SOME ASPECTS OF OBJECTIVE TESTING IN MATHEMATICS WITHIN THE FIELD OF FURTHER EDUCATION

by

NEVILLE KEITH UPTON, M.A., A.F.I.M.A

A Master's Dissertation submitted in partial fulfilment of the requirements for the award of the degree of M.Sc. in Mathematical Education of the Loughborough University of Technology, December 1977.

Supervisor: P. E. LEWIS, M.Sc., Ph.D.

ABSTRACT

The dissertation reviews the role of objective testing in mathematics at institutes of further education in the United Kingdom. The possibilities of usefully expanding this role are also discussed.

Within typical further education classes, students exhibit a wider variation in age, maturity, and mathematical ability than is seen in school classes; because of this, testing in colleges is more important, and serves a greater range of purposes, than in schools. The variation in mathematical ability is particularly pronounced in courses where mathematics is a service subject, and any means of rapidly locating areas of weakness early in the course are most valuable. In many colleges, the bulk of the mathematics teaching is of this type, and it may be partly for this reason that further education teachers of mathematics show at least as much interest in objective testing as do those of any other subject.

The discussion of the potential role of objective testing with further education students in mathematics is based largely on the published findings of prominent researchers in educational assessment methods. The literature, however, covers the wide field of education generally, and evidence based on the writer's own experience at Birmingham Polytechnic is therefore included; a brief account is also given of the practices and attitudes at certain other colleges.

Suggestions are offered regarding the use of objective tests at the beginning of, and throughout, each year of a course, and proposals are also made for introducing such methods into the formal end-of-session examinations, where at present they appear to be little

used. The complete replacement of conventional methods of examining is not suggested, but rather a combination of the two so as to exploit the various strengths of each method.

ACKNOWLEDGMENTS

I wish to thank Dr. Lewis for the valuable comments and suggestions he has made during the writing of the dissertation.

I am also grateful to the eight institutes of further education who have given information about their experiences with objective testing in mathematics, and to those of my colleagues at Birmingham Polytechnic who have co-operated in the work of evaluating multiple-choice tests in action; similar thanks are due to the staff at Loughborough University of Technology who administered one of my tests to some of their undergraduates.

Special words of thanks are due to my wife, Ruth, for her patience and help during the preparation of this submission.

CONTENTS

INTRODUCTION

Objective tests can be briefly defined as those in which no judgement is called for in the marking. One of the subjective elements usually present in examinations is thus avoided by the use of objective tests, but the writing and/or selection of the items (questions) is still subjective. The nature of objective tests, however, makes it possible to compile these from banks of items, each having its previous performance recorded. The examiner's own subjectivity therefore need not lead him to use tests which are too easy or too difficult; the level of facility, at least, can be objectively determined. The efficiency of an item in distinguishing between candidates of differing abilities can also be estimated. The subjective effect is likely to be further reduced in the case of an examiner using a bank, by the very fact of his being able to choose items written by others.

Of the several types of objective testing available, "multiple-choice" is the most widely used, and the dissertation deals almost exclusively with this type. In these tests, the candidate selects one of several (typically four) responses to a stem (question), knowing that there is one and only one correct response.

In further education, there is a considerable variety of courses calling for differing amounts of mathematics (taken here to include statistics and computing); further, within these courses there is a large amount of variation in the age, attitude and background of the students. For these reasons, frequent testing seems to be more necessary than in other educational establishments. Testing is carried out in F.E. for four distinct purposes, as follows:-

1. Selection tests are given to potential students, to help in deciding whether, and in what course, to enrol them;
2. Induction tests are given to whole classes soon after enrolment, and give an indication of the best way to treat the syllabus, and of the need for any remedial work;
3. Progress tests during the course help to monitor the learning process, and in fact contribute directly to that process;
4. Attainment tests at the end of each year assess the effectiveness of the individual's learning during the year.

Objective tests in mathematics are being increasingly used for the first three of these purposes, but are much less common in sessional examinations. In fact, I have yet to find any course at Birmingham Polytechnic or any of the local technical colleges where objective items are included in the internal sessional examination paper in mathematics. One reason for this is that external bodies such as the various joint committees for national certificates seem generally reluctant to depart from the tried and trusted conventional ("essay"-type) examination, and colleges naturally wish their own internal non-assessed examinations to serve as a foretaste of the final assessed one.

The dissertation is presented in six chapters, as follows:-

Chapter I contains a brief outline of the origin of objective tests, the philosophy behind their adoption in the field of educational assessment, and the various types of item available. This is followed by an explanation of the various numerical measures used to describe the characteristics both of individual items and of complete tests. These measures are dealt with in some detail as they play an important part in the interpretation of test results and in

the technique of selecting items from banks.

In Chapter II, an account is given of a typical procedure for objective testing in institutes of further education; this includes a description of item banking, since this system for pooling test material is essential to an adequate supply of items. Both the selection, and the modification, of items in the light of data obtained from their use are also discussed.

In Chapter III, educational objectives are briefly discussed and the characteristics of essay-type tests described in outline. Essay-type and objective testing methods are then compared and the conclusions presented in table form. Finally, some observations are made on short-answer questions, which represent a compromise between the two contrasting types featured in the table.

Chapter IV consists of an account of the experience at Birmingham Polytechnic with multiple-choice tests in mathematical subjects. The correlation between the results of such tests and the subsequent performance of the students is discussed, and some observations are also made on the correlation between the characteristics of the items of one such test when used with different groups of students. Finally, there is a report on some experiments carried out at Birmingham relating to open-book examinations, and to the advantage which may be enjoyed by test candidates when they have been taught throughout the year by the examiner while other candidates have been taught by a different lecturer.

To enable a comparison to be made between the methods and findings at Birmingham Polytechnic and those elsewhere, questionnaires were sent to a number of colleges in England contributing to the Manchester Objective Testing Item Bank. Replies were received from

eight of these, and the information given forms the basis of Chapter V, in which attitudes in general to objective testing are discussed.

Chapter VI comprises a summary of the dissertation and the conclusions to be drawn therefrom, together with predictions concerning the future of objective testing in further education mathematics courses.

Each chapter is provided with its own bibliography, the sources being listed in the order in which reference is made to them; the main bibliography is in alphabetical order of the authors' names.

To conclude this introduction, a list of abbreviations and their meanings is given below, together (where appropriate) with the subsection in which they are defined.

ONC	Ordinary National Certificate	
HNC	Higher National Certificate	
HND	Higher National Diploma	
FV	Facility value	(1.3.1)
ID	Index of discrimination	(1.3.2)
KR	Kuder-Richardson reliability factor	(1.3.3)
s	Sample standard deviation	
σ	Population standard deviation	
$\hat{\sigma}$	Estimate of σ made from sample	
r	Product-moment correlation coefficient	

CHAPTER I

THE ORIGIN AND NATURE OF OBJECTIVE TESTS AND THE EVALUATION OF THEIR CHARACTERISTICS

1.1 Origin

The use of numerous short questions in examinations, marked more or less mechanically, is at least seventy years old. In 1904, at the request of the French Ministry of Education, Alfred Binet began using such questions in an attempt to measure the educability of pupils. The multiple-choice question, in which several responses are provided and the candidate has to select the correct one, appeared shortly after this. The American Arthur Otis was one of its first exponents; his methods were used in 1917 by the U.S. Army to assist in the selection and classification of their men. Objective marking and wide coverage of subject matter were recognized as major advantages in both of these applications.

Many forms of objective test have since been used, all having the feature of speedy and reliable marking which requires no academic skill on the part of the marker. Sometimes the scoring calls for grading according to the numerical proximity of the candidate's answer to the correct one, but even here the judgement is exercised by the examiner and not the marker; the criteria are built into the marking scheme. In the last two decades, advantage has been taken of the advances made in computing by writing programs which mark the scripts mechanically and (perhaps more important) provide a detailed statistical analysis of the results.

In spite of this powerful advantage, and others described later, objective testing has not entirely, or even largely, replaced the

more traditional kind. Most GCE and CSE boards now incorporate a number of objective items in certain papers, and this is bound to lead to increasing use of the method in schools. It is right that the introduction of a new and controversial method of assessment should be gradual, and accompanied by experiment and research, and it is clear that objective testing can never examine all the facets which can be tested by essay-type questions; style of presentation is one such facet, and others are discussed in Chapter III. Nevertheless, the caution in the attitude of some educationalists to objective testing seems to me excessive. This may be the result of the emphasis placed on the harmful effects of poorly designed items, and of the imbalance which can occur in objective tests between the various skills being tested. While these dangers do exist, I suspect that they would be found to be at least as prevalent in essay-type examinations had these been subjected to as much intensive research as have objective tests.

Another factor which may be hindering the growth of objective testing is the belief - widespread among the lay public and even students, less so among experienced teachers - that objective tests are "easier" than essay ones. It is argued that a candidate needs more skill to be able to compose a good answer to a question than merely to select the correct response from those presented by the examiner; and if he has no idea at all, there is always the chance of guessing correctly. There is in fact little evidence to support the view that objective tests are too easy; indeed, it is my experience that the mean scores are usually lower than those obtained on essay tests judged to be of equal difficulty. It is true that

guesswork can be more fruitful in objective than in essay tests, and ways of minimising the effect of this are discussed in Chapter II.

1.2 Nature of objective and short-answer tests

In the following discussion, one subsection is devoted to each of the four main types of such tests in common use.

1.2.1

The multiple-choice type described in the introduction has become established as the most popular in the objective field, and is generally regarded as superior to all others. Bonney Rust[1] thinks it probable that "the multiple-choice test, with a minimum of four choices, is the most widely used and educationally respectable form of objective testing". Ebel[2] states that "multiple-choice items are currently the most highly regarded and widely used form of objective test item". Unlike Bonney Rust, however, Ebel considers that good items can be written with "only two or three alternatives". Gronlund[3] also considers the multiple-choice item the most widely applicable and useful form of objective item, and points out that "the use of a number of plausible alternatives makes the results amenable to diagnosis".

To appreciate fully the reasons for these opinions it is necessary to consider one of the tables featuring in the standard methods for presenting and analysing the results of multiple-choice tests, and then to compare this type of objective test with others. One of the most important of the tables used to display multiple-choice test results shows for each item the number of candidates choosing each response. Not only does this indicate the proportion of students choosing the correct response, but it also reveals the

extent to which they are attracted to each of the false responses (or distractors). The latter will have been chosen by the item writer in the light of known weaknesses in students knowledge, and the amount of attention which should be given to the correction of these misapprehensions (possibly with the class which has taken the tests and certainly with future classes) can be judged from this table. No other type of item can give such precise and compact information about the performance of a class on the test, and the chief reasons for failure to select the correct response. Because of this feature, satisfactory selection, induction and progress tests can in general be composed entirely of multiple-choice items in most subjects (and certainly in mathematical ones); attainment tests can also benefit from their inclusion, although part of such tests will have to consist of essay-type questions.

Multiple-response items, although superficially similar to multiple-choice, have more in common with the true/false type and are dealt with in the next subsection.

1.2.2

The true/false item has a stem consisting of a factual statement, and the candidate has to decide whether it is true or false. Items of this kind clearly have the greatest possible vulnerability to guessing. They are a special case of the multiple-choice item with two responses, and Ebel was quoted earlier as believing that good items with as few as two choices could be written. When the choices are true or false, however, Ebel refers to a loss in discrimination and an increase in ambiguity and misunderstanding. True/false items are no better at identifying students weaknesses

than short-answer questions (discussed below) and are less resistant to guessing.

Another type of objective item which has come into use is the multiple-response. This resembles multiple-choice, but the number of correst responses to an item is not restricted to one; it can be any number from zero to the number of responses presented to the candidate. Scoring presents a difficulty here. If each item carries one mark, a student who indicates, say, all but one of the correct responses is awarded a zero mark the same as a student who did not indicate a single one of them. On the other hand, if a mark is provided for each correct response, then a four-choice multiple response item is the same as four true/false items, with the disadvantages described above. Ebel considers that if "the statements were presented and scored as independent true-false statements, they would yield more detailed and reliable information concerning the examinee's knowledge than they can do in multiple-response form".

Multiple-response items therefore seem to be inferior to multiple-choice, and the scoring and the analysis of the results are more complicated. They are therefore not considered in this dissertation, although it should be recorded that certain professional bodies (the Institute of Medical Laboratory Sciences, for instance) use them in their examinations, as do the Open University.

1.2.3

Matching items provide a way of testing a number of pieces of knowledge with a single item. The item consists of two lists of words, formulae, dates or such other elements as the subject demands; to avoid the possibility of the last match being determined solely by elimination, the lists usually contain different numbers of elements.

Each element in the first list has to be matched with one in the second; sometimes an element can be used more than once. These items can be completely objective, and they offer little chance of successful guessing. A table showing all possible responses would however be impracticably large, and of negligible use except with the very largest classes; for instance, if the first list contained five elements and the second six, such a table would contain 30 columns.

1.2.4

Short-answer questions, and especially the "completion" type in which the candidate supplies a number, symbol, word or short phrase, are sometimes called objective although not in general conforming to the definition of mechanical marking; possibly this terminology arises from the property which such questions have in common with truly objective items - brevity, and the consequent possibility of wide syllabus coverage in a short time. It is therefore desirable to give some attention to this method of testing at this point.

Some short-answer questions are objective, but those requiring verbal answers are unlikely to be; synonyms are common in the English language and not unknown in mathematics, so some skilled judgement is called for in the marking. Truly objective completion questions have one advantage over multiple-choice items, however; the chance of successful guessing is less in the former. It is the ease with which the extent of the popularity of wrong responses can be measured and displayed with multiple-choice tests which give then a telling advantage over the short-answer type.

1.2.5

My conclusion from the foregoing discussion is that no type of

objective or short-answer test can have an overall efficiency greater than that of the multiple-choice.

1.3 Evaluation of items and tests

Objective tests, and especially those comprising multiple-choice items, are usually scored dichotomously; that is to say, each item is given either one mark or none. Such scoring would be quite inappropriate in the conventional examination, where each question requires a long answer, often subdivided, and typically marked out of 20 - the so-called essay-type question. Dichotomous scoring facilitates the calculation of three important measures which not only allow the test scores to be interpreted more precisely, but can also be useful when modifying tests for future use. These measures are described in the following three sections.

1.3.1

The Facility Value (FV) of an item is the proportion of candidates who gave the correct response to that item. Because this is equal to the mean raw score obtained by the class for this item, an equivalent figure could be determined for an essay question, or indeed for any item regardless of whether the scoring is dichotomous or not. The advantage of dichotomous scoring can be seen by considering an item for which FV = 40%. With dichotomous scoring, 40% of the candidates must have given the correct response to this item while 60% failed to do so. In an essay-type test, no such definite conclusion can be reached from a facility value of 40%. In one extreme case, 40% of the candidates could have scored full marks on the item and the rest zero - as described for the dichotomously scored case. In the other extreme, each candidate could have scored exactly 40% on this item. Generally, the situation would be some-

where between these extremes. The average success of the class in answering the question is indicated by FV in any test, but with non-dichotomous scoring this value gives no information on whether the item is detecting differences between candidates or not.

If an item is to be stored for future use (as in an item bank), its value is greater if its FV when used with a stated course is recorded alongside the item.

1.3.2

The Index of Discrimination of an item is a measure of how efficient it has been at distinguishing between the stronger and the weaker candidates as indicated by the results of the other items. The strict definition of this index is the correlation coefficient (see Appendix A) over the class taking the test between the scores on the item under consideration and those on all other items in the test. Because of the large amount of calculation required in obtaining this for every item, various simplifications have been devised. In 1939 an article by Truman Kelley[4] in the Journal of Educational Psychology included a proposal for basing the calculation on only a few scores obtained by the strongest and weakest candidates. Kelley showed that under certain conditions there was an optimum size for the upper and lower groups as a proportion of the number of candidates. Small numbers would produce a large difference in ability between the two groups but would allow large sampling errors within the groups; large number would reduce these sampling errors at the expense of the between-groups difference. The most reliable value for discrimination was shown to result from taking upper and lower groups of 27 percent of the total group, and Kelley further claimed that even when the conditions he had stated

did not apply, the choice of 27 percent was "ordinarily the most serviceable".

Of the simplified forms of discrimination index which have been based on Kelley's proposal, it is likely that the one now in the widest use is that due to A.P. Johnson[5], published in 1951 (again in the Journal of Educational Psychology). This index, denoted henceforth by ID, is defined as:-

$$\frac{n_c(U) - n_c(L)}{n_t(U)}$$

where n_c is the number of correct responses to the item given by the group specified and n_t is the total number of candidates in the group; U and L are equal-sized groups taken respectively from the upper and lower ends of the rank-order list of candidates. It will be seen that dichotomous scoring is implied in this definition, as otherwise the phrase "correct response" would lack precision. It is universal practice to use the rank order obtained from the results of the entire test; it is impracticable to obtain a different rank order for use with each item, as would be necessary if that item were to be excluded from the scores being ranked. This amount of approximation is seen to be acceptable when it is recognized that the index is in any case internal to the test, in that the rank order is not determined by any external assessment.

Ebel warns against the intuitive feeling that upper and lower groups of say 33 percent are better than 27 percent because of the larger size, or that 25 percent groups are better because of the larger difference between the abilities of the groups. He does however say that "Although ... groups of 27 percent are best, they are

not really much better than groups of 25 or 33 percent would be". It is in fact common practice, and one adopted in my department, to use groups of one-third rather than 27 percent when the number of candidates is less than about 20; the reduction of within-group sampling error is considered more important than the retention of large between-groups differences.

In selection tests, and others where the main aim is to establish a rank order, the discriminating power is at least as important as the facility value in determining the suitability of an item. Items with ID values of less than about 0.3 are usually regarded as unsuitable for such tests. For instance, items with ID = 0 are allowing the weaker candidates to score as many marks as the stronger ones, and so are not helping to rank the candidates correctly. Most authors agree on the critical ID value being around 0.3, but many pay little attention to a point which is equally important, namely that ID is of no relevance if the object of the test is to find whether the class as a whole has an adequate grasp of the subject matter. In this situation, items with facility values of 0 or 100% may give valuable information about the subject matter they are testing, but will all have ID = 0. Ebel is one author who draws a distinction between "relative achievement" tests (in which rank order, and hence ID, are important) and "content mastery" tests (in which they are not). He considers that the emphasis on the former type "seems reasonably well justified" because of the lower reliability achieved by the latter as a result of the retention of items with low discrimination. I believe that the varied background of F.E. students in this country gives more relevance to the content mastery test than Ebel has found in the U.S.

Values of ID less than 0.2 are quite common, and even the minimum of -1 is not unknown. In small classes, such values are likely to be due to sampling variation, but they warrant attention when they occur with large classes (or recur with several smaller ones). Sometimes the reason appears to be a peculiarity of the subject matter which allows some weak candidates to arrive at the correct response by superficial reasoning. The educational value of items of this sort is too great for them to be discarded merely to raise the average level of discrimination of the test; they draw attention to possible pitfalls. Some examples of this kind are given in Appendix C.

Where ID is below about 0.2, and especially where it is negative, it is instructive to examine the numbers choosing each of the distractors in multiple-choice items and note how these are distributed between the U and L groups as defined earlier. (It is not, however, practicable to arrange the data so that this can be done unless using a computer program offering this facility.) A distractor gaining much support from the U group has clearly revealed an area where further explanation and practice are necessary (assuming of course that the item has been well written).

The index of discrimination is usually treated in the literature as a characteristic of the item alone; certainly its dependence on the type of student taking the test is much less obvious than with FV. Ebel points out that ID is subject to sampling error, and that in the case of a test given to a small class this sampling error can be considerable. Where tests are given to only a few candidates, therefore, too much attention should not be paid to low ID values unless these are seen to be consistently low over several groups of candidates. On the other hand, an item for which a high ID value

has been established when used in one test will not necessarily discriminate so well if it is transferred to another test which differs in objectives, subject matter or overall facility. ID is not therefore a function of the item alone.

1.3.3

Since measures of facility and discrimination apply to the individual item, measures which evaluate a complete test must now be considered. The most important of these measure "test reliability". The reliability of a test is the extent to which the results are reproducible, perfect reproducibility being impossible because of the variability present in all measurements, including test scores. A score must be regarded as an estimate of a notional "true score".

Some measures of reliability use the analysis of variance technique (see Appendix A) to estimate how much of the variability between scores arises from the differences between candidates as distinct from that which occurs as a result of random errors in the responses made (such as those due to careless reading by the candidate or imperfect writing of the item). This approach leads to the following as the test reliability factor:-

$$1 - \frac{\sigma_e^2}{\sigma^2}$$

where σ_e^2 and σ^2 denote respectively the (population) error variance and total variance. In a perfectly reliable test σ_e^2 will be zero as there will be no random errors, so that the reliability factor will be unity; in an unreliable test this error variance will account for nearly all of the variability, making the variance ratio (i.e. the second term) so near to unity that the reliability

factor approaches zero.

Reliability factors derived from variance considerations are discussed fully by a number of authors, notably Ebel and Thorndike[6]. One of the most widely used forms is that given by the Kuder-Richardson[7] formula 20, namely

$$KR = \frac{n}{n-1}\,\frac{s^2 - \Sigma pq}{s^2}$$

where n = the number of items

s = the (sample) standard deviation of students' scores

p = the proportion of candidates answering an item correctly

q = 1 - p

Σ pq = the summation of the product pq over the n items.

Dichotomous scoring is implicit in this formula. The formula was first published in the September 1937 issue of Psychometrika in an article, "The Theory of the Estimation of Test Reliability", by G.F. Kuder and M.W. Richardson. Thorndike describes the formula as "the most generally useful of the formulas for estimating reliability from the relationship of total test variance to item variance"; Ebel states that this and the related Kuder-Richardson formula 21 "have become widely accepted as a basis for estimating test reliability".

The derivation of the Kuder-Richardson formula involves a number of assumptions which in practice are only partly justified. The user should not therefore be surprised at obtaining KR values which are negative; this can happen quite easily with small classes.

As an alternative to analysis of variance, the correlation coefficient can be used to measure reliability. The reliability factor of a test can be defined as the correlation between the scores obtained by a class on that test and the scores they would obtain on

a notional test which is equivalent to the actual one in facility, discrimination and reliability. The difficulties of measuring such a quantity in practice are obvious, and of the various methods which have been proposed only one will be described; this is the "split-half" method. Here the test is split into two sub-tests of equal size and the correlation determined between the sets of scores yielded by these sub-tests. This technique is sometimes used in the simple form stated, but it is clear that the decision as to how to divide the test is subjective; even with only 10 items there are 126 possible ways, and with 20 items there are 92,378. One way of eliminating this subjectivity is to obtain the arithmetic mean of the correlation coefficients arising from all sub-test pairs which could be formed from the given test. Clearly direct calculation is impracticable, but fortunately it is not necessary. It has been shown jointly by Nuttall and Willmott[8] that Kuder-Richardson formula 20 gives the mean of all possible split-half correlation coefficients. Because of assumptions which are seldom true in practice, KR is again only an approximation to this mean correlation coefficient, but the latter is useful to users of KR as an alternative means of explaining its significance.

1.3.4

These three measures - facility, discrimination and reliability - form the basis of records of item and test characteristics. All three are best regarded as only approximate guides, and especially so when based on tests given to fewer than about 20 candidates; this constraint applies less to FV than to D and KR.

To complete the treatment of methods of evaluation, another

useful statistic must be described. This is the "standard error of measurement", defined by

$$SE = s\sqrt{(1 - r)}$$

where s is the (sample) standard deviation of scores and r is reliability factor obtained as a variance ratio. SE can be used to define confidence limits for the "true score" of a candidate. This concept arises from viewing a candidate's score as one of an infinite number of measurements which could have been made of his ability in the field being tested; his true score will be the mean of all these measurements, and the latter will be normally distributed about that mean - that is to say, the scores of all the tests will follow the Gaussian curve of error (see Appendix A). While confidence limits found in this way will be familiar to many, a more generally understood measure is the "probable error", given by

$$PE = 0.6745s\sqrt{(1 - r)}$$

The true score is then as likely as not to lie within the interval bounded by the limits

$$\text{observed score} \pm PE$$

The concept of standard or probable error is useful in emphasising the inherent uncertainty in any test scores. In fact, since probable error can be explained so easily, it is a competitor to the reliability factor as a means of conveying to users of tests some information on how reproducible their results are. A further advantage of probable error is that it depends more on the test itself than on the group tested, whereas the latter has a significant effect on the reliability factor. Probable error is however the

less satisfactory of the two as an indicator of test reliability; F.M. Lord[9] has shown that it depends almost entirely on the number of items in the test, and is but little influenced by their characteristics. For this reason, it has not been included amongst the main characteristics described in sub-sections 1.3.1 to 1.3.3. 1.3.5

The recording of values for facility, discrimination and reliability allows future users of objective items to have an indication of their characteristics, and to use their results as a more precise measurement of the candidates' performances. (The values are also useful when modifying items for later use, and this point is further developed in the next chapter.)

Although it is possible to obtain measures of facilities, discrimination and reliability for essay tests, this is rather more difficult than with objective ones, and in practice it is seldom done other than by the major public examining bodies.

With these three measures described, we can now proceed to the consideration of objective tests in action; this is the topic of the next chapter.

BIBLIOGRAPHY FOR CHAPTER I

1. Bonney Rust, W. (1973). Objective Testing in Education and Training. Pitman.

2. Ebel, R.L. (1972). Essentials of Educational Measurement. Prentice-Hall.

3. Gronlund, N.E. (1965). Measurement and Evaluation in Teaching. MacMillan.

4. Kelly, T.L. (1939). "The selection of upper and lower groups for the validation of test items", Journal of Educational Psychology, 30, 17-24.

5. Johnson, A.P. (1951). "Notes on a suggested index of item validity", Journal of Educational Psychology, 62, 499-504.

6. Thorndike, R.L. (1951). "Reliability". In Educational Measurement, edited E. Lindquist. American Council on Education.

7. Kuder, G.F., and Richardson, M.W. (1937). "The theory of estimation of test reliability", Psychometrika, 2, 151-60.

8. Nuttall, D.L., and Willmott, A.S. (1972). British Examinations. National Foundation for Educational Research in England and Wales.

9. Lord, F.M. (1957). "Do tests of the same length have the same standard error of measurements?", Educational and Psychological Measurement, 17, 501-21; and

(1959). "Tests of the same length do have the same standard error of measurements", ibid, 19, 233-39.

CHAPTER II

OBJECTIVE TESTS IN PRACTICE; ITEM BANKS

2.1 General

The very small amount of writing which an objective item requires of a candidate makes it possible to set a large number of items in a short time; a test of 20 items, for instance, can be conducted well within a one-hour class period without even a slow student feeling deprived of time. This feature of objective testing makes it possible to give informal tests to classes quite frequently (at least twice a term, say) without interfering too much with the teaching programme, and to cover large sections of the syllabus in the process. Provided full advantage is taken of its objective nature, the marking of a test can also be carried out quickly, without any excessive demands on the teacher's time.

2.2 A typical test procedure

A number of methods are available for setting and marking tests, recording results, and informing the students in a way which forms part of the teaching process. The procedure used by some colleagues and myself in the Department of Computer Studies and Mathematics at Birmingham Polytechnic will be described, as this is reasonably typical of objective testing practices in further education generally; unless otherwise indicated by the context, "Birmingham" henceforth should be taken as referring to this department.

2.2.1

At Birmingham, each candidate is given a question paper on which he records his responses. The rubric states that there is one and

only one correct response to each item, and advises students to avoid blind guessing, but to guess intelligently if not sure of the right response; it states that there might be a small mark penalty for wrong responses but none for an omission. By each item there is a set of boxes lettered A, B, C and D (and occasionally E, but five-response items are seldom used), and the candidate places a tick in the box of his choice. At the end of the question paper there is another set of boxes marked with the item numbers; the candidate completes this by entering for each item the letter of his chosen response, or an X (for "omit") in each box where he has not chosen a response. Having the responses recorded in two different ways not only facilitates the punching of the computer cards, but also provides a cross-check which is most useful when the question paper has not been clearly marked.

2.2.2

The examiner will have made himself a marking template consisting of a copy of the test paper mounted on card and with a hole punched in the correct response box for each item. Using a red ball pen, he draws circles on the scripts through the template, each circle thus surrounding either a tick or an empty box. For each candidate, he records on the front of the script the number of correct responses, the number of wrong responses, and the number of omissions.

2.2.3

The raw score is of course the number of correct responses, but if the examiner wishes to apply a "guessing correction" he will deduct from this one-third of the number of items for which incorrect

responses were given (or one-quarter if there were five responses per item). The justification for this is that for any items where a student guesses blindly, the expected score, as a percentage of the number of such items, will be 25% (for the 1 in 4 where his guesses were right) minus one-third of 75% (for the 3 in 4 where his guesses were wrong) - namely zero, as justice demands. It is usual to point out to classes, when intending to use this correction, that students stand to gain by guessing blindly between two or three responses if they can confidently reject the other one or two; if they can see a reason for preferring one of the possible responses to the others, then their chance of gaining is further increased. Guessing in this way is different from the completely blind guessing advised against in the rubric.

Application of this correction in practice almost invariably reduces the mean score and increases the standard deviation. It often has little or no effect on the rank order, and so ID is largely unaffected. FV is completely unaffected. Since the scoring is no longer dichotomous, the correct formula for finding the reliability factor is more involved than the Kuder-Richardson one. The computer programme used at Birmingham only gives the KR value whether guessing correction is used or not. The effect on KR of using this correction arises wholly from its tendency to increase the standard deviation; this effect is to give a higher KR value, which is in keeping with the greater discrimination between candidates afforded by the guessing correction. Because of the variable nature of KR with classes of fewer than about fifteen, the rough indication given by KR with guessing correction has so far been considered adequate.

There is some controversy over whether or not a guessing correction should be applied. One argument sometimes used against it is that a wrong response is thereby assumed to be a guess, whereas it might be a considered judgement, although incorrect. This objection is not valid; it could equally be argued that a correct response is assumed to be a considered judgement whereas it might be a guess. The reasoning behind the correction is probabilistic and the uncertainty of any test score should be recognized; the guessing correction is a logical attempt to obtain an unbiassed estimate of the true score, and without such a correction the score is likely to be biassed upwards by the effect of guessing.

Ebel[1] discusses guessing correction very fully. He points out that the correction seldom makes much difference to the rank order; that the chance of blind guessing giving a "respectable" score is extremely small; and that rational (as distinct from blind) guesses can provide useful information on candidates and so should not be discouraged. To these observations of Ebel's can be added the thought that guesswork plays a part in most real-life decision-making, and so it would be undesirable to try to ban it in tests and unrealistic to try to enforce such a ban. On the whole, however, Ebel concludes that there are no strong arguments for or against the use of a guessing correction.

Gronlund[2] considers the idea of instructing candidates to answer all items, and stating that wrong answers will not be penalized; this prevents the bold guesser from having an unfair advantage over the more cautious student who might otherwise feel inhibited from guessing when he is not sure. He does recognize

the objection of some teachers that this system could encourage blind guessing, which is educationally undesirable. He accepts as a sound compromise the use of a guessing correction together with a warning of this and a recommendation to use reasoned guessing as distinct from wild guessing (exactly as is done at Birmingham). Gronlund reaches a conclusion similar to that of Ebel given above, but argues against the correction in "power" tests (when time is virtually unlimited), while not opposing it in "speed" tests (where some candidates do not have time to consider all items and so might tend to guess blindly if there is no penalty for wrong responses). At Birmingham, however, the preference is for power tests with guessing correction.

Referring to objective tests on U.S. Army trainees at the State University of Iowa in 1944, Ebel points out that the highest reliability factors were obtained when no guessing correction was applied but the candidates had been told otherwise and advised to avoid blind guessing. He concedes that such deception can only be effectively practised once or twice on one class - a restriction which makes it virtually useless for progress tests, and tends to support our preference for using the guessing correction.

2.2.4

Another departure from dichotomous scoring is open to the examiner; this is differential weighting. Objective items vary in difficulty, in length of time required, and in the depth at which they test the candidate's potential or attainment. It is therefore not obvious why it is common practice to give each item the same maximum mark of 1, and not surprising that research workers

have given attention to the question of whether to depart from this practice by varying the maximum marks of items.

Ebel considers that differential weighting, like guessing correction, has little effect and that there are no strong arguments for or against it; he quotes Wilks and Aiken as reaching similar conclusions. Sabers and White are reported as finding "... not only that there is little to be gained from weighted scoring, but also that, from the point of view of test construction, weighted scoring is probably not worth the effort. The same advantage can be gained by adding more items or by selecting only the best items from a larger pool. From the administrative point of view, unweighted scoring saves time and offers fewer possibilities for errors in calculating the scores; in addition, the resulting raw scores are probably easier to interpret".

The question of objectivity is raised by the suggestion that more items can be set. If weighting is used, the decision as to which items to weight and by how much is essentially subjective - if, in an attempt to use an objective basis for weighting, this is based on FV or ID, the effect is likely to be merely an increased dispersion with no change in rank order. Choosing more items on the topics to which the examiner would otherwise give greater weight is only slightly less subjective than weighting existing items, although the inclusion of more, even subjectively chosen, items will tend to raise the reliability of the test; there are therefore solid grounds for preferring the "extra item" policy to that of weighting.

The reduced possibility of errors in scoring when weighting is

not used is of only marginal importance when using a computer program which has a weighting option; mistakes in hand scoring will be revealed by the computer.

The final point quoted from Sabers and White, that raw scores are easier to interpret, takes on further significance when using the Kuder-Richardson formula to find the reliability factor; this formula assumes dichotomous scoring. When weighting is used at Birmingham, an item given a weight of n is treated as n items, and the Kuder-Richardson formula will treat its score as that of n dichotomous items although no score between 0 and n (exclusive) is possible. The KR value is therefore subject to the same weakness as when guessing correction is used, although the same defence holds in both cases - the higher KR value resulting from the higher dispersion is consistent with the greater discrimination.

The "extra item" policy, without differential weighting in the scoring, is the one favoured at Birmingham.

2.2.5

After marking the scripts, the examiner arranges for the students' responses to be punched onto computer cards, together with test details which of course include the correct responses. A feature of the program is that the examiner may also have a "match mark" punched on each student's card, and the program will then calculate and print the coefficient of correlation between these match marks and the marks given by the computer. This facility can be used in either of two ways, as follows:-

(a) If the examiner enters as match marks the scores he has given as a result of his template marking, then the correlation should be 1. If it is not, there has been a mistake in his marking or in the

punching; if it is 1, his marking is almost certainly correct, although a punching error which has led to a wrong response being shown for a candidate who has given another wrong response will not be detected. As a result of the use of this check, it has been found at Birmingham that marking errors are rare and punching errors virtually non-existent.

(b) The examiner may use as his match marks any other numerical assessment of the class - the scores obtained, say, at an induction test, or on last year's sessional examination. He may even wish to obtain the correlation with scores obtained in a different subject. (The assessment need not be a score at all; it can be a rank order.)

The cards are run as data on a Fortran program stored on the Polytechnic's computer; details of the program, and a specimen print-out, are included in Appendix B.

The use of the correlation coefficient facility, although almost standard practice at Birmingham, is in fact optional, and is called for by using the letter C on one of the data cards; other options (listed for convenience in the order, and under the letter, required by the program) are

P : changes the proportion in upper and lower groups from 1/3 to 27%

Y : eliminates items specified by the examiner

W : weights items as specified by the examiner

L : lists candidates' results as many times as the examiner wishes (one list being given if the option is not exercised)

n : applies guessing correction of -1/n per wrong response; n = 3 for the usual four-choice-per-item test

The options W and L have been described. The elimination option Y is useful when certain items are not relevant for the class taking the test or have proved generally unsatisfactory; alteration of the key card and the item numbering is thus obviated.

The print-out comprises nine tables, and a specimen is given in Appendix B, together with an explanation of the tables. Chief among the values given are:

FV, ID, and response analysis for each item, and
mean and standard deviation of the scores,
reliability factor, probable error, and correlation coefficient.

Nearly all mathematics classes at Birmingham meet once a week, so at the class meeting following the test the teacher can return the marked scripts and also draw attention, using the computer print-out, to any special points, such as items with a low FV, or distractors which attracted a disproportionately large response. The students are allowed to keep the corrected scripts, and this is believed to make the tests a valuable part of the teaching process as well as a source of feed-back to the teacher. The scripts form a permanent record relating to a large part of the syllabus, showing each student his response (ticked) and the correct response (red circle).

2.2.6

The retention of the scripts by the students means that whenever a test is re-used there is a possibility that some of the candidates will already have seen a corrected version of this. Some of my colleagues have misgivings about this lack of total security, but the majority view (which is mine also) can be summed up as follows:-

(a) Unless the scripts can be kept by the students, so as to augment their lecture notes, the amount of class time and staff effort consumed by the tests would not be justified.

(b) The large number of items per test, and the small amount of contact between students on different courses (many being part-time), make the possibility of any significant leakage of information rather remote.

(c) Even if some students do gain better scores than others as a result of leaked information, this is of no great importance as the marks are not used for continuous assessment of individuals.

(d) A plentiful supply of items makes possible the provision of so many equivalent tests that the risk of leakage is negligible; there would be too many tests for students to gain significantly by parrot-like learning of the correct responses. This is believed to be true even if the tests were part of continuous assessment or of sessional examinations, where the need for fairness is paramount.

2.3 Item Banks

These considerations lead to the conclusion that colleges using objective tests extensively need to be able to call on a great many items. This is a problem, as the writing of good items is not easy; not only is skill and experience needed, but some form of vetting (or "shredding") is required. Ambiguity is more serious than in essay questions, where a candidate's misunderstanding can be identified and allowed for. Each distractor must be definitely wrong. It must not be possible for a candidate with no knowledge of the topic being tested to reject some responses from purely logical considerations.

An item that avoids all these pitfalls must still test something useful. A co-operative effort by item-writers is necessary, and with the increase in specialisation in so many further education courses, this usually implies inter-college liaison. The logical outcome of this is the item bank.

2.3.1

The practice of large organizations (such as the National Foundation for Educational Research) maintaining a pool of items from which tests could be extracted, or new tests constructed, is not new, but the whole concept received a fresh impetus in 1966. The Schools Council asked the N.F.E.R. to investigate the introduction of item banks suitable for use with 16-year-old candidates; the intention was that teachers could draw on the bank in setting school-based examination for the Certificate of Secondary Education.

The project is of little direct relevance to this dissertation, but it may well have contributed to the increasing interest in item banking. Of special importance in the field of further and higher education is the formation of less formal banks, each specializing in one of the main disciplines (mathematics, physics, chemistry, etc.) and allowing for the pooling of items written by a number of institutions. These banks have one vital feature in common with the nationally established ones - the wide availability of a great many items, catalogued so as to facilitate location of subject areas and each with its performance recorded.

Suitability for storage in banks is not peculiar to objective items; essay-type questions can be banked in a similar way. But an objective item generally covers a narrower field of subject area than an essay-type question and so can more easily be given a

catalogue reference; and further, performance data is more easily found for objective items, and is more informative.

2.3.2

My department is a member of the Manchester Objective Testing Item Bank, based on Manchester Polytechnic and dealing with mathematical subjects. Its method of operation is typical of banks which depend on inter-college co-operation.

The bank receives from its member colleges a steady flow of their items, which have (preferably) been used already so that FV and ID values can be supplied, together with the number of candidates, the course they attend, the number choosing each response, and the correct ("key") response. A panel of teachers study the items and supporting data, and each that is considered satisfactory is entered in the bank under a number which includes the Dewey decimal classification for the topic which it covers. Items which are not immediately acceptable are either modified and entered, or returned to the writer with the reason for rejection and (if appropriate) a request for it to be re-submitted after alteration. Bank members are sent all new items at suitable intervals, so that their holdings are kept up to date. Since a bank will confine itself to one subject, each member "college" tends to be a department.

A department participating in a bank thus has at its disposal far more items than it could reasonably expect to write on its own. Apart from the obvious advantages - economy of effort, wide choice of items, infrequent repetition and hence reduced security risk - the cross-fertilization effect is beneficial; other people's items often show a different emphasis or technique, and not only are these available for use but their study tends to stimulate more inventive

item-writing.

Each item is typed on a separate sheet of A4 paper, with the response grid adjacent to the item, and a table giving item analysis and other data including the correct response is at the bottom of the page. This arrangement enables items to be grouped in any order and photocopied four or five to a page, each analysis table of course being concealed by the sheet bearing the next item. Strict adherence to "house style" by members typing items for submission to the bank is therefore important, but the resultant saving when they use the bank is ample reward. Our own item writing proceeds continuously, but items from the bank form an increasing proportion of our tests. We send the bank about 12 items per term, and our holding of bank items is at present about 1000. About one-third of these have no immediate application in the polytechnic, but the remainder represents nearly ten times as many items as we have written ourselves.

In using items from the bank, a balance has to be struck between two extremes. On the one hand, security problems and the effect of teacher subjectivity could be minimized by making up each test anew by using random numbers to select items from each subject area; this however would entail a great deal of work for both academic and clerical staff, and would prevent meaningful comparison of mean scores, mean ID, and mean KR between classes and between tests. On the other hand, one test could be compiled for each stage of each course and used year after year; this would merely transpose the merits and demerits of the former strategy. At Birmingham the following pattern is emerging as a suitable compromise. Pairs of tests are made up, tested in class use, and modified where necessary to make

each test equivalent in mean score and mean ID with its "partner", and a random choice made between the two when a class becomes due for that test. These tests are updated at intervals of about two years, not more than approximately one-fifth of the items being replaced or improved at any one time; security can be further improved without loss of comparability by interchanging similar items between partner tests at the same time.

It can be said with certainty that the existence of item banks offers colleges a powerful new means of assessment. The greater frequency of testing which the banks make possible helps to reduce the tension which tests induce in some students, so that the information fed back to the teacher is more accurate than would be expected with infrequent testing; it is at least possible also that apprehensive students may approach their sessional examinations in a more relaxed frame of mind if they have been exposed to a number of informal tests during the year.

2.3.3

It was stated in the previous section that items should preferably have been used in class tests before submission to the bank, so that performance data can be provided. In my opinion this should be a strict requirement, but in practice some items appear in banks without such data and so may never have been used in tests at all.

The use of items in class tests before submission serves another purpose; it allows items to be improved in the light of the information given by computer analysis of the kind described in section 2.2.5. It follows that items modified in this way should be used again in tests, and their performance noted, before being sent to the

bank. The technique of modifying items in this way is one of the subjects of the next section.

2.4 Selection and revision of multiple-choice items

With the results of the first use of an objective test in front of him, the item-writer can consider modifications. Facility values and indices of discrimination may influence his decisions, but the item analysis (showing how effective each distractor has been) will usually play a greater part.

The criteria to be adopted both for revising items, and choosing them from a bank, will depend on the purpose of the test, and to illustrate this some examples from the literature are discussed below.

2.4.1

Opinions differ on the question of the best arrangement of facility values. Fraser and Gillam[3] dispute the claim that "for maximum spread of results a facility value of 0.5 is required for each item", arguing that this could result in the candidates being divided into two equal groups, one group having a zero score and the other 100% each. This is a highly theoretical possibility and I do not think it merits serious consideration. The authors are on firmer ground in advocating the inclusion of a few easy items (FV over 85%) "to allow candidates to make a confident start". I can also see an argument for including some very difficult items, to extend the more capable candidates and to promote a discussion of some of the more advanced aspects of the subject. The policy of having a roughly normal distribution of FV with a wide range is therefore defensible for progress tests, where the items must be generally acceptable to the students if the learning process is to be assisted by their use.

Ebel on the other hand argues against such variability in FV, and demonstrates that a large spread in FV reduces the variance of the scores. Using items whose facility values for 300 college students were already known, he assembled three tests and obtained the following results:-

Test	Range(s) of FV	Mean score (%)	Standard deviation of scores (%)	Reliability factor
1	40-55 %	58	17	0.485
2	10-85 %	50	14	0.426
3	10-30 % & 80-90 %	50	10	0.013

The poor reliability of test 3, which had a bi-modal, wide-range, distribution of FV, shows such tests to be highly unsuitable for selection or attainment purposes, and not really suitable for induction tests, where an item with an inherently high, or low, FV will give a false impression of the class's knowledge on that topic. For progress tests, however, a low reliability is less of a disadvantage since the correction of mistaken ideas is more important than the individual score.

My conclusion from Ebel's experiment is therefore than when compiling tests for purposes other than measuring progress, it is preferable to select items which have facility values around 50%, while the index of discrimination should not be less than about 0.3 (see subsection 1.3.2).

A further conclusion, and one of relevance to the question of revising items, is that sound items are worthy of inclusion in a

bank subject to the following conditions:-

(a) The facility value, whether or not in the vicinity of the popular value of 50%, must be stated along with the level of the course with which it was obtained.

(b) The index of discrimination, whether high or low, must be stated.

(c) The item must be well-written in the sense that all its distractors are evoking a reasonable amount of support from the candidates.

Users of a bank formed in this way can then choose items having whatever FV and ID values are appropriate to the users' current needs.

2.4.2

From the previous subsection, it is only to be expected that Ebel's policy in revising items is to aim at a facility value close to 50%, an index of discrimination of at least 0.3, and distractors which are chosen by roughly equal numbers of candidates (that is, about 17% of those answering four-choice items). He presents five items which initially performed unsatisfactorily, and describes the attempts (usually successful) to produce the desired characteristics. Two of these have been selected as relevant to this dissertation; although one of them is not mathematical, the principles involved are the same.

(a) I (Original form)

What, if any, is the distinction between climate and weather?

A: There is no important distinction.

B: Climate is primarily a matter of rainfall, while weather includes many other natural phenomena.

C: Climate pertains to longer periods of time than weather.

D: Weather pertains to natural phenomena on a local rather than a national scale.

Results: FV = 36% ID = 0.13

		Item analysis			
Responses (correct one starred)		A	B	C*	D
Frequencies	Overall	7	84	73	36
	Upper half	1	33	43	23

Both FV and ID were considered too low, and the uneven distribution between distractors caused concern. Using the median score obtained by 200 high-school students on the entire test as a criterion, 33 of the 84 choosing response B had exceeded the median score, as had over half of those choosing D (23 out of 36). With a view to improving this situation, distractors B and D were modified as follows:-

II (Modified form)

B: Climate is primarily a matter of rainfall while weather is primarily a matter of temperature.

D: Weather is determined by clouds, while climate is determined by winds.

Results: FV = 62% ID = 0.58

		Item analysis			
Responses (correct one starred)		A	B	C*	D
Frequencies	Overall	24	28	124	24
	Upper half	2	3	91	4

The changes have been highly effective in achieving the desired results, and no doubt the revised item II is more suitable than I for use in selection and attainment tests. Discrimination is high, and the distractors were chosen by only 9 of the students with above-median scores. However, I suggest that, for induction and progress testing, I is the better item; the distractors represent more reasonable beliefs than those in II, and the fact that a number of students with above-median scores did not know they were wrong indicates the need for the correction of these beliefs.

(b) I (Original form)

What is the maximum mechanical advantage obtainable with a single fixed pulley and a rope that will break under a load of 500 pounds?

A: 1

B: 2

C: 500

D: 100

Results: FV = 12% ID = 0.22

		Item analysis			
Responses (correct one starred)		A*	B	C	D
Frequencies	Overall	23	50	78	49
	Upper half	22	20	38	20

Although the index of discrimination is a little low, the distractors are working fairly evenly, and all were marginally more popular with students having below-median scores on the whole test. The item writer, however, considered the facility value to be far

too low, and attributed this to "the abstract nature of the concept" of mechanical advantage. He therefore rewrote the item in more concrete terms, thus:-

II (Modified form)

A workman lifts planks to the top of a scaffold by pulling down on a rope passed over a single fixed pulley attached to the top of the scaffold. The rope will break under a load of 500 pounds, and the workman weighs 200 pounds. What is the heaviest load the workman can lift with the pulley?

A: 100 pounds

B: 200 pounds

C: 400 pounds

D: 500 pounds

Results: FV = 38% ID = -0.07

		Item analysis			
Responses (correct one starred)		A	B*	C	D
Frequencies	Overall	7	77	55	61
	Upper half	1	35	32	32

Ebel attributes the negative discrimination now obtained to the fact that not only is II easier than I, but the correct response is much more obvious to the weaker students than it was in I (42 against 1) and only slightly more so to the better ones (35 against 22). He describes the problem situation of II as "just complex enough to mislead the good students, while being fairly simple on a superficial basis to the poor students".

By implication, then, Ebel condemns the revised version, but he has not referred to its much altered nature. Both I and II require a general understanding of mechanical advantage and a realisation that the breaking strength of the rope is irrelevant in the situations described; to answer II, however, the candidate needs to know that the greatest effort which can be applied to the rope below the pulley is the workman's weight, but on the other hand he does not need to understand the technical term "mechanical advantage". Abstract it may be, but this latter concept, together with its terminology, is of fundamental importance and a proper subject for testing.

It seems therefore that I is highly suitable for selection, induction and progress testing, and not entirely out of place in an attainment test in spite of its low FV; and that II should only be used in progress tests, where its negative discrimination and uneven distribution of response to distractors would not constitute serious drawbacks - indeed they may help to stimulate discussion and rectify misconceptions. In other types of test the modified item II may lead the college to make wrong decisions about how to deal with individual students or with the teaching of the class.

2.4.3

The policy at Birmingham is to revise items only if this is necessary either to remove ambiguities and similar flaws or to rectify a grossly uneven response to the distractors. Far from trying to change items with high or low facility values, we regard these as giving valuable information; it is a fact however that items we have written tend to have an <u>average</u> FV of around 50%. As stated in subsection 2.4.1, items revised (if necessary) in this way

are considered suitable for banking, and seem to be acceptable to the Manchester Objective Testing Item Bank.

Sufficient has now been said on the subject of objective testing for this method to be compared with essay-type testing, and this is done in the next chapter.

BIBLIOGRAPHY FOR CHAPTER II

1. Ebel, R.L. (1972). Essentials of Educational Measurement. Prentice-Hall.

2. Gronlund, N.E. (1965). Measurement and Evaluation in Teaching. MacMillan.

3. Fraser, W.G., and Gillam, J.N. (1972). The Principles of Objective Testing in Mathematics. Heinemann.

CHAPTER III

COMPARISON OF OBJECTIVE AND ESSAY-TYPE TESTS

3.1 General

In the debate on objective testing, some of the claims in its favour have been over-enthusiastic; it has even been suggested that this method of assessment is so superior to essay-type testing that it can supersede the latter. On the other hand, opponents of objective testing have condemned it as being unable to test any but the most basic skills; Professor L.R.B. Elton of the University of Surrey has quoted one critic as claiming that multiple-choice tests "sap the strength and vitality of a nation"!

One topic which needs to be considered before the relative merits of the two types of test can be analysed is that of educational objectives and their classification, since testing should be carried out in the light of the objectives which the course is intended to achieve.

3.2 Educational objectives.

The classification of educational objectives is too large a subject to be discussed in detail in a dissertation on objective testing in further education; therefore only those aspects which are relevant to the main theme are dealt with here.

3.2.1

In his "Taxonomy of Educational Objectives", Bloom[1] sets out six major areas of skills which can be furnished or developed by education; he names these objectives as follows:-

Knowledge

Comprehension

Application

Analysis

Synthesis

Evaluation.

There is a certain amount of overlap between these areas, and this has a bearing on the writing of examination papers, whether essay-type or objective. An examiner should have an idea of the relative importance to be attached to these skills in the course for which the paper is being set, and should try to reflect this in his examination. In addition to the natural overlap between the skills listed in Bloom's classification, there is a further complication in that different candidates may use different combinations of skills in solving the same problem, or even in answering the same multiple-choice item. The more advanced the level of work, the greater this difficulty becomes. It is thus less easy to classify test material in further education than in, say, junior schools.

Another consideration is the varied background of further education students. To take an extreme case, one student may be able to answer a certain question purely from his experience of that topic, so that he is using knowledge and comprehension; another student, never having encountered such a problem, may need to use synthesis and evaluation, and possibly some of the other four skills, in answering the same question. This applies both to essay-type questions and to multiple-choice items. The test constructor in further education can therefore only have a very general idea of the skills being tested by each question.

Partly for these reasons, many authorities have shortened Bloom's list by omitting and/or combining some objectives. The Joint

Matriculation Board,[2] for instance, suggest for advanced level science subjects a list which combines the last three of Bloom's objectives into one class, namely "Evaluation and investigation". The City and Guilds of London Institute find it sufficient to use only the first three of Bloom's objectives.

Many other variations are in common use, but the one given below seems to cover the needs of test constructors in further education and will be adopted throughout this chapter. It is used by London University and the Associated Examining Board in their combined scheme for advanced level economics.

Recall of factual knowledge and understanding

Application

Analysis and evaluation.

3.2.2

To illustrate the classification described above, and the associated difficulties, three multiple choice items are now presented and discussed.

(a) The following expressions relate to data from a sample of n values. Which is the best estimator of the variance of the population?

A: $\frac{\Sigma x^2}{n} - (\bar{x})^2$ B: $\frac{(\Sigma x)^2}{n} - \frac{\Sigma x^2}{n^2}$

C: $\frac{\Sigma (x - \bar{x})^2}{n - 1}$ D: $\frac{(\Sigma x)^2}{n - 1}$

Basically, this item tests recall of knowledge. A candidate who remembers the definition of variance and the need to divide the sum of squares by n - 1 when estimating this parameter will recognize

option C as the correct response. One who can only recall the more commonly used computational version, namely

$$\frac{n\sum x^2 - (\sum x)^2}{n(n-1)}$$

may proceed by eliminating options A and B on the grounds of an incorrect divisor, and D because it does not contain the sum of squared values; he may then confirm his reasoning by expanding the form given in option D. This student will then be, to a limited extent, using application (of logic and algebra).

(b) If three of the roots of a certain quintic equation (with real coefficients) are known to be complex, then

A: there are no other complex roots

B: there are no real roots

C: there are two other roots, one real and one complex

D: nothing can be said about other roots without further information.

Since it is unlikely that a candidate will remember the specific case of a quintic equation with only one real root, he will have to use his knowledge of the number of roots which a polynomial equation has, and of the fact that in this example complex roots can only occur in pairs; logical application of this knowledge is then required, leading to option C as the correct response. It is possible for a student not aware of these two facts but with some knowledge of algebraic equations and complex numbers to deduce the solution, thus using analysis (in the sense used by Bloom).

(c) A certain hypothesis is being tested for statistical significance. If this hypothesis is in fact true, the results of rejecting it could be disastrous. Which of the following levels

of significance would be the most appropriate in this test?

A: 0.1% B: 1% C: 5% D: 25%

In this item, the candidate has to deduce that the correct response is the level which is least likely to lead to rejection; this calls for the application of logic. Analysis and evaluation of the implications of a sample result which differs from the null hypothesis, and of the associated probability, are also required. (Knowledge and understanding of the terminology and principles used in significance testing are of course equally necessary.)

These examples show the hierarchical nature of classifications of objectives. Complexity increases with progress through the list, and at a given level any skills appearing earlier in the list may be included. The above items have been used at Birmingham with a number of courses of varying mathematical ability. The individual classes have been too small for the indices of discrimination to show any meaningful pattern, but facility values are less susceptible to sampling variation, and they can be averaged. With only three items, even the FV pattern may be fortuitous, but it is recorded below as being typical of results in general; items testing the higher skills tend to be more difficult.

Item	Highest skill tested	Facility values (%)		
		Range: From	To	Mean
(a)	Recall of knowledge	28	60	46
(b)	Application	10	50	30
(c)	Analysis	15	38	28

3.3 Essay-type tests

Objective tests have already been described, but before comparing them with the essay type, the implications of using the latter as a method of assessment are briefly discussed. Some writers classify tests as objective, essay and problem. In this dissertation, the essay type is taken to include the problem type. This is because the answers to most problem-type questions include some descriptive material, and even those which do not (such as a straightforward calculation or proof in mathematics) cannot be scored dichotomously and so have much more in common with essay-type questions than with objective items.

Most of what has been written in earlier chapters applies to educational testing in general; this comparison however deals primarily with mathematical subjects.

3.3.1

An essay-type test usually consists of questions whose answers require between about 15 and 45 minutes each. The answer may indeed be an essay in the everyday sense of the word, but in mathematics it is more likely to be in several parts, consisting of calculations, proofs, interpretations, conclusions and possibly graphs or other diagrams. A considerable amount of writing is therefore required, and so if a reasonable amount of subject matter is to be covered by the candidate the examination will last for at least $1\frac{1}{2}$ hours and more likely for 3 hours; the number of answers required is usually between three and ten, five being a typical figure.

If the questions are of the "structured" type, each part will be more difficult than the preceeding one, and its solution may depend on at least partial success in previous parts.

3.3.2

In marking essay-type questions it is usual for a detailed scheme to be used, allocating a certain maximum score to each part and sometimes subdividing this allocation. Such a scheme makes possible reasonably consistent and reproducible marking, except where the candidate has used an unexpected method or has misunderstood the question. In practice, many answers present difficulties by not conforming to the method to which the scheme relates. Here then are the two areas in which subjectivity enters into marking such tests - the construction, and the interpretation, of the marking scheme.

3.3.3

Because of the depth to which essay-type questions test knowledge and ability, it is normal practice to allow the candidate to choose which questions he answers - a typical example being five out of eight. If the marking reveals significant differences in difficulty between questions, it is possible to allow for this by reducing the weighting of marks given to the harder questions - but only if all questions have to be attempted; the element of choice present in most essay-type examinations prevents any such adjustment.

Closely related to this point is the loss of comparability between scripts; there are 56 different ways of choosing five questions out of eight, for instance, and so with a class of 20 (a fairly typical size in further education) all the candidates might have answered different sets of questions. In fact, Bonney Rust has pointed out that two candidates might submit scripts which, as a result of the choice offered and the fact that they may not have attempted the maximum permitted number of questions, do not overlap

at all; the examiner is then trying to give comparable marks on the basis of non-comparable examinations.

One effect of allowing choice is therefore that values for FV, ID and reliability lose much of their significance; each may be quite different from the value which would have been obtained had it been obligatory to answer every question. (As mentioned in sub-section 1.3.5, these statistics are seldom evaluated for essay-type tests except by the major examining boards.)

Because of limitations of time, an essay-type examination is unlikely to cover the whole of the syllabus to which it relates, and an important external effect of allowing choice of questions is that this coverage is further curtailed, certainly for the individual and often for the whole examination (since some questions tend to be avoided by nearly all the candidates).

3.3.4

Essay-type questions can be and usually are made to call for some inventiveness and initiative on the candidate's part; they test his ability to present answers clearly and concisely, and they allow his style of presentation to be assessed.

The testing of inventiveness, initiative and style is often desirable. Inventiveness can only be tested with difficulty in objective tests, by the use of complex items. Initiative cannot be tested at all by objective items since however difficult such an item might be it does not require the candidate to <u>initiate</u> the response, but only to choose the correct one. Style is clearly beyond the scope of objective testing.

In the more advanced work found in further education compared

with schools, the testing of clarity and style of expression is generally desirable, although it has to be borne in mind that in craft and technician courses this may be relatively unimportant; measuring these skills may be unfair to people who are efficient operatives, capable of making correct decisions, and who only seldom need to communicate with others. Most students in further education, however, are preparing for or starting in careers which will certainly call for some report-writing, correspondence, etc.

Unfortunately, this particular superiority of the essay-type test is to some extent vitiated when the marking of the scripts is considered; the areas in which essay testing excels are the very ones in which the marking is most subjective. This is well known, and only two instances (both reported by Bonney Rust) will be given. In 1936, Hartog and Rhodes[3] published the results of investigations into examinations. When the scripts of 210 history examinees were re-marked over a year later, but by the same examiners, the final judgement (pass, fail or credit) altered for 92 (nearly half) of them. In another check, seven examiners marked an English script consisting of an essay and a precis; their percentage marks ranged from 28 to 80, distributed more or less symmetrically about a mean of 51.

3.3.5

From the comments already made in this section, it would seem likely that essay-type tests would have a lower reliability than objective ones. This expectation is borne out by the experience of the major examining bodies. The Joint Matriculation Board,[2] in its pamphlet "Examining in Advanced level science subjects of the GCE" (1970), defines essay-type questions as those "in which the candidate is allowed a large degree of freedom to select the material to

be used in answering the question and is required to marshal his thoughts and present them in a clear and logical manner". The Board states that a "basic difficulty with conventional essay-type questions is that they cannot give a high reliability, that is they do not measure attainment very accurately or consistently. Considerable efforts are made to increase the reliability of marking ..., but some subjective element must remain in assessment; moreover the small number of questions .... must also reduce reliability."

3.4 Identification of educational objectives

The range of educational objectives whose attainment can be assessed by an essay-type question is, by the very nature of the latter, far wider than is the case with an objective item. The advantages of this feature of the essay type have been referred to in 3.3.4, but it also has its disadvantages to the examiner; the identification and quantification of the objectives are almost impossible.

The Joint Matriculation Board[2] says of objective testing "The identification of the abilities tested is easier for these questions than it is for questions which require long, complex answers. It is therefore possible to design a paper in which specific weightings are given to the various abilities".

3.5 Tests for different purposes

In the introduction, tests in further education were classified according to their purposes. The relative suitability of objective and essay-type tests for these purposes is now considered.

3.5.1

For selection, induction, and progress tests in mathematics, essay-type questions have in general little to offer in comparison

with objective items; in the limited time usually available for conducting and marking such tests, objective items provide a much larger subject coverage, afford easier marking and higher reliability, and can provide better discrimination. There are however certain areas, even in mathematics, where initiative, clarity and style need to be tested throughout a course, and not merely at the end of the session; statistics and data-processing are two examples, as interpretation and report-writing might constitute a large part of the overall objective. In such courses it might be necessary to include some essay-type questions in these tests, but this is not typical of mathematical testing.

3.5.2

Attainment tests serve a wider purpose than those considered in 3.5.1. They have to assess the extent to which students have profited from the course, and their suitability for further studies, the award of a qualification, or advancement in their careers.

For these reasons, initiative, clarity and style of presentation need to be assessed. In spite of their subjectivity, essay-type questions must continue to be included in examinations in all subject, including mathematics. Nevertheless it seems that some objective items should also be included in attainment tests, especially in mathematics, the balance between essay and objective types depending of course on the precise nature of the course. The presence of good objective items is certain to increase both test reliability and the possibility of covering the full syllabus.

3.5.3

An example is now given of the way in which essay-type questions

and multiple-choice items can be made to complement each other in attainment testing, instead of essay-type questions being used exclusively.

The following essay-type question is typical of the structured version sometimes used in attainment tests at Higher National Certificate and similar levels.

"Estimates of the purchasing power of a certain currency are made at annual intervals. In the following table of extracted values, t represents the number of years after 31st December 1960 and p represents the purchasing power scaled so that the 1964 value is 100.

t	4	5	7	8	9
p	100	91.8	78.1	71.8	66.0

(a) Assuming a uniform percentage rate of depreciation of the currency, convert the data to a form which follows a linear law and plot the points on graph paper.

(b) Draw the straight line which appears by eye to match the points best, and obtain the equation of this line.

(c) Using the equation found in (b), calculate the estimated purchasing powers at the end of (i) 1966, and

(ii) 1975.

(d) Discuss the confidence which can be placed in each of the two estimates found in (c). Refer in your answer to the errors inherent in the data, and to any departures from the law you have found; where confidence intervals can be given for any variations, explain briefly how these can be calculated."

Parts (a), (b) and (c) of this question test mainly recall and application, and so could be replaced by multiple-choice items; in the composite question set out below, (i), (ii) and (iii) are suggested as these replacements.

Part (d) calls in addition for analysis and evaluation. The candidate is expected to distinguish between uncertainty arising from the errors of measurement on the one hand and departures from the law determined by the data on the other. Using the standard error of the regression coefficient found by the least-squares method, confidence limits can be attached to the former uncertainty but not to the latter. The unreliability of the estimates found in part (c) will be greater for 1975 than for 1966 because the former relies on extrapolation, but the candidate will be expected to explain that the uncertainty arising from possible change in the depreciation rate after 1969 is greater than that which results from errors in the estimates in the table, and cannot be quantified. Multiple-choice items cannot test the candidate's ability to realise these points on his own initiative or to discuss them clearly. Part (iv) of the composite question below therefore tests these abilities using the conventional essay methods.

The composite question proposed as a replacement for the structured one is as follows:-

"(i) A certain currency depreciates at a uniform rate of k% per annum; its value at time t = 0 is represented by p_0. Which of the following gives the value of the currency at time t?

A: $p_o(1 - \frac{k}{100} t)$ B: $p_o e^{-(k/100)t}$

C: $-p_o \ln \frac{k}{100} t$ D: $p_o t^{-k/100}$

(ii) A graph of y against x is linear, and passes through the points (4, 4.605) and (9, 4.205). The equation of the line is:-

A: $y = -12.5x + 4.925$ B: $y = 4.605 - 0.08x$

C: $y = 12.5x + 4.605$ D: $y = 4.925 - 0.08x$

(iii) If $\ln q = -0.1\ t + \ln 40$, which of the following best represents the value of q when t = 6?

A: 1.13 B: 3.09 C: 10.0 D: 22.0

(iv) For the period 1964-69 inclusive, the estimated purchasing power p of the pound is known for the end of each year except 1966; the law of depreciation has been found from these data.

Discuss the confidence which can be placed in estimates of p made from this law for the end of 1966 and the end of 1975. Refer in your answer to the errors inherent in the data, and to any departure from the law you have found; where confidence intervals can be given for any variation, explain briefly how these can be calculated."

This is described as a composite question because in the actual test it would be preferable to group the objective items in one section and the essay questions in another; such an arrangement would be less confusing to the candidates and would also facilitate the marking.

The object of replacing the original (essay-type) question is two-fold; the reliability of the scoring of parts (i) to (iii) in the composite question is greater than that of parts (a) to (c) in the original, and the same skills are tested as efficiently in the replacement question, and in a shorter time. The original question is a little longer than average and would require a time allowance of about 40 minutes. The three objective items (i) to (iii) would require a total time allowance of about 8 minutes, and the essay-type part (iv) needs about 12 minutes. Thus the composite question needs only about half as much time as the original.

Assuming that a fixed time (say three hours) is scheduled for testing purposes, it is clear that a combination of the two methods in the test would allow a more thorough examination of the subject area. If the test constructor decided that reliability was of great importance, the extra time ($1\frac{1}{2}$ hours in the case described) could be devoted to further objective items; if the emphasis is rather to be on the testing of the higher skills and of inventiveness, initiative and style, then essay-type questions could be added to make use of the extra time. For general purposes, it is probable that a mixture in a ratio close to that used in the composite question (one essay to three objective) would prove to be the optimum.

Similar considerations apply to the scoring. Again, unless there are strong reasons for a different arrangement, the dichotomous scores of the objective items could be weighted on the basis of each item being equal in importance to about one-third of one essay-type question. In the case of the composite question presented above, eight minutes spent answering the objective items could

then earn as many marks in the aggregate score as twelve minutes devoted to the essay question; this is justified by the greater reliability of the objective scores.

There is one skill which is tested in the original question and not in the replacement; this is the "motor" skill required in plotting the graph and reading values from the scaled axes. The classification of educational objectives adopted in this dissertation does not include motor skills, because they do not play a sufficiently large part in mathematical studies at the further education level. Furthermore, the presence of a graphical determination in a structured question has the disadvantage that a seemingly trivial error in this part may completely distort the treatment of the subsequent parts of the question and so further reduce the poor reliability of subjective scoring. If the drawing and interpretation of graphs feature in the syllabus, this work can be covered by short questions in the essay section of the examination; the importance of graphs can thus be recognized without interfering with the assessment of other skills. (Some examiners in subjects such as physics and biology are inclining to a similar attitude towards laboratory work; their practical examinations concentrate on the motor skills, while calculations and interpretations arising from observations are tested on the theory papers.)

The conclusions drawn from this discussion are that there are few if any courses in mathematical subjects in which the sessional attainment tests cannot benefit from the inclusion of some multiple-choice items, and that the proportion of such items can be adjusted by the examiner in the light of the aims and objectives of the course.

3.6 Comparison of properties of essay-type and objective tests.

It is now possible to present these properties in table form, rating each type of test on a scale ranging from "excellent" to "bad". Many of these ratings are of course subjective, but references are given so that the reasoning on which each is based may be considered by the reader if he wishes.

In all groups except the third (efficiency in testing various educational objectives), the references are to earlier sections; the reasoning relating to the third group has been left until after the table has been presented.

3.6.1

TABLE OF COMPARISONS

	Property under comparison	Essay-type test	Objective test	References
1	Ease of construction of satisfactory test	good	poor	1.1, 2.3
2	Capacity for wide coverage of subject	poor	exc	2.1, 3.3.3
3	Efficiency in testing:-			
	(a) knowledge and understanding	exc	exc	3.6.2
	(b) application	exc	good	3.6.2
	(c) analysis and evaluation	good	fair	3.6.2
4	Efficiency in testing inventiveness, initiative and style	good	bad	3.3.4
5	Possibility of identifying and weighting educational objectives	poor	good	3.4
6	Suitability for			
	(a) selection tests	bad	exc	1.2.1, 3.5.1
	(b) induction tests	fair	good	1.2.1, 3.5.1
	(c) progress tests	poor	exc	1.2.1, 2.1, 3.5.1
	(d) attainment tests	good	fair	1.2.1, 3.5.2
7	Immunity from successful blind guessing	good	fair	1.1, 2.2.3
8	Possibility of amendment in light of results	fair	good	1.3.2, 2.3.4
9	Direct contribution to learning process	good	exc	2.2.5
10	Suitability for banking	fair	exc	2.3.1
11	Ease and speed of marking	bad	exc	1.1, 2.2.1-5, 3.3.2
12	Reproducibility of marks (reliability)	poor	exc	1.1, 3.3.2, 4, 5

TABLE OF COMPARISONS (continued)

Property under comparison	Essay-type test	Objective test	References
13* Usefulness of values obtained for			
(a) FV	fair	good	1.3.1, 5
(b) discrimination	good	good	1.3.2, 5
(c) reliability	good	good	1.3.3, 5
14 Possibility of useful analysis of results	fair	exc	1.2.1, 2.2.5
15 Possibility of useful comparison of results with those previously obtained	poor	good	1.3.5
16 Freedom from security risks	poor	good	2.2.6, 2.3.2

* In 13, it is assumed that no choice of questions is allowed; if such choice is allowed, then all three ratings for essay-type tests should be "bad"; subsection 3.3.3 refers to this.

Distribution of ratings:-

	Excellent	Good	Fair	Poor	Bad
Essay	2	8	5	6	2
Objective	9	9	3	1	1

The clear inference from this table is that neither essay-type nor objective testing has an overall superiority over the other in every area of educational assessment; in many examinations, a balance should be struck between the two.

3.6.2

Groups 3 and 4 in the table of comparisons have been treated separately, implying that classification of abilities into (a) knowledge and understanding, (b) application, and (c) analysis and evaluation, is independent of whether or not inventiveness, initiative and style are being tested. This is not quite the case; the abilities in group 4 of the table are different from knowledge and understanding, and have more in common with analysis and evaluation than with application.

It has been argued in subsection 3.3.4 that objective methods are not very suitable for testing inventiveness, and incapable of testing initiative and style. For similar reasons, the difficulty of writing sound objective items increases as we move from (a) to (c) in group 3; items which test analysis and evaluation can be as good as those testing knowledge, but they are inherently more difficult to write and the scope for writing them is more restricted. One such item has been given in an earlier section under reference 3.2.2(c). The literature contains little on the relative suitability of the various educational objectives for testing by objective methods. The work of N. Wilson[4] is one exception; dealing exclusively with mathematical learning, Wilson classifies objectives as follows:-

Comprehension of new material

Problem solving

Ability to follow and construct proofs

Invention of tentative intuitive solutions.

He states that "normally only analytic aspects of proof may be tested in the normal objective tests, which do not lend themselves to testing ability to synthesize a proof", and (about the invention of

solutions) "it is very difficult to set objective questions in this area, but some attempts have been made". Since his classification shows a hierarchy of complexity similar to the one adopted in this dissertation, his comments are seen to support the ratings given in my table.

3.7 Short-answer questions

The table in subsection 3.6.1 deals with types of questions which are in a sense at opposite ends of the spectrum of assessment methods. This marked contrast suggests an intermediate method of assessment - the short-answer question. This was mentioned in subsection 1.2.2 and stated to be not truly objective as judgement is usually needed in the marking. The comparison of essay-type and objective tests would not however be complete without some further observations on the short-answer question.

3.7.1

In answering a short-answer question, the candidate has to supply a number, symbol, expression or short phrase. This type of test represents an attempt to combine the advantages of essay-type and objective testing. The advantages of the short-answer questions are classified below according to their similarities to those of the other two types:

In common with essay-type:-	In common with objective:-
Ease of preparation.	Brevity allows wide syllabus
Need for initiation of answer.	coverage.
Immunity from guessing.	Standard answer sheets facili-
Testing of vocabulary.	tate marking.

The disadvantages of short-answer questions are similarly classified:

In common with essay-type:-	In common with objective:-
Subjective marking.	Style cannot be tested.

3.7.2

Some comments by Ebel refer to other disadvantages, not shared with essay-type or objective testing:-

"The short-answer item .... is inordinately popular and tends to be used excessively in classroom tests. It is easy to prepare ..... It has the apparent advantage of requiring the examinee to think of the answer .... Some studies have shown a very high correlation between scores on tests composed of parallel short-answer and multiple-choice items, when both (tests) are intended to test the same knowledge or ability .... Accurate measures of how well a student can identify correct answers tend to be somewhat easier to get than accurate measures of his ability to produce them ....

"The disadvantages of the short-answer form are that it is limited to questions that can be answered by a word, phrase, symbol, or number and that its scoring tends to be subjective and tedious. Item writers often find it difficult to phrase good questions on principles, explanations, applications or predictions that can be answered in a word or phrase and that can be answered satisfactorily by only one specific word or phrase".

Ebel then goes on to point out that the quality of a test depends more on the weightings given to the various aspects of achievement, and on the quality of the items, than on the type of item chosen.

3.7.3

Short-answer questions in mathematics are hardly ever used at

Birmingham. Whenever a departure from essay-type questions is considered desirable, multiple-choice items have always seemed to be the appropriate choice.

3.8 Conclusion from the comparison

There seems to be no sound educational reason for the preponderance of essay-type questions which is still apparent in many, possibly most, examinations in further education. Some objective items could with advantage be included in nearly all examinations in mathematical subjects, the exact proportion depending on the objectives of the course and of the test.

The short-answer question seems to have no clear advantage over either the essay or the objective variety.

BIBLIOGRAPHY FOR CHAPTER III

1. Bloom, B.S. (1956). Taxonomy of Educational Objectives: Cognitive Domain. Longmans.

2. Joint Matriculation Board (1970). Examining in Advanced Level Science Subjects of the G.C.E. J.M.B.

3. Hartog, P., and Rhodes, E.C. (1936). An Examination of Examinations. MacMillan.

4. Wilson, N. (1970). Objective Tests and Mathematical Learning. Oliver and Boyd.

CHAPTER IV

EXPERIENCE WITH MULTIPLE-CHOICE TESTS AT BIRMINGHAM POLYTECHNIC

4.1 General

Multiple-choice tests have been increasingly used in the Department of Computer Studies and Mathematics, Birmingham Polytechnic, since 1972. They have as yet played no part in selection and little in the assessment of attainment. Their main role has been in induction and progress testing, and in the latter case they have been used both as a means of teaching and as a method of informal continuous assessment.

This chapter comprises a report on student reaction and the correlation between test results and the subsequent progress of the students.

4.2 Student reaction

With one exception (reported in subsection 4.2.2), no attempt has been made to assess student opinion, but any indications from students of their views on objective testing have been noted.

The attitude of students in further education towards methods of teaching and assessment is even more important than the attitude of school pupils. The students are adults; many of them are employed (sometimes in responsible positions); and they consider that they are paying for the course (through fees, rates and taxes and often by accepting "trainee" earnings). They are prepared to accept the guidance of the staff up to a point, but in the long run the way the course is conducted must be reasonably acceptable to the students if it is to succeed. For this reason, even the limited

experience at Birmingham is thought to deserve some detailed treatment in this dissertation.

4.2.1

The general opinion of students on objective tests seems to be favourable; no hostility towards them has been noted, and classes show rather more interest in post-test discussions of the pattern of their responses to multiple-choice items than of the results of essay-type tests. There is often a lively discussion when a student seeks to defend his choice of a distractor; usually this ends with the student recognizing that he was mistaken, but occasionally it transpires that the item is lacking in clarity and the students then help in correcting this. When this happens, the class is serving as a "shredding session" - an activity undertaken by a panel of teachers when there are enough interested staff available; some such check on the test constructor's work is highly desirable. So long as the tutorial role of the test is seen as its chief one, students do not resent such shortcomings in the test; for assessment purposes it is of course easy to eliminate the offending item and re-run the data on the computer. (To keep the matter in perspective, it should be recorded that only three such incidents have occurred in about 150 items written at Birmingham.)

The permanent record of the correct responses to items covering a large subject area is appreciated by the students as a valuable adjunct to their lecture notes; the fact that each student has a record of his misapprehensions is recognized as a help in preparation for subsequent examinations.

4.2.2

In one course (second-year Higher National Certificate in Mathematics, Statistics and Computing), multiple-choice tests in mathematics and statistics were given in a tutorial period about once a fortnight; the intention was partly to help in their learning and revision and partly to assess the various tests which had been written.

Some of the students took exception to this arrangement and lodged a complaint that too much time was being given to testing and that many of the items were "trick" questions. The course tutor immediately offered to withdraw all testing from the tutorial period, whereupon another, equally vociferous, section of the class objected to this proposal on the grounds that the tests were the most useful part of the tutorial!

The following points were then put to the class. Their corrected scripts should be useful to them; the distractors represented mistakes which were often made, and so the items were "trick questions" only in the sense that life's decisions often resemble "trick questions" to which it is easy to make the wrong response; the objective test scores were confidential and did not feature in their assessments or reports and so could not be harmful to them in any way. After a private discussion among the class, their spokesman reported a unanimous decision that the objective tests should continue at the same frequency.

4.3 Correlation with other assessments

Although objective and essay-type tests measure somewhat different abilities, it is to be expected that students who do well in

one type of test will do reasonably well in the other within the same group of subjects (in this case, mathematics and statistics); the correlation between scores on the two types, if not high, should at least be positive. This means that the sample correlation coefficient should significantly exceed zero.

4.3.1

Thirteen students in the first year of a Higher National Diploma course in engineering subjects were given a 14-item test on basic mathematics, made up as follows:-

Topic	Number of items
Approximations	4
Factorisation	3
Logarithms	3
Simple arithmetic	2
Circular measure	1
Differentiation (function of a function)	1
	14

The complete test is included in Appendix C. Summary statistics of the results are given in the table below. Test D1 is the sessional first-year three-hour essay-type examination in mathematics and statistics, and D2 a similar one at the end of the second year. Although no scaling was applied to these marks, the mean and standard deviation are close enough to permit a simple unweighted average to be taken; summary statistics of these mean scores are given in the column headed D1/2.

Birmingham HND engineering students

Test	Objective	D1	D2	D1/2
Mean	65.3	65.7	65.6	65.6
Estimated population standard deviation	10.9	14.7	13.4	11.6

The table below gives for each relevant pair of tests the product-moment correlation coefficient, the t value and the level of significance (see Appendix A).

Birmingham HND engineering students

Pair	Correlation coefficient	t	Level of significance
Objective and D1	0.43	1.6	15%
Objective and D2	0.66	2.9	1%
Objective and D1/2	0.66	2.9	1%
D1 and D2	0.35	1.2	25%

Thus correlation is not high, but it is worthy of note that the lowest value is that between the two sessional examinations D1 and D2 (not altogether surprising since these are on different syllabuses), while the correlation between the objective test and the average of the sessional examinations is one of the highest. On this small sample, a short objective test on basic mathematics has proved a better predictor of D2 performance than has the D1 three-hour examination. While it certainly cannot be inferred that this will always be the case, it is not surprising that simple objective tests such as the Vernon's graded test in mathematics and arithmetic are

very efficient for selecting students for GCE ordinary and advanced level courses.

In subsection 1.3.3 it was stated that with small classes KR can have low (even negative) values even when the test is quite reliable; this is especially so with short tests like the 14-item one used here. In fact KR for this test was 0.08, and yet the results do seem to be quite reliable.

4.3.2

Another objective test in use at Birmingham is one of 50 multiple-choice items on basic mathematical subjects, ranging from elementary arithmetic to polynomial equations. The test appears in Appendix C and is composed as follows:-

Topic	Number of items
Simple arithmetic	4
Simple algebra	7
Mensuration	3
Approximations	6
Exponents	3
Logarithms	3
Theory of equations	4
Differentiation	6
Integration	4
Trigonometry and geometry	8
Co-ordinate geometry	2
	50

This test was designed mainly for induction purposes, and has proved useful both in locating weaknesses in the students' knowledge and in stimulating class-room discussion of points where revision is seen to be needed.

One interesting use of the test was with a first-year class of 20 full-time students working for a polytechnic diploma in Estate Management and Surveying. The mathematics syllabus in this course consists mainly of statistics, but includes some work on logarithms and graphs (for use in economics); the trigonometry needed in their surveying is taught in classes in the latter subject. The official sessional (assessed) examination in mathematics therefore had little in common with the 50-item objective test, but the correlation between the two sets of scores is quite high; this is noteworthy because it was only because of the presence of the small amount of non-statistical material that the objective test in basic mathematics was used with the class. The results for the 19 students who sat the sessional examination are given below.

Birmingham Estate Management students (19)

Statistics of percentage scores \ Test	Basic mathematics objective test (50 items)	Sessional essay-type examination
Mean	24	50
Estimated population standard deviation	20	27

The correlation coefficient between the two sets of scores is 0.71; with t equal to 4.19, this is highly significant (0.1%).

Again there is a reasonably close association between the scores, suggesting that skill at basic mathematics indicates an ability to learn the somewhat different, and (to most of the students) new, subject of statistics.

4.4 Correlation between different student groups

In each of the two investigations just discussed, we have compared the scores of one group of students when taking different tests; we now examine the results achieved on one test by two groups of students.

The 50-item basic mathematics test, after the elimination of 13 items not considered appropriate, was given to 19 undergraduates studying Transport Management and Planning at Loughborough University of Technology. The composition of the resulting 37-item test is given below.

Topic	Number of items
Simple arithmetic	4
Simple algebra	5
Mensuration (of cube)	1
Approximations	5
Exponents	3
Logarithms	3
Complex roots of cubic equation	1
Differentiation	4
Integration	3
Trigonometry and geometry	7
Equation of straight line	1
	37

For comparison purposes, the results of the Estate Management students from Birmingham were re-run with the same 13 items eliminated. The results of both groups on the 37-item test are tabulated below.

37-item objective test in basic mathematics

Statistics of percentage scores \ Group	Birmingham (Estate Management) (20 students)	Loughborough (Transport Management) (19 students)
Mean	30	62
Estimated population standard deviation	24	18

The difference between the mean scores is not surprising, since the minimum entry qualification for the estate management course includes only an Ordinary Level GCE pass in mathematics. A positive correlation between the scores would of course be expected, and the value was 0.61; with a t value of 4.56 this is highly significant (0.1%). Although not very high, this coefficient is thus high enough to establish a definite relationship.

The behaviour of the index of discrimination is quite different, and bears out an earlier comment (1.3.2) that ID is not a function of the item alone, but is affected both by the nature of the other items in the test and by the group being tested. The latter point is clearly displayed in this case, where the coefficient of correlation between the discrimination indices in the two courses is only 0.13; this gives a t value of 0.80, and fails (even at the modest 25% level) to establish any correlation at all. This is not to say that the

discriminating power of the test as a whole varies greatly from one course to the other. With the Birmingham Estate Management students, mean ID was 0.37, and with the Loughborough Transport Management students it was 0.33. Reliability was high in both cases, KR being 0.93 and 0.90 respectively.

4.5 Experiments using multiple-choice tests

The observations made so far in this chapter relate specifically to objective testing, and are based on normal usage of this method in the teaching work of my department (together with the data obtained from Loughborough University of Technology). This section gives the results of two experiments in which objective tests were used to investigate aspects of educational assessment in general - open-book examinations and the subjectivity of the examiner.

4.5.1

The part which memory should be allowed to play in determining a candidate's performance has received surprisingly little attention from educational psychologists. It has been, and generally still is, standard practice for examination candidates to have to answer the questions from memory - apart from such obvious exceptions as tables, standard integrals, physical constants, etc., and the recent allowance of tables of useful formulae in certain cases. It is argued in support of this policy that examinations should test, among other things, the candidate's ability to understand and remember a large amount of information and to reproduce some of this under pressure in the examination room; this pressure includes the need to reproduce material from memory in a limited time.

An argument against this emphasis on memory is that it lacks realism. Except in examinations, reference to documents is not only permitted but strongly encouraged; decision-making without reference to reliable sources is clearly undesirable. Calling for answers from memory in examinations favours the candidate with a good (though possibly short-term) memory at the expense of one with a poor memory even though the latter may have as good a grasp of the subject and a greater ability to find what he needs in his notes or books (because his poor memory may have forced him to develop a more organized system).

An alternative approach is to set questions which test the higher skills (such as application, evaluation and synthesis), and to allow unlimited reference to literature during the examination. This type of question is thought to be more difficult to set than those which test memory, a belief which may explain the predominance of memory-type examinations. In my view, too much is made of this argument. A question which mainly tests recall in a memory-type examination can still be useful in an open-book examination; it finds whether the candidate can locate and recognize the information - an ability which should not be despised, especially if a number of reference books are required. In many memory-type questions, the formula required is so difficult to remember that it is given in the question; this assistance (which is unrealistic - a research worker is not usually told in advance the exact method to use) need not be given if the examination is of the open-book type.

My main reasons for preferring open-book examinations in general are that they reflect real life more faithfully, remove some of the strain from preparation for the test (especially for candidates not

endowed with a good memory), and increase the credibility of the assessment process. Sufficient pressure still remains; the time factor, the competitive nature of examinations, and the mere fact of being assessed, are enough to put candidates on their mettle. For these reasons, most objective items which I have written are intended for open-book use; this means that there has been no attempt to avoid difficult formulae, since students are expected to be able to find these in their notes. The few exceptions appear in special tests designed to measure ability in mental arithmetic, in which neither tables nor calculators are permitted.

It may be thought (and often is, by students) that the open-book type of examination, whether or not more fair, must be easier than the memory type. My own experience over the last decade, using open-book tests whenever permissible, has been that the pass rate, and the distribution of scores, do not seem to differ appreciably from what would have been expected from memory-type tests. In the absence of controlled experiments, such an impression is bound to be subjective; but with essay-type tests, experimentation has not been possible at Birmingham, since student numbers are generally small and time is limited. With informal objective tests and the co-operation of teachers with parallel classes, however, it has been possible to obtain some quantitative evidence.

Two multiple-choice tests on statistics, each of 20 items and designed to be equivalent in subject coverage and difficulty, were first tried with the same class of 11 students (Second year HNC in Mathematics, Statistics and Computing). The tests are given in Appendix C, and the results with this class are given below.

	Mean Score (%)	Estimated population standard deviation of % scores
Test X	31	16
Test Y	30	15

These results were taken to indicate that the tests were reasonably equivalent. The reason for the low scores is thought to be that the class had covered the test topics in the previous year; further, the items were written by me but the class had been taught statistics by another teacher, and this fact raises an issue discussed later in this chapter.

In the experiment, both tests were given to a first year statistics class in a Higher National Certificate course in Medical Laboratory Sciences; I had taught the subject to this class, which had 16 students. Some of the topics on the tests were not on the syllabus for this course, and it was necessary to eliminate three items from Test X and one from Test Y. The former test was administered in the open-book form for which it was designed, but the latter had to be answered from memory in this experiment. The results were as follows:-

Test	X (open book)	Y (from memory)
Mean score (%)	47.4	33.0
Estimated population standard deviation of % scores	29.2	23.9
Number of items	17	19

It will be seen that both the mean and standard deviation are

smaller in the test which had to be answered from memory. The relative dispersion of scores was not reduced by this restriction, however; the ratio of standard deviation to mean value is 0.62 and 0.72 respectively for tests X and Y.

The ratio of the mean scores, 47.4 : 33.0 or 1.44 : 1, suggests a 44% advantage to candidates allowed to refer to books and notes; the difference between the means is significant at about the 4% level (see Appendix A). The magnitude of this difference at first caused some surprise, but a probable explanation was soon seen. Both tests were originally written for open-book use, but the students had to rely on memory when answering Test Y; had the test been prepared for this sort of use, it would not have asked for detailed consideration of such relatively complicated formulae as that for correlation. Students are encouraged to look up such matters in their notes before applying them to problems, and so a memory-type question would be more likely to give the formula and ask about its evaluation or application.

The only conclusion to be drawn from this experiment is that items which test recall are more difficult if reliance has to be placed upon memory rather than reference to documents. It does not suggest that open-book examinations are too easy (47% is not an excessively high mean score) and it does not prove that testing a candidate's memory has any special merit. In my opinion, these observations apply equally to essay-type and objective tests when used for the assessment of progress and attainment. In the case of selection and induction tests, however, it would be unreasonable to expect students to come equipped with the relevant notes and

text-books; in general, they would not be sufficiently aware of the subject matter of the test to be able to select the appropriate material. Fairness in this situation is best achieved by not allowing reference to documents, tables or calculators.

4.5.2

The second experiment was designed to investigate the subjective nature of testing. As mentioned in the introduction, the writing (or selection) of objective items is, unlike their marking, a subjective activity. It is possible that the test constructor will subconsciously include items which over-represent topics which particularly interest him; it is also possible that his teaching will tend to stress these same topics. In these circumstances the students he has taught would be expected to obtain higher scores in the test than similar students taught by another teacher of equal proficiency.

In this investigation, the results of statistics tests X and Y previously described were compared with those obtained when a parallel class took essentially the same tests, Y again being answered from memory. The two classes were both part of the same HNC course in Medical Laboratory Sciences; Class A was taught by a colleague and Class B (the subject of the open-book experiment already discussed) was taught by myself. The treatment of statistics differed slightly between these classes because of the requirements of the optional subjects. Class A was chosen to participate in this experiment because their statistics teacher has frequently taught classes parallel with my own, and our examination results in such cases have been similar.

Throughout the year, each of us taught his class as usual with these courses, and without relating his teaching to the objective tests. At the end of the session the results of both classes in all the course subjects were noted, and were found not to differ significantly. The results in the statistics tests are given below, the following abbreviations being used:-

m = mean of percentage scores

$\hat{\sigma}$ = estimated population standard deviation of percentage scores

n = number of candidates

i = number of items

Test \ Class	A	B
X	m = 40.2 $\hat{\sigma}$ = 22.9 n = 12 i = 17	m = 47.4 $\hat{\sigma}$ = 29.2 n = 16 i = 17
Y	m = 17.9 $\hat{\sigma}$ = 17.2 n = 13 i = 13	m = 33.0 $\hat{\sigma}$ = 23.9 n = 16 i = 19

The treatment of statistics with Class A differed slightly from the standard syllabus, causing as many as seven items to be eliminated from Test Y. The results are therefore not strictly comparable, but there is at least an indication that the expected advantage

enjoyed by Class B through having been taught by the test constructor was more pronounced with Test Y than with Test X. This might be because Test X was of the open-book variety; topics which had been covered with Class A but not given as much stress as with Class B could have been found in the notes by Class A students during Test X, thus reducing any advantage to Class B.

The difference between the mean scores obtained on Test X is not statistically significant (see Appendix A). This means that the hypothesis that being taught by the test constructor does not confer an advantage cannot be rejected on the evidence provided by Test X; it does not in any way confirm the hypothesis.

In Test Y, the difference between the mean scores is marginally significant (at about 7%), but as pointed out earlier this comparison is not valid because of the different items eliminated for the two courses. This objection can be overcome by considering only those items which were left in the tests for both classes. The tests being compared are then identical, but sampling error is increased by having fewer items - 15 in Test X and 12 in Test Y. The between-class differences are not then significant for either of the tests, or for the combination of them. For this reason, only the mean percentage scores are given in the following table; they still show Class B to have achieved the higher scores.

Mean Percentage Scores

Test \ Class	A	B
X (15 items)	41.5	50.0
Y (12 items)	19.6	31.3

I do not know whether any deeper research has been undertaken into this aspect of the subjectivity of examining; the point is of no direct concern to the large examining bodies, who of course administer examinations which are externally set. This small experiment marginally supports the common-sense view that the examiner's students are likely to have an advantage over other students taking the same examination. The results emphasise the desirability of the precautions normally taken with internal examinations; when more than one class and one teacher are involved, the setting of questions should be shared among the staff carrying out the teaching. This prevents any unfair advantage being enjoyed by one group of students. If any choice of questions is allowed (unusual in objective tests but commonplace elsewhere), the subjective effect may however operate unequally, and so reduce the reliability of the scores. This can happen when some candidates avoid questions set by their own teacher(s), and so constitutes another argument in favour of requiring all questions to be attempted (others were discussed in subsection 3.3.3).

If it is impossible to avoid the situation where some candidates will have to answer a number of questions not set by their teacher while other candidates will not, consideration should be given to reducing any unfairness by making the examination open-book, or at least providing a generous amount of information by means of a standard reference sheet.

This chapter has dealt with experience at Birmingham Polytechnic, and has included suggestions based partly on that experience. The opinions of other colleges have also been sought on many of the points discussed, and these are described in the next chapter.

CHAPTER V

ATTITUDES TOWARDS MULTIPLE-CHOICE TESTING IN MATHEMATICS AT OTHER BRITISH COLLEGES

5.1 General

To complete this study, the views and practices of other colleges are considered. About fifteen institutes of further education (referred to henceforth as "colleges") in Great Britain contribute to the Manchester Objective Testing Item Bank, which as far as is known is the only large such venture in the country catering for mathematics; the other contributors are universities. Fourteen of the colleges associated with the Manchester bank were approached for information, and eight responded. It has been assumed that most college experience in writing objective items is concentrated in these institutes, and no approach was made to other colleges.

Another useful source of information is Polymaths - a new evening course for mature students described more fully in section 5.3. The 17 colleges at present offering this course conduct multiple-choice tests at the rate of five items per week for much of the academic year, but these items are written by the Polymaths Course Production Team and not by the college staff; several of the colleges are however contributors to the Manchester bank. The experiences of these colleges with objective testing are relevant to this study and are therefore included.

5.2 Survey of Manchester item bank members

A questionnaire (see Appendix D) was sent to fourteen member colleges and responses were received from eight of these. Since in general the person completing the questionnaire can be assumed to be

the one most involved in objective testing, it has seemed appropriate to add my own response, on behalf of my department at Birmingham Polytechnic. The following tables therefore relate to a total of nine member colleges; the starred number in each category is the one which includes the Birmingham response.

Tests were classified as follows:-

A Selection

B Induction

C Progress

D Attainment

In this survey the terms "objective" and "multiple-choice" are synonymous, as the Manchester bank does not deal with any objective items other than the multiple-choice type.

5.2.1

Table showing extent of usage of objective tests

(number of colleges)

Type \ Usage	Often	Sometimes	Never
A	1	0	8 *
B	1 *	2	6
C	7 *	2	0
D	2	2 *	5

This table demonstrates the popularity of objective tests for assessing progress, and also seems to reflect the fact that mathematics departments are not always free agents in the matter of selection or attainment tests. It is a little surprising that objective tests are

unpopular for induction purposes except at Birmingham; perhaps there is less need elsewhere than here, where syllabus content and student background seem to be constantly changing.

In subsequent tables, numbers in brackets show how many of the colleges are frequent users of the type (A, B, C, or D). Colleges recording "no strong feelings" have been omitted.

5.2.2

The first two questions were designed to ascertain feelings about the best arrangements of facility values; some workers favour making these as near as possible to 50% while others prefer a wide range of FV's with an average of 50%. These and subsequent questions took the form of statements to which the respondents were asked to give their attitudes.

Attitudes towards arrangements of facility values

"Each item should have a facility value between about 40% and 60%."

Type	Agree	Disagree
A	2 * (1)	1 (0)
B	0 (0)	2 (0)
C	1 (1)	5 * (3*)
D	1 * (1)	3 (1)

Ebel's findings, discussed in Chapter II, are that close grouping of FV around 50% results in a greater spread of scores and a higher reliability factor; this suggests that the statement under discussion here is correct for types A and D - a belief evidently not shared by my colleagues in other colleges.

"Items should have facility values which are roughly normally distributed about a mean of approximately 50% and lie between limits of approximately 10% and 90%."

Type	Agree	Disagree
A	1 (0)	2 * (0)
B	2 (0)	1 (0)
C	5 * (3*)	1 (1)
D	3 (0)	2 * (1)

The support shown for this statement in connection with progress tests is in line with the conclusions reached in Chapter II. Opinion is roughly equally divided on the applicability of this arrangement of FV's to the other types of test, although it seems unsuitable for A and D. It is appropriate for type B so long as it is possible to compare the FV's obtained in the test with those previously established for the items; without this proviso, wrong decisions may be taken on the basis of extreme FV's which are in fact normal for the items in question.

5.2.3

It was suspected that some test constructors may place undue emphasis on the numbers of candidates choosing distractors, and modify items in an attempt to equalise these numbers. The following statement was put forward in order to test this belief.

Attitudes towards responses to distractors

"The responses of candidates who choose distractors (i.e. wrong responses) should be roughly equally distributed between these

distractors."

Type	Agree	Disagree
A	4 (0)	1 (0)
B	4 (0)	1 (0)
C	5 (2)	1 (0)
D	4 (1)	1 (0)

The statement is reasonable enough, but in view of the support it has received a word of warning is called for. Erroneous impressions in mathematics are often quite distinct, well-known to teachers, and easily exposed by suitable distractors. Wrong notions of equal importance may not be equally widespread, however, and unequal response to distractors therefore does not prove that the item is at fault. Over-zealous attempts to equalise these responses may reduce the diagnostic power of a well-written item, and this may be the reasoning behind the one disagreement recorded. The Birmingham response of "no strong feelings" was made because we would wish to change distractors which received little or no support, but would go no further towards seeking equality of response.

5.2.4

The retention of corrected scripts by students was discussed in Chapter II, and reasons for this policy prevailing at Birmingham were stated. In the questionnaire, two statements were made, one being the opposite of the other; this was done so that the reasons for and against allowing scripts to be kept could be stated. The overall result is summarized below.

Attitudes towards retention of scripts by students

"Students should be allowed to retain their corrected scripts."

Type	Agree	Disagree
A	1 (0)	4 (1)
B	2 * (1*)	3 (0)
C	6 * (4*)	3 (2)
D	1 (0)	4 (1)

With selection and attainment tests there is a majority against students keeping their scripts, presumably for security reasons. In the case of progress tests, the majority of those with strong views were in favour; it is in fact surprising that as many as 3 colleges are opposed to the retention of progress test scripts by students. In my opinion, the educational arguments in favour of this policy far outweigh the security considerations.

5.2.5

The acceptability to students of educational methods, including testing, was stated in Chapter IV to be even more important in further education than in schools. A question was put in the questionnaire to see whether objective testing was proving as interesting to students elsewhere as it seemed to be at Birmingham.

Attitudes towards student interest in objective test results

"Students are more interested in the results of objective tests than in those of other tests."

Type	Agree	Disagree
A	2 (1)	2 (0)
B	3 * (1*)	2 (0)

Type	Agree	Disagree
C	5 * (3*)	1 (0)
D	2 (0)	2 (0)

Only in connection with progress tests is there any marked support for this suggestion. The fact that the only disagreement comes from colleges where objective testing is not often used may be significant, but whether the opinion is the effect or the cause of the infrequency of use is open to question!

In retrospect, the statement is seen to be ambiguous. It was intended to refer to the interest shown in FV's and item analysis when the test results are discussed with the students, but could have been taken as meaning interest in the scores obtained.

5.2.6

Reasons were given in Chapter IV (subsection 4.5.1) for my preference for open-book examinations (whether of the essay or objective kind), with certain exceptions, such as selection tests. The opportunity was taken in this survey to test the reaction of other colleges to this preference, at least when applied to objective testing.

Attitudes towards open-book objective testing

"Objective testing is more realistic, and the results more reliable, if candidates are allowed to refer to text-books and notes during the test."

Type	Agree	Disagree
A	1 (0)	2 (1)
B	1 (0)	2 (0)
C	4 * (2*)	3 (3)
D	3 * (0)	1 (0)

It is interesting that the support for open-book methods is relatively stronger for attainment than for progress tests. Since books can be used for homework, course work and projects, the only obvious justification for not allowing them in progress tests would be to make these a preparation for memory-type sessional examinations; if these examinations are to be open-book, it is difficult to see any reason for holding progress tests of the memory type.

However, for purposes other than testing progress, most of the responses showed no strong feelings about the use of books and notes during tests.

5.2.7

In the discussion of test procedure in Chapter II, reference was made to the question of whether wrong responses should attract negative marks; Ebel was quoted as seeing no compelling reasons for or against this approach, whereas Gronlund's treatment of the subject showed a slight preference for the use of negative marks. This was the subject of the last part of the questionnaire.

Attitudes towards the award of penalty marks

"Multiple-choice testing is fairer, and the results more reliable, if a small negative mark (say - 1/3) is awarded for wrong responses, as a correction for guessing."

Type	Agree	Disagree
A	3 * (0)	1 (0)
B	3 * (1*)	2 (0)
C	4 * (3*)	4 (3)
D	3 * (0)	2 (1)

Here the even division of opinion is consistent with Ebel's views, namely that it is of little importance whether a guessing correction is applied or not.

5.2.8

Even in this small sample of nine colleges, unanimity was not achieved or even approached on any issue for any of the four categories. In almost every case, over half of the colleges had no strong feelings. Clearly no uniformity of practice is yet to be seen within the member colleges of the Manchester bank under these headings.

5.3 Objective testing in the Polymaths course

A recent innovation in further education is a one-year evening course in mathematics for mature students who lack the qualifications and/or confidence to take a degree course in the subject; this is the Polymaths course. There are no entry requirements, and successful completion of the course is accepted by the Council for National Academic Awards and the Open University as qualifying the student for entry into a degree or honours degree course in mathematics.

The course started in 1974 with about 200 students in 9 colleges, and by 1977 both of these numbers had approximately doubled. There are at present 17 colleges (polytechnics and others) running Polymaths. A noteworthy feature is the use of standard text-books written specifically for the purpose by the Polymaths Course Production Team.

It has been stated by R.C. Adams and D.J.G. James[1] that multiple-choice tests are "ideal for mature students in a variety of ways". Such students, it is suggested, perform badly in formal

three-hour examinations, partly because of a perfectionist approach which results in too much time being spent on the first question; there is a more relaxed approach to objective tests of a few items each. The authors offer no reason for the latter; I suggest that the reason may be the greater frequency of these tests, made possible by the small amount of time required for each item and of course by the small number of items. In each teaching week a five-item multiple-choice test is held, the items being written by the Polymaths Course Production Team. The total score for these tests accounts for 500 out of a total of 1,200 marks for the year's course. Every fifth week there is a periodic test which includes essay-type and short-answer questions but no (strictly) objective items. The team report a correlation coefficient in the region of 0.75 between the objective test and periodic test marks. The aim of some of the objective items is described as the elimination of "commonly occurring misconceptions before they take hold".

I close this chapter with the observation that Polymaths is the most recent mathematics course to be established in this country on a national basis, and it seems significant that objective testing, used for assessing both progress and attainment, forms an essential part of the structure of the course. Adams and James state that "virtually all students agree that they need the stick of weekly objective tests . ..". Referring again to objective tests, the authors later say that "by general consent (these) have been an outstanding success. They were originally designed to overcome the fears and poor performance of adult students in conventional 3-hour written examinations by providing an alternative means of

assessment. In practice their significance has been much broader. They have acted as the mortar holding the parts together. They provide the discipline of a routine ....; running through an earlier week's objective test items is a most effective teaching device to consolidate and tie up the loose ends of a topic whose essentials have been mastered ..... Moreover they do the basic assessment job into the bargain."

BIBLIOGRAPHY FOR CHAPTER V

1. Adams, R.C. and James, D.J.G. (1977). "Polymaths - an opportunity for a second start in mathematics for mature students", Bulletin of the Institute of Mathematics and its Applications, 13, 197 - 201.

CHAPTER VI

SUMMARY AND CONCLUSIONS

6.1 General

In this chapter the more important of the conclusions reached in the dissertation are first summarized and then used as the basis of a discussion of the present and future roles within further education of objective testing in mathematics. Because the superiority of multiple-choice testing over other objective forms is generally accepted, this chapter concentrates on the former type.

6.2 Summary

6.2.1

It has been seen that the multiple-choice type has the following advantages over essay-type testing:-

(a) The marking is more reliable, and can be carried out wholly (by use of special cards) or partly (as described in Chapter II) by computer.

(b) Because a multiple-choice item can be read and answered by the candidate in so short a time, the subject coverage is much greater than with an essay-type test of the same duration.

(c) The way in which the candidate's responses are recorded makes it possible (or indeed easy, if a suitable computer program is available) to obtain the characteristics of the whole test, and of each item. This feature makes the banking and re-use of items more feasible than with essay-type questions, and the large number of items in a test reduces the security risk involved in their re-use.

(d) The method of recording responses referred to in (c) also

facilitates the precise identification of weaknesses both in a class as a whole and in individual students. If the students can keep their scripts, these form a useful adjunct to their text-books and lecture notes by revealing their mistakes.

6.2.2

When compared with essay-type questions, multiple-choice items have the following disadvantages:-

(a) Sound multiple-choice items are more difficult to write at all levels. (This point has to be taken in conjunction with the ease of banking such items, which helps to offset this disadvantage.)

(b) It is much more difficult to test the higher skills with multiple-choice items than with essay-type questions.

(c) It is impossible to assess style, clarity, inventiveness and the presentation of logical arguments by means of multiple-choice testing.

6.2.3

My conclusion from these considerations is that multiple-choice items can be used in all of the four test categories of selection, induction, progress and attainment, but that they need to be supplemented by essay-type questions in attainment tests, and occasionally in the other three categories as well. In mathematical subjects the amount of essay-type testing which is required is not usually very great in selection, induction and progress testing, and is less than in most other subjects. (The difficulty of testing the higher skills with multiple-choice items of course becomes increasingly important as the level of the work advances.)

6.3 Present role

It is difficult to obtain a complete picture of the extent to which objective testing in mathematics is being used in further education. The colleges to which I wrote are all associated with the Manchester Objective Testing Item Bank, and so presumably more active than the average in this type of testing. Not all of even these colleges replied, and had I undertaken the much greater task of writing to every college in England and Wales, the proportion replying would probably have been much less than half; furthermore, the replies received would have constituted a sample which, although large, was very variable with respect to involvement in objective testing. The sample which I obtained is probably reasonably typical of colleges actively interested in objective testing, but not necessarily typical of the country as a whole.

Of the 9 colleges which returned questionnaires, all used multiple-choice items in progress tests; only 1 used them in selection tests, 3 in induction tests, and 4 in attainment tests. The impression I have gained from informal talks with other further education staff (mainly in the Midlands) is that very little selection testing, and not very much induction testing, takes place in colleges by <u>any</u> method. It seems that at present attainment tests (with two exceptions discussed below) seldom include any objective items; although four of the Manchester bank member colleges reported that they used objective items for this purpose, they may not be representative of the country as a whole.

The Polymaths course, referred to in Chapter V, provides an interesting example of multiple-choice tests being used in the combined role of progress <u>and</u> attainment testing, in that the marks of the weekly tests form part of the continuous assessment. The staff

at the colleges running this course do not write these items, and so are not necessarily fully involved in objective testing; if however the benefits are as great as is claimed by Adams and James in the article quoted in section 5.4, these colleges may be encouraged by the student reaction to use objective tests with other courses also, if they are not already doing so.

Similarly, many colleges prepare students for G.C.E. examinations, some of which include objective items in mathematics (although these are at present mainly confined to the Ordinary Level papers). Those teaching G.C.E. may be less involved in objective testing than those concerned with Polymaths, since unlike the latter they will not be able to see the analysis of their students' performance on these items. It may safely be assumed however that they will prepare students for these examinations by setting at lease some progress tests of the objective type; they may write them themselves or they may purchase tests which are offered commercially.

On the other hand, neither the Institute of Mathematics and its Applications nor the British Computer Society use objective items in their own examinations. Neither do the other large examining bodies catering for further education, at least as far as mathematics is concerned. In the Midlands, for instance, the Birmingham-based Union of Educational Institutions uses objective items in its examinations for engineering crafts, science, horticulture and electronics, but restricts its mathematics papers to essay-type questions. The City and Guilds of London Institute follows a similar policy; while using objective items extensively in other subjects, it does not do so to any great extent in mathematics. It may be that these

bodies believe conventional examinations in mathematical subjects to be sufficiently objective because of the very nature of the subjects.

Many examinations in further education are set internally; about half of these are formally assessed on behalf of the validating bodies such as the Joint Committees for National Certificates and Diplomas. The Institute of Mathematics and its Applications appoints the assessors for these in Mathematics, Statistics and Computing. The Institute favours experiments with new methods of examining, and some (but not many) HNC papers are known to have included objective items.

Most mathematics examination papers in further education relate however to courses in which mathematics is only a part, and here again objective items seldom appear. The following are suggested as possible reasons in the minds of the teachers:-

(a) Reluctance to change their methods.

(b) Belief that assessors, and/or the department responsible for the course, may oppose the use of objective testing.

(c) Fear that students may do badly because the teachers are prevented from marking such items leniently, whereas they can exercise discretion in essay-type tests.

(d) Fear that objective tests may be too easy compared with conventional ones.

(e) Recognition of the difficulty of writing sound items, coupled with an unawareness of the existence of item banks and commercial tests.

The present role of objective tests in mathematics in further

education would therefore appear to be mainly confined to progress testing; in view of the various ways in which they can be used, however, it seems probable that many if not most colleges use them at least to some extent.

Turning now to my department at Birmingham Polytechnic, there is a teaching staff of 22 and about one-third of us frequently use objective testing (one to three tests each per month), about one-third occasionally (one to three tests each per year), and the rest not at all. Five years ago objective testing was virtually unknown here, and I feel that its use is likely to continue to increase.

6.4 Future role

It has been strongly argued in this dissertation that objective items (and particularly the multiple-choice variety) can be used to good purpose in virtually all forms of mathematics tests, including formal examinations, and that it is unfortunate that their advantages are not being more fully exploited in further education. In particular, their potentialities in attainment tests when used in conjunction with essay-type questions were discussed in subsection 3.5.3. I would therefore like to see an increase in the part played by objective testing in mathematics. The various ways in which such an increase might be expected to take place are reviewed in this section.

6.4.1

If the national examining bodies mainly concerned with further education were to introduce objective testing in mathematics into their papers, this step towards more reliable assessment of candidates would tend to encourage progress testing along the same lines. There is however no indication that such a development is likely in

the near future; as suggested in section 6.3, these bodies seem to be satisfied with essay-type examinations in mathematics.

The Polymaths course, however, is in a similar category, being a nationwide activity. As stated in Chapter V, the number of colleges involved in this venture is increasing, and objective tests are used weekly for the joint purpose of measuring progress and attainment - and to assist in the learning process. Here is one area where some of the looked-for expansion is likely to take place. Birmingham Polytechnic is in fact one of the colleges at present proposing to run this course.

6.4.2

Another development likely to encourage the growth of objective testing is the move to place all courses at present leading to C.G.L.I. qualifications and National Certificates and Diplomas under the control of the newly-formed Technician Education Council. All such syllabuses are having to be reconstructed in far greater detail than before, and written so that every part has its objectives clearly stated and each syllabus is related clearly to the aims and objectives of the course. Since the identification of educational objectives and skills is considered to be easier when these are being assessed by objective items, there is considerable emphasis on this kind of testing. Examinations will not be set by the T.E.C., but there is likely to be some pressure on examiners to include objective items. It should be recorded that the whole T.E.C. scheme has involved a number of further education staff and others, in a great deal of work; not all of these are convinced of the necessity for this effort, and the scheme has aroused much controversy

and some hostility. While there is certain to be an increase in the amount of objective testing in colleges as a result of the T.E.C. scheme, this method of assessment might suffer some temporary unpopularity with opponents of the scheme. Nevertheless I believe that the overall effect will be beneficial to educational assessment.

6.4.3

National bodies responsible for examinations in further education are not the only agencies whereby change in methods can be brought about; certain others are considered in this section.

The Manchester Objective Testing Item Bank has already been mentioned. By providing a means of communication and co-operation between colleges the bank is assisting in the interchange of ideas on objective testing as well as of the items themselves, and by holding conferences which are advertised throughout the field of further (and higher) education it contributes to the spread of these ideas. Although objective testing was in use at Birmingham before we were aware of the bank, our activity in this field has increased considerably as a result of becoming a member department.

Until about 1960 most further education staff had no teacher training, but an increasing number have now attended courses at Technical Teachers' Colleges. All of these colleges include the theory and practice of objective testing in their curricula, and some are very enthusiastic about this technique. This should facilitate the introduction of objective testing into colleges, and I consider the slow rate of progress in this respect to be surprising, and unfortunate.

Many teachers at colleges are studying for Open University

degrees, and so will be experiencing this form of testing as candidates. A typical O.U. mathematics student will take one objective test of between 20 and 30 items per month. Items can be of the multiple-choice, multiple-response, or true/false type, and all are computer-marked. For security reasons the scripts are not returned, but at the request of a number of students the solutions are now issued for the benefit of those who have made a note of their responses. Again, teachers with experience of this sort could help in the spread of this method at their colleges, or even introduce it themselves, but so far this has not taken place to the extent which might have been expected.

6.4.4

It seems appropriate to conclude this dissertation with a reference to the latest developments in objective testing in mathematics at Birmingham Polytechnic. It was stated earlier that the method was only occasionally used here in attainment tests, and never in the formal sessional examinations. After the earlier chapters had been written, however, the opportunity was taken to introduce objective items into the continuous assessment scheme recently incorporated in the H.N.C. Mechanical and Production Engineering course. Out of a total of 6 hours testing in the first year, the first hour was devoted to a 25-item multiple-choice test; about half the items were from the Manchester bank. There were 67 candidates and the reliability factor was 0.82. (See Appendix C for this test.)

Since the scores will eventually be aggregated with those of conventional examinations in mathematics, and compared with those

in other subjects, some scaling was necessary. The items were of the four-option type, and since the guessing correction was applied the theoretical range of scores obtained from the computer was from -33.3% to 100%. The final score was obtained by adding 25% to three-quarters of the computer score, converting the theoretical range to the conventional 0 - 100%. The scaled scores had a mean of 49% and a standard deviation of 15%; the distribution was approximately Normal, and the range was from 15% to 85%. These results are therefore quite suitable for combining with those of other tests.

The other development concerns the Diploma course in Estate Management and Surveying. The first-year examinations, which are set by us but externally assessed, include 5 questions on statistics of which the candidate has to answer 2 or 3. We have submitted a paper in which one of the questions consists of 7 multiple-choice items; and this question is compulsory; the internal examination board has approved this arrangement and it is thought unlikely that the assessors will object.

The indications are of a slow but steady growth of objective testing throughout further education in the United Kingdom, of which the above examples are likely to be fairly typical. In spite of my disappointment at the slow pace, it is probably better for such changes to come about voluntarily, as a result of teachers hearing about developments elsewhere, rather than under pressure, and with possible misgivings about the innovations.

BIBLIOGRAPHY

Adams, R.C., and James, D.J.G. (1977). "Polymaths - an opportunity for a second start in mathematics for mature students", Bulletin of the Institute of Mathematics and its Applications, 113, 197 - 201.

Bloom, B.S. (1956). Taxonomy of Educational Objectives: Cognitive Domain. Longmans.

Bonney Rust, W. (1973). Objective Testing in Education and Training. Pitman.

Ebel, R.L. (1972). Essentials of Educational Measurement. Prentice-Hall.

Fraser, W.G., and Gillam, J.N. (1972). The Principles of Objective Testing in Mathematics. Heinemann.

Gillam, J.N. - see Fraser, W.G.

Gronlund, N.E. (1965). Measurement and Evaluation in Teaching. MacMillan.

James, D.J.G. - see Adams, R.C.

Johnson, A.P. (1951). "Notes on a suggested index of item validity", Journal of Educational Psychology, 62, 499 - 504.

Joint Matriculation Board (1970). Examining in Advanced Level Science Subjects of the G.C.E. J.M.B.

Kelly, T.L. (1939). "The selection of upper and lower groups for the validation of test items", Journal of Educational Psychology, 30, 17 - 24.

Kuder, G.F., and Richardson, M.W. (1937). "The theory of estimation of test reliability", Psychometrika, 2, 151 - 60.

Lord, F.M. (1957). "Do tests of the same length have the same standard error of measurements?", Educational and Psychological Measurement, 17, 501 - 21; and

(1959). "Tests of the same length do have the same standard error of measurements", ibid, 19, 233 - 39.

Nuttall, D.L., and Willmott, A.S. (1972). British Examinations. National Foundation for Educational Research in England and Wales.

Richardson, M.W. - see Kuder, G.F.

Rust, W. Bonney - see Bonney Rust, W.

Thorndike, R.L. (1951). "Reliability". In Educational Measurement, edited E. Lindquist. Americal Council on Education.

Willmott, A.S. - see Nuttall, D.L.

Wilson, N. (1970). Objective Tests in Mathematical Learning. Oliver and Boyd.

Completed questionnaires were received from:

Blackburn College of Technology and Design

Brighton Polytechnic

Lanchester Polytechnic, Coventry

Leicester Polytechnic

Liverpool Polytechnic

Manchester Polytechnic

North Staffordshire Polytechnic, Stoke-on-Trent

Wigan College of Technology

APPENDIX A

Statistical Methods

References in brackets show where each method is first referred to in the main chapters.

A.1 Statistical concepts

A measure which is fundamental to nearly all statistical techniques is the standard deviation. This is generally the most useful measure of the natural dispersion of a quantity; its value within a sample is denoted in this dissertation by s, and defined by

$$s = \sqrt{\frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}}$$

where x_i is a member of the sample of n values and $\bar{x}$ is the sample mean, namely

$$\frac{\sum_{i=1}^{n} x_i}{n}$$

s^2 is the sample variance.

The standard deviation, σ, of the population from which the sample was taken is usually unknown; its best estimate is $\hat{\sigma}$, where

$$\hat{\sigma} = s\sqrt{\frac{n}{n-1}}\,.$$

A.1.1 (1.3.2)

The product-moment correlation coefficient of a sample of observations of two variables x and y is given by

$$r = \frac{\sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\left\{ \sum_{i=1}^{n} (x_i - \bar{x})^2 \sum_{i=1}^{n} (y_i - \bar{y})^2 \right\}}}$$

As the sample size n approaches infinity, r approaches the population correlation coefficient, ρ .

A positive value for the coefficient indicates that an increase in x tends to be accompanied by an increase in y; a negative one, that the changes are of opposite sign.

Perfect positive correlation is indicated by $\rho = 1$ and perfect negative correlation by $\rho = -1$; if the variables are mutually independent, then $\rho = 0$. Values of r which are not equal to zero may or may not indicate some degree of mutual dependence. See A.2.2.

A.1.2 (1.3.3)

Analysis of variance is a technique for identifying the source of differences in a variable which is capable of more than one classification. That part of the variation which results from differences between classes under one classification can be isolated and compared with any other, and especially with the residual; the latter is the variation remaining after the elimination of the differences between all the classifications being considered.

A.1.3 (1.3.4)

The Gaussian curve of error, now more usually known as the Normal distribution, applies to the distribution of the errors in most kinds of measurements (and to that of many other variables). The probability density, ϕ, of normally distributed variable x with mean μ and standard deviation σ is given by

$$\phi(x) = \frac{1}{\sigma\sqrt{(2\pi)}} e^{-\frac{1}{2}\left\{\frac{x-\mu}{\sigma}\right\}^2}$$

Thus the probability of x exceeding any critical value X is equal to $\int_X^{\infty} \phi(x)\, dx$. This definite integral cannot be found analytically, and tables are available which, in conjunction with a transformation which standardises x, allow the probability to be read. Such tables show that the probability of a value exceeding the mean by more than 0.6745 σ is one-quarter; because of the symmetry of the curve, the probability of a value differing from the mean by more than 0.6745 σ is one-half.

A.2 Significance testing

This technique consists of putting forward a statement, the Null Hypothesis, for the purposes of testing its validity by means of probability considerations.

If on this hypothesis a result such as the observed one, or a result even more inconsistent with the hypothesis, is unlikely to occur (that is to say, has a probability less than say α%), then the hypothesis is rejected "at the α% level of significance". This means that the risk of being mistaken in rejecting the hypothesis on this basis is not more than α%. A 5% significance level is frequently adopted. If the probability calculated from the observations is not less than α%, the hypothesis is not rejected; this is not the same as accepting it, which would imply that the evidence tended to establish the truth of the hypothesis.

A.2.1 (4.5.1)

Where two sample means $\bar{x}_1$ and $\bar{x}_2$ are being compared with a view to deciding whether or not they differ significantly, the null

hypothesis is that the population mean of $(\bar{x}_1 - \bar{x}_2)$ is zero. The population standard deviation of this variable is called the standard error, and its estimate is denoted by est(SE). Provided the distributions of x_1 and x_2 do not differ too much from the Normal and the samples are not too small, it can be assumed that

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\text{est(SE)}}$$

where $\phi(t)$ has an integral which is available in tables. Unlike the Normal probability integral, this depends on sample size and the tables have to provide information accordingly; this is done by giving t values for various probabilities (significance levels) and various numbers of "degrees of freedom", the latter being the sum of the two sample sizes, reduced by 2.

A.2.2 (4.3.1)

In testing correlation, the t distribution referred to above can be used provided the null hypothesis is that $\rho = 0$. In this case the probability of r numerically exceeding the observed value is that of t numerically exceeding

$$R\sqrt{\frac{n - 2}{1 - R^2}} .$$

n is the number of pairs of values in the sample (giving n - 2 "degrees of freedom") and R is the observed sample correlation coefficient.

APPENDIX B

Computer marking and analysis

B.1 The program.

The Fortran program used at Birmingham has been developed from one written by Swansea University. In its present form it comprises about 600 statements, and is stored on disc for speed of access.

B.2 The print-out

The following example shows the results obtained when the basic mathematics test was given to a group of Loughborough University students, as described in section 4.4; the test itself can be seen in Appendix C (subsection C.2.2).

Details of the nine tables, with comments on this set of results, are as follows:-

Table.

1. The "option code" AC Y 3 shows that correlation, elimination, and a guessing correction of - 1/3 for each wrong response, were called for (see 2.2.4). The Loughborough students are identified by number only (third column). A different format was used for this test, making template marking impossible; no match marks were therefore available, but correlation was called for as otherwise some of the data on Table 9 would have been lost. (It will be noticed that when operated in this way the program sets the correlation to zero; this is to avoid an attempt to divide by zero.)

   The correct responses are printed across the top of the table, aligned with the students' responses. For ease of

reference, the match marks are converted to percentages and so listed in the penultimate column.

2. Here the marks given by the computer are presented in the form of a frequency table; when grouped, these marks are seen to follow an approximately normal distribution with a slight negative skewness; the mean and median are each approximately 23 (out of 37) and the mode 24.

3. Summary of sample statistics.

4. In this table the students are arranged in rank order with the computed scores given in two forms - first, out of the equivalent number of items, and second, as a percentage. The key is listed above each student's responses, in case it is desired to cut up this list and distribute the individual results.

5. The response analysis is given in the usual form, with the correct response given in brackets after the item number; "X" is the response code for an item omitted by the student.

6. Here the response analysis has the frequencies grouped into upper, middle, and lower thirds (or 27%, 46% and 27% if required).

7. In the item analysis, correct responses are grouped as above, and followed by the facility value and the index of discrimination (see 1.3.1 and 1.3.2). It is noteworthy that no negative ID's appeared in this application of the test. ID was zero for three items, but two of these had FV's of 0 and 100% respectively.

8. This table gives the Kuder-Richardson reliability factor,

the mean of all the item ID's, and the probable error. (See 1.3.3 and 1.3.4.)

9. The first part of this table gives an assessment of the test results in percentage terms, including the population parameters.

The second part gives the correlation coefficient between match and computed marks and also shows some of the intermediate values in its determination.

BASIC MATHS. LUT TRANSPORT MANAGEMENT & PLANNING. GUESS-CORRECTED. MAY 1977 TABLE 1

AC Y 3

DATA PROVIDED BY TEACHER:-

	KEY BASIC MATHS			DCAACBCABDBCDBABDBABBDBDCCBDCADDDCDACCAADCBBDACCAC	%	
		50 ITEMS				
1	LUT/PVB	01	0	DAAAABCAADBCXXDBBCDCBDXXXXBBBBABAXXXXCADBCBBXAXAXX	0	1
2	LUT/PVB	02	0	DDAABBCABDBCXXBBDCXBBBXXXXXDCDXCXXXXXCACDCDCXAXCXX	0	2
3	LUT/PVB	03	0	DCAABBCABDCCXXABDBACBDXXXXBBCADAAXXXXCAADCBBXAXCXX	0	3
4	LUT/PVB	04	0	DCAAABCAXDBCXXABDBAXBDXXXXBDDADCBXXXXCAADCBBXAXCXX	0	4
5	LUT/PVB	05	0	DCAAAABBCDXCXXABBBABBDXBXXBBDADBBXXXXCABBCBBXAXCXX	0	5
6	LUT/PVB	06	0	DCABAAAAADBCXXABDBABBDXAXXBDDAABDXXXXCADBDBXXAXCXX	0	6
7	LUT/PVB	07	0	DCAADBCABDBBXXDBBCADBDXXXXBDDADCAXXXXCBADCBBXAXDXX	0	7
8	LUT/PVB	08	0	DCABBBCABDBCXXABDBABBDXBXXBDCDDAXXXXXXXXDBDACXBXCXX	0	8
9	LUT/PVB	09	0	CCDABBCACAXXXXBXDXDCBXXXXXBDDXAAXXXXXDXXBCABXDXCXX	0	9
10	LUT/PVB	10	0	DCAABBCAXDBCXXABDBABBDXDXXBDCCDAAXXXXCAADCBBXAXCXX	0	10
11	LUT/PVB	11	0	DCAAXBCABDBCXXDBBXXDBDXXXXCDDAXBAXXXXXXABXXXXXXCXX	0	11
12	LUT/PVB	12	0	DCABBBCABDBCXXABDBACBDXXXXBCCAACXXXXXDAADCBBXAXCXX	0	12
13	LUT/PVB	13	0	DCAACBCABDBCXXABDBABBDXAXXBDCADBAXXXXCAADDBBXAXCXX	0	13
14	LUT/PVB	14	0	DCAABBCABDXCXXBBDXABBBXBXXBDDADAAXXXXCAABCBBXXXCXX	0	14
15	LUT/PVB	15	0	DCAADBCABDDXXXBBDADBBDXDXXBDDADBDXXXXAAABCBBXAXCXX	0	15
16	LUT/PVB	16	0	DCABCBCABDBCXXDBDDXCBDXAXXBDDDACAXXXXDAADCBBXAXCXX	0	16
17	LUT/PVB	17	0	DCAABBCABDBCXXABDBABBDXBXXBDCADCDXXXXCAADCBBXAXCXX	0	17
18	LUT/PVB	18	0	DCAABBCABDBCXXABDAABBDXXXXBDDAACAXXXXCAXDCBBXXACXX	0	18
19	LUT/PVB	19	0	DCAABBCABDCCXXABDBABBDXXXXBDDADCXXXXXCXADCBBXAXCXX	0	19

HENCEFORWARD ALL SCORES ARE THOSE COMPUTED AFTER ANY ELIMINATION, WEIGHTING AND/OR GUESSING CORRECTION.

AN ITEM GIVEN A WEIGHT OF N IS COUNTED AS N ITEMS.

BASIC MATHS. LUT TRANSPORT MANAGEMENT & PLANNING. GUESS-CORRECTED. MAY 1977 TABLE 2

AC Y 3

37 ITEMS

X	FREQ
*	****
33	1
32	1
31	1
30	0
29	2
28	1
27	0
26	2
25	0
24	1
23	1
22	0
21	2
20	1
19	3
18	0
17	0
16	0
15	1
14	0
13	1
12	0
11	0
10	0
9	0
8	0
7	1

BASIC MATHS. LUT TRANSPORT MANAGEMENT & PLANNING. GUESS-CORRECTED. MAY 1977 TABLE 3

AC Y 3

SAMPLE DATA OF SCORES

NUMBER OF STUDENTS =	19
MINIMUM SCORE =	7
MAXIMUM SCORE =	33
NUMBER OF ITEMS =	37
MEAN =	22.877
VARIANCE =	42.833
STANDARD DEVIATION =	6.545

BASIC MATHS. LUT TRANSPORT MANAGEMENT & PLANNING. GUESS-CORRECTED. MAY 1977 TABLE 4

AC Y 3

KEY BASIC MATHS	DCAA CBCA BDBC DBAB DBAB BDBD CCBD CADD DCDA CCAA DCBB DACC AC	/37	%	
LUT/PVB 17	DCAA BBCA BDBC XXAB DBAB BDXB XXBD CADC DXXX XCAA DCBB XAXC XX	33	89	
KEY BASIC MATHS	DCAA CBCA BDBC DBAB DBAB BDBD CCBD CADD DCDA CCAA DCBB DACC AC	/37	%	
LUT/PVB 13	DCAA CBCA BDBC XXAB DBAB BDXA XXBD CADB AXXX XCAA DDBB XAXC XX	32	86	
KEY BASIC MATHS	DCAA CBCA BDBC DBAB DBAB BDBD CCBD CADD DCDA CCAA DCBB DACC AC	/37	%	
LUT/PVB 10	DCAA BBCA XDBC XXAB DBAB BDXD XXBD CCDA AXXX XCAA DCBB XAXC XX	31	83	
KEY BASIC MATHS	DCAA CBCA BDBC DBAB DBAB BDBD CCBD CADD DCDA CCAA DCBB DACC AC	/37	%	
LUT/PVB 19	DCAA BBCA BDCC XXAB DBAB BDXX XXBD DADC XXXX XCXA DCBB XAXC XX	29	77	
KEY BASIC MATHS	DCAA CBCA BDBC DBAB DBAB BDBD CCBD CADD DCDA CCAA DCBB DACC AC	/37	%	
LUT/PVB 04	DCAA ABCA XDBC XXAB DBAX BDXX XXBD DADC BXXX XCAA DCBB XAXC XX	29	77	
KEY BASIC MATHS	DCAA CBCA BDBC DBAB DBAB BDBD CCBD CADD DCDA CCAA DCBB DACC AC	/37	%	
LUT/PVB 03	DCAA BBCA BDCC XXAB DBAC BDXX XXBB CADA AXXX XCAA DCBB XAXC XX	28	76	
KEY BASIC MATHS	DCAA CBCA BDBC DBAB DBAB BDBD CCBD CADD DCDA CCAA DCBB DACC AC	/37	%	
LUT/PVB 18	DCAA BBCA BDBC XXAB DAAB BDXX XXBD DAAC AXXX XCAX DCBB XXAC XX	26	70	
KEY BASIC MATHS	DCAA CBCA BDBC DBAB DBAB BDBD CCBD CADD DCDA CCAA DCBB DACC AC	/37	%	
LUT/PVB 12	DCAB BBCA BDBC XXAB DBAC BDXX XXBC CAAC XXXX XDAA DCBB XAXC XX	26	69	
KEY BASIC MATHS	DCAA CBCA BDBC DBAB DBAB BDBD CCBD CADD DCDA CCAA DCBB DACC AC	/37	%	
LUT/PVB 15	DCAA DBCA BDDX XXBB DADB BDXD XXBD DADB DXXX XAAA BCBB XAXC XX	24	65	
KEY BASIC MATHS	DCAA CBCA BDBC DBAB DBAB BDBD CCBD CADD DCDA CCAA DCBB DACC AC	/37	%	
LUT/PVB 14	DCAA BBCA BDXC XXBB DXAB BBXB XXBD DADA AXXX XCAA BCBB XXXC XX	23	63	
KEY BASIC MATHS	DCAA CBCA BDBC DBAB DBAB BDBD CCBD CADD DCDA CCAA DCBB DACC AC	/37	%	
LUT/PVB 16	DCAB CBCA BDBC XXDB DDXC BDXA XXBD DDAC AXXX XDAA DCBB XAXC XX	21	58	

KEY	BASIC	MATHS	DCAA	CBCA	BDBC	DBAB	DBAB	BDBD	CCBD	CADD	DCDA	CCAA	DCBB	DACC	AC	/37	%
LUT/PVB		07	DCAA	DBCA	BDBB	XXDB	BCAD	BDXX	XXBD	DADC	AXXX	XCBA	DCBB	XAXD	XX	21	58
KEY	BASIC	MATHS	DCAA	CBCA	BDBC	DBAB	DBAB	BDBD	CCBD	CADD	DCDA	CCAA	DCBB	DACC	AC	/37	%
LUT/PVB		06	DCAB	AAAA	ADBC	XXAB	DBAB	BDXA	XXBD	DAAB	DXXX	XCAD	BDBX	XAXC	XX	20	54
KEY	BASIC	MATHS	DCAA	CBCA	BDBC	DBAB	DBAB	BDBD	CCBD	CADD	DCDA	CCAA	DCBB	DACC	AC	/37	%
LUT/PVB		08	DCAB	BBCA	BDBC	XXAB	DBAB	BDXB	XXBD	CDDA	XXXX	XXXD	BDAC	XBXC	XX	19	52
KEY	BASIC	MATHS	DCAA	CBCA	BDBC	DBAB	DBAB	BDBD	CCBD	CADD	DCDA	CCAA	DCBB	DACC	AC	/37	%
LUT/PVB		05	DCAA	AABB	CDXC	XXAB	BBAB	BDXB	XXBB	DADB	BXXX	XCAB	BCBB	XAXC	XX	19	50
KEY	BASIC	MATHS	DCAA	CBCA	BDBC	DBAB	DBAB	BDBD	CCBD	CADD	DCDA	CCAA	DCBB	DACC	AC	/37	%
LUT/PVB		02	DDAA	BBCA	BDBC	XXBB	DCXB	BBXX	XXXD	CDXC	XXXX	XCAC	DCDC	XAXC	XX	19	50
KEY	BASIC	MATHS	DCAA	CBCA	BDBC	DBAB	DBAB	BDBD	CCBD	CADD	DCDA	CCAA	DCBB	DACC	AC	/37	%
LUT/PVB		11	DCAA	XBCA	BDBC	XXDB	BXXD	BDXX	XXCD	DAXB	AXXX	XXXA	BXXX	XXXC	XX	15	41
KEY	BASIC	MATHS	DCAA	CBCA	BDBC	DBAB	DBAB	BDBD	CCBD	CADD	DCDA	CCAA	DCBB	DACC	AC	/37	%
LUT/PVB		01	DAAA	ABCA	ADBC	XXDB	BCDC	BDXX	XXBB	BBAB	AXXX	XCAD	BCBB	XAXA	XX	13	36
KEY	BASIC	MATHS	DCAA	CBCA	BDBC	DBAB	DBAB	BDBD	CCBD	CADD	DCDA	CCAA	DCBB	DACC	AC	/37	%
LUT/PVB		09	CCDA	BBCA	CAXX	XXBX	DXDC	BXXX	XXBD	DXAA	XXXX	XDXX	BCAB	XDXC	XX	7	19

BASIC MATHS. LUT TRANSPORT MANAGEMENT & PLANNING. GUESS-CORRECTED. MAY 1977 TABLE 5

AC Y 3

ITEM		A	B	C	D	E	X
1	(D)	0	0	1	18	0	0
2	(C)	1	0	17	1	0	0
3	(A)	18	0	0	1	0	0
4	(A)	15	4	0	0	0	0
5	(C)	4	10	2	2	0	1
6	(B)	2	17	0	0	0	0
7	(C)	1	1	17	0	0	0
8	(A)	18	1	0	0	0	0
9	(B)	2	13	2	0	0	2
10	(D)	1	0	0	18	0	0
11	(B)	0	13	2	1	0	3
12	(C)	0	1	16	0	0	2
			THIS ITEM ELIMINATED				
			THIS ITEM ELIMINATED				
15	(A)	11	4	0	4	0	0
16	(B)	0	18	0	0	0	1
17	(D)	0	4	0	15	0	0
18	(B)	2	10	3	1	0	3
19	(A)	13	0	0	3	0	3
20	(B)	0	11	5	2	0	1
21	(B)	0	19	0	0	0	0
22	(D)	0	2	0	16	0	1
			THIS ITEM ELIMINATED				
24	(D)	3	4	0	2	0	10
			THIS ITEM ELIMINATED				
			THIS ITEM ELIMINATED				
27	(B)	0	17	1	0	0	1
28	(D)	0	3	1	15	0	0
29	(C)	0	1	7	11	0	0
30	(A)	13	1	1	3	0	1
31	(D)	6	0	0	11	0	2
32	(D)	5	6	8	0	0	0
33	(D)	9	2	0	3	0	5
			THIS ITEM ELIMINATED				
			THIS ITEM ELIMINATED				
			THIS ITEM ELIMINATED				
			THIS ITEM ELIMINATED				
38	(C)	1	0	13	3	0	2
39	(A)	14	1	0	0	0	4
40	(A)	12	1	1	3	0	2
41	(D)	0	8	0	11	0	0
42	(C)	0	0	15	3	0	1
43	(B)	2	15	0	1	0	1
44	(B)	0	15	2	0	0	2
			THIS ITEM ELIMINATED				
46	(A)	14	1	0	1	0	3
			THIS ITEM ELIMINATED				
48	(C)	1	0	17	1	0	0
			THIS ITEM ELIMINATED				
			THIS ITEM ELIMINATED				

BASIC MATHS. LUT TRANSPORT MANAGEMENT & PLANNING. GUESS-CORRECTED. MAY 1977 TABLE 6

AC Y 3

ITEM		A			B			C			D			E			X		
		U	M	L	U	M	L	U	M	L	U	M	L	U	M	L	U	M	L
1	(D)	0	0	0	0	0	0	0	0	1	6	7	5	0	0	0	0	0	0
2	(C)	0	0	1	0	0	0	6	7	4	0	0	1	0	0	0	0	0	0
3	(A)	6	7	5	0	0	0	0	0	0	0	0	1	0	0	0	0	0	0
4	(A)	6	4	5	0	3	1	0	0	0	0	0	0	0	0	0	0	0	0
5	(C)	1	1	2	4	3	3	1	1	0	0	2	0	0	0	0	0	0	1
6	(B)	0	1	1	6	6	5	0	0	0	0	0	0	0	0	0	0	0	0
7	(C)	0	1	0	0	0	1	6	6	5	0	0	0	0	0	0	0	0	0
8	(A)	6	7	5	0	0	1	0	0	0	0	0	0	0	0	0	0	0	0
9	(B)	0	1	1	4	6	3	0	0	2	0	0	0	0	0	0	2	0	0
10	(D)	0	0	1	0	0	0	0	0	0	6	7	5	0	0	0	0	0	0
11	(B)	0	0	0	4	5	4	2	0	0	0	1	0	0	0	0	0	1	2
12	(C)	0	0	0	0	1	0	6	5	5	0	0	0	0	0	0	0	1	1
THIS ITEM ELIMINATED																			
THIS ITEM ELIMINATED																			
15	(A)	6	3	2	0	2	2	0	0	0	0	2	2	0	0	0	0	0	0
16	(B)	0	0	0	6	7	5	0	0	0	0	0	0	0	0	0	0	0	1
17	(D)	0	0	0	0	1	3	0	0	0	6	6	3	0	0	0	0	0	0
18	(B)	0	2	0	6	2	2	0	1	2	0	1	0	0	0	0	0	1	2
19	(A)	6	5	2	0	0	0	0	0	0	0	1	2	0	0	0	0	1	2
20	(B)	0	0	0	4	4	3	1	2	2	0	1	1	0	0	0	1	0	0
21	(B)	0	0	0	6	7	6	0	0	0	0	0	0	0	0	0	0	0	0
22	(D)	0	0	0	0	1	1	0	0	0	6	6	4	0	0	0	0	0	1
THIS ITEM ELIMINATED																			
24	(D)	1	2	0	1	1	2	0	0	0	1	1	0	0	0	0	3	3	4
THIS ITEM ELIMINATED																			
THIS ITEM ELIMINATED																			
27	(B)	0	0	0	6	7	4	0	0	1	0	0	0	0	0	0	0	0	1
28	(D)	0	0	0	1	0	2	0	1	0	5	6	4	0	0	0	0	0	0
29	(C)	0	0	0	0	0	1	4	1	2	2	6	3	0	0	0	0	0	0
30	(A)	5	6	2	0	0	1	1	0	0	0	1	2	0	0	0	0	0	1
31	(D)	0	4	2	0	0	0	0	0	0	6	3	2	0	0	0	0	0	2
32	(D)	2	1	2	1	2	3	3	4	1	0	0	0	0	0	0	0	0	0
33	(D)	3	4	2	1	0	1	0	0	0	1	2	0	0	0	0	1	1	3
THIS ITEM ELIMINATED																			
THIS ITEM ELIMINATED																			
THIS ITEM ELIMINATED																			
THIS ITEM ELIMINATED																			
38	(C)	0	1	0	0	0	0	6	4	3	0	2	1	0	0	0	0	0	2
39	(A)	5	6	3	0	1	0	0	0	0	0	0	0	0	0	0	1	0	3
40	(A)	6	5	1	0	0	1	0	0	1	0	1	2	0	0	0	0	1	1
41	(D)	0	0	0	0	3	5	0	0	0	6	4	1	0	0	0	0	0	0
42	(C)	0	0	0	0	0	0	5	6	4	1	1	1	0	0	0	0	0	1
43	(B)	0	0	2	6	7	2	0	0	0	0	0	1	0	0	0	0	0	1
44	(B)	0	0	0	6	6	3	0	0	2	0	0	0	0	0	0	0	1	1
THIS ITEM ELIMINATED																			
46	(A)	6	5	3	0	0	1	0	0	0	0	0	1	0	0	0	0	2	1
THIS ITEM ELIMINATED																			
48	(C)	0	0	1	0	0	0	6	6	5	0	1	0	0	0	0	0	0	0
THIS ITEM ELIMINATED																			
THIS ITEM ELIMINATED																			

BASIC MATHS. LUT TRANSPORT MANAGEMENT & PLANNING. GUESS-CORRECTED. MAY 1977 TABLE 7

AC Y 3

DISTRIBUTION OF CORRECT RESPONSES; NUMBER OF OMISSIONS IN BRACKETS.

ITEM	UPPER	MIDDLE	LOWER	TOTAL	FV %	ID
1	6(0)	7(0)	5(0)	18(0)	94.74	0.17
2	6(0)	7(0)	4(0)	17(0)	89.47	0.33
3	6(0)	7(0)	5(0)	18(0)	94.74	0.17
4	6(0)	4(0)	5(0)	15(0)	78.95	0.17
5	1(0)	1(0)	0(1)	2(1)	10.53	0.17
6	6(0)	6(0)	5(0)	17(0)	89.47	0.17
7	6(0)	6(0)	5(0)	17(0)	89.47	0.17
8	6(0)	7(0)	5(0)	18(0)	94.74	0.17
9	4(2)	6(0)	3(0)	13(2)	68.42	0.17
10	6(0)	7(0)	5(0)	18(0)	94.74	0.17
11	4(0)	5(1)	4(2)	13(3)	68.42	0.00
12	6(0)	5(1)	5(1)	16(2)	84.21	0.17
		THIS ITEM ELIMINATED				
		THIS ITEM ELIMINATED				
15	6(0)	3(0)	2(0)	11(0)	57.89	0.67
16	6(0)	7(0)	5(1)	18(1)	94.74	0.17
17	6(0)	6(0)	3(0)	15(0)	78.95	0.50
18	6(0)	2(1)	2(2)	10(3)	52.63	0.67
19	6(0)	5(1)	2(2)	13(3)	68.42	0.67
20	4(1)	4(0)	3(0)	11(1)	57.89	0.17
21	6(0)	7(0)	6(0)	19(0)	100.00	0.00
22	6(0)	6(0)	4(1)	16(1)	84.21	0.33
		THIS ITEM ELIMINATED				
24	1(3)	1(3)	0(4)	2(10)	10.53	0.17
		THIS ITEM ELIMINATED				
		THIS ITEM ELIMINATED				
27	6(0)	7(0)	4(1)	17(1)	89.47	0.33
28	5(0)	6(0)	4(0)	15(0)	78.95	0.17
29	4(0)	1(0)	2(0)	7(0)	36.84	0.33
30	5(0)	6(0)	2(1)	13(1)	68.42	0.50
31	6(0)	3(0)	2(2)	11(2)	57.89	0.67
32	0(0)	0(0)	0(0)	0(0)	0.00	0.00
33	1(1)	2(1)	0(3)	3(5)	15.79	0.17
		THIS ITEM ELIMINATED				
		THIS ITEM ELIMINATED				
		THIS ITEM ELIMINATED				
		THIS ITEM ELIMINATED				
38	6(0)	4(0)	3(2)	13(2)	68.42	0.50
39	5(1)	6(0)	3(3)	14(4)	73.68	0.33
40	6(0)	5(1)	1(1)	12(2)	63.16	0.83
41	6(0)	4(0)	1(0)	11(0)	57.89	0.83
42	5(0)	6(0)	4(1)	15(1)	78.95	0.17
43	6(0)	7(0)	2(1)	15(1)	78.95	0.67
44	6(0)	6(1)	3(1)	15(2)	78.95	0.50
		THIS ITEM ELIMINATED				
46	6(0)	5(2)	3(1)	14(3)	73.68	0.50
		THIS ITEM ELIMINATED				
48	6(0)	6(0)	5(0)	17(0)	89.47	0.17
		THIS ITEM ELIMINATED				
		THIS ITEM ELIMINATED				

BASIC MATHS. LUT TRANSPORT MANAGEMENT & PLANNING. GUESS-CORRECTED. MAY 1977 TABLE 8

AC Y 3

XMN = 22.88

SUM PQ = 5.46

$$REL = \frac{37}{36} * \frac{42.833 - 5.463}{42.833}$$

$$= 0.90$$

MEAN DISCRIMINATION = 0.32

PROBABLE ERROR = 3.8

FOR ABOUT HALF THE CANDIDATES, THE TRUE PERCENTAGE SCORES WILL DIFFER FROM THOSE GIVEN BY LESS THAN THE PROBABLE ERROR.

BASIC MATHS. LUT TRANSPORT MANAGEMENT & PLANNING. GUESS-CORRECTED. MAY 1977 TABLE 9

AC Y 3

NUMBER OF ITEMS IN STANDARD TEST =	50	
NUMBER AFTER ELIMINATION AND WEIGHTING =	37	
NUMBER OF STUDENTS =	19	
MEAN OF PERCENTAGE SCORES =	61.8	
STANDARD DEVIATION OF PERCENTAGE SCORES =	18.2	(POPULATION)
SUM(SCORE) =	435	
SUM(TEACHER) =	0	
SUM(SCORE*SCORE) =	10758	
SUM(TEACHER*TEACHER) =	0	
SUM(TEACHER*SCORE) =	0	
VARIANCE(SCORE) =	42.833	(SAMPLE)
VARIANCE(TEACHER) =	0.000	(SAMPLE)
COVARIANCE =	0.000	
CORRELATION COEFFICIENT =	0.000	

PROGRAM RUN ON 22/12/77

APPENDIX C

Some multiple-choice tests

C.1 General

All the tests referred to in Chapter IV are reproduced here, together with a test which includes items of a more searching nature than usual. Some items with abnormally low indices of discrimination, as discussed in subsection 1.3.2, are also given.

C.2.1

The 14-item test on basic mathematics referred to in subsection 4.3.1 is given below.

ELEMENTARY MATHEMATICS TEST

NAME (Surname first) .....

COURSE ..... DATE .....

There is <u>one</u> correct response to each item; place a tick, ✓, in the appropriate box. Avoid blind guessing; there is a small penalty for wrong responses but not for omissions.

Reference may be made to notes, but <u>not</u> to calculators, slide-rules or tables.

1. $\sqrt{0.004}$ =

A.	0.0063 approximately	B.	0.02	A ☐
C.	0.063 approximately	D.	0.2	B ☐
				C ☐
				D ☐

2. $\sqrt{0.742}$ is approximately equal to

A. 0.0861 B. 0.272

C. 0.861 D. none of A, B, and C

A ☐
B ☐
C ☐
D ☐

3. 5.2449 when rounded to two decimal places is equal to

A. 5.20 B. 5.24

C. 5.25 D. 5.30

A ☐
B ☐
C ☐
D ☐

4. $1/\pi^2$ is approximately equal to

A. 0.010 B. 0.032

C. 0.10 D. 0.32

A ☐
B ☐
C ☐
D ☐

5. $68^2 - 66^2 =$

A. 168 B. 188

C. 208 D. 268

A ☐
B ☐
C ☐
D ☐

6. $a^3 - b^3 =$

A. $(a - b)(a^2 + b^2)$ B. $(a + b)(a^2 - ab + b^2)$

C. $(a - b)(a^2 + ab + b^2)$ D. $(a - b)(a^2 + 2ab + b^2)$

A ☐
B ☐
C ☐
D ☐

7. $x^4 - y^4 =$

A. $(x^2 + y^2)(x + y)(x - 6)$

B. $(x + y)^2(x - y)^2$

C. $(x - y)(x^3 + y^3)$

D. none of A, B, and C

A ☐
B ☐
C ☐
D ☐

8. $\log_{10} 0.001 =$

A. -2

B. $\bar{2}.0$

C. $\bar{3}.0$

D. $\bar{3}.1$

A ☐
B ☐
C ☐
D ☐

9. $\log_4\left(\frac{1}{64}\right)$

A. -16

B. -4

C. -3

D. $\frac{1}{3}$

A ☐
B ☐
C ☐
D ☐

10. $\log_{10}(100a) =$

A. $100 \log_{10} a$

B. $1 + \log_{10} a$

C. $2 + \log_{10} a$

D. $\log_{10}(a^2)$

A ☐
B ☐
C ☐
D ☐

11. $(-0.01)^2 =$

A. -0.001

B. 0.001

C. -0.0001

D. 0.0001

A ☐
B ☐
C ☐
D ☐

12. $24 \times 0.00352 =$

A. 0.015 approximately
B. 0.08 approximately
C. 0.15 approximately
D. 0.8 approximately

A ☐
B ☐
C ☐
D ☐

13. 3 revolutions can be expressed as

A. 6 right-angles
B. 1000^{o} approximately
C. 10 radians approximately
D. 40 radians approximately

A ☐
B ☐
C ☐
D ☐

14. $\frac{d}{dx} \surd (2x) =$

A. $\frac{1}{2} \surd (\frac{2}{x})$
B. $\surd (\frac{2}{x})$
C. $\surd (\frac{1}{x})$
D. $\frac{1}{2 \surd (2x)}$

A ☐
B ☐
C ☐
D ☐

C.2.2

The following is the 50-item basic mathematics test referred to in subsection 4.3.2. A shortened version of this test was used at Loughborough University, as described in section 4.4; an analysis of the results of this test is given in Appendix B.

BASIC MATHEMATICS TEST

NAME (Surname first) .....

COURSE ..... DATE .....

There is one correct response to each item; place a tick, ✓ , in the appropriate box. Avoid blind guessing; there is a small penalty for wrong responses but not for omissions.

Reference may be made to notes, but not to calculators, slide-rules or tables.

1. $\frac{1}{0.02}$ =

A. 0.05 B. 5 C. 20 D. 50

A ☐ B ☐ C ☐ D ☐

2. $(-0.03)^2$ =

A. -0.09 B. -0.06 C. 0.0009 D. 0.1732

A ☐ B ☐ C ☐ D ☐

3. $\sqrt{810}$ =

A. 28.5 approximately B. 90 C. 405 D. none of A, B, and C

A ☐ B ☐ C ☐ D ☐

4. 15.2449 when rounded to two decimal places is represented by

A. 15.24 B. 15.25 C. 1.52×10^1 D. 0.15×10^2

A ☐ B ☐ C ☐ D ☐

5. 165.849 when rounded to two significant figures is represented by

A. 165.85 B. 170

C. 1.7×10^2 D. 1.66×10^2

A ☐
B ☐
C ☐
D ☐

6. 0.163502 when rounded to three decimal places is represented by

A. 0.163 B. 0.164

C. 1.635×10^{-1} D. none of A, B, and C

A ☐
B ☐
C ☐
D ☐

7. $\frac{2}{3}$ =

A. 0.6 correct to 1 decimal place B. 0.66 correct to 2 decimal places

C. 0.7 correct to 1 decimal place D. none of A, B, and C

A ☐
B ☐
C ☐
D ☐

8. $\frac{15.1}{9} \times 2$ =

A. $\frac{15.1}{4.5}$ B. $\frac{30.2}{18}$

C. $\frac{7.55}{9}$ D. $\frac{15.1}{18}$

A ☐
B ☐
C ☐
D ☐

9. $\sqrt{(10^{3.2})}$ =

A. 5.657 B. 39.81

C. 61.5 D. 1585

A ☐
B ☐
C ☐
D ☐

10. $\frac{1}{2} \div \frac{1}{4} =$

A. $\frac{1}{8}$ B. $\frac{1}{6}$

C. $\frac{1}{2}$ D. 2

A ☐
B ☐
C ☐
D ☐

11. $\log_4 1024 =$

A. 4 B. 5

C. 6 D. 7.071 approximately

A ☐
B ☐
C ☐
D ☐

12. $\frac{1}{a} - \frac{1}{b} =$

A. $a - b$ B. $\frac{a - b}{ab}$

C. $\frac{b - a}{ab}$ D. $\frac{ab}{b - a}$

A ☐
B ☐
C ☐
D ☐

13. $\sqrt{(x^2 + y^2)}$

A. $x + y$ B. $x + y - 2xy$

C. $(x + y)/2$ D. none of A, B, and C

A ☐
B ☐
C ☐
D ☐

14. $x^2 - y^2 =$

A. $(x - y)^2 + 2xy$ B. $(x + y)(x - y)$

C. $(x + y)^2 - 2xy$ D. none of A, B, and C

A ☐
B ☐
C ☐
D ☐

15. $x^2 + y^2 =$

A. $(x + y)^2 - 2xy$ B. $(x + y)(x - y)$

C. $(x - y)^2 + 4xy$ D. $(x + y)^2$

A ☐
B ☐
C ☐
D ☐

16. $ax + by + bz =$

A. $b(ax + y + z)$ B. $ax + b(y + z)$

C. $a(x + by + bz)$ D. $b(\frac{a}{b}x + y + bz)$

A ☐
B ☐
C ☐
D ☐

17. $\frac{8a}{8b} =$

A. 8 B. $8(\frac{a}{b})$

C. 1 D. $\frac{a}{b}$

A ☐
B ☐
C ☐
D ☐

18. $1.3\log_{10}x =$

A. $x^{1.3}$ B. $\log_{10}(x^{1.3})$

C. $(\log_{10}x)^{1.3}$ D. $\log_{10}(1.3x)$

A ☐
B ☐
C ☐
D ☐

19. $\log_e(\frac{1}{e}) =$

A. -1 B. 0

C. 1 D. e

A ☐
B ☐
C ☐
D ☐

20. $\frac{x}{y} \div 3 =$

A. $\frac{3x}{3y}$ B. $\frac{x}{3y}$

C. $\frac{x}{y/3}$ D. none of A, B, and C

A ☐
B ☐
C ☐
D ☐

21. $1000^0 =$

A. 0 B. 1

C. 10 D. 1000

A ☐
B ☐
C ☐
D ☐

22. $(a^x)^y =$

A. $(ya)^x$ B. a^{x+y}

C. $y(a^x)$ D. a^{xy}

A ☐
B ☐
C ☐
D ☐

23. If one root of the equation $2x^2 - 9x + c = 0$ is 6.13, the other root is

A. -6.13 B. -1.63

C. -1 D. 0

A ☐
B ☐
C ☐
D ☐

24. Two of the roots of a cubic equation with real coefficients are 3 and (6 + j5). The third root is

A. real B. imaginary

C. -3 D. 6 - j5

A ☐
B ☐
C ☐
D ☐

25. Two of the roots of a quartic equation with real coefficients are 8 and (2 - j27). Which of the following statements applies to the other two roots?

A. Both are real
B. Both are complex
C. One is real, one complex
D. Without knowing the equation, nothing can be said

A ☐
B ☐
C ☐
D ☐

26. Three of the roots of a quintic equation with real coefficients are complex. Which of the following statements applies to the other two roots?

A. Both are real
B. Both are complex
C. One is real, one complex
D. Without further information, nothing can be said

A ☐
B ☐
C ☐
D ☐

27. $\frac{d}{dx}(x^n) =$

A. nx
B. nx^{n-1}
C. nx^{n+1}
D. $\frac{x^{n+1}}{n+1}$

A ☐
B ☐
C ☐
D ☐

28. $\frac{d}{d\theta}(\cos 2\theta) =$

A. $\sin 2\theta$
B. $-\sin 2\theta$
C. $2\sin 2\theta$
D. $-2\sin 2\theta$

A ☐
B ☐
C ☐
D ☐

29. $\frac{d}{dx}\left(\frac{u}{v}\right) =$

A. $\frac{du/dx}{dv/dx}$

B. $v\frac{du}{dx} + u\frac{dv}{dx}$

C. $\frac{1}{v}\frac{du}{dx} - \frac{u}{v^2}\frac{dv}{dx}$

D. none of A, B, and C

A ☐
B ☐
C ☐
D ☐

30. $\frac{d}{dx}e^{(x^2)} =$

A. $2xe^{(x^2)}$

B. $2e^x$

C. $x^2e^{(x^2)}$

D. none of A, B, and C

A ☐
B ☐
C ☐
D ☐

31. $\int 2xdx =$

A. x^2

B. 2

C. $2 + c$

D. $x^2 + c$

A ☐
B ☐
C ☐
D ☐

32. $\int e^{(x^2)}dx =$

A. $e^{(x^2)} + c$

B. $\frac{1}{2}e^{(x^2)} + c$

C. $\frac{e^{(x^2)}}{2x} + c$

D. none of A, B, and C

A ☐
B ☐
C ☐
D ☐

33. $\int \frac{dx}{3x} =$

A. $\ln(3x) + c$

B. $\frac{1}{3}\ln x$

C. $\ln(ax) + c$

D. none of A, B, and C

A ☐
B ☐
C ☐
D ☐

34. If $\frac{dy}{dx} = y$, then $y =$

A. $\frac{y^2}{2} + c$

B. $\frac{x^2}{2} + c$

C. ae^x

D. $e^x + c$

A ☐
B ☐
C ☐
D ☐

35. If $\frac{dy}{dx} = \frac{d^2y}{dx^2} = 0$ at the point (h, k), then (h,k)

A. is a point of inflexion

B. is a maximum

C. is a minimum

D. could be any of these three

A ☐
B ☐
C ☐
D ☐

36. Which of the following is a necessary and sufficient condition for a point of inflexion?

A. The gradient is a maximum or a minimum

B. The second derivative is zero

C. The curvature is zero

D. None of A, B, and C

A ☐
B ☐
C ☐
D ☐

37. Which of the following is a correct statement of the value of π ?

A. $\frac{22}{7}$

B. 3.141 to 4 significant figures

C. 3.142 to 4 significant figures

D. 3.143 to 4 significant figures

A ☐
B ☐
C ☐
D ☐

38. $\frac{\sin 2x}{\cos 2x} =$

A. tan x

B. 2 tan x

C. tan 2x

D. 2 tan 2x

A ☐
B ☐
C ☐
D ☐

39. $\cos^2\theta + \sin^2\theta =$

A. 1 B. $(\cos\theta + \sin\theta)^2$

C. $\sin 2\theta$ D. $\cos 2\theta$

A ☐
B ☐
C ☐
D ☐

40. $\tan 30^\circ =$

A. $\frac{\sqrt{3}}{3}$ B. $\frac{\sqrt{3}}{2}$

C. $\sqrt{3}$ D. $\frac{1}{2}$

A ☐
B ☐
C ☐
D ☐

41. If $\cos\theta = \frac{1}{2}$ then $\theta =$

A. 30° or 150° B. 60° or 120°

C. $n180^\circ + 30^\circ$ D. $n360^\circ \pm 60^\circ$

A ☐
B ☐
C ☐
D ☐

42. 1 radian is equal to

A. 1 revolution B. 60°

C. $\frac{180^\circ}{\pi}$ D. none of A, B, and C

A ☐
B ☐
C ☐
D ☐

43. 1 right-angle is equal to

A. $\frac{\pi}{4}$ radians B. $\frac{\pi}{2}$ radians

C. π radians D. 2π radians

A ☐
B ☐
C ☐
D ☐

44. 4π radians =

A. 1 revolution B. 2 revolutions

C. 4 right-angles D. 400°

A ☐
B ☐
C ☐
D ☐

45. The sum of the angles of an n-sided polygon is

A. $n\pi/2$ B. $2\pi n$

C. $(n - 1)\pi$ D. $(n - 2)\pi$

A ☐
B ☐
C ☐
D ☐

46. Which of these equations represents a straight line?

A. $x + 2y + 3 = 0$ B. $\frac{1}{x} + \frac{1}{y} = \frac{1}{2}$

C. $xy = 4$ D. $y = \frac{1}{x + 2}$

A ☐
B ☐
C ☐
D ☐

47. Which of these equations represents a parabola?

A. $x^2 = y^2 + 1$ B. $y = x^{1/3}$

C. $y^2 + y = 3x$ D. $y = x + 1$

A ☐
B ☐
C ☐
D ☐

48. The volume of a cube with an edge of 3mm is

A. $3\ mm^3$ B. $9\ mm^3$

C. $27\ mm^3$ D. $81\ mm^3$

A ☐
B ☐
C ☐
D ☐

49. A sphere 2 metres in diameter has a surface area of approximately

A. $12.6\ m^2$ B. $34\ m^2$

C. $50\ m^2$ D. $67\ m^2$

A ☐
B ☐
C ☐
D ☐

50. Which of the following solids has the greatest volume?

A. Pyramid, base 30 mm^2, height 90 mm

B. Cone, base 30 mm^2, height 91 mm

C. Cylinder, diameter 10 mm, height 14 mm

D. Cube, edge 10 mm

A ☐
B ☐
C ☐
D ☐

C.2.3

The two 20-item statistics tests referred to in subsections 4.5.1 and 4.5.2 are given below. Test Y was answered from memory in the experiment.

Statistics

Test X

NAME: COURSE: DATE:

For each item, there is one correct response; place a tick in the box appropriate to your choice of response.

Avoid blind guessing; there will be a small penalty for wrong answers but not for omissions.

Calculators and tables are not required for this test, but reference to notes and text-books is permitted.

1. 12.4, 13.9, 15.0, 15.6, 15.7, 19.7, 28.4

One position measure of the above sample has a value of 15.6; this position measure is the:

A. geometric mean

B. median

C. arithmetic mean

D. mid-range

A ☐
B ☐
C ☐
D ☐

2. 4.9, 5.8, 6.1, 6.9, 7.0, 7.2, 7.7, 7.9, 8.1, 8.6, 9.0, 10.1, 10.7

The standard deviation of the above sample is:

A. 5.8 B. 3.8

C. 1.59 D. 0.20

A ☐
B ☐
C ☐
D ☐

3. When sample values of the variable x are transformed by the relationship $X = \frac{x - 35}{0.2}$, it is found that $\bar{X} = 1.8$.

The sample mean $\bar{x}$ is equal to:

A. 35.36 B. 0.36

C. 44.0 D. 9.0

A ☐
B ☐
C ☐
D ☐

4. Which of the following is a correct expression for sample variance?

A. $\frac{\Sigma x^2}{n} - (\bar{x})^2$ B. $\frac{(\Sigma x)^2}{n} - \frac{\Sigma x^2}{n^2}$

C. $\frac{\Sigma (x - \bar{x})^2}{n - 1}$ D. $\frac{(\Sigma x)^2}{n - 1}$

A ☐
B ☐
C ☐
D ☐

5. If a small sample only is available and it is desired to minimise the effect of any extreme values, the best measure of dispersion to use is the:

A. median B. range

C. standard deviation D. interquartile range

A ☐
B ☐
C ☐
D ☐

6.

x	2.0	2.1	2.2	2.3	2.4	2.5	2.6
f	1	1	3	5	9	7	6

The above distribution is:

A. positively skewed
B. negatively skewed
C. symmetrical
D. Poissonian

A ☐
B ☐
C ☐
D ☐

7. The mode, median and mean value of a frequency distribution are all found to equal -16. The distribution must be:

A. normal
B. symmetrical
C. artificially contrived
D. binomial

A ☐
B ☐
C ☐
D ☐

8. A certain event has a very low constant probability of occurring, but a great many opportunities per month. The monthly results for one year are noted; the resulting distribution is:

A. normal
B. rectangular
C. Poissonian
D. unpredictable

A ☐
B ☐
C ☐
D ☐

9. If a fair coin is tossed three times and shows heads the first two, the probability of third throw showing heads is:

A. 1
B. $\frac{2}{3}$
C. $\frac{1}{2}$
D. $\frac{1}{3}$

A ☐
B ☐
C ☐
D ☐

10. The probability of two events both occurring is the product of their separate probabilities if, and only if, the events are:

A. mutually independent
B. mutually exclusive
C. exhaustive
D. mutually exclusive and exhaustive

A ☐
B ☐
C ☐
D ☐

11. The curve which encloses between itself the x axis, and two ordinates, an area equal to the probability of a value of x lying between these ordinates, is a curve of:

A. probability density
B. relative frequency
C. cumulative frequency
D. frequency

A ☐
B ☐
C ☐
D ☐

12. The maximum ordinate on a cumulative relative frequency curve is:

A. the sample size
B. unity
C. the mode
D. the median

A ☐
B ☐
C ☐
D ☐

13. 95% confidence limits are required for a certain quantity, but a large pilot sample gives an interval which is about twice as large as the investigators wish. Further observations should have a sample size of:

A. half
B. a quarter
C. twice
D. four times

that of the pilot sample.

A ☐
B ☐
C ☐
D ☐

14. If an event has a probability of $\frac{1}{4}$, and a sample of 27 trials is taken, the mean and standard deviation of the number of trials resulting in the event are respectively:

A. 4, 2.25
B. 6.75, 2.25
C. 6.75, 0.25
D. 20.25, 4

A ☐
B ☐
C ☐
D ☐

15. To help decide whether a distribution could be normal, a six-cell χ^2 test is used. The number of degrees of freedom is:

A. 6
B. 5
C. 4
D. 3

A ☐
B ☐
C ☐
D ☐

16. In testing a sample mean against a hypothetical population mean μ_0, the observed t value is -4.7, and the critical t value at the chosen significance level is 2.2. The conclusion is that:

A. the population mean is greater than μ_0
B. the population mean is less than μ_0
C. The results are not significant
D. a larger sample should be taken

A ☐
B ☐
C ☐
D ☐

17. Which of the following formulae is correct? (s denotes sample standard deviation; σ denotes population standard deviation).

A. $t = \frac{x - \mu}{\sigma}$

B. $u = \frac{(x - \mu)\sqrt{n}}{s}$

C. $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$

D. $t = \frac{\bar{x} - \mu}{s/\sqrt{(n - 1)}}$

A ☐
B ☐
C ☐
D ☐

18. If failure to reject a false null hypothesis could have disastrous results, an appropriate level of significance would be:

A. 0.1%
B. 1%
C. 5%
D. 25%

A ☐
B ☐
C ☐
D ☐

19. The correlation coefficient between two characteristics of a sample of eight subjects is determined. The number of degrees of freedom is:

A. 6
B. 7
C. 14
D. 15

A ☐
B ☐
C ☐
D ☐

20. The correlation coefficient for a small set of data is calculated as 1.6. The conclusion is that:

A. a close positive correlation exists
B. there is unlikely to be any correlation
C. additional data is required
D. a mistake has been made in the calculations

A ☐
B ☐
C ☐
D ☐

Statistics

Test Y

NAME: COURSE: DATE:

For each item, there is one correct response; place a tick in the box appropriate to your choice of response.

Avoid blind guessing; there will be a small penalty for wrong answers but not for omissions.

Calculators and tables are not required for this test, but reference to notes and text-books is permitted.

1. 5, 11, 12, 12

The median of the above set of numbers is:

A. 8.5 B. 10 C. 11 D. 11.5

A	
B	
C	
D	

2. A sample frequency distribution of the integer r is given below:

r	0	1	2	3	4
f	1	3	8	3	1

One of the dispersion measures of the sample has a value of 1; this dispersion measure is the:

A. range B. mean deviation C. standard deviation D. interquartile range

A	
B	
C	
D	

3. 68.8, 69.3, 66.3, 69.0, 67.5, 71.1

In finding the standard deviation of the above sample of values of x, using a slide-rule or logarithms only, it would be good practice to:

A. sum the squares of the differences between each number and the sample mean

B. use the transformation $X = \frac{x}{0.1}$

C. use the transformation $X = \frac{x - 68}{0.1}$

D. use the transformation $X = \frac{x - 68}{10}$

A ☐
B ☐
C ☐
D ☐

4. The following expressions relate to data from a sample of n values. Which is the best estimator of the variance of the population?

A. $\frac{\Sigma x^2}{n} - (\bar{x})^2$

B. $\frac{(\Sigma x)^2}{n} - \frac{\Sigma x^2}{n^2}$

C. $\frac{\Sigma (x - \bar{x})^2}{n - 1}$

D. $\frac{(\Sigma x)^2}{n - 1}$

A ☐
B ☐
C ☐
D ☐

5. If a small sample only is available and it is desired to minimise the effect of any extreme values, the best measure of position to use is the:

A. median

B. arithmetic mean

C. interquartile range

D. variance

A ☐
B ☐
C ☐
D ☐

6.

x	0.5	0.6	0.7	0.8	0.9	1.0	1.1
f	1	2	3	4	3	2	1

The distribution represented by the above table is:

A. positively skewed

B. symmetrical

C. negatively skewed

D. binomial

A ☐
B ☐
C ☐
D ☐

7. A test which assumes a normal distribution of sample means is being used in a case where the variate is not normally distributed. This procedure is:

A. always valid
B. never valid
C. acceptable provided a large sample is used
D. acceptable so long as the sample is truly random

A ☐
B ☐
C ☐
D ☐

8. If n is a digit taken from a table of random numbers, then as the number of digits approaches infinity the probability density histogram of n tends to the form of a:

A. rectangle of height 0.1
B. rectangle of height 1
C. rectangle of height 10
D. normal distribution curve

A ☐
B ☐
C ☐
D ☐

9. If a fair coin is tossed three times, the probability of its showing heads once only is:

A. $\frac{1}{8}$
B. $\frac{1}{3}$
C. $\frac{1}{2}$
D. $\frac{3}{8}$

A ☐
B ☐
C ☐
D ☐

10. The probabilities of two independent events are respectively p and q. The probability of neither event occurring is:

A. $1 - p - q$
B. $1 - pq$
C. $(1 - p)(1 - q)$
D. Dependent on whether the events are mutually exclusive

A ☐
B ☐
C ☐
D ☐

11. On a graph of the cumulative probability curve of a variate x, $p(a < x < b)$ is given:

A. by the area under the curve between the ordinates x = a and x = b

B. by the difference between the ordinates at x = a and x = b

C. by the difference between the abscissae at x = a and x = b

D. approximately by the mean of the ordinates at x = a and x = b

A ☐
B ☐
C ☐
D ☐

12. The gradient of a cumulative frequency curve is at all points:

A. increasing

B. positive

C. less than unity

D. non-negative

A ☐
B ☐
C ☐
D ☐

13. A normally distributed variate has a standard deviation of 15. In a sample of 1000 values, 950 of these lie within an interval symmetrically disposed about the population mean and having a range of approximately:

A. 90

B. 60

C. 45

D. 30

A ☐
B ☐
C ☐
D ☐

14. An event has a constant probability of 0.01. The standard deviation of the number of occurrences in samples consisting of 400 trials is approximately:

A. 0

B. 1

C. 2

D. 4

A ☐
B ☐
C ☐
D ☐

15. The difference between the means of samples of size 18 and 20 respectively is to be tested by means of the t distribution. The number of degrees of freedom is:

A. 38 B. 37

C. 36 D. 19

A ☐
B ☐
C ☐
D ☐

16. In testing the difference between two sample means, the t value obtained by standardising the difference $\bar{x}_1 - \bar{x}_2$ is -2.2; the critical t value at the chosen level of significance is 4.7. The conclusion is that:

A. $\mu_1 > \mu_2$ B. $\mu_1 < \mu_2$

C. $\mu_1 = \mu_2$ D. μ_1 may equal μ_2

A ☐
B ☐
C ☐
D ☐

17. The difference between the means of samples of size 10 and 15 respectively is to be tested by means of the t distribution. The correct procedure involves finding:

A. the difference between observations taken in pairs

B. the ration of the estimated population variances

C. the mean of all 25 observations

D. the standard deviation of all 25 observations

A ☐
B ☐
C ☐
D ☐

18. If rejecting a <u>true</u> null hypothesis could have disastrous results, which level of significance would be most appropriate?

A. 0.1% B. 1%

C. 5% D. 25%

A ☐
B ☐
C ☐
D ☐

19. Which formula is correct for the sample correlation coefficient?

A. $$r = \frac{n\Sigma xy - \Sigma x \Sigma y}{\sqrt{\left\{\left[n\Sigma x^2 - (\Sigma x)^2\right]\left[n\Sigma y^2 - (\Sigma y)^2\right]\right\}}}$$

B. $$r = \frac{\frac{\Sigma xy}{n-1} - \bar{x}\bar{y}}{\sqrt{\left[\left(\frac{\Sigma x^2}{n-1} - \bar{x}^2\right)\left(\frac{\Sigma y^2}{n-1} - \bar{y}^2\right)\right]}}$$

C. $$r = \frac{\frac{1}{(n-1)}\sum(x - \bar{x})(y - \bar{y})}{s_x s_y}$$

D. $$r = \frac{\frac{1}{n}\sum(x - \bar{x})(y - \bar{y})}{\hat{\sigma}_x \hat{\sigma}_y}$$

A ☐
B ☐
C ☐
D ☐

20. A report states that "the correlation coefficient between the two variables was found from the sample data to be 0.92; this is statistically significant". The statement:

A. is correct

B. is incorrect because the coefficient is not significant

C. is correct if the sample is very large

D. cannot be judged without knowing both the null hypothesis and the sample size

A ☐
B ☐
C ☐
D ☐

C.2.4

The 25-item test referred to in subsection 6.4.4 as having been used for continuous assessment is given below. It is included mainly on account of items 12, 14, and 17, which show how the multiple-choice type of item can be made to test ability to a greater depth than is usual with this type.

CHRISTMAS TEST IN MATHEMATICS

A1 Mechanical/Production Engineering,
December 1977.

Time allowed:- One hour. Use of calculators, tables, books and notes is <u>not</u> permitted, but the standard reference sheet may be used.

Attempt all 25 questions. Avoid blind guessing; there is a small penalty for wrong answers, but none for omissions.

Name:- ..... Date:- .....

There is <u>one</u> correct response for each item; place a tick, ✓ , in the appropriate box.

1. The differential coefficient of $\sin'(x^2)$ with respect to x is:-

A. $2 \cos (x^2)$ B. $2x \cos (x^2)$

C. $2 x^2 \cos (x^2)$ D. none of A, B, and C

A	
B	
C	
D	

2. If $y = \sqrt{(2x)}$, then $\frac{dy}{dx} =$

A. $\frac{1}{2}\sqrt{(2/x)}$ B. $\sqrt{(2/x)}$

C. $\sqrt{(1/x)}$ D. $\frac{1}{2\sqrt{(2x)}}$

A	
B	
C	
D	

3. The derivative of $x^2 \cos(2x)$ with respect to x is:-

A. $-4x \sin(2x)$

B. $2x\left[\cos(2x) - x \sin(2x)\right]$

C. $x\left[2\cos(2x) - x \sin(2x)\right]$

D. $2x\left[\cos(2x) + x \sin(2x)\right]$

A ☐
B ☐
C ☐
D ☐

4. If $y = \dfrac{x}{1 + \sin x}$, then $\dfrac{dy}{dx} =$

A. $x \cos x + 1 + \sin x$

B. $\dfrac{1 + \sin x + x \cos x}{(1 + \sin x)^2}$

C. $\dfrac{1 + \sin x - x \cos x}{(1 + \sin x)^2}$

D. $\sec x$

A ☐
B ☐
C ☐
D ☐

5. If $f(\theta) = \cos(2\theta)$, which of the following represents $f'(\theta)$?

A. $\sin(2\theta)$

B. $-\sin(2\theta)$

C. $2 \sin(2\theta)$

D. $-2 \sin(2\theta)$

A ☐
B ☐
C ☐
D ☐

6. Which of the following represents the derivative of $\ln(\sin 2x)$ with respect to x?

A. $2 \sec 2x$

B. $\sec 2x$

C. $2 \cot 2x$

D. $\cot 2x$

A ☐
B ☐
C ☐
D ☐

7. Which of the following represents the derivative of $e^{\tan x}$ with respect to x?

A. $e^{\tan x}$

B. $e^{\sec^2 x}$

C. $\sec^2 x \, e^{\tan x}$

D. $\sec^2 x \, e^{\sec^2 x}$

A ☐
B ☐
C ☐
D ☐

8. If a is a constant, which of the following represents $\frac{d^2y}{dx^2}$ when $y = a\,e^{-3x}$?

A. $-9y$ B. $-3y$

C. $9y$ D. none of A, B, and C

A ☐
B ☐
C ☐
D ☐

9. If $y = 3 \cosh 2x$, which of the following represents the value of $\frac{dy}{dx}$ when $x = 4$?

A. $3 \sinh 8$ B. $3 \cosh 8$

C. $6 \cosh 8$ D. $6 \sinh 8$

A ☐
B ☐
C ☐
D ☐

10. The gradient of the graph of $\tanh x$ against x is:-

A. $\cosh^2 x$ B. $\frac{\cosh 2x}{\cosh^2 x}$

C. $1 + \tanh^2 x$ D. $1 - \tanh^2 x$

A ☐
B ☐
C ☐
D ☐

11. In the Maclaurin series expansion of the general function $f(x)$, the fourth term is:-

A. $\frac{f'''(0)}{3!}\,x^3$ B. $\frac{f'''(0)}{3}\,x^3$

C. $\frac{f''''(0)}{4!}\,x^4$ D. $\frac{f''''(0)}{4}\,x^4$

A ☐
B ☐
C ☐
D ☐

12. The following is an attempt to evaluate the sine of 0.1° by a Maclaurin series:-

(1) $\sin(0.1^{\circ}) = 0.1 - \frac{0.001}{3!} + \frac{0.00001}{5!} - \ldots\ldots$

(2) $\sin(0.1^{\circ}) = 0.09983$ correct to 5 decimal places.

Which of the following statements is true about stages (1) and (2)?

A. (1) and (2) are both correct

B. (1) can be corrected by multiplying each term in the series by $\frac{\pi}{180}$ but there is a further mistake in (2)

C. (1) is incorrect but there is no further mistake in (2)

D. (1) is correct but (2) is incorrect

A ☐
B ☐
C ☐
D ☐

13. $\int_0^1 e^{2x} dx =$

A. $e^2 - 1$

B. $\frac{1}{2} e^2$

C. $2(e^2 - 1)$

D. $\frac{1}{2}(e^2 - 1)$

A ☐
B ☐
C ☐
D ☐

14. Consider these statements:-

(1) $\int x^3(2 + x^4)^5 \, dx = \frac{1}{24}(2 + 4^4)^6 + C$

(2) $\int \frac{\ln x}{x} \, dx = \frac{1}{2}(\ln x)^2 + C$

Which combination of these is valid?

A. (1) but not (2)

B. Both

C. (2) but not (1)

D. Neither

A ☐
B ☐
C ☐
D ☐

15. $$\int \frac{dx}{\sqrt{(a^2 + x^2)}} =$$

A. $\frac{1}{a} \sinh^{-1} (\frac{x}{a}) + C$

B. $\sinh^{-1}(\frac{x}{a}) + C$

C. $\tan^{-1}(\frac{x}{a}) + C$

D. $\frac{1}{a} \tan^{-1}(\frac{x}{a}) + C$

A ☐
B ☐
C ☐
D ☐

16. If the substitution $x = \sin\theta$ is used to transform the integral $\int\sqrt{(1 - x^2)}\, dx$ into $\int f(\theta)\, d\theta$, then $f(\theta) =$

A. $\cos\theta$

B. $\cos^2\theta$

C. $\cos^3\theta$

D. $\sin\theta\cos\theta$

A ☐
B ☐
C ☐
D ☐

17. Consider the following stages in an attempt to obtain $$\int \frac{x^2}{\sqrt{(x^2 - 1)}}\, dx :-$$

(1) Let $x = \cosh\theta$
Then $dx = \sinh\theta\, d\theta$
and $\sqrt{(x^2 - 1)} = \sinh\theta$

(2) The integral becomes $\int \cosh^2\theta\, d\theta$

Which of the following statements is true?

A. There are mistakes in both (1) and (2)

B. (1) is correct but (2) is not

C. (1) is incorrect but there is no further mistake in (2)

D. Both stages are correct

A ☐
B ☐
C ☐
D ☐

NOTE. In the remaining items, $j = \sqrt{(-1)}$ and the letters a, b, c and d are real numbers.

18. If $z_1 = 2 + j3$ and $z_2 = 3 + j2$, then the product $z_1 z_2 =$

A. $12 + j13$ B. $5 + j5$

C. 13 D. $j13$

A ☐
B ☐
C ☐
D ☐

19. The product $(a + jb)(c + jd) =$

A. $(ac + bd) - j(ad + bc)$ B. $(ac - bd) - j(ad + bc)$

C. $(ac - bd) + j(ad + bc)$ D. $(ac + bd) + j(ad - bc)$

A ☐
B ☐
C ☐
D ☐

20. If $x = 3 + j2$, then $\frac{1}{z} =$

A. $3 - j2$ B. $\frac{3 - j2}{3 + j2}$

C. $2\sqrt{13} + j2\sqrt{13}$ D. $\frac{3}{13} - j\frac{2}{13}$

A ☐
B ☐
C ☐
D ☐

21. If $x = j5$ and $y = 2 + j2$, then $\arg(xy) =$

A. $\frac{\pi}{4}$ B. $\frac{3\pi}{4}$

C. $\frac{5\pi}{4}$ D. $\frac{7\pi}{4}$

A ☐
B ☐
C ☐
D ☐

22. If $z = -3 - j\sqrt{3}$, then:-

A. $|z| = 12$ and $\arg(z) = \frac{7\pi}{6}$

B. $|z| = 2\sqrt{3}$ and $\arg(z) = \frac{\pi}{6}$

C. $|z| = 2\sqrt{3}$ and $\arg(z) = -\frac{5\pi}{6}$

D. $|z| = -12$ and $\arg(z) = -\frac{5\pi}{6}$

A ☐
B ☐
C ☐
D ☐

23. If $z = \frac{1 + j}{j^5}$, then arg(z) =

A. $\frac{\pi}{4}$ B. $-\frac{\pi}{4}$

C. $\frac{3\pi}{4}$ D. $-\frac{3\pi}{4}$

A ☐
B ☐
C ☐
D ☐

24. The polar form of $\sqrt{2} - j\sqrt{2}$ is:-

A. $2e^{-j\pi/4}$ B. $2e^{j3\pi/4}$

C. $\sqrt{2}e^{-j\pi/4}$ D. $\sqrt{2}e^{j\pi/4}$

A ☐
B ☐
C ☐
D ☐

25. If z is represented by a point in the first quadrant of the Argand diagram, which quadrant contains the point representing $\frac{z}{j}$?

A. First B. Second

C. Third D. Fourth

A ☐
B ☐
C ☐
D ☐

C.3 Low discrimination

In the main dissertation, reference was made to low ID values in subsection 1.3.2.

Some items which have frequently shown indices of discrimination of 0.2 or less are given below. On such items, students with low scores on the test as a whole perform almost as well as, or even better than, those with high scores. In making the selection, those with very high or very low facility values have been excluded, since such items will inevitably have near-zero discrimination.

This selection of examples showing poor discrimination was made from about 120 items which have had sufficient use to yield reliable information; only four qualify for inclusion in this appendix.

In each table the correct response is marked with an asterisk. Figures in brackets denote the responses made by candidates in the upper third of the test scores. Since all classes taking the test were in HNC or similar courses, it is considered unnecessary to give details of each course.

C.3.1

Item:-

$$\int e^{(x^2)} dx =$$

A. $e^{(x^2)} + c$

B. $\frac{1}{2} e^{(x^2)} + c$

C. $\frac{e^{(x^2)}}{2x} + c$

D. none of A, B, and C

Analysis:-

Class	Number of students	A	B	C	D*	FV(%)	ID
1	10	0(0)	0(0)	3(2)	4(1)	40	- 0.67
2	20	3(2)	0(0)	3(2)	4(0)	20	- 0.14
3	24	7(1)	1(1)	9(3)	6(3)	25	0.13

(ID with other classes: 0.4, 0)

Few of the more able students chose A. One-third of all the candidates chose C, and nearly half of these were in the upper third; treatment of the 2x in the denominator of the suggested integral as though it were a constant is seen to be a common error. About 70% of those giving the correct response were not in the upper third. It is possible that the weaker students, aware of their limitations, were more prepared than the others to make the cautious approach of differentiating the suggested integrals.

C.3.2

Item:-

Which of these equations represents a straight line?

A. $x + 2y + 3 = 0$ B. $\frac{1}{x} + \frac{1}{y} = \frac{1}{2}$

C. $xy = 4$ D. $y = \frac{1}{x + 2}$

Analysis:-

Class	Number of students	A*	B	C	D	FV(%)	ID
1	14	10(4)	1(0)	1(0)	1(1)	71	0.20
2	18	9(4)	4(0)	4(2)	1(0)	50	0.17
3	24	14(5)	7(2)	0(0)	2(1)	58	0.13

(ID with other classes: 0, 0.29, 0.5)

Here few of the more able students chose any of the three distractors. Unlike the previous item, this one has achieved a poor discrimination because of the relatively large number of weaker students who recognised the linear nature of the equation in A - 20 out of the 37 in the lower two-thirds.

C.3.3

Item:-

The following is an attempt to evaluate the sine of 0.1^{o} by means of a Maclaurin series:-

(1) $\sin(0.1^{o}) = 0.1 - \frac{0.001}{3!} + \frac{0.00001}{5!} - \ldots\ldots$

(2) $\sin(0.1^{o}) = 0.09983$ correct to 5 decimal places.

Which of these statements about stages (1) and (2) is true?

A. (1) and (2) are both correct

B. (1) can be corrected by multiplying each term in the series by $\frac{\pi}{180}$ but there is a further mistake in (2)

C. (1) is incorrect but there is no further mistake in (2)

D. (1) is correct but (2) is incorrect

Analysis:-

Class	Number of students	A	B	C*	D	FV(%)	ID
1	20	1(0)	3(2)	5(1)	3(0)	25	0
2	24	6(2)	1(0)	8(2)	1(0)	33	- 0.13
3	67	12(5)	5(1)	15(3)	12(2)	22	- 0.09

(ID with other class: 0)

Although C was overall the most popular response, it was marginally less so than A with the stronger students. Students who know of the importance of using radians in calculus seem less aware

of the status of the radian as the fundamental unit of angle.

C.3.4

Item:-

A normally distributed variate has a standard deviation of 15. In a sample of 1000 values, 950 of these lie within an interval symmetrically disposed about the population mean and having a range of approximately:

A. 90 B. 60

C. 45 D. 30

Analysis:

Class	Number of students	A	B*	C	D	FV(%)	ID
1	29	3(1)	10(4)	5(0)	9(4)	35	0
2	10	4(1)	1(0)	3(1)	1(0)	10	0
3	13	3(2)	5(1)	3(0)	8(1)	38	0
4	16	0(0)	5(2)	2(0)	7(3)	31	0.2

(ID with other class: 0.43)

D proved to be the most popular both overall and with the stronger students; the range of the interval bounded by $\mu \pm 1.96\ \sigma$ has been taken as about 2 σ rather than 4 σ .

C.3.5

It is again mentioned that although these items are not suitable for use in selection and attainment tests because of their poor (even negative) discrimination, their ability to reveal common misconceptions qualifies them for inclusion in induction and progress tests.

APPENDIX D

The questionnaire sent to other colleges associated with the Manchester Objective Testing Item Bank and discussed in section 5.2 is given below.

QUESTIONNAIRE ON OBJECTIVE TESTS IN MATHEMATICAL SUBJECTS

Please complete by placing a tick in the appropriate box.

I use objective tests for the purposes stated with the following frequencies:-

	Often	Sometimes	Never
A Selection of students before enrolment			
B Induction (to assess the ability of a class)			
C Progress testing (during session)			
D Attainment testing (at end of session)			

In writing, selecting, and using objective items for the above purposes A to D, my attitudes are shown in the following tables:-

1. Each item should have a facility value between about 40% and 60%.

Attitude / Purpose	Agree	No strong feelings	Disagree
A Selection			
B Induction			
C Progress			
D Attainment			

2. Items should have facility values which are roughly normally distributed about a mean of approximately 50% and between limits of approximately 10% and 90%.

Attitude / Purpose	Agree	No strong feelings	Disagree
A Selection			
B Induction			
C Progress			
D Attainment			

3. The responses of candidates who choose distractors (i.e. wrong options) should be roughly equally distributed between these distractors.

Attitude / Purpose	Agree	No strong feelings	Disagree
A Selection			
B Induction			
C Progress			
D Attainment			

4. Students should be allowed to retain their corrected scripts, so as to supplement their lecture notes.

Attitude / Purpose	Agree	No strong feelings	Disagree
A Selection			
B Induction			
C Progress			
D Attainment			

5. To maintain the security of tests, students should not be allowed to retain their corrected scripts.

Attitude / Purpose	Agree	No strong feelings	Disagree
A Selection			
B Induction			
C Progress			
D Attainment			

6. Students are more interested in the results of objective tests than in those of other tests.

Attitude / Purpose	Agree	No strong feelings	Disagree
A Selection			
B Induction			
C Progress			
D Attainment			

7. Objective testing is more realistic, and the results more reliable, if candidates are allowed to refer to text-books and notes during the test.

Attitude / Purpose	Agree	No strong feelings	Disagree
A Selection			
B Induction			
C Progress			
D Attainment			

8. Objective testing of the multiple-choice kind is fairer, and the results more reliable, if a small negative mark (say -1/3) is awarded for wrong responses, as a correction for guessing.

Attitude / Purpose	Agree	No strong feelings	Disagree
A Selection			
B Induction			
C Progress			
D Attainment			

I would like / do not wish to receive a summary of the results of this survey.

(Please delete as appropriate)

Many thanks for completing this questionnaire; now please return it to

Neville Upton
Department of Computer Studies and Mathematics
Birmingham Polytechnic
Franchise Street
Birmingham B42 2SU

Name and college:-
(Optional)