AN INVESTIGATION INTO OBJECTIVE TESTING IN SCIENCE IN A SCOTTISH SECONDARY SCHOOL

by

S. Friel

A Thesis in part fulfilment of the requirements for the degree of Master of Science of the University of Glasgow

Chemistry Department
June, 1976

AN INVESTIGATION INTO OBJECTIVE TESTING IN SCIENCE IN A SCOTTISH SECONDARY SCHOOL

This thesis surveys the literature on objective testing with particular reference to work done in the area of multiple choice testing.

The history of the Objective Testing movement is traced from the intelligence tests constructed by Binet in France to the present day and includes a survey of the use of multiple choice tests in Science examinations in British schools today.

The experimental work in the thesis was carried out on first and second year pupils in a comprehensive school who were following the Integrated Science Syllabus. Several aspects of the use of multiple choice tests were investigated beginning with an examination of the effect of changing initial responses on multiple choice tests. It was found that pupils gained marks by making these changes more often than they lost marks. This was found to be true for both boys and girls.

To investigate the effect of rearranging the responses within a multiple choice tests three rearrangements of items were made and the results of the rearranged test compared with the original test. The following rearrangements were investigated:-

(a) Reversing the order of the distractors while keeping the key in the same position;

(b) Placing the key in one of the last two positions in a five choice item, when previously it had been in one of the first three positions;

(c) /

(c) Altering the position of the most plausible distractor to a position immediately before the key.

No significant results were obtained except when the most plausible distractor was placed immediately before the key.

To investigate the effect of item arrangement within multiple choice tests, two forms of the same test were constructed, one having the items arranged in the easy to difficult order and the other having the items arranged in the difficult to easy order. The marks were slightly higher for the difficult to easy form of the test.

A preliminary investigation, into a scoring method which allowed for partial knowledge, was carried out and the results suggested that it may be possible to adopt it for use in schools.

CONTENTS

A C K N O W L E D G E M E N T S

I would like to thank the following people for their assistance during the course of my research:-

(1) Mr. J. Brennan, my headmaster, for allowing me to carry out the work within the school and for allowing me the time off to attend the University when required.

(2) The Renfrew division of the Strathclyde Education authority for financial assistance and leave of absence from school to complete the work of the thesis.

(3) My Supervisors, Professor D.W.A. Sharp and Dr. A.H. Johnstone for their help and encouragement during the course of the work, with particular thanks to Dr. Johnstone for giving so generously of his time and advice.

(4) Mrs. M. Carter for her expertise in the typing of the thesis.

(5) The Computer Departments of Glasgow University and Jordanhill College of Education for processing the examination results.

June, 1976

A B S T R A C T

This thesis surveys the literature on objective testing with particular reference to work done in the area of multiple choice testing.

The history of the Objective Testing movement is traced from the intelligence tests constructed by Binet in France to the present day and includes a survey of the use of multiple choice tests in Science examinations in British schools today.

The experimental work in the thesis was carried out on first and second year pupils in a comprehensive school who were following the Integrated Science Syllabus. Several aspects of the use of multiple choice tests were investigated beginning with an examination of the effect of changing initial responses on multiple choice tests. It was found that pupils gained marks by making these changes more often than they lost marks. This was found to be true for both boys and girls.

To investigate the effect of rearranging the responses within a multiple choice test three rearrangements of items were made and the results of the rearranged test compared with the original test. The following rearrangements were investigated:-

(a) Reversing the order of the distractors while keeping the key in the same position;

(b) Placing the key in one of the last two positions in a five choice item, when previously it had been in one of the first three positions;

(c) Altering the position of the most plausible distractor to a position immediately before the key.

No significant results were obtained except when the most plausible distractor was placed immediately before the key.

To investigate /

To investigate the effect of item arrangement within multiple choice tests, two forms of the same test were constructed, one having the items arranged in the easy to difficult order and the other having the items arranged in the difficult to easy order. The marks were slightly higher for the difficult to easy form of the test.

A preliminary investigation, into a scoring method which allowed for partial knowledge, was carried out and the results suggested that it may be possible to adopt it for use in schools.

CHAPTER 1

INTRODUCTION

CHAPTER 1

INTRODUCTION

Multiple choice items are currently the most widely used form of objective test item. They are adaptable to the measurement of many important educational outcomes - knowledge, understanding, judgement, ability to solve problems, to recommend appropriate action, to make predictions. However, like most forms of assessment they are not without their critics.

The most general criticism of multiple choice questions is that they punish the highly able student while rewarding the less able student. Barzun(1) was one of the earliest exponents of this view. He has observed that it is the best student who goes down one grade and the next best who goes up as a result of multiple choice tests. One of the reasons the critics give why multiple choice items discriminate against the first-rate student is that such questions are frequently ambiguous. Hoffman(2) suggests that difficult questions which are meant to distinguish among the better students are difficult for this reason. In order to select the so-called 'best' answer, the student must often try to guess what the examiner wanted when he constructed the item. This procedure, Hoffman contends, puts the first-class student in a 'mental strait-jacket'. On the other hand, the student who is willing to conform to the 'one best answer' norm is rewarded with a higher score .

In this introductory chapter research findings on the following aspects of multiple-choice testing will be reviewed:-

1. The Effect of Guessing
2. The Assessment of Partial Knowledge
3. The Effect of Initial Response changing
4. /

4. The Effect of Item order alteration
5. The Optimum Number of Alternatives
6. The Position Response Set.

1. <u>The Effect of Guessing</u>

Almost since the introduction of multiple choice tests into the field of educational measurement, teachers have been concerned about the guessing factor. To prevent guessing, correction formulae of varying degrees of complexity were developed. All were variations of the original method of subtracting the number of wrong responses from the number correct, ignoring omitted items, and considering the remainder as the significant number. More recently Traub, Hambleton and Singh(3) suggested the addition of some fractional number of points for omitted items as a more direct and effective method of eliciting the desired test-taking procedure. They presented evidence of higher inter-form reliability and an "increase in the number of omitted items for the 'reward for omits' instructions as compared with the 'penalty for wrongs' instructions". It would seem reasonable that more examinees would prefer the direct positive reinforcement of omissive behaviour approach to the less direct negative approach.

In another study, Waters and Waters(4) compared these two approaches and a third 'rights only' scoring method, in terms of -

a) how well each method is liked by examinees;
b) the correlations of scores obtained under each method with test and course performance criteria.

All the examinees were given the same instructions for answering the items. In addition, three sets of instructions concerning the scoring of the tests were given. The scoring procedures (in terms of the weights for 'rights', 'wrongs' and 'omits', respectively) were /

were $(1,0,0)$, $(0,-\frac{1}{4},0)$ and $(1,0,\frac{1}{5})$.

In general, they found that more items were omitted under both the 'bonus for omits' and 'penalty for wrongs' instructions than under the 'rights only' instruction. No significant differences were found in the correlation of scores obtained under the instructions with test and course performance criteria.

However, the sections of the test taken under the 'bonus for omits' instructions were liked better than those taken under the 'penalty for wrongs' instructions. In a similar study Taylor(5) found that there was no significant effect attributable to instructions in a multiple choice mathematics examination.

Those advising against the use of correction formulae have used two major arguments. Little(6)(7) and Edgington(8) have argued that correction formulae should not be used because they treat mistakes produced through misinformation as though they occurred by chance; others, Ruch et al(9) and Jackson(10) have argued that correction formulae do not alter the rank assigned to students sufficiently to warrant their continued use. They point to the high correlations between corrected and uncorrected scores as evidence to support their contention. In addition, Gulliksen,(11) Frederiksen(12) and Lord(13) suggest that as pupils' responses to items are non-random, even when they guess, no correction for random guessing is necessary as long as the test is of an appropriate level of difficulty, and is not markedly speeded, and the instructions do not direct the pupil to guess wildly.

The majority of the research into guessing has been done in America and has been restricted mainly to students in the field of Education and Psychology. In Britain there has been some work done into guessing in school science examinations.

Mathews /

Mathews[14] sees guessing as less of a problem than it is sometimes made out, he further adds that random guessing is rare; candidates who know any chemistry at all will attempt to assess the merits of alternative answers. He concludes by saying that this latter sort of speculation is not without some merit, as without the intelligent guesser science would never have got off the ground. This view is supported by Hudson[15] who also emphasises the place of inspired guesswork in scientific research. Handy and Johnstone [16] maintain that answers to multiple choice questions per se cannot show the line of thought that led the student to choose a particular response, and in a further investigation[17] found that most correct answers were selected validly, the number of students guessing blindly on any one particular question beihg very small.

In general, a course examination is set in order to measure the present achievement of students - their knowledge of course content, or their ability to use the knowledge they have gained in relevant problem situations. If a penalty for guessing instruction is added to an ordinary multiple choice examination, a new set of decisions faces the individual student. For example, such questions as "How certain am I of this answer?", "What is the probability of my being correct?", "If the probability is four-to-one that I am correct, then am I better off to guess or omit the item?". Ability to 'figure the odds', subjective levels of confidence in own judgement, and willingness to gamble or to take a chance, are among the variables which might be expected to influence the result of a test where there is a correction for guessing.

2. The Assessment of Partial Knowledge

Multiple choice items can probably be made more effective if they are modified to give the examinee partial credit for partial knowledge. The related problems /

problems of guessing and partial knowledge have stimulated quite a lot of consideration by examiners who are dissatisfied by the limited amounts of information conveyed by responses to multiple choice items. Two methods have been proposed as ways of assessing partial knowledge.

a) Differential weighting of response alternatives;
b) Confidence Testing.

a) The effects of assigning different weights to different options on multiple choice items have been investigated both theoretically and empirically by Davis and Fifer[18] and Aitken.[19] One result of allowing partial credit for options that are not absolutely correct is an increase in the total score variance and, consequently, test reliability. However, Davis and Fifer[18] have found that this procedure does not appear to alter the test validity significantly. Willey[20] claimed several advantages for a special kind of five-option item which he labelled as a "three-decision-multiple-choice item". The examinee's task on this type of item is to indicate which one of the five options is definitely correct and which two options are definitely incorrect. In scoring the item, three points are given if the examinee indicates that the correct response is 'definitely correct', two points if he does not select the correct response at all, and no points if he indicates that the correct response is 'definitely incorrect'. Effectively, then, the three-decision item requires the examinee to sort five options into three categories, an item being scored 3, 2 or 0 points depending on the category into which the examinee places the correct response. Willey claimed that this method appeared to favour the conscientious examinee over the one who was superficial and impulsive. An added benefit was that standard multiple choice items constructed according to accepted principles of item construction were used.

Bernardson /

Bernardson[21] compared the three-decision multiple choice test and the conventional multiple choice test in terms of their relationship to a criterion of academic achievement (the average mark obtained by each student in his courses for the year). His results showed that the three-decision multiple choice test was no improvement on the conventional multiple choice test in terms of the relationship of the test scores with academic achievement. Previous research by Coombs et al[22] with a similar technique for assessing partial knowledge reached the same conclusions. They used aptitude and intelligence as their reference variables but failed to find a higher correlation between the experimental scores and the reference variables than between the conventional scores and the reference variables. It is possible that questions with five responses might not be the best form of the three-decision multiple choice test, because with five responses most students can find at least two which seem definitely wrong. As a result, good students and poor students alike are not likely to lose points by designating a response which is actually correct as one of the definitely wrong responses. The three-decision multiple choice test with questions having four responses seems like a better form because the examinee has to make a judgement about each of the responses. This is true even though he only marks three of the responses because the one response he does not mark has to be the one he thinks is most likely to be correct if his choice of the correct response should happen to be wrong.

Arnold and Arnold[23] have produced a scoring procedure for multiple choice examinations which allows for partial knowledge and also allows the examiner to control the expected gain due to guessing. This procedure is based on elementary games theory and in it the examinees are instructed to give the smallest set which they are confident contains the correct response. A penalty will be received if the correct or 'best' response /

response is marked out as an incorrect response. This procedure then allows the examinee to receive credit for partial knowledge, the credit being greater the more 'incorrect' responses that he can eliminate.

b) One valid criticism of multiple choice tests is that they do not represent problematic situations very well. If one wishes to sample subject responses related to incomplete information or uncertain knowledge or degree of preference, questions calling for single correct responses are perhaps not appropriate. Asking a student to select a single response compels him to commit himself entirely to a position which he may not be able to do on the basis of his experiences and information. He may not express his certainty over the responses offered him.

Rippey(24) has suggested a system for responding to and scoring multiple choice tests which asks students to express their distribution of preference for options as well as their certainty in that distribution. He states that one of the options available to the test constructor when confronted by partial knowledge and uncertainty is to adapt conventional items to confidence-scoring procedures or use intrinsic items.

Intrinsic items require a distribution of belief over the options on a multiple choice test and do not have unique correct responses. Intrinsic items reflect realistically situations which require choices among a finite number of responses none of which is uniquely correct. Such situations are common and characterise the boundaries of fields of knowledge, and conditions of incomplete information. Rippey has devised a test which contains such items (see Appendix 1). Hambleton et al(25) have compared the reliability and validity of the two methods of assessing partial knowledge, namely, differential weighting of response alternatives and confidence testing which they regarded as having the /

the same goal - the assessment of partial knowledge. Results showed that the confidence testing procedure yielded the most valid and the least reliable scores. Of the two sets of differential weights used the first set produced scores with higher validity, but the second set produced scores with greater reliability. Conventional Testing (used as the baseline) produced scores with the least validity and with as much reliability as the scores yielded by the second set of differential weights.

Although the results suggest that both experimental procedures, and in particular the confidence-testing procedure, may be the means to improve validity, such a conclusion cannot be given any substantial degree of confidence for two main reasons - the small sample sizes tested and the non-significance of the differences in validity.

Much current dissatisfaction with conventional scoring stems from the feeling that it seems inappropriate to require a student to pick a single response when all of his information and intuition tells him that he is not that sure of himself. Current practice does not allow a student to make use of his knowledge of his uncertainty, and it might be argued that, at times, a knowledge of our uncertainty may have considerable value.

At least a modified testing procedure might teach a student to play a different kind of game than he has been learning to play under the usual multiple choice test situation. He could have a greater choice in expressing how he feels about each item. He could be penalised more for making a dogmatic choice than for honestly assessing his uncertainty.

Intrinsic items seem appropriate in exploring strategies of students confronted with problematic situations /

situations or incomplete information. They require that attention be given to the agreement of criterion group responses, thus almost forcing test constructors into the question of validity. But most important, it seems that some items of this sort should be included in examinations simply because they suggest that not all questions worth asking have single, impeccably defined answers.

3. The Effect of Initial Response Changing

Students have frequently asked teachers whether it is advantageous to change initial answers on multiple-choice tests provided time permits a second consideration of at least some of the items. Many students have received the reply that first impressions are the best; others have been told that a second consideration of an item is likely to be better than the first. In one investigation Matthews[26] reported that a majority of students questioned on the issue believed it better not to make such changes, possibly indicating that they had been previously so advised. Many authors give advice on changing responses. Morgan and Deese[27] after telling students to look carefully for errors, state: "When you re-read your examination paper, you will probably be tempted to change some of your responses. We have some sound advice on this point. If you feel strongly that a response should be changed, change it. On the other hand, if you waver between two responses, not being able to make up your mind, do not change the response you set down originally." Dresel and Jensen[28] concur by saying: "Do not change any of your responses unless you find you have made an obvious error." . Frederick[29] says: "Your first thought is generally best. This is a very good rule to follow in taking a multiple choice test ...... If the student has time to think, he may forget the broader aspect which the teacher meant him to take and get mixed up by details." Ehrlich[30] tells students not to /

to panic late in the test: "You take a real risk of ruining correct answers." Armstrong[31] also states: "If there is any doubt, leave your first response."

In a check of student opinion, Lynch and Smith[32] found that 75% of students agreed that re-reading tests and changing responses based on reconsideration of items would tend to lower their scores.

Since both students and written sources strongly favour not changing responses, it is reasonable to expect that a scientific basis for these opinions exists. In fact, literature dating vack to 1928 indicates just the opposite. Lehman,[33] Lowe,[34] Matthews,[26] Berrein[35] and Reille and Briggs[36] report that changing responses (probably based on reconsideration) tends to raise scores. It is this latter view which will be investigated later in this thesis.

4. Item Order Arrangement

It is generally accepted practice for examiners to arrange the items in a multiple choice test in order of increasing difficulty. One obvious reason for this practice is that it increases the probability that an examinee will succeed on the earlier items and thereby gain confidence for the more difficult items later in the test. However, tests are not always constructed in this way.

There is some evidence that the order of items in a multiple choice test has an effect on test performance. MacNicol[37] found that when items were ordered from difficult-to-easy, the mean number of correct responses on the test was significantly lower than the mean number of correct responses obtained when the items were ordered in one of two other ways; from easy-to-difficult and at random. These results were obtained for a test administered under essentially power conditions.

Flaugher /

Flaugher et al[38] discovered an item order effect in their work but under testing conditions where one would more likely expect the effect to occur. They found that when easy items appeared later in a test they were not reached by some examinees. In other words, if the test is speeded it is clear that the difficult-to-easy item order would disadvantage slow students since they would not have a chance to answer the easier items. Brenner[39] and Sax and Cromach[40] have found that item arrangement influenced test performance significantly but this only held true for speed not power tests.

In contrast to these four studies, several other researchers Huck and Bowers,[41] Marso,[42] Monk,[43] and Munz and Smouse[44][45] have failed to observe an item order effect.

Three points should be made in respect of these latter investigations. Firstly, the researchers do not report whether or not the students actually took the items in roughly the same order that they were presented in the test. While it would normally be reasonable to assume that students attempt items in the same order which they are presented, this may not be the case when items appearing early in the test are on the difficult side. Secondly, there is no information on the variation of item difficulties in their tests. If the range of item difficulties is small enough it would be surprising to observe an item order effect. Thirdly, it is not at all clear from the descriptions of their research how important the students perceived the test to be.

One plausible explanation of the item-order effect, if it exists at all, has been offered by Mollenkopf.[46] He argued that fatigue and pressure to finish could account for the poorer performance on easy items when they appeared later in the test.

Another /

Another and perhaps more interesting possibility is that personality characteristics of individual students hinder their performance on items in the difficult-to-easy order. One such personality characteristic which has been found to influence test performance is anxiety.

The existence of this so called item-order-effect will be investigated later in this thesis.

5. The Optimum Number of choices

Tversky[47] has presented mathematical proof that given a fixed number of choices on a multiple choice test, the use of three choices at each choice point will maximise the discrimination and power of the test. Empirical support for this proof could be of considerable practical importance, since in constructing multiple choice tests examiners commonly prefer items with more than three choices in the expectation that by doing so their tests will be more discriminating.

Unfortunately, very few empirical studies have been carried out which are directly relevant to Tversky's proof. Pressey,[48] reported that in a psychology course taught with auto-instruction devices, three choice items were more efficient than four choice items in measuring gains in knowledge, however, neither time nor the number of alternatives was systematically controlled. Ebel and Williams[49] found that by deliberately discarding the least discriminating distractor in a four choice item they could obtain discrimination indices and reliabilities which were about equal to those for four choice items. They also found that students could complete three choice items more rapidly than four choice items.

In a study involving aviation cadets, Zimmerman and Humphrey's[50] found that by eliminating the most poorly discriminating distractors in five choice items, they /

they could improve the reliability of the test and furthermore the time per item on the test was reduced considerably.

Similar work by Ramos and Stenon[51] using French and Spanish reading tests came to the same conclusion.

In a study involving double choice items for measuring achievement on psychology courses, Smith[52] discovered that such items were highly reliable, but he did not make any direct comparisons between them and other types of items.

On a theoretical level Ebel[53] has predicted, according to a formula he developed, that an appreciable increase in the reliability of a multiple choice item will occur when the number of choices is increased from two to three, a smaller increase when four choices are used, and a still smaller increase beyond that point.

Although suggestive, none of these studies provides direct evidence of the empirical validity of Tversky's proof. In order to gather such evidence Costin[54] carried out a study in more controlled conditions than the previous studies. He found that the mean discrimination indices for three choice items were considerably higher than for four choice items, test reliabilities following the same pattern. In a further study, Costin[55] found these measurements to be slightly higher. These findings do not demonstrate, of course, that any three choice test will automatically be more discriminating than any four choice test. For one thing the study was restricted to items which measured knowledge of empirical generalisations. If the same procedures were used to measure higher levels of cognitive achievement - the kinds of tasks which Bloom[56] and his associates have classified under the rubric of 'skills' (e.g. interpreting and applying knowledge) - then different results might occur. It should also be kept /

kept in mind that these two studies compared only two types of multiple choice item. Further investigations have to be carried out which would expand these comparisons by including two choice items and perhaps five choice items as well.

6. The Position Response Set

Several investigators of position response sets (i.e. the tendency to select a particular alternative when the correct response is not known) in multiple choice tests have yielded contradictory results. Cronback(58) concluded that multiple choice tests are relatively free from position response sets. He analysed questions from psychological tests and found no evidence for the response set. However, McNamara and Weitzman(59) report that a tendency does exist among examinees to favour certain response positions to the neglect of others. Their findings suggest that the difficulty level of a multiple choice item is influenced by the position to which the correct response has been assigned. They found that for five choice items, those items having correct responses in the fourth position are the most difficult, those with correct responses in the second and third positions are the easiest, and the first and fifth positions are of moderate difficulty.

Their results for four choice items, show that the third position is the most difficult and that difficulty increases from the first to the third position, and decreases with the fourth. This finding is interpreted as being in agreement with the results for five choice items in that the penultimate position was always found to be the most difficult. This would suggest that both the first and the last choice in an item are more outstanding than the middle three (or two as the case may be), these inner choices become less noticeable and are less likely to be selected. Another /

Another possible explanation offered by McNamara and Weitzman is that an examinee going through the list, without making a choice, is more likely to select the last choice rather than go through the list for a second time. This, however, is not sufficient to explain why the penultimate position was always found to be more difficult than the rest.

On the other hand, Marcus[(60)] has found that multiple choice tests are relatively free of position preferences and suggests that it is the position of the most plausible distractor that more logically accounts for any significant response bias than does a positional preference.

Jessell and Sullins[6] in a related investigation to study the effects on reliability and performance of keyed response option sequencing, that would be extremely unlikely under a random model, found that even the most extreme of these forms (fourteen consecutive items with the same keyed option) did not differ significantly from an 'ideal' arrangement with respect to performance and reliability. They concluded that efforts by test constructors to follow an elaborate scheme of random permutations in cyclic order, or any other elaborate plan for arranging the correct response, was unwarrented.

In view of the fact that much of the research into multiple choice testing has been done in America mainly with university level students of Education and Psychology, it is the aim of this thesis to investigate the relevance of these findings to multiple choice testing in school science at the present time.

The results of the investigation should be of interest to multiple choice test constructors in science - because of the fact that more pupils of a wider range of ability are being presented for the S.C.E. examinations in science. It will, therefore, be of value /

value to find out to what extent these American research findings apply to younger pupils in a comprehensive school situation. To this end the following aspects of multiple choice testing will be investigated.

1. A brief survey of the history and use of objective tests and the extent to which multiple choice questions are used in school science examinations at the present time.
2. The effect of altering initial responses on overall test performance.
3. The Item-Order effect.
4. The effect on multiple choice test statistics of altering the positions of keys and distractors within items.
5. A method of scoring multiple choice items based on partial knowledge.
6. In all the above topics reference will be made to the relative performances of boys and girls.

A P P E N D I X 1

Rippey's[24] Examination using Intrinsic Items

The student is given a description of an experiment to read, then he is given the following:-

a) You select the correct answer with certainty on those items which have unique correct responses.
b) You express your honest degree of confidence on those items which do not have unique correct responses.
c) You do not choose your numbers randomly. If you have no preference for any of the responses, give each response the same numerical value, i.e. 1,1,1.

If you like all of the responses equally well, mark the answer sheet as follows:-

a. 3
b. 3
c. 3

If you like none of the responses, mark the answer sheet as follows:-

a. 0
b. 0
c. 0

If you like just the second answer and none of the others mark -

a. 0
b. 9
c. 0

If you like some answers better than others, respond with a digit from 0 to 9 which will express your odds on that answer. For example, if you like answer 'c.' three times as much as answer 'b.' mark -

a. /

a. 0
b. 3
c. 6

Items which call for a distribution of confidence can be computer scored by means of a Euclidian Scoring Function.

CHAPTER 2

THE HISTORY OF THE OBJECTIVE TEST MOVEMENT AND THE PRESENT DAY SITUATION IN SCIENCE SUBJECTS

C H A P T E R 2

THE HISTORY OF THE OBJECTIVE TEST MOVEMENT AND THE PRESENT DAY SITUATION IN SCIENCE SUBJECTS

2.1 The origin of the objective test movement may be traced to the Intelligence Tests pioneered by Binet around the turn of the century in France. In 1910 the Binet tests crossed the Atlantic with the translation and adaptation done by Goddard, director of a training school for retarded children. Soon afterwards, Terman at Stanford University gave to one of his post graduate Students, Otis, the assignment of adapting the Binet test to the children in America. The Binet test consisted of free-response exercises or observation of responses to assigned activities and had to be administered invividually. Otis replaced the individual test items with multiple choice items in which the subject was to pick the correct answer from among several choices. This modification permitted the test to be given to large groups of children.

There are several significant points in this adaptation of the Binet test:-

(a) The test could be given to a large group.
(b) It was objective.
(c) It could be scored reliably.
(d) Collection of the data and statistical treatment made possible the study of groups and inter-group comparisons.

The next impetus to objective tests came as a result of World War I and the development of what came to be known as the Army Alpha Test, which solved the problem of testing quickly and reliably millions of servicemen. The factors which gave the initial impetus to the development of objective tests was present to an even /

even larger extent in the American educational system after World War II, with the rapid expansion in the education system especially at secondary school level.

2.2 Within this educational environment, educators in many fields in the schools were experimenting with and using objective tests. It was at this time also that nationwide testing organisations were founded including the American Council on Education, the College Entrance Exam Board, and the Carnegie Foundation for the advancement of Teaching. In 1947 these three organisations merged to found the Educational Testing Service. The E.T.S. provides current tests and develops new tests. It conducts research into all aspects of testing and calls upon substantial numbers of teachers to advise it upon curriculum aspects of testing. Much of this testing is in the form of objective-type tests because they can so obviously be checked, pre-tested and analysed. But E.T.S. is not by any means restricted to objective testing. It aims to cover all areas and all forms of testing, including essay-style testing. A separate division of E.T.S. provides an Evaluation and Advisory Service. This department stands ready to advise on the current availability of suitable tests for a given purpose; on the interpretation of test marks derived from class or training centre sources; and on the methods by which an educational institution, or a training organisation could produce its own testing scheme.

2.3 Mention must be made at this point of the work of Sir Godfrey Thomson in Scotland. Sir Godfrey was one of the early pioneers in test construction and his Moray House Tests set standards which have been followed by many others both at home and abroad. The tests get their name from the fact that the University Department in which Sir Godfrey worked was housed in the same building as the Moray House College of Education in Edinburgh. Although the tests are no longer produced in /

in Moray House the name is retained as a kind of trademark. Moray House Tests are actually produced by the Godfrey Thomson Unit for Educational Research.

In fact, when in the autumn of 1931 the Scottish Council for Research in Education Mental Survey Committee decided to give a Group Intelligence Test to all the 1921 age group of Scottish children, but found that the expense of buying published tests was too great, Sir Godfrey Thomson placed the Moray House Tests at the disposal of the committee, and test Number 12 was chosen. As however, it had been designed to spread out the able more than the dull children, it was also decided to supplement it by picture tests given with oral instructions. The complicated marking system was also modified.

There were two events of significant importance in the field of objective testing which were held in 1967.

2.4 In May 1967 an International Workshop was held in West Berlin to discuss the possibilities and limitations of educational testing. Twenty-eight countries sent representatives, who came from all five continents. Educationalists in all twenty-eight countries were considering reforms in their methods of examining candidates. The major countries especially interested in objective testing included America, Japan, Sweden, Argentine, Chile, Germany and the Netherlands.

2.5 In the same year, the International Association for the Evaluation of Educational Achievement, published its first report based on objective testing. The council of this association had been formed in 1960. Its first project was to attempt to establish evaluation techniques which would be acceptable internationally. It was hoped by this methodological approach to be able in future to undertake more meaningful comparative /

comparative education studies in such problem areas as failure at school and the internationally comparative meaning of examination results.

The Association decided to undertake a study of a measure of achievement in mathematics for children aged 13+ in twelve different countries including Scotland, England, Japan and Israel. It was obvious that with countries of such widely different educational systems, social structures and languages, extreme care would have to be taken about the style of the test to be administered to the selected children in the twelve countries. In the event, the tests were objective in style, were of the multiple choice type and contained five choices. This study was the first thoroughly researched undertaking into the measurement of achievement of children in a given subject from widely differing educational systems. The Association has now undertaken similar internationally comparative studies of achievements in other subjects such as science and reading comprehension. The Science test was given to three age groups corresponding roughly to the late primary, early secondary and late secondary stages as we know them in this country. The Science test as given, for example, to 14 year olds in each of the nineteen participating countries consisted of two multiple choice papers each of forty items with a time limit of forty-five minutes. The tests contained a variety of questions on Chemistry, Physics and Biology together with a few questions on the Earth Sciences. The results showed that there were different levels of achievement in science among the participating countries, and that this may be due, at least in part, to different emphasis being put on the different branches of Science within these countries.

The International Baccalaureate, after several years experiment, held its first full examination in May 1970. The following subjects included an objective test component: Economics, Biology, Mathematics and the main /

the main foreign languages.

2.6 Recent Developments in England

Until 1964 there was almost no application of objective testing at the secondary or tertiary levels of education. The simpler forms of objective test were well known to teachers at the primary level, because such tests formed part of the secondary school selection tests then taken towards the end of primary school courses (at the age of 11+).

Special studies of objective-style testing at the secondary or tertiary education levels were undertaken by the National Foundation for Education Research and by some university departments of education, but still no general use was made of this work by the examining boards for the General Certificate of Education.

But in 1964 the Secondary Schools Examinations Council published "The Certificate of Secondary Education : An Introduction to Objective Type Examinations". The survey was written by Professor Vernon and gave a well-balanced account of objective-type testing and indicated support for its use, especially at the level of the Certificate of Secondary Education.

Because the first examinations for the C.S.E. were well under way when the Schools Council first published its introduction to objective-type testing, there was no immediate response among the fourteen regional examining boards. The publication, however, was widely disseminated in schools, colleges and within examining boards throughout the country. Almost immediately following the publication there was a rapid development of interest by some examining boards covering the examinations for the General Certificate of Education - notably the University of London University Entrance and School Examinations Council and the Associated Examining /

Examining Board.

These bodies involved substantial numbers of teachers in their advisory and subject committees. These teachers formed the nucleus of the group set up to study appropriate applications of objective testing to their own subject areas. First reactions were sceptical and critical, but the prospect of greater precision in measurement, of application to wider areas of mental process, and the professional expertise which the Boards' administrative staffs built up, all began to change attitudes.

In June 1969 the University of London University Entrance and Schools Examinations Council introduced operationally a multiple choice test as a component part of the examination for the General Certificate of Education Ordinary level paper in Chemistry. The remaining component consisted of a Chemistry test on conventional lines.

By 1971 almost every Board examining for the General Certificate of Education was studying closely the need to introduce an objective test component into their examination structure. The Associated Examining Board, in addition to its joint project with the University of London in the use of objective tests in Economics, had working groups studying the introduction of objective testing at Ordinary Level for Chemistry, English, Mathematics, Physics, Religious Knowledge, Statistics and History. The Joint Matriculation Board published in July 1970 a booklet giving examples of objective tests and encouraging their use in science subjects of their General Certificate of Education examinations.

2.7 In order to find out what results these and other investigations have had the following questionnaire was sent to all the G.C.E. and C.S.E. Examination Boards /

Boards in England at the beginning of 1975. The questionnaire was aimed specifically at science examinations.

(a) In which science examinations is objective testing used?
(b) What weighting is given to it?
(c) How many items are there in each paper?
(d) On the basis of what pre-testing are individual items selected?
(e) Could you provide me with any relevant research findings you have made?
(f) What general conclusions have you reached concerning their use?

Results

Associated Examination Board

Subject	Level	Weighting	No. of Items	Pre-test
Chemistry	O	50%	60	Satisfactory performance in pre-test involving 300-400 candidates.
Physics	O	50%	60	
Physical Science	O	40%	60	
Integrated Science 'A'	O	60%	45	No formal pre-testing at the moment.
Integrated Science 'B'	O	30%	30	

N.B. An objective test component has been included in new syllabus proposals for Advanced Level chemistry and physics.

University of London

Subject	Level	Weighting	No. of Items	Pre-test
Chemistry	O	40%	70	Pre-test paper taken by sufficient candidates to ensure that valid statistical analysis of the results can be carried out.
Physics	O	40%	-	
Biology	O	40%	60	
Chemistry	A	30%	50	

University of Cambridge Local Examinations Syndicate

Subject	Level	Weighting	No. of Items	Pre-test
Chemistry	O	27%	30-40	Carried out on school candidates by the test and development unit.
Biology	O	27%	-	
Chemistry	A	15%	-	
Biology	A	15%	-	
Physics	A	15%	-	
Social Biology	A	15%	-	

Oxford and Cambridge Schools Examination Board

Subject	Level	Weighting	No. of Items	Pre-test
Chemistry	O	40%	40	
Chemistry (Phys. with Chemistry)	O	40%	30	Pre-tested well in advance in schools preparing candidates in that subject.
Physics (Nuffield)	O	50%	50	
Physics (Nuffield) with Chemistry	O	50%	30	
Chemistry	A	40%	50	
Physics	A	45%	50	
Physics (Nuffield)	A	23%	40	

Oxford Delegacy of Local Examinations

Subject	Level	Weighting	No. of Items	Pre-test
Physics	A	30%	50	Pre-tested on representative sample of 200 candidates.

Joint /

Joint Matriculation Board

Subject	Level	Weighting	No. of Items	Pre-test
Chemistry	O	40%	40	Pre-tested on samples of approximately 350.
Physics	O	30%	-	
Physical Science	O	40%	-	
Science (Double)	O	40%	-	
Physics with Chemistry	O	40%	-	
General Science	O	40%	-	
Chemistry	A	28%	-	
Physics	A	38%	50-60	
Physical Science	A	31%	-	

Welsh Joint Education Committee

Subject	Level	Weighting	No. of Items	Pre-test
Physical Science	O	10%	10	Not disclosed.

The Southern Universities Joint Board for School Examinations do not make use of objective tests in any of their science examinations.

Observations

1. Objective testing is well established in the Physical Sciences at both 'O' and 'A' level. This is not so for the Biological Sciences.
2. The weighting given to the objective part of an examination is fairly constant within boards at a given level. In each case the average weighting given to 'O' level examinations is higher than that given at 'A' level. There is also a slight variation in weighting between examination boards.
3. Where information is available this also applies to the number of items in the test.

4. /

4. Pre-testing is done on a large enough scale to ensure an adequate number of satisfactory items.

Certificate of Secondary Education

The West Yorkshire and Lindsey Regional Examining Board

Subject	Weighting	No. of Items	Pre-test
General Science	12%	12	With the exception of the joint J.M.B. biology exam, formal pretesting is not arranged.
Biology	10%	8	
Chemistry	12%	10	
Physics	40%	40	
Biology (with J.M.B.)	25%	60	

Yorkshire Regional Examinations Board

All science subjects	variable	variable	Only a very little done.

South Western Examinations Board

Chemistry	-	10	Only pre-tested questions are in General Science 'A' (sample of up to 400).
General Science 'A'	-	20	
General Science 'B'	-	12	

Metropolitan Regional Examinations Board

Biology	25%	30-35	Draft papers are pre-tested as such. There is no pre-testing of individual items.
Human biology	25%	30-35	
Chemistry (2 syllabuses)	25%	30-35	
Physics (2 syllabuses)	25%	30-35	
Environmental Science	25%	30-35	
Natural Science	25%	30-35	

The South-East /

The South-East Regional Examinations Board

Subject	Weighting	No. of Items	Pre-test
Biology	15-20%	70	Two papers containing 60-70 items in each subject are pre-tested the previous year. Approximately 400 papers sent out to schools in other regions with appropriate syllabuses.
Chemistry	15-20%	60	
Chemistry(Nuffield)	15-20%	60	
Integrated Science	15-20%	60	
Physics	15-20%	60	
Physics (Nuffield)	15-20%	60	
Science of Living	15-20%	70	
Rural Studies	15-20%	60	

Associated Lancashire Schools Examining Board

Biology	-	9	Pre-testing not done.
Human and Social Biology	-	20	
Chemistry	-	19	
General Science	-	21	
Physics	-	20	

North Regional Examinations Board

This board uses multiple choice objective questions only to a very limited extent. They therefore regretted that they could not supply answers to the questionnaire.

East Anglian Examinations Board

This board was unable to supply any information because of the number of syllabuses involved.

Middlesex Regional Examining Board

Objective testing has not been used by this board in C.S.E. Science examinations.

2.8 Observations /

2.8 Observations

(1) Where objective tests are used they appear to be well established in both the Physical and Biological Sciences.

(2) The weighting given to the objective part of the examination is fairly consistent within examination boards. Where information is available, the variation, with one exception is only 11%, which is considerably lower than the differences at either 'O' or 'A' level.

(3) The number of items is reasonably consistent within boards, but there is a wide variation between boards.

(4) In more than half the examination boards no formal pre-testing of items takes place.

2.9 Conclusions

Apart from Biology at G.C.E. 'O' and 'A' level, objective tests are a well established part of school Science examinations in England. There is naturally a variation between the different examination boards as regards the weightings to be given to these objective tests, since within the members of any one board there is bound to be a difference of opinion as to how objective tests are to be used.

The difference in average weightings that occurs between the C.S.E. and G.C.E. 'O' and 'A' level examinations can be explained by the nature of the examination. The low percentage weighting given to the objective component of C.S.E. examinations being due to the fact that a proportion of the overall marks, in certain cases, are given for teacher assessment, and in 'A' level a certain proportion of the marks are given for a practical examination. One point worth noting is the apparent lack of pre-testing in some of the C.S.E. examination boards. This may be a matter of economy and/or insufficient numbers of pupils taking the examinations.

2.10 The position in Scotland is much clearer than in England because there is only one examination board responsible for examinations in school Science subjects. At the present time there is an objective test paper in both Chemistry and Physics at both ordinary and higher grades. Objective tests are shortly to be introduced into examinations in Biology and Anatomy, Physiology and Health. Here there is a similarity with England in that objective tests in the Physical Sciences were established some years before those in the Biological Sciences. As regards pre-testing in Scotland, all questions go through a rigorous shredding and pre-testing procedure before they are accepted for use in the S.C.E. examinations.

In general, the objective test paper contains questions of varying difficulty designed to test knowledge, comprehension, application, and the ability to analyse and evaluate. The paper tests at different ability levels all the educational objectives of the syllabus amenable to written tests except those -

(a) where recall rather than recognition is required;
(b) where partial marking is desirable;
(c) where the ability to be tested requires more lengthy response and a higher number of marks.

CHAPTER 3

THE EFFECT OF ALLOWING PUPILS TO CHANGE THEIR FIRST CHOICE IN MULTIPLE CHOICE EXAMINATIONS

CHAPTER 3

THE EFFECT OF ALLOWING PUPILS TO CHANGE THEIR FIRST CHOICE IN MULTIPLE CHOICE EXAMINATIONS

Introduction

The purpose of this chapter is to investigate the effect of allowing pupils to change their first choice in multiple choice science examinations. Each of the multiple choice items used consisted of one correct response and four distractors. The number of items in each test was, with one exception, twenty-five. The response-changing behaviour is to be investigated with respect to:-

(a) The variability between tests;
(b) The variability among pupils;
(c) The types of change made;
(d) Sex differences;
(e) The overall test performance as related to response changes;
(f) The relationship of item difficulty to response changing;
(g) The effect of item position on response changing.

Procedure

The tests were given to first and second year pupils in a comprehensive school over a period of two years. The tests were administered in the week immediately after completing the relevant section of the Integrated Science Syllabus.[61] There was no time limit put on the tests, therefore all the pupils completed every item and there was sufficient time for any response alterations to be made.

After every test the same procedure was followed. Each paper was examined for evidence of response changes as indicated by erasures or score-outs. When an item response was changed, it was recorded in one of three categories:-

(a) /

(a) an incorrect to a correct response;
(b) a correct to an incorrect response;
(c) an incorrect to another incorrect response.

The instances in which any given item was changed more than once were too few to be of any consequence. When this occurred, if any of the responses made prior to a final incorrect response were correct, this was listed as a "correct to an incorrect response".

If all responses made prior to the final response were in error, then there was no problem; all these responses were considered as "first responses" and were recorded either "first response incorrect; final response correct" or "first response incorrect; final response incorrect".

The results are shown in the following tables.

Table 3.1

Response Changes by Sex
Test on section one of the Integrated Science Syllabus given to all first year pupils in session 1974/75.

	Changes by Males (N=91)		Changes by Females (N=130)		Total Changes	
Type of Response Change	N	%	N	%	N	%
Incorrect to Correct	69	65.1	79	60.8	148	62.7
Correct to Incorrect	9	8.5	19	14.6	28	11.9
Incorrect to Incorrect	28	26.4	32	24.6	60	25.4
Total	106	100.0	130	100.0	236	100.0

Table 3.2

Response Changes by Sex
Test on section one of the Integrated Science Syllabus given to all first year pupils in session 1975/76.

Type of Response Change	Changes by Males (N=94)		Changes by Females (N=101)		Total Changes	
	N	%	N	%	N	%
Incorrect to Correct	70	58.3	79	59.9	149	59.1
Correct to Incorrect	22	18.3	18	13.6	40	15.9
Incorrect to Incorrect	28	23.4	35	25.5	63	25.0
Total	120	100.0	132	100.0	252	100.0

Table 3.3

Response Changes by Sex
Test on section four of the Integrated Science Syllabus given to all first year pupils in session 1974/75.

Type of Response Change	Changes by Males (N=88)		Changes by Females (N=108)		Total Changes	
	N	%	N	%	N	%
Incorrect to Correct	57	60.0	80	49.1	137	53.1
Correct to Incorrect	11	11.6	34	20.9	45	17.4
Incorrect to Incorrect	27	28.4	49	30.0	76	29.5
Total	95	100.0	163	100.0	258	100.0

Table 3.4

Response Changes by Sex
Test on section four of the Integrated Science Syllabus given to all first year pupils in session 1975/76.

Type of Response Change	Changes by Males (N=94)		Changes by Females (N=93)		Total Changes	
	N	%	N	%	N	%
Incorrect to Correct	60	49.6	76	51.4	136	50.6
Correct to Incorrect	24	19.8	30	20.3	54	20.1
Incorrect to Incorrect	37	30.6	42	28.3	79	29.3
Total	121	100.0	148	100.0	269	100.0

Table 3.5

Response Changes by Sex
Test on section ten of the Integrated Science Syllabus given to all second year pupils in the session 1974/75.

Type of Response Change	Changes by Males (N=105)		Changes by Females (N=101)		Total Changes	
	N	%	N	%	N	%
Incorrect to Correct	71	48.6	60	46.9	131	47.8
Correct to Incorrect	27	18.5	26	20.3	53	19.3
Incorrect to Incorrect	48	32.9	42	32.8	90	32.9
Total	146	100.0	128	100.0	274	100.0

Table 3.6

Response Changes by Sex
Test on section ten of the Integrated Science Syllabus given to all second year pupils in the session 1975/76.

Type of Response Change	Changes by Males (N=90)		Changes by Females (N=130)		Total Changes	
	N	%	N	%	N	%
Incorrect to Correct	60	54.1	68	41.5	128	46.5
Correct to Incorrect	19	17.1	33	20.1	52	18.9
Incorrect to Incorrect	32	28.8	63	38.4	95	34.6
Total	111	100.0	164	100.0	275	100.0

Table 3.7

Response Changes by Sex
Test on section twelve of the Integrated Science Syllabus given to all second year pupils in the session 1974/75.

Type of Response Change	Changes by Males (N=97)		Changes by Females (N=81)		Total Changes	
	N	%	N	%	N	%
Incorrect to Correct	50	36.8	62	42.5	112	39.7
Correct to Incorrect	31	22.8	29	19.9	60	21.3
Incorrect to Incorrect	55	40.4	55	37.6	110	39.0
Total	136	100.0	146	100.0	282	100.0

Table 3.8

Response Changes by Sex

Test on section twelve of the Integrated Science Syllabus given to all second year pupils in the session 1975/76.

Type of Response Change	Changes by Males (N=87)		Changes by Females (N=124)		Total Changes	
	N	%	N	%	N	%
Incorrect to Correct	48	33.8	69	37.7	117	36.0
Correct to Incorrect	34	24.0	51	27.9	85	26.1
Incorrect to Incorrect	60	42.2	63	34.4	123	37.9
Total	142	100.0	183	100.0	325	100.0

Comments

To see if there is any difference in response changing between successive years of pupils on the same test, the relevant pairs of tests, e.g. Integrated Science Syllabus section one test in session 1974/75 and in session 1975/76 were compared.

(a) Test one - 1974/75 and 1975/76

There was a slight decrease in the proportion of pupils making positive changes (i.e. from an incorrect response to a correct response) with an accompanying increase in negative changes (i.e. from a correct response to an incorrect response). Neutral response changing (incorrect to another incorrect response) stayed relatively constant.

(b) Test four - 1974/75 and 1975/76

Again there was a decrease in the proportion of pupils making positive changes with a corresponding increase in negative changes. The proportion of pupils making neutral changes also remained constant.

(c) Test ten - 1974/75 and 1975/76

There was a variation in the proportions of pupils making all three types of change, notably the proportion of /

of boys making positive changes rose by 5.5% and the proportion of girls making positive changes decreased by about 5.5%.

(d) Test twelve - 1974/75 and 1975/76

For the third time there was a decrease in the proportion of pupils making a positive change and increase in the proportion making negative changes. The proportion of pupils making neutral changes was again fairly constant.

In an attempt to explain these variations in the various types of response change, the degree of difficulty of each of the tests was calculated, i.e. the average facility value of all the items in a test. The results are shown in Table 3.9.

Table 3.9

Degree of Difficulty of Examinations

Section Number	Degree of Difficulty (%)
1^a	67.4
1^b	71.4
4^a	57.0
4^b	55.1
10^a	58.0
10^b	55.4
12^a	42.9
12^b	40.5

a denotes 1974/75 examinations
b denotes 1975/76 examinations

By consideration of these degrees of difficulty it can be seen that in each case, with the exception of the section one tests, there is an increase in difficulty with successive years. This increase in difficulty could account for the observed decrease in the /

the proportion of positive response changes, and the observed increase in negative response changes. This apparent increase in examination difficulty could be due to a slightly less intelligent pupil population, however, it is not large enough to alter the overall picture significantly, but it must still be considered.

In the case of section one, any interpretation of the results must take into account the fact that it was the first test of its type that the pupils had encountered, and as such would invariably produce unreliable results. This would not of course apply to later tests, because by that time a certain amount of examination sophisticationwould have been developed. It would appear then that a relationship may exist between the degree of difficulty of an examination and the number of positive and negative response changes.

Considering the results for all eight examinations the proportion of positive (i.e. from an incorrect to a correct response) response changes ranges from 36.8% to 65.1% and the negative (i.e. from a correct to an incorrect response) response range is from 8.5% to 27.9%.

An examination of the literature on response changing shows a similar pattern.

Table 3.10 /

Table 3.10

Reported Frequencies of Response Changing

Type of Response	Lynch D.O.[32] Smith B.C.	Briggs L.J.[36] Reile P.J.	Jarret[62] R.F.	Matthews[26] C.O.	Foote R.[63] Belinky C.
	%	%	%	%	%
Incorrect to Correct	56.5	49	70.6	53	55
Correct to Incorrect	26.9	23	16.8	21	22
Incorrect to Incorrect	16.6	28	12.6	26	24

A closer examination of these investigations shows no mention of the degree of difficulty of each of the tests used. Therefore, if a relationship did exist between degree of difficulty and type of response change, it may be a factor in the variations shown in these independent researches.

In order to test this hypothesis the Pearson's Product Moment Correlation coefficient was calculated for the degree of examination difficulty against -

(a) the proportion of positive response changes;
(b) the proportion of negative response changes;
(c) the proportion of neutral response changes.

The results are shown in Table 3.11.

Table 3.11

Coefficients of Correlation (Pearson Product Moment)

Type of Response Change	Coefficient of Correlation
Positive (incorrect to correct)	+ 0.92
Negative (correct to incorrect)	- 0.88
Neutral (incorrect to incorrect)	- 0.74

Conclusion

From the results it can clearly be seen that a relationship does exist between the degree of difficulty of a test and the proportions of the different types of response change. This relationship can be used, at least in part, to explain the differences in these proportions both in the literature and the work described in this chapter.

To obtain an overall picture of the response changing pattern the results for all of the sections were combined and are shown in Table 3.12.

Table 3.12 /

Table 3.12

Total Response Changes by Sex - All sections

Type of Response Change	Changes by Males (N=746)		Changes by Females (N=868)		Total Changes	
	N	%	N	%	N	%
Incorrect to Correct	485	49.6	573	48.0	1058	48.7
Correct to Incorrect	177	18.1	240	20.1	417	19.2
Incorrect to Incorrect	315	32.3	381	31.9	696	32.1
Total	977	100.0	1194	100.0	2171	100.0

Conclusions

Since the aim of changing a response on a test is to improve performance, those who chose to change items on these tests did in fact improve their scores more frequently than did they reduce their scores. Of a total of 2,171 response changes almost half of these resulted in a correct final response; of those changes that produced an incorrect final response it was more likely that the first response too was incorrect (ratio approximately 1.6 : 1.0). Since a change from an incorrect response to another incorrect response does not alter test performance, it can be said that the chances were about 5 : 2 that a change of response would produce an increase rather than a decrease in test score.

Sex differences with respect to the three types of response change were small. Girls were more likely than boys to make a "correct to incorrect" change, whereas boys were more likely to make a "incorrect to correct" response change and to make a "incorrect to incorrect" response change. Girls were slightly more inclined to make response changes than the boys, whereas 53.7% of the sample were girls they made 55% of all the response changes.

Relationship /

Relationship between Overall Test Performance and Response Changes

To study the possible relationships between quality of performance on the test and the changing of responses, the position of each pupil was determined as being in the highest, middle, or lowest one third of the class distribution. This procedure was applied to the results for tests on sections one, four, ten and twelve of the Integrated Science Syllabus. The results are reported for individual sections and collectively.

Table 3.13 /

Table 3.13

Response Changes and Quality of Overall Performance - Section one, 1975/76

Type of Response Change	Mean Number of Response Changes						Total	
	Highest Third		Middle Third		Lowest Third			
	M(41)	F(29)	M(31)	F(43)	M(22)	F(29)	M(94)	F(101)
Incorrect to Correct	0.88	0.43	0.61	0.84	0.73	0.48	0.75	0.76
Correct to Incorrect	0.15	0.03	0.26	0.28	0.36	0.21	0.23	0.19
Incorrect to Incorrect	0.15	0.03	0.39	0.40	0.45	0.59	0.30	0.35
Total	1.18	0.99	1.26	1.52	1.54	1.28	1.28	1.30

Observations

(1) The tendency to change responses is greatest among the poorest and average pupils. For boys, the tendency to change responses is greatest among the poorer pupils and least among the best pupils. For girls the tendency is greatest among the average pupils and least among the best pupils (see Total row).

(2) Among girls who made "Incorrect to Correct" changes, the best pupils were the most active changers and the poorest pupils were the least active. Among boys making this kind of change the best pupils made the most changes, the least number of changes being made by the average pupil.

(3) Among those pupils who made "Correct to Incorrect" changes, girls in the average and lowest groups made respectively 9, and 7, times more changes than those in the best group. Among boys, most changes of this nature were made by the poorest group, the least changes being made by the best group.

(4) /

(4) Although the "Incorrect to Incorrect" change had no effect upon total test performance it is noted that the tendency to make this change is greatest among the poorest pupils and least among the best pupils. This is true for both sexes.

Table 3.14 /

Table 3.14

Response Changes and Quality of Overall Performance - Section four, 1975/76

Type of Response Change	Mean Number of Response Changes						Total	
	Highest Third		Middle Third		Lowest Third			
	M(25)	F(32)	M(36)	F(31)	M(30)	F(33)	M(91)	F(90)
Incorrect to Correct	1.00	0.81	0.58	0.93	0.47	0.60	0.64	0.83
Correct to Incorrect	0.12	0.19	0.33	0.32	0.30	0.33	0.26	0.30
Incorrect to Incorrect	0.20	0.44	0.42	0.45	0.53	0.39	0.40	0.46
Total	1.32	1.44	1.33	1.70	1.30	1.32	1.30	1.59

Observations

(1) The tendency to change responses is greatest among the average pupils and least among the best and poorest pupils. This is true for both sexes.

(2) Among girls who made "Incorrect to Correct" changes, the average pupils were the most active changers and the poorest pupils the least active. Among boys making this kind of change the best pupils are the most active and the poorest pupils least active.

(3) Among those who made "Correct to Incorrect" the average and poorest pupils made most changes. This is true for both sexes.

(4) Among those boys who made "Incorrect to Incorrect" changes the poorest pupils were the most active, the best pupils making the fewest changes of this type. The best and average girls made slightly more of these changes than did the poorest girls.

Table 3.15

Response Changes and Quality of Overall Performance - Section ten, 1975/76

	Mean Number of Response Changes							
	Highest Third		Middle Third		Lowest Third		Total	
Type of Response Change	M(35)	F(45)	M(36)	F(37)	M(19)	F(48)	M(90)	F(130)
Incorrect to Correct	0.83	0.73	0.75	0.51	0.26	0.31	0.68	0.52
Correct to Incorrect	0.09	0.27	0.28	0.24	0.32	0.25	0.21	0.25
Incorrect to Incorrect	0.23	0.38	0.50	0.51	0.32	0.58	0.38	0.49
Total	1.15	1.38	1.53	1.26	0.90	1.14	1.27	1.26

Observations

(1) The tendency to change responses is greatest among the average boys and the best girls and least among the poorest boys and girls.

(2) The tendency to make "Incorrect to Correct" changes is greatest among the best pupils and least among the poorest pupils. This is true for both sexes.

(3) The tendency to make "Correct to Incorrect" changes is greatest among the poorest boys and least among the best boys. Among girls there appears to be no difference in the tendency to make changes of this type.

(4) The tendency to make "Incorrect to Incorrect" changes is greatest among the average boys and the poorest girls. The tendency to make these changes is least among both the best boys and girls.

Table 3.16

Response Changes and Quality of Overall Performance - Section twelve, 1975/76

Type of Response Change	Mean Number of Response Changes						Total	
	Highest Third		Middle Third		Lowest Third			
	M(33)	F(40)	M(25)	F(37)	M(28)	F(47)	M(86)	F(124)
Incorrect to Correct	0.67	0.68	0.56	0.59	0.39	0.93	0.55	0.56
Correct to Incorrect	0.45	0.33	0.20	0.57	0.50	0.38	0.40	0.42
Incorrect to Incorrect	0.45	0.43	0.68	0.65	0.43	0.53	0.67	0.53
Total	1.57	1.44	1.44	1.81	1.82	1.34	1.62	1.51

Observations

(1) The tendency to change responses is greatest among the average girls and poorest boys. It is least among the average boys and poorest girls.

(2) Among pupils who made "Incorrect to Correct" changes, the best pupils were the most active changers. The poorest pupils were the least active in this respect. This is true for both sexes.

(3) Among boys who made "Correct to Incorrect" changes the poorest pupils were the most active changers and the average pupils the least active. Among girls making this type of change the average pupils were the most active, and the best pupils the least active changers.

(4) The most "Incorrect to Incorrect" changes were made by the poorest boys and the average girls whereas the least changes of this type were made by the best pupils of both sexes.

In order to obtain the overall pattern the results for all the sections were combined. The result is shown in Table 3.17.

Table 3.17

Response Changes and Quality of Overall Performance - All sections, 1975/76

Type of Response Change	Mean Number of Response Changes						Total	
	Highest Third		Middle Third		Lowest Third			
	M(134)	F(146)	M(128)	F(148)	M(99)	F(157)	M(361)	F(451)
Incorrect to correct	0.84	0.77	0.63	0.71	0.46	0.44	0.66	0.64
Correct to Incorrect	0.20	0.22	0.27	0.35	0.37	0.30	0.27	0.29
Incorrect to Incorrect	0.25	0.34	0.48	0.50	0.59	0.53	0.43	0.46
Total	1.29	1.33	1.38	1.56	1.42	1.27	1.36	1.39

Conclusions

The trends in the variation of results in the separate sections are confirmed by this overall investigation.

(1) The tendency to change responses is greatest among the average girls and least among the poorest girls.

(2) For all pupils the tendency to make "Incorrect to Correct" changes declines from the best to the poorest.

(3) For boys the tendency to make "Correct to Incorrect" response changes increases from the best to the poorest. This pattern does not exist for girls.

(4) For boys, the tendency to make "Incorrect to Incorrect" response changes increases from the best to the poorest. This pattern does not exist for girls.

In general, all response changes made by boys follow a regular pattern from best - average - poorest. The pattern for girls is not so regular.

Relationship of Item Difficulty to Response Changing

For the analysis relating to item difficulty, only the data from the tests on Sections one, four, ten and twelve (1975/76) were used. Using Pearson's Product Moment Correlation formula, coefficients were calculated for the correlations between item difficulty and the number of pupils altering their response to that item. The results are shown in Table 3.18.

Table 3.18

Section Number	Correlation Coefficient
One	- 0.30
Four	+ 0.13
Ten	- 0.35
Twelve	- 0.15

Conclusion

Although there is a negative correlation of - 0.30 and - 0.35 for sections one and ten respectively, the result is not statistically significant at either the 1% or 5% levels of significance. What it does indicate is that the easier an item is the more likely it is that pupils would change their responses to that item, but only to a slight degree. However, taking all the correlations together, their variation is such that no firm conclusions can be drawn about the relation of item difficulty to response changing.

The Effect of Item Position of Response Changing

In order to investigate the effect of item position on response changing the number of response changes made in each one-fifth of the test was calculated, commencing with the first one-fifth of the answers and ending with the last one-fifth of the answers. The results are shown in Table 3.19. The data was taken from tests on Sections one, four, ten and twelve (1975/76).

Table 3.19 /

Table 3.19

Item Position and Response Changing

Item Position	No. of Items altered
First 20%	556
Second 20%	385
Third 20%	341
Fourth 20%	358
Fifth 20%	513

Conclusion

Fewer revisions were made upon items placed in the middle of the test than upon items appearing at the beginning and the end of the test. The larger proportion of revisions at the beginning and the end of the test may be due to various factors.

Those at the beginning may be due to what is known as 'examination nerves' or because at that point pupils may be distracted by supervisor's instructions and other pupils' questions.

Those alterations at the end may be due to increasing 'fatigue' or increasing carelessness. They are certainly not due to pressure of time as no time limit was placed on the examination.

It must be recognised, of course, that some changes on items placed near the beginning of the test may have actually been made near the close of the test period.

General Conclusion

From the results of this investigation it can be seen that pupils do in fact gain marks by altering their first choice in multiple choice Science examinations. This is in direct contrast to the long held belief that first choices are indeed the best and should not be /

be altered. What are the implications of these findings in terms of future multiple choice tests? Should a change be made in the instructions at the beginning of a multiple choice test paper which would encourage pupils to change responses when they think that their initial choice is incorrect? Or, should the situation be left as it is at present when the test instructions merely show pupils how to change a response when the need arises? This latter instruction does not in fact encourage pupils to alter responses.

I feel that since these findings were based on this latter type of instruction and pupils gained marks by making changes, then no alteration should be made to the test instructions, as by so doing other factors may emerge which could result in a different number of alterations which would produce different results.

In conclusion, then, pupils should only be instructed as to how to alter a response should the need arise and not be encouraged to make changes by a rewording of the instructions.

CHAPTER 4

THE EFFECT OF ITEM ORDER ON MULTIPLE CHOICE TEST PERFORMANCE

CHAPTER 4

THE EFFECT OF ITEM ORDER ON MULTIPLE CHOICE TEST PERFORMANCE

Introduction

An examination of the relevant literature revealed that conflicting results had been obtained when investigating the effect of item order on multiple choice test performance. MacNicol,[37] found that when items were ordered from difficult to easy, the mean number of correct answers on the test was significantly lower than the mean number of correct answers when the items were ordered from easy to difficult. These results were obtained for a test administered under essentially power conditions. Similar findings were made by Flaugher[38] et al, but their results were obtained when the test was speeded. In contrast several other researchers[42][43][44][45] have found no item order effect.

Consequently, the purpose of the present study is to -

1. Investigate the effect, under power testing conditions, of item order on test performance. It is expected that a difficult to easy order of items would give lower scores than would the reverse order.
2. Investigate the effect of item order on the relative performance of the top, middle and bottom thirds of the examination.

(1) Procedure

Two thirty-item multiple choice tests (four choices per item) were constructed and given to all first and second year pupils following the Integrated Science course. Two forms of each test were constructed, the first having the items arranged in the easy to difficult order (in terms of facility value obtained from /

from a pre-test on a similar sample of pupils), and the second arranged in the difficult to easy order. The tests covered the current session's work up to the date of the examination, i.e. sections one, two, three and four of the Integrated Science Syllabus for first year pupils, and sections ten, twelve and fourteen of the Integrated Science Sullabus for second year pupils. Each test was given to two matched groups of approximately equal numbers of pupils. The testwas without time limit, i.e. a power test. The groups were matched according to Verbal Reasoning Quotient (V.R.Q.).

Results

Table 4.1

Average marks obtained by pupils on both forms of the First year Test

	Easy to difficult form	Difficult to easy form
Boys (N=45)	17.83	18.06
Girls (N=50)	15.66	17.06
Total (N=95)	16.68	17.53

The t-test of significance was carried out on the results shown in Table 4.1 with the following results.

Table 4.2

Analysis of results by t-test

Pupils	Level of Significance
Boys	Not significant
Girls	10-20%
Total	20%

Observations

There was a slightly significant improvement in average performance in the difficult to easy form of the test for the total number of pupils and for the girls /

girls taken alone. This improvement is contrary to what might be expected.

Table 4.3

Average marks obtained by pupils on both forms of the second year test

	Easy to difficult form	Difficult to easy form
Boys (N=46)	16.57	17.20
Girls (N=59)	15.28	15.36
Total (N=105)	15.83	16.07

The t-test of significance was carried out on the results shown in Table 4.3 with the following results.

Table 4.4

Analysis of results by t-test

Pupils	Level of significance
Boys	Not significant
Girls	Not significant
Total	Not significant

Observations

As with the first year pupils there is a slight improvement in average performance in the difficult to easy form of the test. However, this time the improvement is not significant. Even though slight, this improvement is contrary to what might be expected, although regard should be taken of factors such as mental and physical fatigue near the end of the test.

Results (cont'd)

For each form of the test a plot was made of the item difficulty level versus the item position in the test. The results revealed that, except for a few misplacements, the majority of items on each form were in the desired order. (See graphs 5a, 5b, 5c and 5d)

The rank correlations between item position in the test and the position the item should have been in, based /

based on the item difficulty level as estimated from the data of this study are shown in Tables 4.5 and 4.6.

Table 4.5

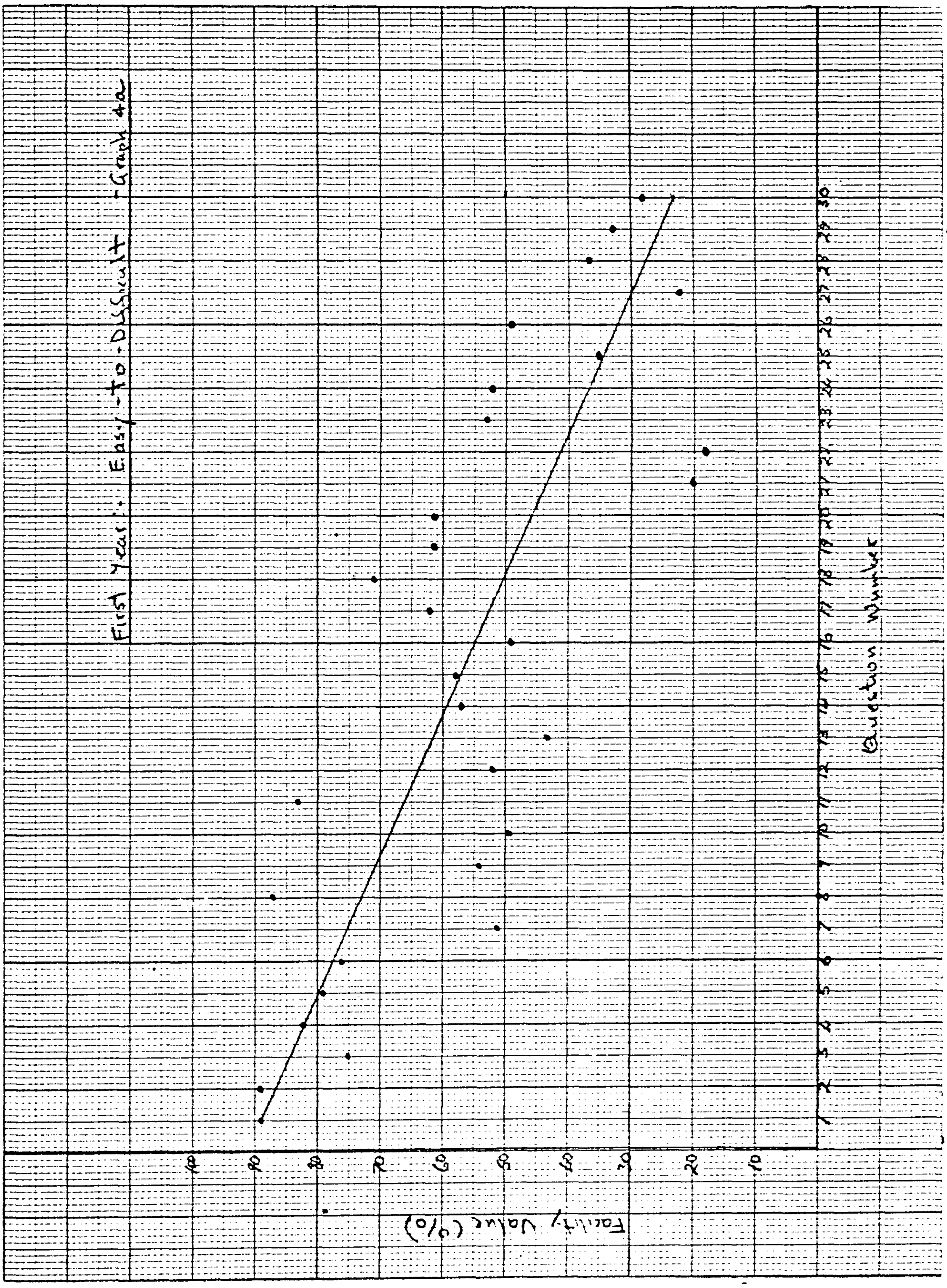
Rank correlations between item position and item difficulty level - first year exam

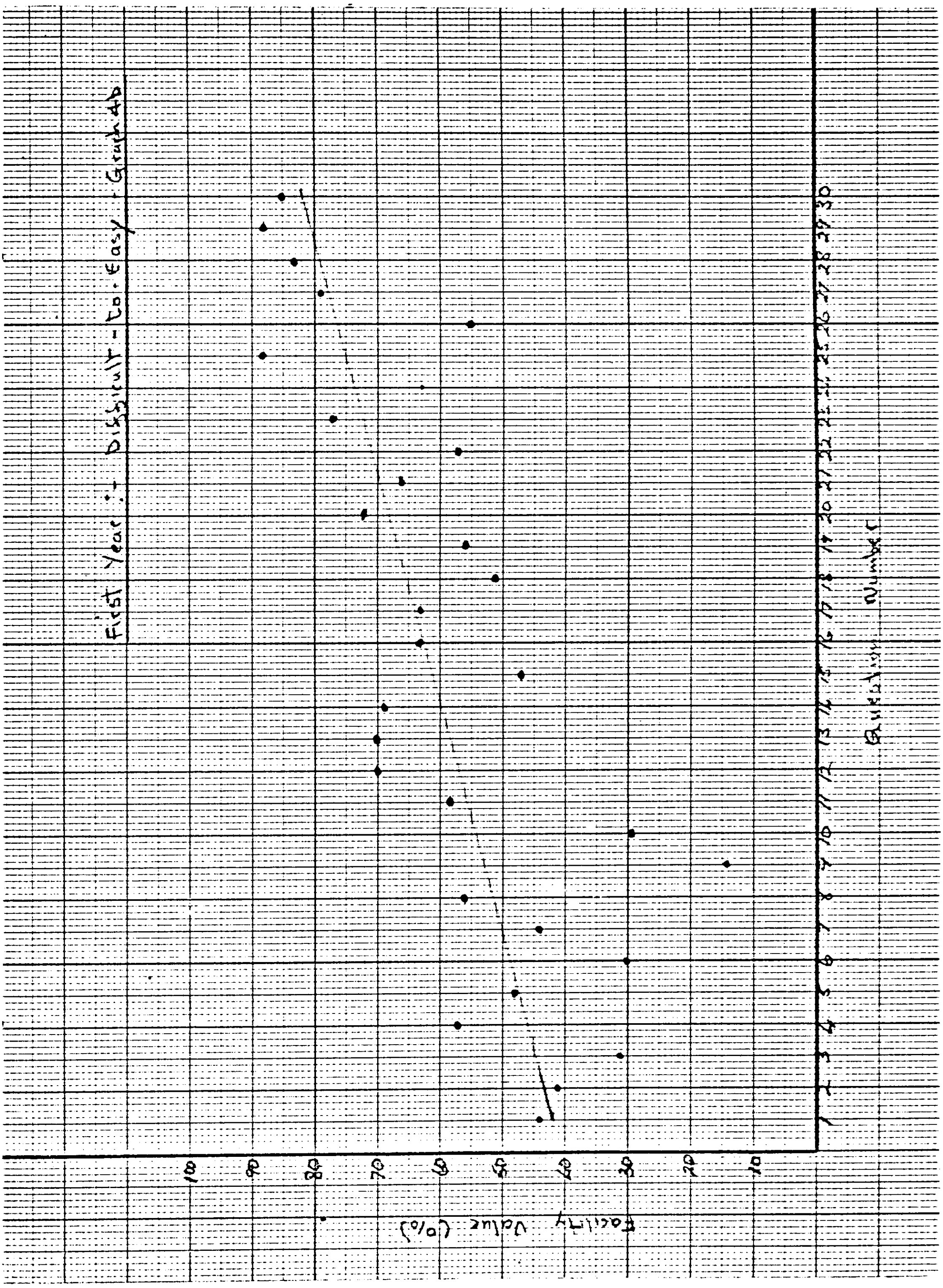
	Easy to difficult	Difficult to easy
Boys	0.75	0.64
Girls	0.76	0.79
Total	0.76	0.76

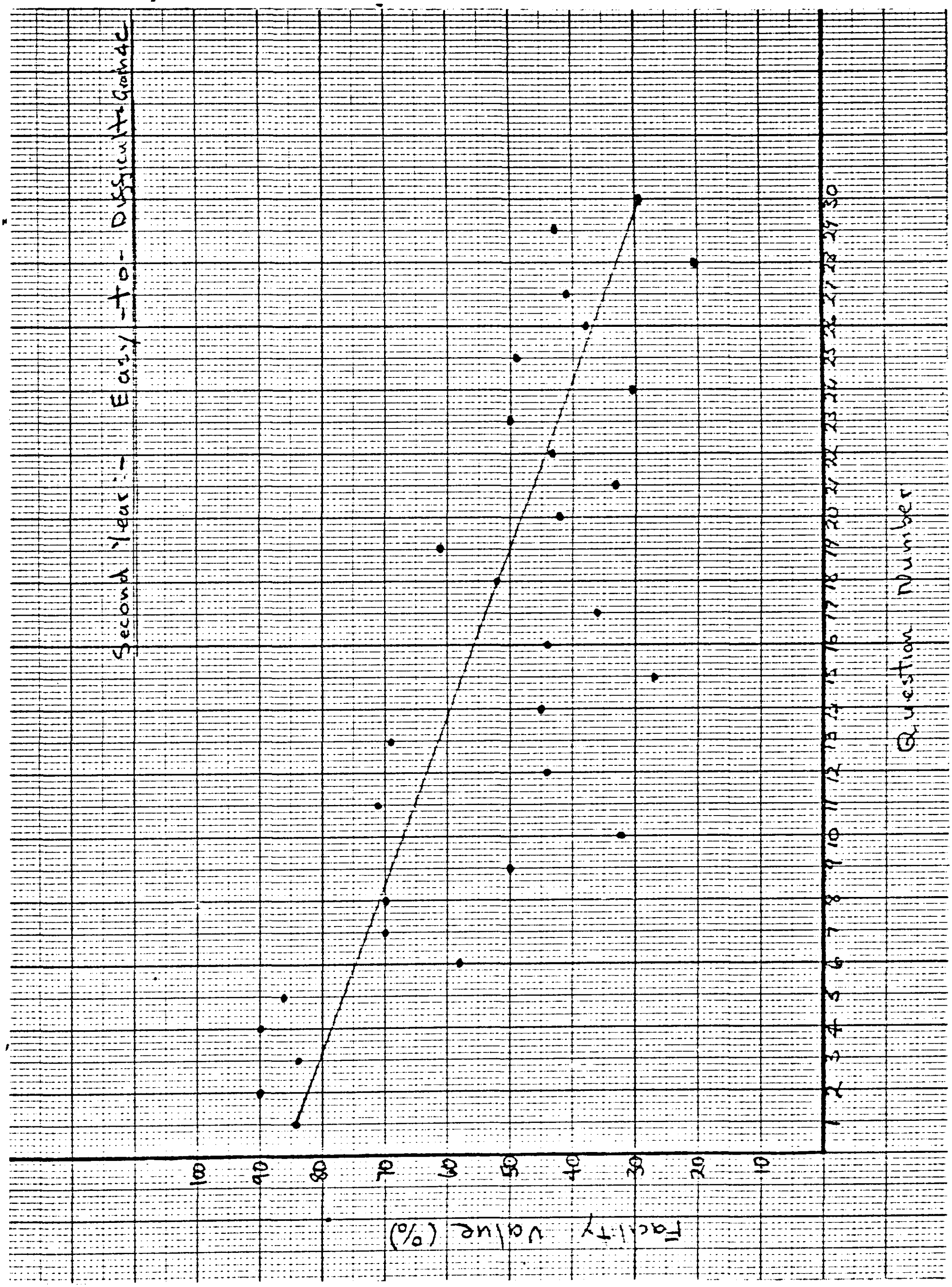
Table 4.6

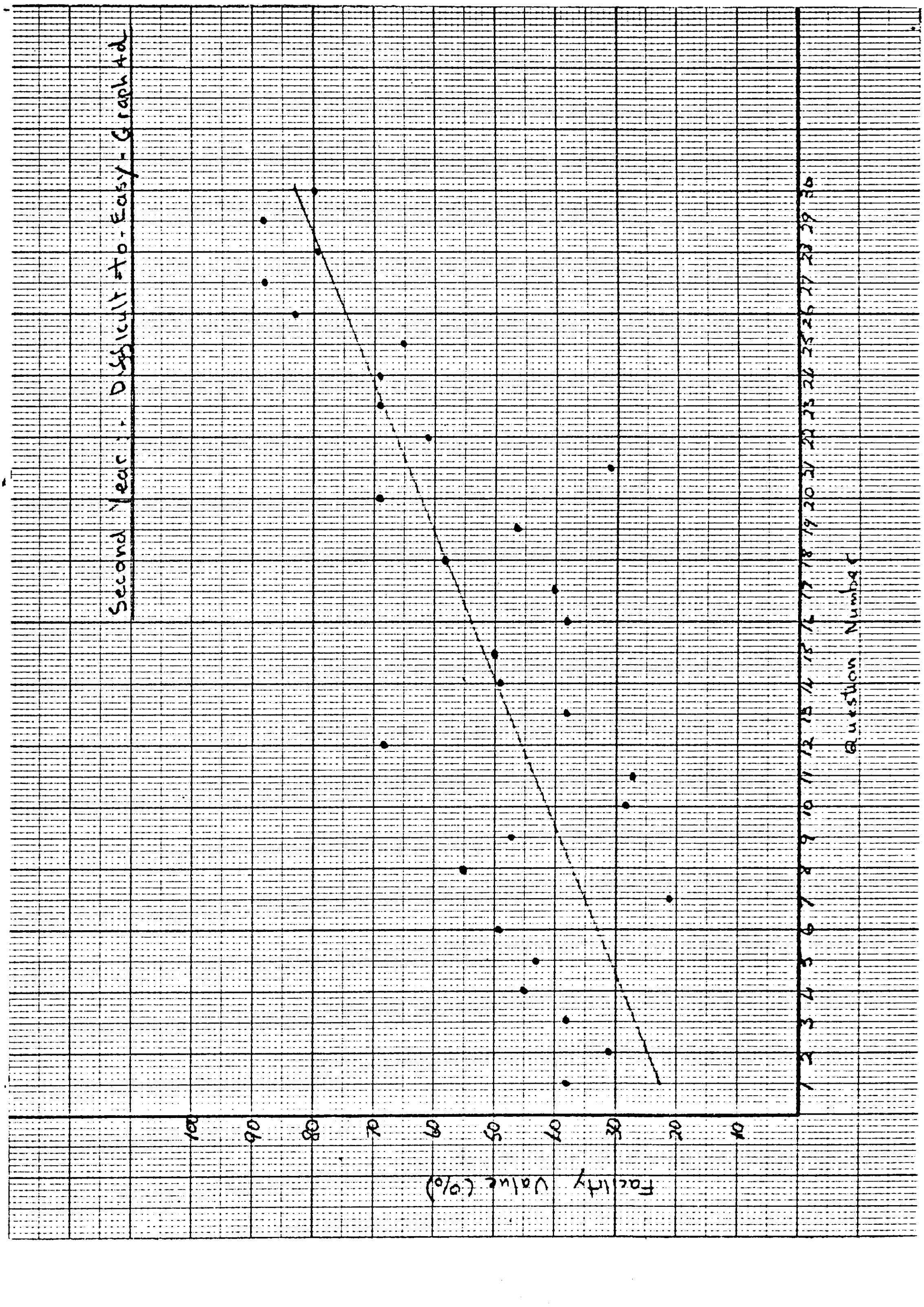
Rank correlations between item position and item difficulty level - second year exam

	Easy to difficult	Difficult to easy
Boys	0.72	0.74
Girls	0.74	0.72
Total	0.79	0.74


First Year: Easy-To-Difficult - Graph 4a
Question Number
Facility Value (%)
100
90
80
70
60
50
40
30
20
10


First Year :- Difficult - to - Easy - Graph 4b
Question Number
Facility Value (%)
100 90 80 70 60 50 40 30 20 10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30


Second Year :- Easy-to-Difficult Grade
Question Number
Facility Value (%)
100 90 80 70 60 50 40 30 20 10
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30


Second Year : Difficult-to-Easy-Graph 4d
Question Number
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20 21 22 23 24 25 26 27 28 29 30
Facility Value (%)
100 90 80 70 60 50 40 30 20 10

Observations

Whereas a lower rank order correlation might have been expected for the difficult to easy order this was not found to be the case, except for the boys in first year. In general the correlations between the different forms were fairly high and consistent between first and second year examinations. This would suggest that the difficult to easy order test was not proving to be more difficult than the easy to difficult order.

Conclusions

The conclusion that can be drawn on the basis of this investigation is that the scores on the difficult to easy order are on the average slightly, though not always significantly, higher than scores on the easy to difficult order. This is true for this test which is a power test. These results contradict the findings of MacNicol[37] who found that scores were significantly higher in the easy to difficult form of the test.

The importance of this study, and others, to test constructors seems to be the evidence it provides against the practice of making the order of presentation of items different for different examinees to reduce, for example, the chance of cheating, in the kind of situation which could arise in the first two years of a large comprehensive school when a large number of pupils are to be given the same science test over a period of time as part of their continuous assessment programme. It may also be of importance when considering the relative order of the items in a pre-test/post-test situation when the results of the pre-test are to be used to construct a test with particular statistical properties.

It would appear on the basis of currently available evidence that reordering of items on a test in effect produces /

produces a test with different properties than the original. Hence it may be impossible to make valid comparisons of the scores obtained by pupils who take the same test items in a different order. These conclusions apply only to tests without a time limit, it is for further research to investigate the effect of item order on test scores when the test is speeded. When the test is speeded other personality factors such as anxiety may come into play.

(2) Introduction

The results so far discussed refer only to the overall examination performance. In order to probe a bit deeper and by so doing explain why the average scores on the difficult to easy form of the test are higher than those on the easy to difficult form, it is appropriate to consider the effect of item order on the performance of pupils on questions in the most vulnerable positions in the test, namely, the first and the last ten questions. Any effect of item re-ordering should be at a maximum in these positions.

Procedure

The average mark scored by the pupils on the first and last ten questions in each test was computed as well as the average mark for the middle ten questions. The middle ten which should be relatively constant will act as a standard with which the performance of the extreme ten questions can be compared.

Results

Table 4.7 /

Table 4.7

Average marks scored per ten questions - first year boys

	Easy to difficult	Difficult to easy
First ten	7.76	4.40
Middle ten	6.49	6.26
Last ten	3.58	7.40
Total	17.83	18.06

Observations

(1) There was a small, but not statistically significant, increase in the average test score when the items were arranged in the difficult to easy order.

(2) The ten easier items did less well when they were arranged in the difficult to easy order. This difference denoted by the letter 'a' is - 0.36 marks, i.e. a = - 0.36.

(3) The average mark for the ten middle items also decreased when arranged in the difficult to easy order. This difference denoted by the letter 'b' is - 0.23, i.e. b = - 0.23.

(4) The average mark for the ten most difficult items increased when arranged in the difficult to easy order. This difference denoted by the letter 'c' is + 0.82, i.e. c = + 0.82.

Table 4.8

Average mark scored per ten questions - first year girls

	Easy to difficult	Difficult to easy
First ten	6.76	3.52
Middle ten	5.50	6.11
Last ten	3.40	7.43
Total	15.66	17.06

Observations

(1) There was a significant (10-20%) increase in the average test score when the items were arranged in the difficult to easy order.

(2) /

(2) The average mark for the ten easier items increased when arranged in the difficult to easy order, i.e. a = + 0.67.

(3) The average mark for the middle ten items also increased when arranged in the difficult to easy order, i.e. b = + 0.61.

(4) The average mark for the ten most difficult items increased when arranged in the difficult to easy order, i.e. c = + 0.12.

Table 4.9

Average mark scored per ten questions - first year total

	Easy to difficult	Difficult to easy
First ten	7.23	3.52
Middle ten	5.97	6.11
Last ten	3.48	7.43
Total	16.68	17.53

Observations

(1) There was a small significant increase (20% level) in the average test score when the items were arranged in the difficult to easy order.

(2) The average mark for the ten easier items increased when arranged in the difficult to easy order, i.e. a = + 0.20.

(3) The average mark for the middle ten items also increased when arranged in the difficult to easy order, i.e. b = + 0.14.

(4) The average mark for the ten most difficult items increased when arranged in the difficult to easy order, i.e. c = + 0.04.

Table 4.10

Average mark scored per ten questions - second year boys

	Easy to difficult	Difficult to easy
First ten	7.20	4.27
Middle ten	5.15	5.36
Last ten	4.22	7.57
Total	16.57	17.20

Observations

(1) There was a small, but not statistically significant, increase in the average test score when the items were arranged in the difficult to easy order.

(2) The average mark for the ten easier items increased when arranged in the difficult to easy order, i.e. $a = + 0.37$.

(3) The average mark for the middle ten items increased when arranged in the difficult to easy order, i.e. $b = + 0.21$.

(4) The average mark for the ten most difficult items increased when arranged in the difficult to easy order, i.e. $c = + 0.05$.

Table 4.11

Average mark scored per ten questions - second year girls

	Easy to difficult	Difficult to easy
First ten	7.12	3.77
Middle ten	4.75	4.73
Last ten	3.41	6.86
Total	15.28	15.36

Observations

(1) There was a small, but not statistically significant, increase in the average test score when the items were arranged in the difficult to easy order.

(2) The average mark for the ten easier items, decreased when arranged in the difficult to easy order, i.e. $a = - 0.26$.

(3) The average mark for the middle ten items decreased when arranged in the difficult to easy order, i.e. $b = - 0.02$.

(4) The average mark for the ten most difficult items increased when arranged in the difficult to easy order, i.e. $c = + 0.36$.

Table 4.12 /

Table 4.12

Average mark scored per ten questions - second year total

	Easy to difficult	Difficult to easy
First ten	7.15	3.96
Middle ten	4.92	4.98
Last ten	3.76	7.13
Total	15.83	16.07

Observations

(1) There was a small, but not statistically significant, increase in the average test score when the items were arranged in the difficult to easy order.

(2) The average mark for the ten easier items decreased when arranged in the difficult to easy order, i.e. a = - 0.02.

(3) The average mark for the middle ten items increased when arranged in the difficult to easy order, i.e. b = + 0.06.

(4) The average mark for the ten most difficult items increased when arranged in the difficult to easy order, i.e. c = + 0.20.

In order to illustrate the effect of item order on the most vulnerable questions, the following array was constructed showing the mathematical sign in front of the variation in the marks of the different item locations. 'a', 'b' and 'c' are as indicated previously.

Table 4.13

	First year			Second year		
	boys	girls	total	boys	girls	total
a	-	+	+	+	-	-
b	-	+	+	+	-	+
c	+	+	+	+	+	+

Observations /

Observations

(1) In every case the average score for the ten most difficult items has increased when the items were arranged in the difficult to easy order.

(2) The average score for the middle ten items has increased in the totals column each year, but within each total there has been a variation between boys and girls.

(3) The variation between scores has been greatest within the easiest ten items. There is no agreement between years.

In order to estimate the magnitude of the changes described Table 4.14 was constructed.

Table 4.14

Total changes per t en questions - all pupils

	First year			Second year			Grand
	boys	girls	total	boys	girls	total	total
First ten	-0.36	+0.67	+0.20	+0.37	-0.26	-0.02	+0.60
Middle ten	-0.23	+0.61	+0.14	+0.21	-0.02	+0.06	+0.77
Last ten	+0.82	+0.12	+0.04	+0.05	+0.36	+0.20	+1.59

Conclusion

It can be seen from the above figures (Table 4.14) that it is the average score for the ten most difficult items that increases by the largest amount when the items are arranged in the difficult to easy order.

This would be the main factor in the small observed increase in test scores when the items are arranged in such a manner. Why this should be the case is not clear, it could be that at the beginning of a test pupils tend to concentrate more and consequently achieve higher scores on these more difficult items than /

than they would if they were located at the end of the test. Towards the end of the test factors such as fatigue and loss of concentration may result, or there might be an increasing awareness among the pupils that the items are indeed becoming more difficult which would account for the lower value of the difficult items in the first instance when they are arranged in the easy to difficult order.

C H A P T E R 5

THE EFFECT OF THE RESPONSE SET IN MULTIPLE CHOICE TESTS

C H A P T E R 5

THE EFFECT OF THE RESPONSE SET IN MULTIPLE CHOICE TESTS

Introduction

As indicated in Chapter 1 several investigations of the position response set in multiple choice tests have yielded contradictory results. Whereas Cronbach(58) concluded that multiple choice tests are relatively free from position response sets, McNamara and Weitzman(59) report that a tendency does exist among examinees to favour certain response positions to the neglect of others.

The aim of the present study is:-

(1) To test the hypothesis that difficulty is a function of correct choice placement;
(2) To investigate the possible tendency among pupils to favour certain response positions to the neglect of others;
(3) To investigate the hypothesis that alteration of the position of the distractors in a multiple choice item while the key remains in the same position, has no significant effect on the degree of difficulty of a five choice item;
(4) To investigate the hypothesis that alteration of the position of the key from one of the first three positions to one of the last two positions will alter significantly the degree of difficulty of a five choice item;
(5) To investigate the hypothesis that the placing of the most plausible distractor immediately before the key will alter significantly the degree of difficulty of a five choice item.

Method

The tests used in all of these investigations were /

were those given to first and second year pupils following the Integrated Science Syllabus.

For the first two investigations the tests used were those for sections two/three and four given to all first year pupils and sections ten, twelve and fourteen given to all second year pupils during session 1974/75.

In the three remaining investigations the tests used were those on sections one, two/three given to all first year pupils in sessions 1974/75 and 1975/76, and those on sections twelve and fourteen given to all second year pupils in sessions 1974/75 and 1975/76. In addition, the final term examination given to all first year pupils in sessions 1973/74 and 1974/75 was used.

(1) Investigation of the hypothesis that difficulty is a function of correct choice placement.

Results

The total number, and hence the proportion, of responses to each of the five positions was calculated and compared with the expected proportion of responses to each position, this latter proportion being calculated from the distribution of correct responses among the five positions. For example, the expected values in Table 5.1 were calculated in the following way. The test consisted of thirty items, nine of which contained the correct response in position A, therefore the expected percentage of responses in that position would be $\frac{9}{30} \times 100 = 30.0\%$. The expected percentages for the other four positions were calculated in the same way. To see if the differences between these observed and expected values were significant the 'chi-squared' test was applied in each case. The results are shown in Tables 5.1 to 5.6 inclusive.

Table 5.1 /

Table 5.1

The significance between the observed and expected proportion of responses to each of the five choice positions in the section two/three test 1974/75.

Response Position	Observed Number of Responses	Observed Percentage	Expected Percentage	Significance
A	1695	28.2	30.0	Highly significant at 1% level
B	1128	18.7	23.3	
C	864	14.3	10.0	
D	1624	26.9	30.0	
E	719	11.9	6.7	
Total	6030	100.0	100.0	

Table 5.2

The significance between the observed and expected proportion of responses to each of the five choice positions in the section four test 1974/75.

Response Position	Observed Number of Responses	Observed Percentage	Expected Percentage	Significance
A	688	17.1	16.0	Highly significant at 1% level
B	1102	27.5	32.0	
C	541	13.5	20.0	
D	852	21.2	20.0	
E	830	20.7	12.0	
Total	4013	100.0	100.0	

Table 5.3

The significance between the observed and expected proportion of responses to each of the five choice positions in the section ten test 1974/75.

Response Position	Observed Number of Responses	Observed Percentage	Expected Percentage	Significance
A	861	17.2	8.0	Highly significant at 1% level
B	1087	21.7	20.0	
C	1114	22.3	24.0	
D	1189	23.8	36.0	
E	738	15.0	12.0	
Total	4989	100.0	100.0	

Table 5.4

The significance between the observed and expected proportion of responses to each of the five choice positions in the section twelve test 1974/75.

Response Position	Observed Number of Responses	Observed Percentage	Expected Percentage	Significance
A	863	20.0	24	Highly significant at 1% level
B	871	20.1	16	
C	890	20.6	20	
D	1013	23.4	28	
E	685	15.9	12	
Total	4322	100.0	100.0	

Table 5.5

The significance between the observed and expected proportion of responses to each of the five choice positions in the section fourteen test 1974/75.

Response Position	Observed Number of Responses	Observed Percentage	Expected Percentage	Significance
A	680	17.4	25.0	Highly significant at 1% level
B	754	19.3	20.0	
C	1010	25.8	25.0	
D	894	22.8	20.0	
E	577	14.7	10.0	
Total	3915	100.0	100.0	

Table 5.6

The significance between the observed and expected proportions of responses to each of the five choice positions taking all of the section two/three, four, ten, twelve and fourteen test together.

Response Position	Observed Number of Responses	Observed Percentage	Expected Percentage	Significance
A	4787	20.6	20.8	Highly significant at 1% level
B	4942	21.2	22.4	
C	4419	19.0	19.2	
D	5572	23.9	27.2	
E	3549	15.3	10.4	
Total	23269	100.0	100.0	

Observations

When the correction had been made for the size of the sample it was found that there was a highly significant difference between the observed and expected number of responses in each of the tests. Inspection of tables 5.1 to 5.6 show that in every case the proportion of pupils selecting response E was higher than would be expected. This was not true for the other response positions.

These results lend support to the hypothesis that difficulty is a function of correct choice placement. It should be noted, however, that in every test the proportion of correct responses placed at position E was smaller than in any of the other positions. It is for further work using a more balanced distribution of correct responses to test the hypothesis more fully.

(2) Investigation of the possible tendency among pupils to favour certain response positions

Results

In order to determine whether the difficulty of any one position was significantly greater than any other, the percentages of correct responses appearing in each position were calculated (Table 5.8). This index was calculated by dividing the total number of observed correct responses in each position (Table 5.7) by the total number of possible responses in each position (obtained by assuming 100% success in that item).

Table 5.7 /

Table 5.7

Total Number of correct responses in each position

Position	Number of correct responses	Number of Possible correct responses
A	2931	4913
B	3198	5217
C	2439	4499
D	3774	6475
E	1380	2421

Table 5.8

Percentages of correct responses appearing in each of the five positions

Position	Percentage of correct responses
A	59.7
B	61.3
C	54.2
D	58.3
E	57.0

Observations

(1) The placement of choices in the several possible positions in five choice items has some slight effect upon the difficulty level of the questions.

(2) The first two positions have a slightly smaller difficulty level than the last two positions.

(3) The middle position has the greatest level of difficulty. Why this should be the case is not clear, and indeed it contradicts the findings of McNamara and Weitzman who found that questions having the correct response in the fourth position were the most difficult and those with correct responses in the second and third positions were the easiest. In the present investigation the second /

the second position is certainly the easiest but the third position is the most difficult.

(3) Investigation of the effect of altering the distractor positions on the degree of difficulty, while keeping the key in the same position.

Method

The test used for this investigation was the term test given to all first year pupils in sessions 1973/74 and 1974/75. Alternate items on the 1973/74 test were altered by completely reversing the order of all the distractors while keeping the key in the same position. This resulted in twenty altered items and twenty unaltered items. The twenty unaltered items acting as a control with which any variation in the altered items could be compared. This altered form of the test was given as a term test to all first year pupils in session 1974/75. The facility value of each item was used to measure the degree of difficulty. The size of the sample was 197. The results are shown in Table 5.9.

Table 5.9

The effect on the level of difficulty of altering the distractor positions - term tests 1973/74 and 1974/75.

	Average Facility Value * (1973/74)	Average Facility Value * (1974/75)	Significance (t-test)
Unaltered Items	57.6	54.0	Significant at 5% level
Altered Items	58.1	54.6	Significant at 5% level

Observations

Both the altered and the unaltered items show an equally significant decrease in average facility value. Therefore, /

* See Appendix 12

Therefore, it appears that reversing the order of the distractors in this investigation has had no significant effect on the degree of difficulty of the altered items as compared with the unaltered items. In general, the pupils in the 1974/75 test have found both altered and unaltered items to be more difficult, than did the pupils in the 1973/74 test.

The reason for this significant decrease in average facility value for both altered and unaltered items is not clear. Possible contributory factors could be:-

(a) a difference in the quality of the pupils;
(b) a difference in teaching method;
(c) a change in teaching staff.

In this investigation the difference in the quality of the pupils from one session to another was probably not a factor as no significant difference was found between the Verbal Reasoning Quotients between the years. As far as possible pupils were taught by the same method, i.e. worksheets + additional notes, and with one exception the same teachers were involved in teaching the Integrated Science Course in each of the years in question. As it is not always possible to check whether a teacher presents a course the way he states he does, then if any of the above factors are involved in bringing about the decrease in average facility value, it is probably this factor which is responsible.

(4) <u>Investigation of the effect of changing the position of the key to one of the last two positions.</u>

<u>Method</u>

The tests used for this investigation were the tests for section two/three, given to all first year pupils, and section fourteen, given to all second year pupils, following the Integrated Science Syllabus in sessions 1974/75 and 1975/76. The section two/three test /

test contained thirty items, ten of which were altered by changing the position of the key from one of the first three positions to one of the last two positions. The section fourteen test contained twenty items, eight of which were altered in the same way. The twenty unaltered questions from section two/three and the twelve unaltered questions from section fourteen acting as controls with which any variation in the altered items could be compared. The items to be altered were chosen in a random fashion. The altered form of the section two/three test was given to all first year pupils in session 1975/76 and the altered form of the section fourteen test was given to all second year pupils in session 1975/76. The facility value of each item was used to measure the degree of difficulty. The sample sized varied slightly from year to year but were about 200 in each instance. The results are shown in Tables 5.10 and 5.11 for both boys and girls separately and together.

Table 5.10

The effect of changing the position of the key to one of the last two positions - section two/three test, 1974/75 and 1975/76.

	Average Facility Value (1974/75)	Average Facility Value (1975/76)	Significance (t-test)
Unaltered Items (Total)	62.2	65.2	Not significant
Altered Items (Total)	72.4	74.5	Not significant
Unaltered Items (Boys)	68.4	69.0	Not significant
Altered Items (Boys)	75.6	78.2	Not significant
Unaltered Items (Girls)	60.8	61.5	Not significant
Altered Items (Girls)	70.0	70.5	Not significant

Observations

Although there is a small increase in average facility value in each case none of the increases is significant. As a result there is no evidence for the hypothesis /

hypothesis that altering the position of the key from one of the first three positions to one of the last two positions will alter the degree of difficulty of the Test .

Table 5.11

The effect of changing the position of the key to one of the last two positions, section fourteen test, 1974/75 and 1975/76.

	Average Facility Value (1974/75)	Average Facility Value (1975/76)	Significance (t-test)
Unaltered Items (Total)	59.2	58.1	Not significant
Altered Items (Total)	62.4	60.3	Not significant
Unaltered Items (Boys)	59.3	59.7	Not significant
Altered Items (Boys)	61.0	60.4	Not significant
Unaltered Items (Girls)	59.0	56.9	Not significant
Altered Items (Girls)	63.8	60.5	Not significant

Observations

Although there is a small decrease in average facility value (except for the boys' unaltered items) none of the decreases is statistically significant. Again, as with the section two/three test there is no evidence for the hypothesis that altering the position of the key from one of the first three positions to one of the last two positions will alter the degree of difficulty of the Test .

(5) Investigation of the effect of placing the most plausible distractor immediately before the key.

Method

The tests used for this investigation were the tests for section one, given to all first year pupils, and section twelve, given to all second year pupils, following the Integrated Science Syllabus in sessions 1974/75 and 1975/76. The section one test consisted of /

of twenty items, ten of which were altered by placing the most plausible distractor (as determined from the 1974/75 test) immediately before the key. The section twelve test contained twenty-five items, eleven of which were altered in the same way. The ten unaltered questions from section one and the fourteen unaltered questions from section twelve acting as control questions with which any variation in the altered items could be compared. The items to be altered were chosen in a random fashion. The altered form of the section one test was given to all first year pupils in session 1975/76 and the altered form of the section twelve test was given to all second year pupils in session 1975/76. The facility value of each item was used to measure the degree of difficulty. The sample sizes varied slightly from year to year but were about 200 in each instance. The results are shown in Tables 5.12 and 5.13 for both boys and girls separately, and together.

Table 5.12

The effect of placing the most plausible distractor immediately before the key, section one test, 1974/75 and 1975/76.

	Average Facility Value (1974/75)	Average Facility Value (1975/76)	Significance (t-test)
Unaltered Items (Total)	69.4	72.1	Not significant
Altered Items (Total)	65.3	70.6	Sig. at 5% level
Unaltered Items (Boys)	71.8	72.2	Not significant
Altered Items (Boys)	67.0	73.3	Sig. at 1% level
Unaltered Items (Girls)	68.2	71.8	Not significant
Altered Items (Girls)	64.3	68.1	Sig. at 10% level

Observations

There is a significant increase in the average facility value for the altered items when all the pupils are /

are taken together. When the results for the boys are considered separately it can be seen that there is a highly significant increase in the average facility value for the altered items, for girls there is only a slightly significant increase. It would have been expected that when the most plausible distractor is placed immediately before the key it would attract responses away from the key. In fact, the opposite effect is observed, i.e. the key attracts more responses. A possible explanation is that the close proximity of the key and the distractor helps pupils to discriminate more effectively between them and therefore select the correct choice more often than when the key and the most plausible distractor are further apart. The evidence from this investigation appears to confirm the hypothesis that by placing the most plausible distractor immediately before the key, the degree of difficulty of the item will be altered significantly. The degree of difficulty, in this case, being decreased significantly.

Table 5.13

The effect of placing the most plausible distractor immediately before the key, section twelve test, 1974/75 and 1975/76.

	Average Facility Value (1974/75)	Average Facility Value (1975/76)	Significance (t-test)
Unaltered Items (Total)	39.7	39.5	Not significant
Altered Items (Total)	46.9	41.7	Sig. at 5% level
Unaltered Items (Boys)	38.2	40.4	Not significant
Altered Items (Boys)	44.3	43.7	Not significant
Unaltered Items (Girls)	41.7	38.5	Not significant
Altered Items (Girls)	50.1	40.4	Sig. at 1% level

Observations

In this case there has been a significant decrease /

decrease in the average facility value, mainly due to a highly significant decrease in the average facility value for the girls on the altered questions. In this case the results are as would be expected, i.e. the most plausible distractor draws responses away from the key when placed immediately before it. This is confirmed by an increase in the proportion of pupils selecting the most plausible distractor as their correct choice. The evidence from this investigation appears to confirm the hypothesis that by placing the most plausible distractor immediately before the key, the degree of difficulty of the item will be altered significantly, although in this case the degree of difficulty has been increased.

Conclusion

As a result of the investigations carried out in this section of the thesis it can be concluded that the location of the correct response can have a significant effect on the results obtained, and that there is a small, though not statistically significant, tendency among pupils to favour certain response positions. It is for further work to ascertain the magnitude of these effects.

It can also be concluded that alteration of the distractor positions within items has no effect on the difficulty level of that item when the position of the correct response is unchanged, it has also been found that when the position of the correct answer itself is changed to one of the last two positions in an item there is no significant effect on the difficulty level of that item.

However, it has been found that the location of the most plausible distractor can have a significant effect on the difficulty level of items when it is located immediately before the correct response. It would /

would appear that multiple choice test constructors should be aware of this fact especially when selecting multiple choice items from a pre-test - any alteration of the order of the items which involves movement of the most plausible distractor could have a significant effect on the subsequent performance of that item.

CHAPTER 6

A SCORING SYSTEM WHICH ALLOWS FOR PARTIAL KNOWLEDGE

CHAPTER 6

A SCORING SYSTEM WHICH ALLOWS FOR PARTIAL KNOWLEDGE

Introduction

It is clear that when an examinee fails to select the correct option in a multiple choice test, his rejection of the correct option and his choice of an incorrect option have a significance which should be recognised in the marking of the test. Failure to do this has been singled out as a weakness of the multiple choice item.

A more sensitive method of administering a five-option multiple choice test which enables the examinee to earn partial credit for partial knowledge about a question has been developed by Wiley.[20] This method, which Wiley calls the "Three-decision multiple choice test", requires the examinee to designate two options which he feels are definitely wrong as well as designating the option which he feels is the correct one. Three marks are given to a question if the option designated as the correct one is actually the correct option. Two marks are given for a question if the option which is actually correct is not designated as the correct option or one of the definitely wrong options. No marks are given to a question if an option designated as one of the definitely wrong choices is actually the correct option.

The purpose of the present investigation is to compare the results obtained from scoring a multiple choice test using four different scoring procedures.

(a) The conventional multiple choice scoring system, i.e. one mark for a correct choice of option, no marks for an incorrect choice.

(b) The system suggested by Wiley as described.

(c) /

(c) A modification of Wiley's system, devised by the author.

(d) A second modification of Wiley's system devised by the author.

Method

The pupils taking part in the investigation were one hundred and three first year pupils following the Integrated Science Syllabus who were about to take the test on section five of the syllabus. The test consisted of twenty multiple choice items each with five options.

The pupils were instructed to answer the questions on the conventional answer grid by using a vertical line to designate the correct option and to designate the two options which they considered to be most incorrect by a cross. The questions were then scored by four different methods to allow a comparison of the different scoring systems to be made.

(a) The questions were scored using the conventional scoring system, i.e. 1 mark for choosing the correct option, 0 marks for choosing an incorrect option.

(b) The questions were scored as a "Three-decision multiple choice test" with a marks allocation 3, 2, 0 as described in the introduction.

(c) The questions were scored as in 'b' but with a marks allocation of 2, 1, 0.

(d) The questions were scored as in 'b' but with a marks allocation of 2, 1, -1.

Results

The average mark scored by all the pupils under each of the four scoring systems is shown in Table 6.1. In each case the marks have been reduced to a mark out of a maximum of twenty marks in order that comparisons can /

can be made of the various scoring systems. Also shown in Table 6.1 are the percentage increases of the three "Three-decision multiple choice test" scoring systems over the conventional scoring system.

Table 6.1

The average marks obtained by pupils under four scoring systems and the percentage increase of the "Three-decision multiple choice test" scoring systems over the conventional scoring system.

Scoring System	Average Mark (Maximum 20)	Percentage increase over Conventional scoring system
(a) Conventional Scoring System	9.4	-
(b) Wiley's[20] Scoring System	13.5	43.8
(c) Modification 1 of Wiley's System	12.4	32.9
(d) Modification 2 of Wiley's System	10.2	8.5

Observations

(1) The conventional scoring system produced the lowest average mark.

(2) Modification 2 of Wiley's scoring system produced the smallest percentage increase over the conventional scoring system. In this system one mark is deducted for an item in which an option designated as one of the definitely wrong choices is actually the correct option.

(3) Wiley's suggested scoring system 'b' and the first modification of his system are basically the same in that no penalty is introduced for an item in which an option designated as one of the definitely wrong choices is actually the correct choice. The relative weightings have been altered merely to make the addition and consequent averaging /

averaging of individual item scores more managable. Therefore, scoring system 'c' will always produce a lower average mark than scoring system 'b'.

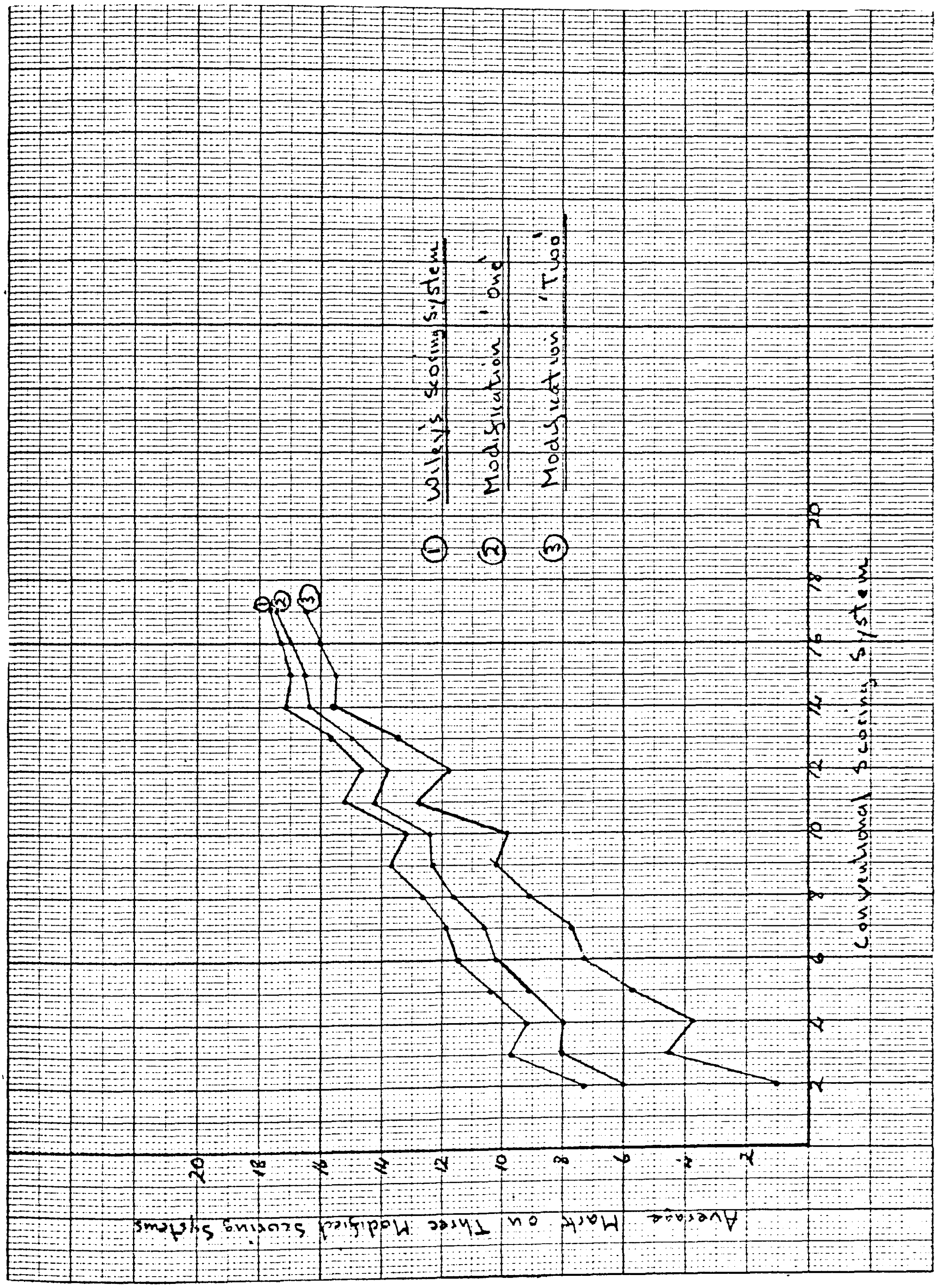
In order to compare the marks on each of the three modified scoring systems with the marks obtained on the conventional scoring system Table 6.2 was constructed which shows the average marks for all of the pupils on each of the three modified scoring systems who scored a particular mark on the conventional scoring system. A graph of conventional test scores against average mark scored on each of the modified scoring systems was also plotted for particular conventional test scores.

Table 6.2

The average mark scored on each of the three modified scoring systems by pupils scoring a particular mark on the conventional scoring system.

Conventional Score	Average Score (Wiley)	Average Score (Modification 1)	Average Score (Modification 2)
2	7.3	6.0	1.0
3	9.7	8.0	4.5
4	9.1	8.0	3.7
5	10.4	9.1	5.7
6	11.5	10.2	7.3
7	11.9	10.6	7.7
8	12.7	11.6	9.1
9	13.7	12.3	10.2
10	13.2	12.4	9.8
11	15.2	14.2	12.8
12	14.7	13.8	11.8
13	15.7	15.0	13.5
14	17.1	16.4	15.5
15	17.0	16.5	15.5
16	17.3	17	16.0
17	17.7	17.5	16.5

Average mark scored on each of the three modified scoring systems by pupils scoring a particular mark on the conventional scoring system.



Observations

By inspection of the results and the graph the following observations can be made:-

(a) The variation of average scores within the three modified scoring systems is greatest when the conventional score has a low value.

(b) The difference between the conventional score and the modified scores is also greatest when the conventional score has a low value.

(c) As the conventional score increases the variation within the modified scoring systems decreases. The difference between the conventional score and the modified scores also decreases.

(d) When the maximum conventional score has been reached the variation within the modified scoring systems, and the difference between them and the conventional score, has decreased to a minimum.

From a consideration of Table 6.2 and the accompanying graph it appears that there are three areas where a closer inspection of the pupils' marks could prove valuable. Namely, the marks scored by the best, average and least able pupils. In order to do this, the marks of the ten best pupils are shown in Table 6.3, the marks of the middle twenty-two pupils in Table 6.4, and the marks of the eleven worst pupils in Table 6.5.

Table 6.3 /

Table 6.3

The marks scored by the ten best pupils under all four marking systems. Integrated Science Syllabus Section five test, 1976.

Conventional Marking system	Wiley's Marking system	Modification 1	Modificaton 2
17.0	17.0	17.3	15.5
17.0	18.3	18.1	17.5
16.0	17.3	17.3	16.0
15.0	16.3	16.8	14.5
15.0	15.7	15.5	13.5
15.0	17.0	16.5	15.5
15.0	17.7	17.3	16.5
15.0	18.3	17.5	17.5
15.0	17.7	17.3	16.5
15.0	16.3	16.8	14.5
Average 15.5	17.2	16.8	15.8

Table 6.4 /

Table 6.4

The marks scored by the middle twenty-two pupils under all four marking systems. Integrated Science Syllabus, section five test, 1976.

Conventional Marking System	Wiley's Marking System	Modification 1	Modification 2
9.0	15.7	14.0	13.5
9.0	13.7	12.5	10.5
9.0	14.3	13.0	11.5
9.0	13.7	12.5	10.5
9.0	17.7	13.0	11.5
9.0	13.0	12.0	9.5
9.0	11.7	11.0	7.5
9.0	13.0	12.0	9.5
9.0	13.7	12.5	10.5
9.0	11.0	10.5	6.5
9.0	13.7	12.5	10.5
10.0	14.7	13.5	12.0
10.0	12.7	12.0	9.0
10.0	14.7	13.5	12.0
10.0	11.3	11.0	7.0
10.0	14.7	13.5	12.0
10.0	12.7	12.0	9.0
10.0	13.3	12.5	10.0
10.0	12.7	12.0	9.0
10.0	12.0	11.5	8.0
10.0	12.7	12.0	9.0
10.0	14.0	13.0	11.0
Average			
9.5	13.5	12.4	10.0

Table 6.5 /

Table 6.5

The marks scored by the eleven lowest pupils under all four marking systems. Integrated Science Syllabus, section five test, 1976.

Conventional Marking System	Wiley's Marking System	Modification 1	Modification 2
2.0	7.3	6.0	1.0
3.0	8.3	7.0	2.5
3.0	8.3	7.0	2.5
3.0	12.3	10.0	8.5
4.0	8.0	7.0	2.0
4.0	10.7	8.5	5.0
4.0	9.3	8.0	4.0
4.0	9.3	8.0	4.0
4.0	8.0	7.0	1.0
4.0	8.0	7.0	2.0
4.0	12.0	10.0	8.0
Average			
3.5	9.2	7.8	3.7

Table 6.6

Average marks scored by the best, middle and lowest pupils under all four marking systems. Integrated Science Syllabus, section five test, 1976.

Type of Pupil	Conventional Marking System	Wiley's Marking System	Modification 1	Modification 2
Best ten	15.5	17.2	16.8	15.8
Middle twenty-two	9.5	13.5	12.4	10.0
Bottom eleven	3.5	9.2	7.8	3.7

Observations

(1) For the best ten pupils there is a small increase in average marks between the conventional and modified scoring systems.

(2) /

(2) For the middle twenty-two pupils the increases in the average marks are larger between the conventional and modified scoring systems, the increase being the smallest in modification 2.

(3) For the bottom eleven pupils there is a large increase in average marks between the conventional and modified systems, with the exception of modification 2.

In order to find out if the three different scoring systems produced a different rank order among the pupils, the rank of each pupil in each of the scoring systems was established and is shown in Table 6.7. The Spearman Rank Order Correlation Coefficients between the conventional scoring system and each of the three modified systems were calculated and are shown in Table 6.8.

Table 6.7

The rank order of each pupil in the conventional, and three modified scoring systems.

Pupil Number	Conventional Score Rank	Rank (Wiley)	Rank (Modification 1)	Rank (Modification 2)
1	25	19	22	18
2	17	19	17	18
3	4	15	10	14
4	52	19	28	18
5	63	74	70	74
6	25	41	35	40
7	63	53	54	53
8	41	35	35	33
9	11	3	3	6
10	93	74	84	74
11	41	61	54	61
12	52	48	49	48
13	11	11	10	10
14	20	29	35	33
15	63	61	64	61
16 /				

Table 6.7 (cont'd)

Pupil Number	Conventional Score Rank	Rank (Wiley)	Rank (Modification 1)	Rank (Modification 2)
16	25	19	22	18
17	25	19	22	18
18	11	3	3	3
19	11	11	10	10
20	3	8	3	6
21	4	19	16	18
22	25	31	28	29
23	41	35	35	33
24	20	35	28	33
25	20	29	22	28
26	4	9	9	8
27	4	5	3	4
28	25	41	35	40
29	11	11	10	10
30	25	31	28	29
31	20	53	44	53
32	4	1	2	1
33	52	41	44	40
34	41	82	70	82
35	1	9	3	8
36	41	35	35	33
37	25	41	35	40
38	11	11	10	10
39	52	48	49	48
40	63	53	54	53
41	63	53	54	53
42	25	19	22	18
43	41	61	54	61
44	52	5	44	40
45	17	19	17	18
46	25	15	17	14
47	4	5	3	4
48	1	1	1	1
49	25	19	22	18
50	25	31	28	29
51 /				

Table 6.7 (cont'd)

Pupil Number	Conventional Score Rank	Rank (Wiley)	Rank (Modification 1)	Rank (Modification 2)
51	83	74	76	74
52	74	78	76	78
53	93	100	98	100
54	52	58	54	58
55	74	70	70	70
56	63	82	76	82
57	63	61	64	61
58	63	82	76	82
59	52	78	70	78
60	25	31	28	29
61	25	15	17	14
62	4	15	10	14
63	63	61	64	61
64	83	91	89	90
65	52	58	54	58
66	93	88	92	90
67	74	78	76	78
68	87	92	92	92
69	52	48	49	48
70	100	98	98	98
71	52	86	76	86
72	63	35	44	33
73	93	96	96	96
74	74	92	89	92
75	52	48	49	48
76	87	92	92	92
77	74	58	64	58
78	74	48	54	48
79	100	98	98	98
80	93	96	96	96
81	17	19	17	18
82	87	90	89	88
83	74	70	70	70
84	41	53	49	53
85	87	70	76	70
86	74	78	76	78
87	/			

Table 6.7 (cont'd)

Pupil Number	Conventional Score Rank	Rank (Wiley)	Rank (Modification 1)	Rank (Modification 2)
87	41	61	54	61
88	41	74	64	74
89	25	41	35	40
90	87	86	87	86
91	74	90	87	88
92	103	103	103	102
93	93	100	98	102
94	93	100	98	100
95	25	41	35	40
96	41	61	54	61
97	83	61	70	61
98	87	92	92	92
99	63	61	64	61
100	83	82	84	82
101	41	47	44	47
102	20	35	28	33
103	100	70	84	70

Table 6.8

Spearman Rank Order Correlation Coefficients between the conventional Multiple choice scoring system and each of the other three systems.

Scoring System	Correlation Coefficient
'Wiley'	0.89
Modification 1	0.95
Modification 2	0.90

Observations

The rank correlations between each of the scoring systems and the conventional scoring system are very high. When the number in the sample is large (N greater than 25) as in this case, the value of the correlation coefficient tends to be unreliable, but, however, it gives /

gives a good indication of the correlation that does exist.

Conclusion

If it is accepted that Wiley's marking system and the first modification of it are essentially the same system (modification 1 is merely a simplified version to facilitate calculations), then the question to be answered is - are either the first or second modifications, as scoring systems which allow for partial knowledge, an improvement on the conventional scoring system which makes no allowance for partial knowledge. The only difference between these two systems is that, whereas the second modification includes a penalty mark, the first modification does not.

In general, the pupils who would be expected to gain most from any marking system which allows for partial knowledge would be those in the middle to bottom range of any distribution of multiple choice test marks, as it is these pupils who have the fewest number of correct responses and consequently a greater number of items which could gain marks if a correction is applied for partial knowledge. On the other hand, since these pupils are not as able as their counterparts at the top end of the distribution there is no guarantee that they will gain significantly from scoring systems which allow for partial knowledge.

The results in Tables 6.3 to 6.6. show that under the first modification of the scoring system the best pupils have gained, on average, 1.3 marks, the middle group of pupils have gained 2.9 marks, and the lowest group of pupils have gained 4.3 marks. This shows that the middle to bottom range of pupils stand to gain most from an application of the first modification of Wiley's scoring system.

Under this first modification, then, pupils can only /

only gain marks or at worst remain on the same mark, since there are no penalty marks. Under the second modification of Wiley's scoring system pupils will have marks deducted from their total mark if an option they designate as one of the definitely wrong choices is actually the correct option. However, their overall mark will only decrease if the number of times this error occurs exceeds the number of instances in which the correct answer is one of the options they have not designated as being definitely wrong or absolutely correct.

From the results, it can be seen that no particular group of pupils made any great gains from this method, in fact, the best pupils only gained 0.3 marks and the lowest pupils gained 0.2 marks, so it would appear that the second modification is of no real advantage to pupils as a measure of partial knowledge, especially to those in the middle to bottom marks range who would stand to gain most from such a marking system. From this point of view, I would say that the first modification of Wiley's scoring system is more desirable as a measure of partial knowledge as any such measure should add to the marks of pupils and not, as happened in some cases, reduce them.

Non-objective Science examinations contain questions that give credit for partial knowledge, whether they be of the essay type with a previously agreed marking scheme or a more structured type of question where the total mark gained is the sum of the marks gained in the various parts of the question which may or may not be related to each other. Therefore, the pupil who has some knowledge of the topic being examined is given credit for some knowledge of that topic.

In multiple choice questions the pupil is given no credit for partial knowledge. The scoring systems described in this investigation give some credit for partial knowledge. Although they only offer a possible fraction /

fraction of a mark per item, the sum of these fractions in a test consisting of a large number of items could have a significant effect on the final mark achieved by the pupil, and give a more accurate representation of that pupil's knowledge at the time of the test than would under the conventional multiple choice marking system.

It has been the author's experience that relatively small correlations exist between pupils' performances on written tests and their performance on multiple choice tests. This could be due to the fact that while pupils are given credit for partial knowledge in written examinations, no such credit is given in multiple choice tests. If this is the case then the correlations between performance on multiple choice tests which give credit for partial knowledge and written tests should be higher, than when no credit is given for partial knowledge on multiple choice tests. It is for further research to investigate this hypothesis.

Although the rank order correlations between the various scoring systems and the conventional multiple choice scoring system are high and could indicate merely a scaling of the marks, it is important that a distinction is made between rank order and raw scores.

The pupils' rank order when obtained using a scoring system which gives credit for partial knowledge may not be altered significantly, but the overall mark obtained is a more accurate representation of his performance in that subject. It must be decided at the outset if the purpose of giving the test is to rank the pupil within his class or peer group or whether the object of the test is to measure accurately his ability in that subject.

Of all the scoring systems which allow for partial knowledge, a system based on the "Three-decision multiple choice" test is the only system which could be adopted as suitable for use with secondary pupils of a wide range /

range of ability. It uses conventional multiple choice questions, the instructions are easily understood by all pupils and it is relatively easy to mark. This type of test takes slightly longer than the conventional multiple choice test to administer although at this time it is not known how the pupils tackled the questions. They may have worked through the test item by item or they may have treated it as a conventional multiple choice test and returned to the beginning to insert crosses against the responses they thought were definitely wrong.

It is for further work to investigate this aspect of the test. It would also prove useful to compare the "three-decision multiple choice test" and the conventional multiple choice test in terms of their relationship to a criterion of academic achievement such as the pupil's science mark for the previous year. The marks for the "Three-decision multiple choice test" should correlate more highly with academic achievement than the marks from the same test scored as a conventional multiple choice test if the "Three-decision multiple choice" test is actually an improvement over the conventional multiple choice test.

C O N C L U S I O N S
A N D R E C O M M E N D A T I O N S

CONCLUSIONS
AND RECOMMENDATIONS

In this thesis several aspects of the use of multiple choice tests in schools have been investigated. The main findings are now summarised:-

(1) An investigation into the effect of changing initial responses on multiple choice tests has shown that, contrary to the popularly held belief, it has proved advantageous for pupils to change their initial responses to multiple choice items if they judge that their initial response was incorrect. This has implications for test constructors, who may wish to examine this finding further with a view to altering the instructions given to pupils in multiple choice tests.

(2) It has been found that the alteration of the response position within an item has no significant effect on the performance of that item except when this has the effect of altering the position of the most plausible distractor. This effect is greatest when the most plausible distractor is moved to a position immediately before the correct answer.

(3) A small item order effect has been observed, but in the opposite direction to that observed by some other investigators in that test performance is improved in the difficult to easy arrangement of the items in the test.

There are two aspects of the use of multiple choice tests where there should be further research:-

(1) To establish what is the optimum number of choices which should be used in multiple choice items. At the present time the number of choices used in multiple choice items is either four or five, with four choices gradually becoming the more popular, because when using five /

five choices it is difficult to obtain four distractors which will all function, therefore, effectively reducing the item to a four choice item. This, in effect, means that some of the time and money spent in writing, selecting and pre-testing the items has been wasted.

In view of the need for economy of time and money it may prove fruitful to investigate the possible use of three choice items which have been suggested by some investigators. In this type of item it would be relatively easy to find two distractors which would function effectively and a test containing this type of item would take less time to construct and administer. Research has shown that the test statistics are not altered significantly by the use of three choice items. For example, Ebel(53) has stated that the expected reliability for a 100 item multiple choice test containing items with three choices is 0.83 compared with 0.86 for a test containing items with four choices.

(2) The construction of a scoring system which allows for partial knowledge. A preliminary investigation has already been carried out, in Chapter 6 of this thesis, into the use of the "three-decision multiple choice" test using items with five choices as a possible method of assessing partial knowledge. It is possible that questions with five choices may not be the best form of the "three-decision multiple choice" test because with five choices most pupils can find at least two choices which seem definitely wrong. The "three-decision multiple choice" test with items having four choices would appear to be a better form because the pupil has to make a judgement about each of the choices, because the one choice he does not mark has to be the one he thinks is most likely to be correct if his choice of the correct alternative should happen to be wrong. In addition, by using items with only four choices the average chance score will be lowered. It is for further research to find out if the four choice /

choice form of the "three-decision multiple choice" test is the better form.

A second method of scoring which allows for partial knowledge is what is known as the confidence-scored multiple choice test in which the examinees are asked to express their distribution of preference for options as well as their certainty in that distribution. The Method of Intrinsic Items, devised by Rippey[24] and described in Appendix 1 is one such method which is worthy of further investigation. Intrinsic items require a distribution of belief over the options in a multiple choice test and do not have unique correct responses and they reflect realistically situations which require choices among a finite number of choices, none of which is uniquely correct.

It is essential that the nature of the objectives to be assessed and the ability of the examinees be taken into account when using scoring procedures which allow for partial knowledge. Whereas the Method of Intrinsic Items could be used with university students, the instructions would be far too complex for school pupils to comprehend, especially in the first two years of the secondary course where there is a wide range of ability among the pupils following the common course. It has been shown in Chapter 6 that the "three-decision multiple choice" test can be used successfully with pupils at this level.

A P P E N D I C E S

APPENDIX 2

TEST FOR INTEGRATED SCIENCE SYLLABUS - SECTION ONE

1. The pulse rate of a boy before exercise was 60 per minute. Which is likely to be his pulse rate after exercise?

   (a) 40 per minute
   (b) 60 per minute
   (c) 200 per minute
   (d) 0 per minute
   (e) 90 per minute

2. To find the weight of water in a carton you must use:-

   (a) the balance once
   (b) the balance twice
   (c) the balance three times
   (d) a measuring cylinder once
   (e) a measuring cylinder twice

3. Fifty cubic centimetres of water was boiled using a yellow flame, then the same volume heated using a blue flame; which one of the following statements is true?

   (a) The yellow flame takes less time.
   (b) The yellow flame is hotter.
   (c) The blue flame takes less time.
   (d) The blue flame takes longer.
   (e) The two flames take the same time.

4. You have a jar of sand and water and you wish to separate them - you would:-

   (a) Shake the mixture.
   (b) Heat the mixture.
   (c) Freeze the mixture.
   (d) Pour the mixture through a filter.
   (e) Pour the mixture through a tube.

5. The weight of an object is measured in -

   (a) /

(a) degrees
(b) grammes
(c) centimetres
(d) millimetres
(e) millilitres

6. When we want to take the temperature of something we use -

(a) a manometer
(b) a barometer
(c) a thermometer
(d) a thermopile
(e) a centigrade

7. A girl was asked to find out what was in a sealed box. She found that when she held it one way, she heard things rolling inside the box. She held it another way and she heard sliding. She shook the box and heard a metallic sound. What do you think was in the box?

(a) Steel balls
(b) Wooden balls
(c) Wooden cylinders
(d) Steel cylinders
(e) Steel cubes

8. A boy was asked to find out as much as possible about what was inside a sealed box. He picked it up, shook it and said, "It is a bath sponge." He was correct. What is your scientific comment on this?

(a) He was a very observant boy.
(b) He was a boy who always guesses.
(c) The box did not rattle so it must have been a sponge.
(d) He had not enough evidence to make that reply.
(e) The box was just the right size for a sponge.

9. /

9. What is the main reason for using paper for filtering?

(a) It is easily cut to shape.
(b) It is full of very small holes.
(c) It tears easily when wet.
(d) It can easily be folded.
(e) It is easy to write on.

10. Using the apparatus available in your science laboratory, what would you do to find the weight of potato crisps in a packet of crisps?

(a) Weigh a full packet of crisps.
(b) Read the weight printed on the packet.
(c) Put the crisps in a measuring cylinder.
(d) Weigh a packet full and weigh a packet empty.
(e) Weigh yourself before and after eating the crisps.

11. An empty beaker weighs 20 g. When it contains 50 cm^3 of a liquid it weighs 60 g. What is the weight of the liquid?

(a) 20 g
(b) 30 g
(c) 40 g
(d) 50 g
(e) 60 g

12. If water in a beaker is heated until it boils, its temperature will be -

(a) 0^oC
(b) 70^oC
(c) 100^oC
(d) 105^oC
(e) 39^oC

13. A thermometer is used for measuring -

(a) heat
(b) weather
(c) therms
(d) temperature
(e) expansion

14. Which of the following would you use to measure the volume of a stone?

(a) A balance and a beaker of water.
(b) A measuring cylinder.
(c) A measuring cylinder and water.
(d) A balance.
(e) A ruler.

15. If someone reported that he had made a new discovery, which upset a well-known scientific theory, you would -

(a) ignore his report.
(b) find if he came from a scientific family.
(c) find out if his experiments can be repeated.
(d) give up the theory.
(e) change the textbooks.

16. A detective was investigating a crime. He walked into a room where the crime had just been committed. He carried a bottle full to the brim with water, poured out the water and then screwed the stopper firmly on and walked out. What was he doing?

(a) He was making the dust settle.
(b) He was making a plaster cast of a footprint.
(c) He was purifying the air.
(d) He was collecting a sample of the air in the room.
(e) He was measuring the rate of flow of water.

17. A stethescope is used for -

(a) looking at steths.
(b) taking your pulse.
(c) looking at tiny creatures.
(d) measuring your breathing rate.
(e) measuring your heartbeat.

Questions 18 and 19 refer to the following:-
John and James dived into the swimming pool. John had a cold /

a cold shower before entering the water. James had a hot shower before entering the water. They could say:-

(a) "The water is cold."
(b) "The water is normal."
(c) "The water is chlorinated."
(d) "The water is warm."
(e) "I wish I had not taken a shower."

18. Which is John likely to say?

19. Which is James likely to say?

20. They both have different answers yet they were both in the same pool. What does this indicate?

(a) That the water in the pool was not well mixed.
(b) That one of them was thick skinned.
(c) That one of them was cold-blooded.
(d) That they should not have had a shower before entering.
(e) That human senses are not very reliable.

APPENDIX 3

TEST FOR INTEGRATED SCIENCE SYLLABUS - SECTION TWO/THREE

1. What is the name given to the swollen part of a worm?

   (a) Saddle
   (b) Segment
   (c) Sheath
   (d) Stirrup
   (e) None of these

2. We classify an animal as an amphibian if -

   (a) it breathes through gills all its life.
   (b) its body is covered with scales.
   (c) it spends early part of its life entirely in water but leaves the water when fully grown.
   (d) it can either live in water or on land.
   (e) it lays eggs.

3. Vertebrates all possess backbones - Man is a vertebrate. Which of the pairs of animals below belong to the Vertebrate group?

   (a) Earthworm and jellyfish
   (b) Sponge and snail
   (c) Crab and spider
   (d) Cat and elephant
   (e) Starfish and leech

4. Which of the following lists is made up <u>entirely</u> of substances that have come from once living material?

   (a) Concrete, brick, lead
   (b) Glass, cement, silver
   (c) Oil, wood, paper
   (d) Rock, clay, snow
   (e) Water, gold, iron

5. /

5. Which of the following animals is invertebrate?

(a) Hamster
(b) Kangaroo
(c) Snail
(d) Snake
(e) Shark

6. Which of the following is a set of vertebrates only?

(a) Horse, worm, salmon, frog
(b) Mouse, rabbit, sheep, tadpole
(c) Lizard, snake, caterpillar, snail
(d) Man, ape, starfish, jellyfish
(e) Frog, hamster, snake, jellyfish

Questions 7 and 8 refer to the following table:-

Set 1	Mouse	Herring	Apple	Moss	Coral
Set 2	Balance	Beaker	Nail	Water	Ink

Select from the following list the correct name for each set

(a) Moving
(b) Living
(c) Dead
(d) Stationary
(e) Non-living

Question 7. All the members of set 1 are?

Question 8. All the members of set 2 are?

Use the following types of energy for questions 9 to 13:-

(a) Kinetic or Movement energy
(b) Light energy
(c) Electrical energy
(d) Potential energy
(e) Chemical energy

Choose /

Choose the energy <u>produced</u> when,

9. A bulb is connected to a cell.
10. A boy lifts a weight from the ground.
11. Electricity is supplied to an electric motor.
12. Coal is burned.
13. An object is dropped from a height.

The following refer to questions 14 to 20:-

(a) Mammals
(b) Reptiles
(c) Amphibians
(d) Fish
(e) Birds

14. A toad belongs to the set of -
15. A python belongs to the set of -
16. You belong to the set of -
17. A shark belongs to the set of -
18. A whale belongs to the set of -
19. A crocodile belongs to the set of -
20. A duck belongs to the set of -
21. When a piece of magnesium is heated, the energy change is:-

(a) Heat energy + chemical energy → light energy
(b) Heat energy + kinetic energy → sound energy
(c) Heat energy + chemical energy → sound energy
(d) Heat energy + potential energy → light energy
(e) Heat energy + light energy → chemical energy

22. Which one of the following is not a form of energy?

(a) Sound
(b) Light
(c) Electricity
(d) Gas
(e) Movement

23. /

23. All energy on the earth comes from -

(a) the sun.
(b) light.
(c) movement.
(d) heat.
(e) air.

24. Living things mainly get energy directly from -

(a) the sun.
(b) water.
(c) heat.
(d) food.
(e) light.

25. Which device, or combination of devices would you use to change light energy into movement energy?

(a) Electric motor
(b) Photo-cell
(c) Battery
(d) Photocell, electric motor
(e) Battery, electric motor

26. Which device, or combination of devices, would be the simplest way of changing chemical energy into light energy?

(a) Battery, bulb
(b) Photocell, bulb
(c) Dynamo, bulb
(d) Steam engine, bulb
(e) Steam engine, dynamo

27. Which of the following is not a form of energy?

(a) Rubbing
(b) Heat
(c) Electricity
(d) Sound
(e) Light

28. When a tin of nails is shaken, the energy change is from:-

(a) /

(a) potential energy to sound energy
(b) sound energy to kinetic energy
(c) chemical energy to sound energy
(d) kinetic energy to sound energy
(e) sound energy to chemical energy

29. When an electric light is switched on the main energy conversion is from electrical energy to -

(a) sound energy.
(b) heat energy.
(c) chemical energy.
(d) light energy.
(e) potential energy.

30. The spring of a toy car was wound up and then it was placed on the bench. Which of the following describes the energy conversion that takes place?

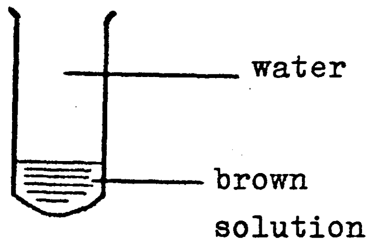
(a) Kinetic energy $\rightarrow$ sound energy
(b) Potential energy $\rightarrow$ kinetic energy
(c) Potential energy $\rightarrow$ heat energy
(d) Kinetic energy $\rightarrow$ heat energy
(e) Potential energy $\rightarrow$ sound energy

A P P E N D I X 4

TEST FOR INTEGRATED SCIENCE SYLLABUS - SECTION FOUR

1. After a few weeks the brown solution will be -

   (a) lower
   (b) higher
   (c) the same height
   (d) crystallised
   (e) blue



2. The process taking place in the above experiment is called -

   (a) dispersion
   (b) filtration
   (c) distillation
   (d) diffusion
   (e) crystallisation

3. When watching Brownian movement in a suspension of smoke in air, through a microscope, the white specks are seen to dance about. This is because -

   (a) the white specks are moving themselves about.
   (b) the white specks are jostled about by colliding with air particles.
   (c) heat from the light bulb was causing the white specks to dance about.
   (d) the air particles all move in the same direction.
   (e) a draught was getting into the smoke cell.

4. If a bottle of perfume is spilled at one end of a room we quickly smell it at the other end because the perfume particles find their way between the air particles. This process is called -

   (a) diffusion
   (b) decomposition
   (c) dilution
   (d) dispersal
   (e) mixing

5. /

5. The states of matter are -

   (a) solid, crystal, gas.
   (b) solid, liquid, gas.
   (c) solid, vapour, gas.
   (d) atoms, molecules, particles.
   (e) mass, length, time.

6. All matter consists of particles called -

   (a) crystals
   (b) elements
   (c) colloids
   (d) compounds
   (e) atoms

7. In solids the particles are -

   (a) tightly packed in neat rank and file order.
   (b) spaced well apart.
   (c) moving about at high speeds in all directions.
   (d) not moving at all.
   (e) close together but moving about slowly.

8. Which of the following substances is a gas?

   (a) Oxygen
   (b) Mercury
   (c) Carbon
   (d) Water
   (e) Smoke

9. A substance which contains only one kind of atom is called -

   (a) a molecule.
   (b) a crystal.
   (c) a compound.
   (d) a mixture.
   (e) an element.

10. The least number of elements which can form a compound is -

   (a) 0
   (b) 1
   (c) 2

11. Hydrogen is an example of -

(a) a compound.
(b) a mixture.
(c) a solid.
(d) an element.
(e) an alloy.

12. Copper sulphate is an example of -

(a) an alloy.
(b) a compound.
(c) an element.
(d) a mixture.
(e) a metal.

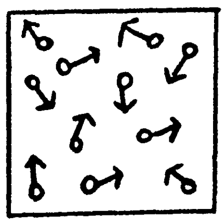
13. Copper sulphate contains -

(a) only copper atoms.
(b) copper, sulphur and oxygen atoms.
(c) only sulphur atoms.
(d) copper and sulphur atoms.
(e) copper and oxygen atoms.

14. If an element Iron combines with an element sulphur to form a compound called iron sulphide, the properties of iron sulphide will be -

(a) similar to those of iron.
(b) similar to those of sulphur.
(c) entirely different from those of iron or sulphur.
(d) a mixture of the properties of iron and sulphur.
(e) more like those of iron than sulphur.

15. The figure in the diagram opposite represents -



(a) the molecules of a solid.
(b) the molecules of water.
(c) the molecules of a liquid.
(d) the molecules of a metal.
(e) the molecules of a gas.

16. In which of the following substances would the particles be closest together -

(a) /

(a) carbon dioxide
(b) water
(c) alcohol
(d) lead
(e) carbon

17. In which of the following materials would the particles have the most kinetic energy?

(a) oxygen at - 20°C
(b) ice at 0°C
(c) water at 20°C
(d) iron at 200°C
(e) mercury at 20°C

18. Which of the following compounds contains three elements?

(a) sulphur trioxide
(b) sodium sulphate
(c) silver iodide
(d) water
(e) sodium chloride

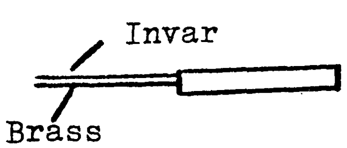
19. Name the elements in Magnesium phosphide:-

(a) Magnesium, phosphorus, oxygen.
(b) Manganese, phosphorus.
(c) Magnesium, potassium, oxygen.
(d) Magnesium, phosphorus.
(e) Manganese, phosphorus, oxygen.

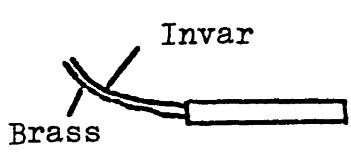
20. On heating -

(a) all substances contract.
(b) all substances expand.
(c) only gases expand.
(d) all solids contract.
(e) some substances contract and some expand.

21.



before heating



after heating

The above /

The above experiment with the bi-metallic strip indicates that -

(a) invar expands at a greater rate than brass.
(b) brass expands but invar does not.
(c) brass expands at a greater rate than invar.
(d) invar expands but brass does not.
(e) brass and invar expand at the same rate.

22. I put my finger on the end of a bicycle pump. When the pump is filled with air I can push the handle down a long way. When the pump is filled with water I cannot push the handle down at all. This suggests that -

(a) water particles are continually on the move.
(b) the spaces between air particles are smaller than the spaces between water particles.
(c) the spaces between air particles are bigger than the spaces between water particles.
(d) there are no spaces between water particles.
(e) there are no spaces between air particles.

23. Air pressure can be measured using -

(a) a thermometer.
(b) a balance.
(c) a hygrometer.
(d) a hydrometer.
(e) a barometer.

24. What should a compound containing the elements Calcium, Nitrogen and Oxygen be called?

(a) Calcium nitride.
(b) Calcium nitrate
(c) Calcium oxide
(d) Calcium oxine
(e) Calcium nitroxide

25. A thin sheet of paper is placed on top of a test tube full of water. When the test tube is inverted the water remains in the test tube because -

(a) /

(a) paper is a solid and is waterproof.

(b) the water acts as a glue and holds the paper on to its surface.

(c) water cannot exert pressure.

(d) air pressure is pressing the paper against the water.

(e) water and air have different densities.

APPENDIX 5

TEST FOR INTEGRATED SCIENCE SYLLABUS - SECTION FIVE

1. A piece of filter paper is soaked in water. It would dry most quickly when -

   (a) placed on a cold flat surface free from draughts.
   (b) placed on a warm surface free from draughts.
   (c) when wrapped up in a polythene bag.
   (d) when folded up into small squares.
   (e) placed on a warm surface with warm air blowing over it.

2. Which of the following would be the most suitable method of obtaining drinking water from sea water?

   (a) Filtration
   (b) Crystallisation
   (c) Distillation
   (d) Chromatography
   (e) Chlorination

3. Chlorophyll cannot be extracted from plants using water as a solvent because water is not a solvent for chlorophyll. This means that chlorophyll -

   (a) will dissolve in water.
   (b) is soluble in water.
   (c) is insoluble in water.
   (d) will only dissolve in water.
   (e) will dissolve in anything but water.

4. To dissolve a solid quickly in water one generally should -

   (a) have large particles, stir and heat.
   (b) have large particles, do not stir but heat.
   (c) have small particles, stir and cool.
   (d) have small particles, stir and heat.
   (e) have small particles, do not stir but heat.

5. When saliva is added to starch the starch is -

(a) changed in colour.
(b) combined with saliva.
(c) broken down to sugar.
(d) unchanged.
(e) broken down to water.

6. Nature demonstrates a complete distillation process - where?

(a) Wet roads → winds → dry roads.
(b) Leaves fall → rotting → fertilizer → new leaves.
(c) Dead Sea → Sun → gets saltier.
(d) Gas rising from marshy ground.
(e) Sea → Sun → Clouds → cold → rain

7. In an experiment you have four different liquids and one solid. You wish to discover which liquid dissolves the solid fastest. In setting up the experiment you should observe four of the following rules. Which one is wrong?

(a) Take the same volume of each liquid.
(b) Take always the same weight of the solid.
(c) Stir only those which require it.
(d) Keep all the liquids at the same temperature.
(e) Give each liquid the same amount of stirring.

8. The above experiment is an example of -

(a) keeping only one variable constant.
(b) keeping all variables constant.
(c) setting up controls.
(d) examining several variables at once.
(e) variable controls.

9. Two powders have become mixed by mistake. They can be separated if -

(a) they dissolve in water.
(b) a liquid can be found which will dissolve both equally well.
(c) /

(c) a liquid can be found which will dissolve one but not the other.
(d) they have different boiling points.
(e) they pass through a sieve.

10. The number of bacteria is kept at a safe level in our water supply by treating the water, at the water works, in one of the following ways.

(a) Boiling it
(b) Adding chlorine to it
(c) adding fluoride to it
(d) Filtering it
(e) Adding air to it

11. What is the name given to the arrangement which consists of tiny droplets of fat suspended in a liquid?

(a) An emulsion
(b) A solution
(c) A saturated solution
(d) A detergent
(e) An acid

12. A pinch of mustard helps vinegar to mix with olive oil. Mustard is known as -

(a) an emulsifying agent.
(b) a solvent.
(c) a solute.
(d) a drying agent.
(e) an enzyme.

13. A pinch of substance Y was stirred into water in a beaker. When a beam of light was shone at the liquid, the part of the beam through the liquid could be distinctly seen. Y could be -

(a) starch
(b) sugar
(c) salt
(d) icing sugar
(e) sand

14. /

14. In carrying out experiments in the laboratory to test for the presence of enzymes in saliva the test tubes were kept in a bath of water at $37^{o}C$. Why was this?

(a) Chemical reactions go better at low temperatures.
(b) Starches break down to sugars at temperatures over $30^{o}C$.
(c) To be as close to body temperature as possible.
(d) This is room temperature.
(e) This is the melting point of starch.

15. When a solid is dissolved in a liquid it is called the -

(a) solvent.
(b) solution.
(c) solute.
(d) ingredient.
(e) saturated solution.

The following information refers to questions 16 to 18.

(a) Solute
(b) Solution
(c) Solvent
(d) insoluble
(e) Saturated solution

Which of these terms describes:-

16. A liquid which extracts the red colouring matter from rose petals?

17. The liquid which remains when hot salt water is cooled down until crystals of salt appear?

18. A red powder which, when stirred with alcohol and filtered all remains on the filter paper?

19. A scientist suspects that some canned peas contain a mixture of two green dyes, X and Y. He crushes the peas /

the peas with water and makes a chromatogram with the liquid. He finds that the chromatogram consists of a single ring of dye X. He tries again using acetone as a solvent, and gets a single ring of dye Y. He then repeats the experiment, using alcohol as a solvent, and obtains a separation of the two dyes in his chromatogram. Which one of the following statements is true?

(a) X is soluble in alcohol but Y is insoluble.
(b) Y is soluble in water but X is insoluble.
(c) Y is soluble in water but insoluble in acetone.
(d) X is soluble in alcohol but insoluble in acetone.
(e) X and Y are both soluble in acetone.

20. Two experiments were set up in beakers of warm water. A consisted of Visking tubing which contained starch solution and B consisted of Visking tubing which contained starch solution and saliva. At the start of the experiment the water in the beakers was tested for starch and sugar. Neither was detected. After a period of time the water in the beakers was tested again and all that could be detected was some sugar in bath B. Which one of the following statements on these experiments is NOT correct?

(a) Starch can be changed to sugar.
(b) Sugar can pass through the Visking tubing.
(c) Starch cannot pass through the tubing.
(d) The warmth of the water had caused the starch to change to sugar.
(e) The saliva had caused the change in the starch.

A P P E N D I X 6

TERM TEST FOR FIRST YEAR PUPILS FOLLOWING INTEGRATED SCIENCE SYLLABUS

1. Which of the following methods would you use to obtain pure salt out of a mixture of salt and sand?

   (a) Pick out the grains of sand with tweezers.
   (b) Pour water on the mixture, heat it, and then leave it to evaporate.
   (c) Dissolve the salt, filter off the sand, evaporate the salt water.
   (d) Wash the mixture under the tap to wash away the sand.
   (e) Dissolve the mixture and distill the solution.

2. Chromatography is a process best described as -

   (a) filtering with filter paper.
   (b) showing whether substances are soluble or otherwise.
   (c) dissolving solid substances.
   (d) separating a liquid into the substances which make up the liquid.
   (e) taking colour photographs.

3. Brine (salt water) → steam. Which one of the alternatives below best fits the above process?

   (a) Crystallisation
   (b) Condensation
   (c) Dissolving
   (d) Distillation
   (e) Evaporation

4. Steam → liquid water. Which one of the alternatives below best fits the above process?

   (a) Evaporation
   (b) Condensation
   (c) Distillation
   (d) Dissolving
   (e) Crystallisation

5. Liquid water → brine. Which one of the alternatives below best fits the above process?

(a) Crystallisation
(b) Condensation
(c) Dissolving
(d) Distillation
(e) Evaporation

6. Brine → Saturated salt solution. Which one of the alternatives below best fits the above process?

(a) Evaporation
(b) Distillation
(c) Dissolving
(d) Condensation
(e) Crystallisation

7. Saturated salt solution → salt crystals. Which one of the alternatives below best fits the above process?

(a) Crystallisation
(b) Condensation
(c) Dissolving
(d) Distillation
(e) Evaporation

8. Salt crystals → Brine. Which one of the alternatives below best fits the above process?

(a) Evaporation
(b) Distillation
(c) Dissolving
(d) Condensation
(e) Crystallisation

9. Brine → pure water. Which one of the alternatives below best fits the above process?

(a) Crystallisation
(b) Condensation
(c) Dissolving
(d) Distillation
(e) Evaporation

10. When a liquid is heated, it will sometimes turn into a gas. When the gas is cooled, it turns into a liquid. These two operations are best described as:-

(a) Evaporation and distillation
(b) Distillation and crystallisation
(c) Evaporation and condensation
(d) Distillation and condensation
(e) Evaporation and crystallisation

11. A boy said that aluminium was lighter than lead - yet another boy said he could prove that was wrong. He produced a block of lead and a block of aluminium and weighed them separately. They both weighed the same. When compared with the lead block, was the aluminium block:-

(a) smaller
(b) bigger
(c) the same size
(d) heavier
(e) lighter

12. Two blocks of the same size were placed on a bench. One was much lighter to pick up, because it was -

(a) more dense.
(b) made of plastic.
(c) made of aluminium.
(d) less dense.
(e) made of wood.

13. Two boys decided to balance a pair of scales by pouring some water on one side, and some mercury on the other side. They knew that 1 cm^3 of water weighed 1 gram and that 1 cm^3 of mercury weighed 13 grams. One boy measured out carefully 13 cm^3 of water and poured it on the left side pan. Then the other boy carefully measured out a quantity of mercury, poured it on the other side and the scales balanced. The quantity of mercury he poured on was:-

(a) /

(a) 1 cm^3
(b) 1 gram
(c) 13 cm^3
(d) 1 square centimetre
(e) 13 cubic grams

14. Which of the timing devices below would you use to find the time taken for an egg to be soft boiled?

(a) A spring swings to and fro 20 times a second and when set swinging will swing for 2 seconds.
(b) A 5 cm length of magnesium burns away completely in 5 seconds.
(c) a 10 cm^3 measuring cylinder fills completely from a dripping tap in 10 minutes.
(d) a 12 cm candle burns away completely in 6 hours.
(e) The sun sets approximately every 24 hours.

15. Which one of the timing devices below would you use to find the time taken for a man's hair to grow 2 cms. in length?

(a) The sun sets approximately every 24 hours.
(b) A 12 cm candle burns away completely in 6 hours.
(c) A 10 cm^3 measuring cylinder fills completely from a dripping tap in 10 minutes.
(d) A 5 cm length of magnesium burns away completely in 5 seconds.
(e) A spring swings to and fro 20 times a second and when set swinging will swing for 2 seconds.

16. Which of the timing devices below would you use to find the time taken for a cricket ball to fall from the ceiling of the laboratory to the floor?

(a) A 5 cm length of magnesium burns away completely in 5 seconds.

(b) /

(b) A 10 cm^3 measuring cylinder fills completely from a dripping tap in 10 minutes.
(c) A 12 cm candle burns away completely in 6 hours.
(d) The sun sets approximately every 24 hours.
(e) A spring swings to and fro 20 times a second and when set swinging will swing for 2 seconds.

17. Which of the timing devices below would you use to find the time taken for a man to walk 12 kilometres at a normal walking pace?

(a) The sun sets approximately every 24 hours.
(b) A 12 cm candle burns away completely in 6 hours.
(c) A 10 cm^3 measuring cylinder fills completely from a dripping tap in 10 minutes.
(d) A 5 cm length of magnesium burns away completely in 5 seconds.
(e) A spring swings to and fro 20 times a second and when set swinging will swing for 2 seconds.

18. Which of the following lists is made up entirely of substances that have come from once living material?

(a) Water, gold, iron
(b) Rock, clay, snow
(c) Oil, wood, paper
(d) Glass, cement, silver
(e) Concrete, brick, lead

19. Vertebrates all possess backbones - Man is a vertebrate. Which of the pairs of animals below belong to the vertebrate group?

(a) Earthworm and jelly fish
(b) Sponge and snail
(c) Crab and spider
(d) Cat and elephant
(e) Starfish and leech

20. /

20. An earthworm belongs to a group of animals known as annelids. The body of the earthworm is divided into sections called segments. Also in this group of animals are leeches, bristle worms and marine worms. Which of the following things would you expect them to have in common?

(a) They are all in black.
(b) They all attach themselves to things by using suckers.
(c) They all live on land.
(d) They all have bodies that are divided up into sections.
(e) They all live in water.

21. Copper and aluminium are among the substances in the world that have never lived. Which one of the following pairs of substances h as never lived?

(a) Charcoal and coke
(b) Leather and wool
(c) Sponge and linen
(d) Wood and wire
(e) Tin and iron

22. Earthworms live in the soil and make narrow burrows. This making of burrows is beneficial to the soil because -

(a) the farmer does not have to make holes in which to put seed and plants.
(b) air and rain penetrate the soil.
(c) the earthworms making the burrows eat too much soil.
(d) the soil becomes hardpacked and flat.
(e) the burrows made huge mounds on top of the soil.

23. Which type of water is the purest?

(a) Rain water
(b) Distilled water
(c) /

(c) River water
(d) Sea water
(e) Tap water

24. A beaker is very carefully filled with water so that it is brim full. If two teaspoonfuls of salt are now carefully added to the beaker the water will -

(a) cause the salt to form lumps.
(b) start to foam.
(c) remain unchanged.
(d) go down.
(e) overflow.

25. What is the name of the process that allows water to be lost from the skin's surface?

(a) Boiling
(b) Evaporating
(c) Distilling
(d) Condensing
(e) Fermenting

26. Oil will float on water and wood will float on oil. Water will float on carbon tetrachloride, but mercury will not float on carbon tetrachloride. For equal volumes of these substances, which one of the answers below is correct?

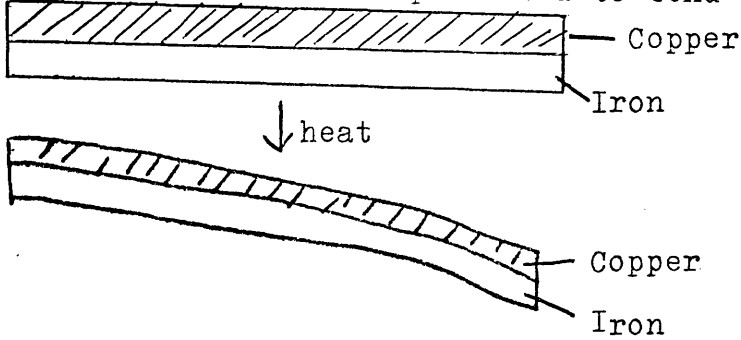
(a) The lightest of all these substances is oil.
(b) Carbon tetrachloride is heavier than mercury.
(c) Wood will sink in mercury.
(d) Mercury is lighter than oil.
(e) Wood is lighter than oil.

27. Which of the following represents the way in which matter expands when heated?

(a) Solids more than liquids, liquids more than gases.
(b) Liquids more than solids, solids more than gases.
(c) /

(c) Gases more than solids, solids more than liquids.
(d) Gases more than liquids, liquids more than solids.
(e) Solids more than gases, gases more than liquids.

28. Two flat strips of copper and iron were stuck together. They were heated with the result that the combined metal strip started to bend as shown.



Which of the following provides the best explanation?

(a) The two metals fused to produce an alloy.
(b) The copper expanded more than the iron.
(c) The iron expanded more than the copper.
(d) The copper started to melt.
(e) The two metals started to come apart.

29. What happens to the particles in a liquid when the liquid expands? The particles -

(a) disintegrate into nothing.
(b) cause the volume to decrease.
(c) move more slowly.
(d) become stationary.
(e) move faster.

30. Old railway lines are not continuous stretches of line - there are spaces in the track at intervals. These spaces are present to allow for expansion of the metal in hot weather. What would happen if there were no spaces, and the metal expanded a lot in hot weather?

(a) The track would buckle and bend.
(b) The two lines would melt together.
(c) /

(c) The track would become shorter.
(d) The trains would travel faster on narrower track.
(e) The track would become too narrow for train wheels.

31. Mercury in a thermometer expands when it is placed under the tongue. What causes the mercury to expand?

(a) Saliva mixes with mercury, and the mercury expands.
(b) Body heat causes the mercury heat to expand.
(c) Pressure from the mouth causes the mercury to expand.
(d) Gases in the mouth cause the mercury to expand.
(e) Pressure from the teeth stretches out the mercury.

32. The best explanation for the use of mercury in thermometers is that it -

(a) can easily be seen.
(b) is a metal.
(c) gives equal divisions.
(d) is cheap.
(e) is a good conductor of heat.

33. When making a thermometer, how would you set about putting a scale on the side of the glass tubing? Place the flask in -

(a) ice
(b) hot water
(c) ice and boiling water
(d) boiling water
(e) water with a thermometer

34. Sulphur is an example of -

(a) an element.
(b) a mixture
(c) /

(c) a sulphate.
(d) a metal.
(e) a compound.

35. Copper chloride is an example of -

(a) a mixture.
(b) an alloy.
(c) an element.
(d) a compound.
(e) a metal.

36. A girl accidentally spilled some iodine on her hands. To remove it she should rub her hands with:-

(a) distilled water
(b) acid
(c) alcohol
(d) water
(e) none of these

37. Milk is made up of tiny droplets of butter fat suspended in water. The droplets of fat do not separate out after a period of time because -

(a) the fat droplets stick to the water particles.
(b) the fat droplets gradually dissolve in the water.
(c) the fat droplets and the water particles repel.
(d) the fat droplets are held suspended by an emulsifying agent.
(e) the particles are too large to form a true solution.

38. What energy change takes place when an electric fire is switched on?

(a) Kinetic $\longleftrightarrow$ Heat
(b) Electrical $\longleftrightarrow$ Kinetic
(c) Chemical $\longleftrightarrow$ Electrical
(d) Electrical $\longleftrightarrow$ Heat
(e) Chemical $\longleftrightarrow$ Heat

39. What energy change takes place when a catapult is released?

(a) Kinetic $\longleftrightarrow$ Sound
(b) Kinetic $\longleftrightarrow$ Potential
(c) Potential $\longleftrightarrow$ Kinetic
(d) Potential $\longleftrightarrow$ Greater potential
(e) None of these

40. 10 cm^3 of a dark purple solution are put into a measuring cylinder and 90 cm^3 of water added, 10 cm^3 of this mixture are put into another measuring cylinder and 90 cm^3 of water are added. By how many times has the original 10 cm^3 of purple solution been weakened?

(a) Ninety-eight times.
(b) Twice
(c) A hundred times
(d) Fifty times
(e) Ten times

APPENDIX 7

TEST FOR INTEGRATED SCIENCE SYLLABUS - SECTION TEN

1. Acid has been spilled on someone's clothes. The best treatment would be to -

   (a) do nothing until it dries.
   (b) send them to the cleaners.
   (c) wash them with vinegar.
   (d) wash them with sodium bicarbonate solution.
   (e) wash them with lots of water.

2. Sodium is stored -

   (a) in a bottle under water.
   (b) in a bottle under oil.
   (c) lying in an open box.
   (d) in a gas cylinder.
   (e) in a dropping bottle.

3. A metal reacts slowly with dilute acid but not with water. The metal is likely to be -

   (a) gold
   (b) mercury
   (c) sodium
   (d) calcium
   (e) zinc

4. Which of the following best explains what happens when sodium is added to water?

   (a) The sodium melts in the water.
   (b) The sodium dissolves in the water.
   (c) The sodium reacts with the water.
   (d) The water absorbs the sodium.
   (e) The sodium floats on the water.

5. The following facts are known about three metals X, Y and Z. Z tarnishes quickly in air while X and Y do not. Z and X react with dilute hydrochloric acid to give hydrogen while Y does not. The order of reactivity, most reactive first, is -

   (a) /

(a) Z X Y
(b) Y X Z
(c) Z Y X
(d) Y Z X
(e) X Y Z

6. A wasp sting is alkaline and so could be countered by dabbing with a solution of -

(a) soap.
(b) salt.
(c) washing soda.
(d) vinegar.
(e) ammonia.

7. These lists of metals are supposed to be in the order of their decreasing vigour of reaction with water. Which one is correct?

(a) Iron - sodium - calcium - magnesium
(b) Magnesium - iron - sodium - calcium
(c) Sodium - magnesium - iron - calcium
(d) Calcium - sodium - iron - magnesium
(e) Sodium - calcium - magnesium - iron

8. All three metals in one of these lists displace hydrogen from dilute acids. Which list is correct?

(a) Lead - magnesium - aluminium
(b) Magnesium - aluminium - iron
(c) Copper - aluminium - iron
(d) Silver - magnesium - aluminium
(e) Mercury - silver - copper

9. A gas lights with a pop and burns with a blue flame. It is likely to be -

(a) nitrogen
(b) hydrogen
(c) air
(d) carbon dioxide
(e) oxygen

10. /

10. A soil sample was found to have pH = 5. It could be neutralised by adding -

(a) a dilute acid.
(b) sand.
(c) a liquid of pH = 7.
(d) Lime (calcium hydroxide).
(e) a concentrated acid.

11. Which of the following metals will displace hydrogen, most readily, from water?

(a) Tin
(b) Magnesium
(c) Lead
(d) Iron
(e) Copper

12. Metal A displaces hydrogen slowly from water. Metal B displaces hydrogen from an acid but not from water. Metal C bursts into flame when placed in oxygen. The order of reactivity, most reactive first, is -

(a) ABC
(b) BAC
(c) CAB
(d) CBA
(e) ACB

13. The pH of a neutral solution is -

(a) 10
(b) 9
(c) 7
(d) 5
(e) 1

14. A boy playing with a bottle from a chemistry set spills some acid on the carpet. The best remedy would be to treat it with -

(a) water
(b) lemon juice
(c) /

(c) vinegar
(d) sodium bicarbonate
(e) soap

15. Water is a -

(a) compound of hydrogen, nitrogen and oxygen.
(b) compound of hydrogen and oxygen.
(c) compound of nitrogen and oxygen.
(d) mixture of hydrogen and nitrogen.
(e) mixture of hydrogen and oxygen.

16. The best way to identify a liquid as being water is to -

(a) taste it.
(b) add calcium.
(c) test it with pH paper.
(d) show that it boils at 100°C.
(e) show that it can put out fires.

17. During a chemical reaction, a gas was given off which turned limewater milky. This showed that the gas was -

(a) hydrogen
(b) oxygen
(c) carbon dioxide
(d) water vapour
(e) nitrogen

18. Which of the following will burn in air to produce water?

(a) Nitrogen
(b) Water vapour
(c) Oxygen
(d) Hydrogen
(e) Carbon dioxide

19. Balloons used for carrying passengers are not normally filled with hydrogen because -

(a) the heat of the sun would burst the balloon.
(b) hydrogen has an unpleasant smell.
(c) /

(c) hydrogen is so light that the balloon would not come down again.
(d) the danger of explosion is too high.
(e) hydrogen is not light enough to lift a man off the ground.

20. 10 cm^3 of dilute acid is placed in each of three beakers X, Y and Z. 10 cm^3 of dilute alkali is added to Y and 20 cm^3 of dilute alkali is added to Z. Which will be the likely series of pH for X, Y and Z in that order?

(a) 7 - 9 - 5
(b) 9 - 7 - 5
(c) 5 - 9 - 7
(d) 9 - 5 - 7
(e) 5 - 7 - 9

21. A test-tube rack has a number of labelled tubes containing solutions of known pH.

L	M	N	P	Q	R
3	4	5	7	8	9

P could be made more alkaline by -

(a) evaporation.
(b) addition of N.
(c) addition of L.
(d) addition of Q.
(e) dilution.

22. Given the following solutions -

	L	M	N	P	Q	R
pH	3	4	5	7	8	9

with which of the following mixtures would it be possible to obtain a neutral solution?

(a) L + M
(b) P + R
(c) N + R
(d) P + Q
(e) Q + R

23. /

23. Some metals can be separated from a mixture of them with gold by dissolving them in dilute hydrochloric acid. For which of the following would this not be possible?

(a) Magnesium
(b) Zinc
(c) Iron
(d) Copper
(e) Tin

24. When tin is placed in water there is no chemical reaction because -

(a) tin is very dense and sinks.
(b) the particles of tin are not small enough.
(c) tin is not active enough.
(d) tin is not stable enough.
(e) tin is a shiny metal.

25. Which of the following statements is true? Hydrogen -

(a) is lighter than air.
(b) does not burn.
(c) supports burning.
(d) is soluble in water.
(e) is heavier than air.

APPENDIX 8

TEST FOR INTEGRATED SCIENCE SYLLABUS - SECTION TWELVE

1. Heating coal out of contact with air produces a gas called:-

   (a) Natural gas
   (b) Coal gas
   (c) Coke gas
   (d) Calor gas
   (e) None of these

2. A night watchman notices that a certain type of coal on his fire gives a sharp, acid smell. This is probably due to the formation of:-

   (a) Sulphur dioxide
   (b) Nitrogen oxide
   (c) Carbon monoxide
   (d) Carbon dioxide
   (e) Ammonia

3. Addition of sulphuric acid to malachite results in:-

   (a) Copper oxide + carbon dioxide
   (b) Copper oxide + sulphur dioxide
   (c) Copper sulphate + sulphur dioxide
   (d) Copper sulphate + carbon dioxide
   (e) Copper + sulphur dioxide

4. A scientist is asked to distinguish between a sample of magnesium carbonate and calcium carbonate. He would do this by means of:-

   (a) A flame test
   (b) The colour difference between the two carbonates
   (c) Reducing them to their metals with carbon
   (d) Adding dilute hydrochloric acid
   (e) Heating them

5. Analysis of a sample of water showed that it contained salts of sodium and calcium. It would be likely to have come from:-

   (a) /

(a) The sea
(b) A reservoir in a limestone region
(c) A reservoir in a chalk region
(d) A river flowing through a region of salt deposits
(e) A river flowing through a region of peat bogs

6. A ship sailing round the coast of Iceland discovers a newly-formed volcanic island a few miles from the mainland. The rocks of which the island is composed will be mainly:-

(a) Metamorphic
(b) Igneous
(c) Sedimentary
(d) Metamorphic and igneous
(e) Metamorphic and sedimentary

7. In New Zealand, hot water springs flowing over rocks composed of silicates -

(a) dissolve them completely.
(b) change them to silica.
(c) change them to silicon.
(d) hardly affect them at all.
(e) make them become much softer.

8. In an article in a newspaper on central heating, mention was made of 'fossil' fuel. This was probably referring to:-

(a) Hydro-electric power
(b) Direct sunlight
(c) Wood and paper
(d) Coal and oil
(e) Atomic energy

9. When iron pyrites (iron sulphide) is heated in air:-

(a) there is no change.
(b) sulphur and iron are produced.
(c) sulphur dioxide and iron are produced.
(d) sulphur and iron oxide are produced.
(e) sulphur dioxide and iron oxide are produced.

10. Which one of the following does not contain the element silicon?

(a) Clay
(b) Sand
(c) Quartz
(d) Glass
(e) Malachite

11. The energy released when peat is burned came <u>originally</u> from -

(a) the sun.
(b) plants.
(c) the soil.
(d) the atmosphere.
(e) the sea.

12. John lit his bunsen burner in the usual way but noticed that the flame showed a green colour. He switched off the gas and tipped the bunsen on its side; little pieces of metal fell from around the top. These pieces of metal were probably:-

(a) Sodium
(b) Calcium
(c) Copper
(d) Aluminium
(e) Potassium

13. An element that is the second most abundant in the earth's crust is used for making glass; this element is:-

(a) Chlorine
(b) Phosphorus
(c) Titanium
(d) Silicon
(e) Sulphur

14. Which of the following elements is the most abundant in the earth's crust?

(a) /

(a) Oxygen
(b) Silicon
(c) Aluminium
(d) Iron
(e) Calcium

15. Which one of the following metals gives an apple green flame when undergoing a flame test?

(a) Calcium
(b) Potassium
(c) Copper
(d) Barium
(e) Sodium

16. Which of the following metals are both found uncombined in nature?

(a) Magnesium, zinc
(b) Magnesium, silver
(c) Copper, silver
(d) Copper, magnesium
(e) Zinc, silver

17. A great deal of energy is released when a certain metal is burned in oxygen. This suggests that the -

(a) metal oxide could be easily broken down by heating.
(b) metal will be found uncombined in nature.
(c) metal oxide would need a lot of energy to free the metal.
(d) metal will be obtained from the oxide by heating with carbon.
(e) metal must be magnesium.

18. After evaporating sea water, a pupil performs a flame test on the residue and finds that the flame is yellow. This suggests that sea water contains -

(a) /

(a) sodium chloride.
(b) sodium salts.
(c) potassium chloride.
(d) potassium salts.
(e) sodium and potassium salts.

19. Rocks which have been formed, then changed in their nature by the effects of great pressure and heat, are called:-

(a) Igneous
(b) Metamorphic
(c) Sedimentary
(d) Carboniferous
(e) Fossilized

20. Marble is a naturally occurring form of -

(a) copper carbonate.
(b) magnesium oxide.
(c) calcium carbonate.
(d) copper oxide.
(e) calcium oxide.

21. Crude oil can be separated by fractional distillation because each fraction has a different -

(a) composition.
(b) density.
(c) boiling point.
(d) melting point.
(e) viscosity.

22. Some soil is shaken up with water in a large jar. The order in which the soil particles settle out is:-

(a) Gravel, coarse sand, fine sand, clay
(b) Clay, gravel, fine sand, coarse sand
(c) Gravel, clay, fine sand, coarse sand
(d) Clay, coarse sand, fine sand, gravel
(e) Coarse sand, fine sand, clay, gravel

23. A piece of Roman pottery dug up in a ruin has a blue-green glaze. This might suggest that Roman potters used glazes containing compounds of:-

(a) /

(a) Sodium
(b) Zinc
(c) Calcium
(d) Lead
(e) Copper

24. Heating coal in a limited supply of air gives a material commonly used in industry for obtaining metals from their ores. This material is:-

(a) Tar
(b) Ammonia
(c) Coal gas
(d) Coke
(e) Sulphur

25. A planet has an atmosphere consisting of nitrogen only - which of the following uncombined elements might be found pure in its crust?

(a) Iron
(b) Lead
(c) Tin
(d) All of these
(e) None of these

APPENDIX 9

TEST FOR INTEGRATED SCIENCE SYLLABUS - SECTION FOURTEEN

1. A dentist is fixing false teeth into a dental plate. If X = incisors, Y = molar, and Z = canines, in which order would he fit them? (front, side, back)

   (a) X Y Z
   (b) X Z Y
   (c) Z X Y
   (d) Y Z X
   (e) Y X Z

2. Which of the following statements refer to incisors? They are -

   (a) used for cutting up food.
   (b) used for grinding up food.
   (c) used for stabbing food.
   (d) sharply pointed in shape.
   (e) grooved on the surface.

3. Which of the following is the best source of protein?

   (a) Bread
   (b) Cabbage
   (c) Chocolate
   (d) Apples
   (e) Fish

4. Which of the following is the best source of starch?

   (a) Bread
   (b) Butter
   (c) Jam
   (d) Egg
   (e) Bacon

5. When a drop of milk was placed on a filter paper, an oily spot was formed. This shows that milk contains -

   (a) /

(a) starch.
(b) glucose.
(c) protein.
(d) fat.
(e) water.

6. Fehling's solution turns red in glucose. A chemist spits (saliva) into a test tube containing starch solution. Drops are removed and tested as shown.

	Fehling's solution	Iodine
0 minutes	blue	blue
5 minutes	pink	pale blue
10 minutes	red	unchanged

The results show that saliva -

(a) turns starch into glucose.
(b) changes glucose into starch.
(c) and Fehling's solution do not react.
(d) and iodine do not react.
(e) must contain glucose.

7. The importance of glucose to humans as a food is the formation of -

(a) oxygen.
(b) carbon dioxide.
(c) energy.
(d) water.
(e) vitamins.

8. The most important foods for growth and repair are -

(a) fats.
(b) proteins.
(c) carbohydrates.
(d) vitamins.
(e) mineral salts.

9. Which of the following foodstuffs would a trainer of top class racehorses use as a balanced diet for his horses?

% Protein /

	% Protein	% Carbohydrate	% Roughage
(a)	80	20	0
(b)	20	80	0
(c)	40	40	20
(d)	20	30	50
(e)	30	20	50

10. During digestion food becomes soluble and passes from the digestive system into the blood. This mainly takes place in the -

(a) mouth.
(b) gullet.
(c) small intestine.
(d) large intestine (colon).
(e) rectum.

11. Which of the following plays no part in our digestion?

(a) Salivary glands
(b) Pancreas
(c) Gall bladder
(d) Appendix
(e) Liver

12. Which of the following is the correct order of parts in the digestive system, starting with the mouth?

(a) Gullet, small intestine, stomach, colon
(b) Gullet, stomach, small intestine, colon
(c) Stomach, gullet, colon, small intestine
(d) Stomach, small intestine, colon, gullet
(e) Small intestine, gullet, stomach, colon

13. While we eat, some digestive fluid is passed into the mouth from the -

(a) gall bladder.
(b) pancreas.
(c) liver.
(d) salivary glands.
(e) gullet.

14. /

14. Which of the following substances is removed from the blood as it passes through the kidney and stored in the bladder?

    (a) Oxygen
    (b) Urea
    (c) Glucose
    (d) Carbon dioxide
    (e) Starch

15. Which of the following is an example of ELIMINATION rather than of EXCRETION ?

    (a) Leaf fall
    (b) Formation of pigments, gums or resin in plants.
    (c) Removal of urea from the blood by the kidneys
    (d) Storing of urine in the bladder
    (e) All of these

16. Identical twins X and Y are found to have a pulse rate of 80. Twin X, after running for five minutes, is found to have a pulse rate of 100. Both twins then blow into beakers of lime water, and twin Y takes far longer to turn the lime water milky. This experiment shows that increased exercise -

    (a) increases heart beat only.
    (b) decreases heart beat only.
    (c) increases heart beat and carbon dioxide output.
    (d) decreases heart beat and carbon dioxide output.
    (e) is not good for twin X.

17. Which of the following would NOT be found in human blood?

    (a) Carbon dioxide
    (b) Oxygen
    (c) Urea
    (d) Sugar
    (e) Roughage

18. Digestion is the breaking down of large food molecules into smaller ones. The main purpose of this is to -

(a) make the food soluble.
(b) use up the digestive enzymes.
(c) make the food slide down the intestine.
(d) make the food insoluble.
(e) break down the roughage.

19. Which of the following is TRUE ? Sweat -

(a) is given off equally all over the body.
(b) is a measure of water intake.
(c) is a measure of body temperature.
(d) consists of water and other waste products.
(e) comes from the veins in the skin.

20. The skeleton of an unknown animal is dug up in a swamp. Which of the following would tell us whether or not it hunted other animals?

(a) The length of its legs.
(b) The size of its body.
(c) The shape of its teeth.
(d) Whether or not it had horns.
(e) Its number of teeth.

APPENDIX 10

EXAMINATION ON SECTIONS ONE, TWO, THREE AND FOUR OF THE INTEGRATED SCIENCE SYLLABUS given to all first year pupils. Easy to difficult item order.

1. Below is a list of five materials.

   Water; air; wood; alcohol; iron

   Which of the following correctly describes these materials?

   (a) One is solid, two are liquid, two are gas
   (b) Two are solid, two are liquid, one is gas
   (c) Two are solid, one is liquid, two are gas
   (d) Three are solid, one is liquid, one is gas

2. A measuring cylinder is used to measure:-

   (a) Height
   (b) Length
   (c) Weight
   (d) Volume

3. To which of the following groups does seaweed belong?

   (a) Vertebrates
   (b) Invertebrates
   (c) Flowering plants
   (d) Non-flowering plants

4. A large cube of rock salt is dissolved in 50 cm^3 water and the final volume is less than the volume of the cube and the water added together. What is the best explanation for this difference?

   (a) Some of the liquid evaporates.
   (b) Air pressure squeezes them into a smaller volume.
   (c) The salt particles go into spaces between the water particles.
   (d) The volumes were not measured accurately.

5. When an electric iron is heating up which of the following energy changes occur in the iron?

   (a) /

(a) Magnetism - electrical - heat
(b) Electrical - heat
(c) Heat - electrical
(d) Electrical - magnetism - heat

6. Which of the following is a group of mammals?

(a) Crocodiles
(b) Frogs
(c) Sharks
(d) Humans

7. The average pulse rate of the pupils of a class is 75 per minute. One girl's is 80 per minute. Which of the following is the best explanation for the difference?

(a) The girl must be odd.
(b) Few people are exactly average.
(c) The girl is very healthy.
(d) The girl's stopclock is not working.

8. Gases expand when heated. The best explanation is that -

(a) the particles become heavier.
(b) the particles begin to move.
(c) the particles move further apart.
(d) the particles move more slowly.

9. Which of the following is a unit of mass?

(a) Millilitre
(b) Centimetre
(c) Gram
(d) Second

10. What volume of liquid would be obtained if 50 cm^3 water and 50 cm^3 alcohol were mixed together?

(a) 50 cm^3
(b) 98 cm^3
(c) 100 cm^3
(d) 102 cm^3

11. /

11. The temperature of the human body is approximately:-

(a) $25^{o}C$
(b) $37^{o}C$
(c) $50^{o}C$
(d) $97^{o}C$

12. If you put your left hand into cold water and your right hand into hot water, then after a short while put them both into warm water, what would you feel?

(a) The left hand warmer than the right.
(b) The left hand colder than the right.
(c) Both hands cold.
(d) Both hands hot.

13. The symbol for the unit of mass is:-

(a) kg
(b) gk
(c) k
(d) mk

14. To which of the following groups do fish belong?

(a) Vertebrates
(b) Invertebrates
(c) Reptiles
(d) None of these

15. A glass stopper was stuck in a glass bottle. To remove it, a cloth was soaked in hot water and wrapped round the neck of the bottle. The stopper was then easy to remove. The correct explanation is that the -

(a) glass was weakened by the heat.
(b) the neck of the bottle expanded more than the stopper.
(c) the stopper contracted.
(d) the neck of the bottle and the stopper expanded equally.

16. /

16. When air is blown into a balloon and the balloon is then heated it gets larger. This is because -

(a) the air particles have received energy and are moving faster.
(b) the air particles have given up energy and are moving about more slowly.
(c) heating increases the number of particles in the balloon.
(d) heating decreases the number of particles in the balloon.

17. One of the ways in which solids differ from gases is that:-

(a) Gases are weightless, solids are not.
(b) All gases are invisible, solids are not.
(c) Solids can dissolve in liquids, gases cannot.
(d) Gas particles are free to move about, solid particles can only vibrate.

18. How do solids, liquids and gases compare when heated equally?

(a) Gases expand more than liquids and solids.
(b) Solids expand more than liquids and gases.
(c) Liquids expand more than solids and gases.
(d) All three expand equally.

19. Which kind of energy is used by a coal-fired power station to produce electricity?

(a) Mechanical
(b) Atomic
(c) Chemical
(d) Electrical

20. In which of the following ways do water particles and steam particles differ?

(a) Steam particles are closer together and move about more slowly.
(b) Steam particles are closer together and move about more quickly.

(c) /

(c) Steam particles are further apart and move about more slowly.
(d) Steam particles are further apart and move about more quickly.

21. A weather balloon filled with helium expands as it rises because -

(a) decreased air pressure causes deflation of the balloon.
(b) increased air pressure causes inflation of the balloon.
(c) decreased air pressure causes inflation of the balloon.
(d) increased air pressure causes deflation of the balloon.

22. What happens to a bar of metal when it is heated?

(a) It gets shorter and thinner.
(b) It gets shorter and thicker.
(c) It gets longer and thinner.
(d) It gets longer and thicker.

23. Equal pressure is applied to the following substances. Which will be compressed most?

(a) Rubber
(b) Silver
(c) Water
(d) Nitrogen

24. Which of the following describes air particles in a room and their movement?

(a) Close together and vibrating
(b) Far apart and moving rapidly
(c) Close together and moving slowly
(d) Far apart and moving slowly

25. When a dynamo runs, the principal energy change is from -

(a) electrical to mechanical.
(b) electrical to light.
(c) /

(c) mechanical to electrical.
(d) light to electrical.

26. Which of the following statements about the structure of matter is correct?

(a) Molecules are made up of atoms.
(b) Atoms are made up of elements.
(c) Elements are made up of compounds.
(d) Atoms are made up of molecules.

27. A small drop of oil is allowed to spread on the surface of clean water contained in a large trough. Which of the following affects the thickness of the oil film formed?

(a) Size of the oil molecules
(b) Size of the original oil drop
(c) Volume of water in the trough
(d) Distance the drop fell before hitting the water.

28. When telephone wires are erected they are left slack to allow for -

(a) contraction in cold weather.
(b) the wind blowing them.
(c) expansion in hot weather.
(d) the electricity heating them.

29. Which of the following forms of energy on Earth can not be traced back to the Sun's energy?

(a) Food energy
(b) Atomic energy
(c) Hydroelectric energy
(d) Fuel energy (coal, oil, gas)

30. A piece of wood held under water and released rises to the surface because -

(a) the water pressure is less near the surface.
(b) gravity does not act under water.
(c) the wood is less dense than water.
(d) the wood weighs less than the water.

APPENDIX 11

EXAMINATION ON SECTIONS TEN, TWELVE, FOURTEEN OF THE INTEGRATED SCIENCE SYLLABUS given to all second year pupils. Easy to difficult item order.

1. Which one of the metals below reacts with water at room temperature?

   (a) Calcium
   (b) Lead
   (c) Copper
   (d) Silver

2. Which part of the digestive system is small and useless?

   (a) Stomach
   (b) Gullet
   (c) Appendix
   (d) Small intestine

3. A boy's pulse and breathing rate were measured before exercise and were 80 beats/minute and 15 breaths/minute respectively. He then ran up and down stairs for 5 minutes after which his pulse and breathing rate were most likely to be:-

   (a) 90 beats/min.; 5 breaths/min.
   (b) 80 beats/min.; 10 breaths/min.
   (c) 70 beats/min.; 18 breaths/min.
   (d) 100 beats/min.; 20 breaths/min.

4. Articles made of the following materials are found in a "treasure hoard". Which would be best preserved?

   (a) Gold
   (b) Tin
   (c) Lead
   (d) Copper

5. A wasp sting is alkaline and so could be counteracted by dabbing with -

   (a) /

(a) soap solution.
(b) iodine.
(c) vinegar.
(d) ammonia.

6. Sailors in olden days suffered from scurvy. This was cured by giving them lime juice. The substance contained in lime juice which gave this cure was:-

(a) vitamin A
(b) vitamin B
(c) vitamin C
(d) vitamin D

7. Sometimes gold coins and ornaments are found which are hundreds of years old. Usually the gold is still bright and shiny; yet if a piece of iron is left unprotected in the open, it will usually rust within a few weeks. The reason gold is still shiny is because -

(a) it is a very unreactive metal.
(b) it is a very heavy metal.
(c) the atoms of gold hold on to each other tightly.
(d) the oxygen in the air helps to keep it shiny.

8. Before being carried round the body oxygen in the blood has to travel from the lungs through the:-

(a) liver
(b) heart
(c) kidneys
(d) duodenum

9. When green copper carbonate is heated:-

(a) It stays green and gives off oxygen
(b) It stays green and gives off carbon dioxide
(c) It turns black and gives off carbon dioxide
(d) It turns black and gives off oxygen.

10. /

10. Which of the following is an element?

(a) Silica
(b) Quartz
(c) Galena
(d) Potassium

11. Dilute acid has been spilled on clothing. What would be the best treatment?

(a) Soak with water
(b) Soak with vinegar
(c) Soak with a dilute alkaline solution
(d) Do nothing until it dries

12. The largest particles found in soil are called:-

(a) sand
(b) gravel
(c) clay
(d) loam

13. Digested food passes from the alimentary canal (food canal) into the blood stream mainly in the -

(a) mouth.
(b) gullet.
(c) stomach.
(d) small intestine.

14. To test that a liquid is pure water you should -

(a) boil it and see if steam comes off.
(b) taste and smell it.
(c) test it with litmus or pH paper.
(d) find its boiling and freezing points.

15. The various kinds of oil in crude oil can be separated by fractional distillation because each fraction has a different:-

(a) composition
(b) density
(c) boiling point
(d) melting point

16. /

16. Food, after digestion, becomes soluble and passes into the blood from the digestive system. This takes place in the -

(a) gullet.
(b) small intestine.
(c) stomach.
(d) rectum.

17. A sample of each of the following solids was mixed with charcoal and heated strongly. Which one would give a metal as the residue?

(a) Sodium sulphide
(b) Aluminium oxide
(c) Lead carbonate
(d) Calcium silicate

18. The movements of the wall of the intestines of an animal -

(a) squeeze the water out of the food.
(b) suck up the digested food.
(c) squeeze out digestive juices onto the food.
(d) squeeze the food along the intestine.

19. Which of the following gives the metals in correct order of reactivity, most reactive first?

(a) Sodium, magnesium, calcium, tin, iron
(b) Calcium, magnesium, sodium, iron, tin
(c) Sodium, calcium, iron, magnesium, tin
(d) Sodium, calcium, magnesium, iron, tin

20. In the body, growth is obtained through the breakdown of:-

(a) fats
(b) proteins
(c) vitamins
(d) carbohydrates

21. Marble is a naturally occurring form of:-

(a) /

(a) calcium carbonate
(b) calcium oxide
(c) aluminium oxide
(d) sodium carbonate

22. A lion runs down its prey and kills it. The teeth used for this killing are:-

(a) canine
(b) incisor
(c) premolars
(d) molars

23. Some vinegar was spilled down an awkward corner of an armchair. It was not possible to use a lot of water to rinse the acid out. Which of the following solutions could be added in a small amount to prevent the acid spoiling the fabric?

(a) Table salt solution
(b) Baking soda solution
(c) Methylated spirit
(d) Lemon juice

24. When strongly heated, a mineral gave off a gas with a sharp, choking smell. The remaining powder gave a greyish metal when heated with carbon. The mineral could be:-

(a) copper oxide
(b) iron sulphide
(c) calcium carbonate
(d) aluminium silicate

25. In which of the following lists do all three metals displace hydrogen from dilute hydrochloric acid?

(a) Magnesium, zinc, aluminium
(b) Copper, zinc, sodium
(c) Silver, magnesium, zinc
(d) Mercury, silver, copper

26. Coal burns with a smokey flame whereas coke does not. This is because of the -

(a) /

(a) tar present in the coal.
(b) carbon present in the coke.
(c) water present in the coal.
(d) carbon present in the coal.

27. Metal X reacts slowly with warm water to give hydrogen. Metal Y does not, but displaces hydrogen from steam. Metal Z is unaffected by steam or water. The correct order of reactivity of these metals, most reactive first, is:-

(a) X Y Z
(b) X Z Y
(c) Y X Z
(d) Z X Y

28. Iron pyrites is a mineral. It is also a -

(a) sulphide.
(b) carbonate.
(c) silicate.
(d) chloride.

29. Calcium reacts with water. In which case will the reaction be most rapid?

(a) A piece of calcium metal with cold water.
(b) A piece of calcium metal with hot water.
(c) Calcium powder with cold water.
(d) Calcium powder with hot water.

30. If red litmus paper is dropped into a beaker of liquid, but does not change colour, which of the following conclusions could be correct?

(a) The liquid is an acid.
(b) The liquid cannot be an acid.
(c) The liquid is an alkali.
(d) The liquid cannot be neutral.

APPENDIX 12

It should be noted that this method of measuring the degree of difficulty of a multiple choice examination has its limitations. When the average Facility Value obtained from the application of the modified forms of the tests are calculated, the effect of equal increases and decreases in individual item facility values can cancel each other out. It is for a further, more rigorous, investigation to take account of the variation in Facility Value of individual items.

REFERENCES

R E F E R E N C E S

1. J. Barzun; Teacher in America (Little-Brown, Boston, 1947)

2. B. Hoffman; The Tyranny of Testing (Crowell-Collier, New York, 1962)

3. R.E. Traub, K. Hambleton and B. Singh; Effects of Promised Reward and Threatened Penalty on Performance of a Multiple Choice Vocabulary Test; Educational and Psychological Measurement; 29 1969 847-861

4. C.W. Waters and L. Waters; Validity and Likability Ratings for Three Scoring Instructions for a Multiple Choice Vocabulary Test; Educational and Psychological Measurement; 31 1971 935-938

5. P.H. Taylor; A Study of the Effect of Instructions in a Multiple Choice Mathematics Test; British Journal of Educational Psychology; 36 1966 1-6

6. E.B. Little; Overcorrection for Guessing in Multiple Choice Test Scoring; Journal of Educational Research; 55 1962 245-252

7. J.B. Little; Overcorrection and Undercorrection in Multiple Choice Test Scoring; Journal of Experimental Education; 35 1966 44-47

8. E.S. Edgington; Scoring Formulae that Correct for Guessing; Journal of Experimental Education; 33 1965 345-346

9. G.M. Ruch and M.H. De Graff; Correction for Chance and "Guess" versus "Do not Guess" Instructions in Multiple Response Tests; Journal of Educational Psychology; 17 1926 368-375

10. R.A. Jackson; Guessing and Test Performance; Educational and Psychological Measurement; 15 1955 74-79

11. H. Gulliksen; Theory of Mental Tests (J. Wiley, New York) p.486

12. N. Frederiksen; The Influence of Timing and Instructions on Co-operative Reading Test Scores; Educational and Psychological Measurement; 12 1952 598-607

13. F.M. Lord; Formula Scoring and Validity; Educational and Psychological Measurement; 24 1964 745-747

14. J.C. Mathews; The Nuffield O-level Chemistry Examination; Educ. in Chem. 4 (1967) 2-10

15. B. Hudson; Guessing in Multiple Choice Tests; School Science Review; 51 1969 175 437-439

16. J. Handy, A.H. Johnstone; Reproducibility in Objective Testing; Educ. in Chem. 10 1973 47-48

17. J. Handy, A.H. Johnstone; How Students Reason in Objective Tests; Educ. in Chem. 10 1973 99-100

18. F.B. Davis, G. Fifer; The Effect on Test Reliability and Validity of Scoring Aptitude and Achievement Tests with Weights for Every Choice; Educational and Psychological Measurement; 19 1959 159-170

19. L.R. Aiken; The Effect on Test Score Variance of Differential Weighting of Item Responses; Psychological Reports; 21 1967 585-590

20. C.F. Willey; The Three-Decision Multiple Choice Test - A Method of Increasing the Sensitivity of the Multiple Choice Item; Psychological Reports; 7 1960 475-477

21. C.S. Bernardson; Comparison of the Three-Decision and Conventional Multiple Choice Tests; Psychological Reports; 20 1967 695-698

22. C.H. Coombs, J.E. Milholland, F.B. Womer; The Assessment of Partial Knowledge in Objective Testing; P.R.B. Technical Research Note 33; Educational Research Unit, University of Michigan, January, 1975

23. J.C. Arnold, P.L. Arnold; On Scoring Multiple Choice Examinations Allowing for Partial Knowledge; Journal of Experimental Education; 39 (1) 1970 8-13

24. R.M. Rippey; Rationale for Confidence Scored Multiple Choice Tests; Psychological Reports; 27 1970 91-98

25. R.K. Hambleton, D.M. Roberts, R.E. Traub; A Comparison of the Reliability and Validity of two Methods for Assessing Partial Knowledge on a Multiple Choice Test; Journal of Educational Measurement; 2 1970 75-82

26. C.O. Matthews; Erroneous First Impressions in Objective Tests; Journal of Educational Psychology; 20 1929 280-286

27. C.T. Morgan, J. Deese; How to Study (McGraw-Hill, New York, 1957)

28. E.P. Dressel, G.C. Jensen; How to Study and Read Successfully (Teachers National Information Service - Longmeadow, Mass., 1955)

29. R.W. Frederick; How to Study Handbook (Appleton-Century, New York, 1938) 345-346

30. E.H. Ehrlich; How to Study and get Higher Marks (Thomas T. Crowell, New York, 1961) 276

31. W.H. Armstrong; Study is Hard Work (Harper and Row, New York) 126

32. D.O. Lynch, B.C. Smith; Item Response Changes : Effects on Test Scores; Measurement and Evaluation in Guidance; 7 1975 220-224

33. H.C. Lehman; Does it Pay to Change Initial Decisions in a True-False Test?; School and Society; 28 1928 456-458

34. M.L. Lowe, C.C. Drawford; First Impressions versus Second Thoughts in True-False Tests; Journal of Educational Psychology; 20 (3) 1929 280-286

35. F.K. Berrein; Are First Impressions Best on Objective Tests?; School and Society; 50 1939 319-320

36. P.J. Reille, L.J. Briggs; Should Students Change Their Initial Answers on Objective-type Tests? More Evidence Regarding an old Problem; Journal of Educational Psychology; 43 1952 110-115

37. K. MacNicol; Effect of Varying Order of Item Difficulty in an Unspeeded Verbal Test; Educational Testing Service, Princeton, New Jersey, 1966

38. R.L. Flaugher, R.S. Melton, C.T. Myers; Item Rearrangement under Typical Test Conditions; Educational and Psychological Measurement; 28 1968 813-824

39. M.H. Brenner; Test Difficulty and Discriminations of Functions of Item Order Difficulty; Journal of Applied Psychology; 48 1964 90-100

40. G. Sax, T.R. Cromack; The Effects of Various Forms of Item Arrangement on Performance; Journal of Educational Measurement; 3 1966 309-311

41. S.W. Huck, W.D. Bowers; Item Difficulty Level and Sequence Effects in Multiple Choice Achievement Tests; Journal of Educational Measurement; 9 1972 105-111

42. R.N. Marso; Test Item Arrangement, Testing Time and Performance; Journal of Educational Measurement; 7 1970 113-118

43. J.J. Monk, W.M. Stallings; Effects of Item Order on Test Scores; Journal of Educational Research; 63 1970 463-465

44. D.C. Munz, A.D. Smouse; Interaction Effects of Item Difficulty and Achievement Anxiety Reaction on Academic Performance; Journal of Educational Psychology; 59 1968 370-374

45. A.D. Smouse, D.C. Munz; The Effects of Anxiety and Item Difficulty Sequence on Achievement Testing Scores; Journal of Psychology; 68 1968 181-184

46. W.G. Mollenkopf; An Experimental Study of the Effects on Item Analysis Data of Changing Item Placement and Test Time Limit; Psychometrika; 15 1950 291-315

47. A. Tversky; On the Optimal Number of Alternatives at a Choice Point; Journal of Mathematical Psychology; 1 1964 386-391

48. S.L. Pressey; Basic Unresolved Teaching - Machine Problems; Theory into Practice; 1 1962 30-37

49. R.L. Ebel, B.G. Williams; The Effect of Varying the Number of Alternatives per Item on Multiple Choice Vocabulary Test Items; Fourteenth Yearbook of the National Council on Measurement used in Education; 1957, 63-65

50. W.S. Zimmerman, L.G. Humphreys; Item Reliability as a Function of the Omission of Misleads; American Psychologist; 8 1953 460-461

51. R.A. Ramos, J. Stern; Item Behaviour Associated with Changes in the Number of Alternatives in Multiple Choice Items; Journal of Educational Measurement; 10 (4) 1973 305-310

52. K. Smith; An Investigation of the Use of "double-choice" Items in Testing Achievement; Journal of Educational Research; 51 1958 387-389

53. R.L. Ebel; Expected Reliabilities as a Function of Choices per Item; Educational and Psychological Measurement; 29 1969 565-570

54. F. Costin; The Optimum Number of Alternatives in Multiple Choice Achievement Tests : Some Empirical Evidence for a Mathematical Proof; Educational and Psychological Measurement; 30 1970 353-358

55. F. Costin; Three Choice versus Four Choice Items : Implications for Reliability and Validity of Objective Achievement Tests; Educational and Psychological Measurement; 32 1972 1035-1038

56. B.S. Bloom (Ed.); Taxonomy of Educational Objectives Handbook I : Cognitive Domain (Longmans Green, New York, 1956)

57. L.J. Cronbach; Response Sets and Test Validity; Educational And Psychological Measurement; 6 1946 475-493

58. W.J. McNamara, E. Weitzman; The Effect of Choice Placement on the Difficulty of Multiple Choice Questions; Journal of Educational Psychology; 36 1945 103-113

59. A. Marcus; The Effect of Correct Response Location on the Difficulty Level of Multiple Choice Questions; Journal of Applied Psychology; 47 1963 48-51

60. J.C. Jessel, W.L. Sullins; The Effect of Key Response Sequencing of Multiple Choice Items on Performance and Reliability; Journal of Educational Measurement; 12 1975 45-48

61. Curriculum Papers No. 7; Science for General Education; (H.M.S.O., Edinburgh, 1969)

62. R.F. Jarret; The Extra Chance Nature of Changes in Student Responses to Objective Test Items; Journal of General Psychology; 38 1948 243-250

63. R. Foote, C. Belinky; It Pays to Switch : Consequence of Changing Answers on Multiple Choice Examinations; Psychological Reports; 31 1972 667-673