Even in a large school of about seven hundred pupils it is unusual to have a group of even twenty candidates for Higher School Certificate

Physics in any one year. It is here that the Examination Boards and Universities could help. They have large numbers of candidates and have had considerable experience in the administration of Practical Tests on a large scale. The validity of newtype practical physics tests must still depend ultimately on the subjective opinions of experienced physicists and teachers but considerable corroborative evidence could be obtained from a large scale experiment using a group of about 300 candidates for an Intermediate B.Sc. or Higher School Certificate examination. An application of the methods of Factor Analysis would give some evidence as to whether the practical and written tests were measuring different outcomes since the factor loadings of each test could be determined. The factors most likely to be involved are the general intelligence or "g" factor the verbal or "v" factor, the numerical or "N" factor, the practical ability of "F" factor of Alexander and the mechanical aptitude or "m" factor of Cox. A reasonable testing programme might consist of the following series of tests,

(a) Two normal written Physics tests (6 hours)
(b) The normal Practical Physics test (3 hours)
(c) A new-type Practical Physics test. (3 hours)
(d) Alexander's Performance Scale (1 hour)
(e) A battery of Cox's tests of Mechanical Aptitude ($1\frac{1}{2}$ hours)
(f) The Bennett-Fry Mechanical Comprehension Test ($\frac{1}{2}$ hour)
(g) One or more reliable verbal Group Intelligence Tests (2 hours)

The amount of time required for actual testing is not unduly large and it should be noted that all the tests involved are group tests with the exception of Alexander's Performance Scale. A more ambitious scheme might involve re-application of tests(b) and (c) after an interval of about six or eight weeks, and the inclusion of an objective written Physics Test such as the Co-operative Physics Test. If the research were sponsored by one of the universities the organisation would be considerably simplified and would only involve in addition to the normal examinations, (a and b,) taken by the candidates, - a programme of six or seven hours group testing and one hour of individual testing. The majority of this testing could be spread over a period of months and there should be no difficulty in obtaining the willing co-operation of the schools or colleges and teachers.

Any experiments which may be applied to the small groups available in a single school or technical college are incapable of producing highly reliable information. It is certain that no vast changes in methods of testing ability in Practical Physics can be expected or justified without experiments involving large and reasonably homogeneous groups, and it is also evident that the Universities and Examination Boards are in an unique position with regard to the initiation and co-ordination of any future investigation.

7. Acknowledgements.

Grateful acknowledgements and thanks are tendered to the following, for assistance,

(1) Mr. W.E. Glister, Headmaster, Chesterfield School, and several members of the staff.

(2) Mr. J. Hall, Chesterfield School, for considerable assistance in administering and preparing the Practical Tests.

(3) Members of the VIth Form who helped in administering the tests, and preparing the apparatus.

(4) The Northern Universities Joint Matriculation Board who kindly made available the correlation coefficients given on page 152.

(5) Mr. J.F. Wood, Lecturer in Physics, King's College Newcastle, for an informative discussion on practical tests.

(6) Mr. D. Smeltzer, Lecturer in Education, King's College, Newcastle for numerous valuable discussions on the work and criticisms which were always constructive.