SOME PERSPECTIVES ON

HIGH SCHOOL

MATHEMATICS EVALUATION

David A. Norton

Papers submitted to the Faculty of Education, University of Cape Town, in partial fulfilment of the requirements for the degree of Master of Education

1982 - 1983

Contents

Paper I : Bloom's Taxonomy of Educational Objectives and their Application to Evaluation in Mathematics : A Critical Approach

Paper II : Confidence Scoring Methods : A Critical Review of the Methods with an Extension of the Application of Confidence-Weighting to Traditional Mathematics Testing

Paper III : The Screening and Selection of Pupils for a Fast Mathematics Class in Standard Seven by Using Tests

Paper IV : An Accelerated Program for a Fast-Paced Mathematics Group in Standard Seven

BLOOM'S TAXONOMY OF EDUCATIONAL OBJECTIVES AND THEIR APPLICATION TO EVALUATION IN MATHEMATICS : A CRITICAL APPROACH

David A. Norton

A paper submitted to the Faculty of Education, University of Cape Town, in partial fulfilment of the requirements for the degree of Master of Education

1982

Abstract

This paper seeks to introduce the taxonomy developed by Bloom and also other taxonomies applied to Mathematics. The behavioural basis is considered before the actual taxonomies are presented giving due cognizance to the influence of behavioural objectives. The paper then demonstrates the way in which a taxonomy can be used in setting examination papers, and also gives a survey of the way Bloom's taxonomy and behavioural objectives have been used in instruction. Philosophical criticisms of objectives and the use of Bloom are considered followed by a survey of empirical findings in connection with both of these. Finally suggestions are firstly made as to how the taxonomy can be used to examine mathematics question papers in South Africa in its albeit subjective manner, and secondly for further research.

Contents

Introduction		p. 1
Chapter 1	The Behavioural Basis of Educational Objectives	p. 2
2	The Taxonomies of Educational Objectives	p. 6
3	Application of Educational Objectives and Bloom's Taxonomy to Instruction	p. 28
4	Criticism of Behavioural Objectives and the Taxonomic Approach	p. 32
5	Empirical Findings on Behavioural Objectives and Bloom's Taxonomy	p. 46
6	Applications of Educational Objectives and Taxonomies to Evaluation in Mathematics in South Africa	p. 54
Conclusion		p. 65
References		p. 66

Figures and Tables

Figures

1. Classification to show Ormell's levels of visibility approach p. 42

2. Madaus et al's Y-shaped structural adaptation of Bloom's Taxonomy p. 44

Tables

1. Comparison of certain mathematical taxonomies p. 14

2. Blueprint applying Avital and Shettleworth's Taxonomy to the Higher Grade Algebra syllabus for Senior Certificate p. 20

3. Scopes suggested proportioning of marks using Wood's taxonomy p. 26

4. Average memory scores of students oriented to different process objectives (from Kunen et al 1981 p. 207) p. 52

5. Completed blueprint for Cape Senior Certificate Higher Grade Mathematics Papers for November 1979, 1980 and 1981 in Algebra p. 56

6. An analysis of Geometry questions in 1978 Senior Certificate Papers (Joubert 1980 p. 54) p. 60

7. Average percentage of marks allocated at various taxonomic levels for Standard 5 to Standard 7 (Smith 1980 p. 86) p. 61

Introduction

The philosophical basis behind the taxonomy of educational objectives as outlined by Bloom in 1956 has been regarded as "the top of a pyramid based on philosophy and empirical research in education : it calls for a functional synthesis of all our knowledge in the field of education and related disciplines" (de Landsheere 1977 p.77), and conversely as suiting "the expression of a pragmatic/materialist ethic, and does not particularly suit the expression of ethics either traditionally, or currently associated with the concept of education" (Ormell 1974 p.3). Thus widely opposing philosophical viewpoints have been propounded and, in addition, rather confusing results of empirical studies have added to the controversy leading for many years to the entrenchment of the viewpoints held by the protagonists. It would, however, seem as stated by Kapfer (1978 preface) that "there now appears to be more concern for exploring ways of redirecting the behavioural objectives impetus of recent years", in order to retain the advantages that have resulted from behavioural studies and the use of taxonomic methods in considering both evaluation and instruction. There is now a greater realisation in education of the truth of Emerson's dictum "that the ends pre-exist in the means", and that objectives, teachers and resources are intimately intermingled and interrelated with learners in the educational process (Davies 1976 p.11).

Chapter 1

The Behavioural Basis of Educational Objectives

In considering Bloom's Taxonomy the first aspect to be discussed must be that of behavioural objectives, which act as a basis for the taxonomy.

In the times of the Romans the term "objective" referred to a pillar which marked some turning-point in a chariot race. As a result, in our usage, an objective does not have to be seen as the final aim marking the end of an activity, but rather as a signpost along the way (Davies 1976 pp. 2 and 3). Objectives are thus developmental representing roads to travel, and thus the achievement of the objectives needs to be planned for continuity with a full appreciation of the developmental stages within that continuity (Taba 1962 p.203). It is therefore necessary to contrast aims at a general level, and objectives which are more concerned with specific outcomes. Aims and objectives must be considered as complementary to each other.

General aims are necessary in that they express the philosophical basis on which a course or a syllabus is based. For example some aims expressed in the Junior Secondary Course Syllabus for Mathematics for the Cape Education Department (1973 p.2) are:

1.1 To give the pupil a clear insight into, and a thorough knowledge and understanding of, those basic mathematical principles which will prepare and equip him for further study in Mathematics and other subjects;

1.4 to arouse in the pupil a love for, and an inter-

est in the further study of Mathematics and related subjects.

General aims such as these although absolutely necessary are too vague to be translated into an active education program. There needs to be a more specific platform of objectives in order to guide the making of curriculum decisions, to decide what content must be covered, how it must be emphasized and which learning experiences must be stressed. (Taba 1962 p. 197). Without a consistent set of objectives there is a very real danger that all evaluation with its concomitant effect on instruction will test the lowest level of learning, and that is knowledge. Not only that, there would be a tendency to test only that type of knowledge which involves recall alone. What then constitutes a good objective?

Objectives have been developed by the behaviourists, and from a behaviourist's point of view (Mager 1962) an objective must contain three main factors:

1. the new learning (ability) that the pupil will be able to demonstrate after the instructional sequence;
2. the standard to which this new ability will be performed, and
3. the conditions or constraints that will pertain at the time of the performance.

Thus in the objectives as stated above great emphasis is placed on the intended outcome, and that there should be a measurable change between what the pupil knew before the particular aspect of work was taught, and what he knows afterwards. This measurable change can be established by considering the pupil's behaviour after he has learnt something, and comparing it with what he

knew before the teaching/learning situation took place. In other words if there is no new or changed behaviour then no learning has taken place. (Wiseman and Pidgeon 1970 p.38). This leads into a definition typifying objectives:

"They are statements of purpose which describe desired student behaviour, and indicate the content through which the behaviour is to be developed." (Ammons 1969 p.911)

When writing behavioural objectives it is thus important to start with a verb which describes the kind of behaviour pupils should be able to show after having completed a section of study, because this places the emphasis directly on the action to be undertaken by the pupils, and also guides the teacher when he comes to prepare the necessary learning experiences required to achieve the objectives (Wiseman and Pidgeon 1970 p.42). For example, the use of the words "to know" is not sufficiently precise in the following stated objective: "to know the properties of the four-sided rectilineal figures."
This should rather read:
"to list the properties of the four-sided rectilineal figures" or "to recognize how to use the properties of the four-sided rectilineal figures in more complex problems."
The reason for this is that the words "to know" being imprecise lead to no real understanding of how measurement of the knowledge can take place, whereas with words like "to list", it is immediately clear what is expected of the pupil as far as both content and level of mastery of that content is concerned. It should be noted that not all responses must be of a written form to satisfy those who are concerned with an approach by

objectives. There are many ways of observing the behavioural change as a result of objectives which for example use such categories as recall, recognize and remember. (Bloom 1971 p. 144)

There are various levels of learning to which objectives can point and the examples given above concern themselves with two such levels. It is this recognition of the various levels of learning and understanding that leads to the idea of a taxonomy. As higher abilities of say analysis are called into play so the writing of objectives becomes more difficult and more complex, with more thought given to the means of testing to see if the objectives have been attained. It is now necessary to consider the taxonomies with special reference to their application in the sphere of mathematics education.

The Domains

Bloom divides the description of human behaviour into three domains: the cognitive, the affective and the psychomotor.

The cognitive domain includes those objectives which deal with the recall and recognition of knowledge and the development of intellectual abilities and skills. The affective domain includes objectives which describe changes in interest, attitude and values, and the development of appreciation and adequate adjustment. The psychomotor domain is the manipulative or motor-skill area.

In considering evaluation the most important of the three is the cognitive. The affective domain does not generally apply to the setting of tests, although certain questions can increase pupils' interest and thus improve response and the subsequent development of cognitive processes. That does not mean to say that the affective domain has no place in mathematics teaching - developing interest is extremely important in any course of instruction. Properties such as awareness and willingness to respond and valuing as developed by Krathwohl et al (1964) should be studied by all teachers in depth so as to make their teaching more effective. These fulfilled affective objectives will have an effect on cognitive objectives as, for example, one's interest in analysis is controlled by the affective part of one's nature - is one interested enough to work for a sufficiently long period in order to fulfil the cognitive objective of analysis? Except for the learning of a few skills in geometry constructions, the psychomotor

domain is of limited importance at the secondary level in mathematics, although of utmost importance in the first years of schooling. Bloom did not develop a taxonomy for this domain, but others, notably Harrow, have. De Landsheere has summarised three such taxonomies (1977).

On the three domains De Landsheere states an important principle very clearly: "It is obvious that the distinction between domains is artificial: man reacts as a whole. This distinction is arbitrary and Bloom and associates stress this without any ambiguity." (1977 p.98). No domain can exist without the others, but it is true that in a particular sphere one of the domains is dominant. In the sphere of evaluation, the cognitive domain is the most important, and has received most attention in the literature.

Bloom's Taxonomy of Educational Objectives : Cognitive Domain

In the first instance Bloom divides the Cognitive Domain into two main outcomes of education viz. Knowledge, and Intellectual Abilities and Skills. These outcomes cover a "multitude of sins" and so Bloom necessarily defines what is meant by each: "Knowledge, as defined here, involves the recall of specifics and universals, the recall of methods and processes, or the recall of a pattern, structure or setting. For measurement purposes the recall situation involves little more than bringing to mind the appropriate material. Although some alteration of the material may be required, this is a relatively minor part of the task. The knowledge objectives emphasize most the psychological process of remembering." (Bloom 1956 p.201)

"Abilities and Skills refer to organised modes of operation and generalised techniques for dealing with materials and problems. The abilities and skills objectives emphasize the mental processes of organizing and reorganizing material to achieve a particular purpose. The materials may be given or remembered." (Bloom 1956 p.204).

Bloom holds that Abilities and Skills require higher order mental processes than does Knowledge. He further divides the abilities and skills into five main classes which he holds form a hierarchy with Knowledge. The main divisions of the hierarchy are as follows:

1.00 Knowledge
2.00 Comprehension
3.00 Application
4.00 Analysis
5.00 Synthesis
6.00 Evaluation

Two comments by Bloom on the above arrangement are pertinent:

"a. Although it is possible to conceive of these major classes in several different arrangements, the present one appears to us to represent something of the hierarchical order of the different classes of objectives.

b. Our attempt to arrange educational behaviours from simple to complex was based on the idea that a particular simple behaviour may become integrated with other equally simple behaviours to form a more complex behaviour. Thus our classification situations may be said to be in the form where behaviours of type A form one class, behaviours of type AB form another class and behaviours of type ABC form still another class." (Bloom 1956 p.18).

Bloom et al had to define each category in order to separate them. Their definitions of the main abilities and skills divisions are given here, but not the subdivisions as these have not proved to be useful. A more detailed breakdown as originally conceived appears in Bloom 1956 pp. 201-207.

"Comprehension represents the lowest form of understanding. It refers to a type of understanding or apprehension such that the individual knows what is being communicated and can make use of the material or idea being communicated without necessarily relating to other material or seeing its fullest implications.
Application is the use of abstractions in particular and concrete situations.
Analysis is the breakdown of a communication into its constituent elements or parts such that the relative hierarchy of ideas is made clear and/or the relations between the ideas expressed are made explicit.
Synthesis is the putting together of elements and parts so as to form a whole.
Evaluation is the use of a standard of appraisal."
(Bloom 1956 pp. 204-207).

At all levels in the taxonomy Bloom gave examples of both intended educational objectives, and the sort of questions that could be asked in the various stages of the hierarchy.

It must be noted that Bloom does not establish whether what is being presented is a hierarchy, but assumes it to be so. It is also assumed that because there is a hierarchy the more complex behaviours include the simple behaviours, and

measurement is based on behaviour. What is being presented is a theory. Bloom and associates made it clear that they believed that there were certain problems with their hierarchy. One of these problems is that the intended behaviours represent social goals imposed upon pupils by their society or culture, and a second that the distinction between the categories is neither sharp nor clear and, therefore, the categories are not mutually exclusive. (De Landsheere 1977 p.104). A third acknowledged problem allied to the second is that attempts to categorise the level of attainment being tested by any given question are attended by the additional difficulty that so much depends on the educational background of the learner.

Bloom and associates were true pioneers in their field stating that it is hoped that what they produced could be useful and suggesting where the taxonomy might be useful. No outrageous claims were made for it. One only wishes that their followers might have been as circumspect, and one can only echo Freudenthal's words when he at last managed to read the original:

"I felt thunderstruck! Rather than the charlatanism I expected on the strength of quotations and applications, so-called, I found a serious decent booklet, though of a quite different kind from what I had expected - in literature one always has to track down the sources." (1978 p.81).

Adaptation's to Bloom's Taxonomy for its use in Mathematics

Bloom et al hoped that the taxonomy would be found to be of value as a means of communication within

the sphere of education, and submitted it in the hope that it would help to stimulate thought and research on educational problems (1956 p.9). This it did not do for a number of years. It had to await the popularisation of behavioural objectives in education in the early sixties, and it was only when the Taxonomy on Affective Objectives appeared in 1964 that real interest was generated in the taxonomies and developments therefrom.

When behavioural objectives were considered, because of the thought actually put into stating what was intended to be achieved by a particular objective, it was realised that most examination questions tested knowledge only, and that the higher abilities were being ignored. With objectives, it was claimed, questions can be written to suit a specification, and an examination can be planned to test the abilities that are felt to be important. (UNESCO 1973 p.120). In applying Bloom's taxonomy to mathematics it was soon realised that alterations had to be made to the various hierarchical levels as Bloom's did not cover all aspects regarded as essential in Mathematics education. In addition the taxonomy was far too complicated for practical purposes with all the subdivisions suggested by Bloom. In 1968, however, at an international meeting of mathematics experts convened in Strasbourg in connection with the Oxford Council of Europe Evaluation Study, Bloom's taxonomy of educational processes was used as a basis to derive a particular taxonomy for mathematics at the academic secondary level. Using Bloom's work as a guide a taxonomy with seventeen subdivisions was developed. (Halls and Humphrey 1968 p.21). Thankfully it would appear as if this taxonomy has since been ignored.

In table 1 on p.14 some mathematical taxonomies are compared. The effect of Bloom's taxonomy on each of them is evident. The following should be noted in connection with the taxonomies shown in the table

1. IEA. This was the model reported in the International Study of Achievement in Mathematics (Husen 1967) which was conducted in twelve participating countries in order to compare examination results. These categories were arrived at after considerable investigation in these countries, and after many difficulties in trying to identify behavioural objectives that would have universal endorsement (Wood 1968b p.89).
2. Avital and Shettleworth (1968). This is a hierarchy suggested by these two authors who had noted that very few items included in Bloom's taxonomy are from mathematics, and thus felt that an adaption to mathematical performance would increase the potential usefulness of a taxonomy to the mathematics teacher in the formulation and evaluation of objectives (pp.4 and 5). To them Comprehension and Application used algorithmic thinking and generalisation while Analysis and Synthesis implied "open search".
3. Wood (1968a). This hierarchy was developed so that a basis whereon Certificate of Secondary Education (CSE) Examinations in England could be compared. Because many of these were internal it was suggested that an Item Bank could be drawn up to help the formulation of a school-based examination which could have national currency. (1968c).
4. NLSMA (1966). This National Longitudinal Study of Mathematical Abilities was drawn up by the Schools Mathematics Studies Group (SMSG) in the

TABLE 1

IEA	AVITAL AND SHETTLEWORTH	WOOD	NLSMA	WILSON
A. Knowledge and Information	A. Knowledge	A. Knowledge and Information	Knowledge	A. Computation
B. Techniques and Skills	B. Comprehension and Application	B. Techniques	Translating	B. Comprehension
C. Translation of Data	C. Analysis and Synthesis	C. Comprehension	Manipulating	C. Application
D. Comprehension		D. Application	Choosing	D. Analysis
E. Inventiveness		E. Inventiveness	Analysing	
			Synthesising	
			Evaluating	

Comparison of certain Mathematical Taxonomies

United States. It is not clear whether this scheme is meant to have taxonomic structure although it has affinities with Bloom's taxonomy (Wood 1968b p.88).

5. Wilson (1971). In his part of the book Formative and Summative Evaluation, J.W. Wilson also formulated a hierarchy for mathematical behavioural response including both the cognitive and affective spheres, although only the former is shown here.

Any comparison between the taxonomies must not be superficial because the meanings that the authors attach to their respective chosen categories must be considered carefully. In most of the taxonomies knowledge and abilities are opposed to each other in that abilities are regarded as belonging to higher order thought processes than knowledge. But, this depends on one's viewpoint as to what constitutes knowledge which to some people is far more than just recognition and recall, and indeed it is argued that to anyone who is seeking to develop abilities, knowledge is an essential tool (MacIntosh 1976 p.12). This confusion is illustrated by the inclusion of both knowledge and ability items under the Comprehension mantle in Wilson's taxonomy although he does place knowledge in the lower part of this level (Wilson 1971 p.646). Avital and Shettleworth (1967 p.8) state that if the objective is given in the form "the student should know....", then knowledge in this sense can refer to at least two abilities which are distinct: the ability to repeat and the ability to use. According to them the first is the only sense in which the category of Knowledge can be used, while the second refers to the higher levels of Comprehension and Application or even to

problem-solving. Despite this explanation there is still a grey area. Whether knowledge should be defined on extremely narrow lines is, as indicated, a matter for debate, but it is important to note that when knowledge is defined on such a narrow basis once any problem, even a complex one using much analysis, has been learnt, all that is required to answer it is recall, no matter how difficult that problem is. It can be argued that it depends on how learning has taken place. One pupil will learn the similar triangle theorem off by heart, and another will learn it by noting the type of proof that must be carried out at the various stages of the theorem. Clearly the first method of learning places the answer to the theorem in the category Knowledge. The second means depends on how one construes the candidate's knowledge of the various stages - these being simple proofs; here one could argue that higher abilities are involved.

It is also clear that what is meant by Comprehension also shows much divergence. Avital and Shettleworth include algorithmic thinking under Comprehension while Wilson includes it under his first heading of Computation where one of the sub-headings reads the ability to carry out algorithms. In the IEA study Comprehension is at level 4, and even though there is no detailed statement as to what is meant by Comprehension it is significant that it is certainly at a higher level than Bloom places it, in that the capacity to analyse problems and to follow reasoning are considered to be part of it. (Husen 1967 p.81). Generally, however, Comprehension objectives are considered to be closely related to Knowledge objectives (Avital and Shettleworth 1967 p.15).

It must be realised that levels in the taxonomy should not be assumed to be reflected by pupils' results in tests. It is not necessarily true that pupils will obtain higher marks on knowledge items than on items involved with comprehension. To assume automatically that students' marks will decline as the cognitive levels become more complex is wrong. It would seem generally true that questions involving higher order abilities will lead to lower results, but this must not be taken as a law from which there is no deviation. Kropp and his associates were the first to make this assumption (Keenen et al 1981 p.203). Bloom held that some questions belonging to the Knowledge category would be of such a difficult nature that low marks would be recorded. It is significant that in designing experiments where the difference in results is assumed between levels, many empirical researchers choose such disparate taxonomic levels as Knowledge and Analysis, for example Barker and Hapkiewicz (1979).

The two European-based taxonomies viz IEA (9 out of the 12 countries taking part were European - the others being Australia, Japan and the United States) and Wood have as a final category - Inventiveness. This does not appear in the American taxonomies nor in Bloom, and indeed Wilson's highest level is Analysis. This seems to reflect a difference in attitude between the pure behaviourist line that was in vogue in America at the time and the more open attitude in Europe. Avital and Shettleworth (Canadians) qualify their Analysis and Synthesis level by stating that the thinking process required at this level is that of open search. To them the difference between generalisation and open search is essentially that between reproductive and

productive thinking. If the student must produce something that is entirely new to him, he will be engaged in higher-level problem-solving i.e. analysis or synthesis (1967 p.19). It is essential that this aspect of inventiveness or open search be included in all mathematical taxonomies, and its absence in Bloom's taxonomy is commented on thus by Freudenthal: "In the catalogue of objectives one looks in vain for such expressions as observation; higher level expressions such as experimenting and designing experiments are also lacking. The authors did not notice at all what an enormous part is played by intelligent observation and intelligent experiment in cognitive development, and how strong the component of educating intelligent observation and intelligent experiment is in school and university instruction of the natural sciences." (1978 p.82). In terms of the traditional examination Freudenthal's argument does not apply, but behaviourists all acknowledge that there are many ways to assess change in behaviour.

Thus, even though broadly speaking the various adapted taxonomies are similar, the one important consideration of inventiveness extends the non-American formulations to make a taxonomy more acceptable to mathematics educationalists.

Despite the criticisms, Bloom's work has deservedly received praise because what he did achieve was to make teachers more aware of educational outcomes other than knowledge, outcomes that should be tested in examinations, and thus highlighting that examinations should not be merely stereotyped basic knowledge tests in how to factorise or how to answer specific problem types. This should definitely

apply to Mathematics Higher Grade papers. Pupils should be challenged to think not by making the questions more complicated algorithmically, but by increasing the level of analysis required so that the student has to bring wider knowledge to bear on the task at hand and then synthesize it into a whole.

Use of a Taxonomy in Mathematics Examinations

To use a taxonomy effectively, a blueprint should be drawn up consisting of a content by process matrix as illustrated in Table 2 using Avital and Shettleworth's Taxonomy as a basis and applying it to an examination on the Higher Grade testing the Algebra syllabus in Standard Ten at South African schools.

In setting the paper the examiner has to decide on the number of marks to give to questions such as:

1. Solve for x : $3^{x+1} = 27$
2. Solve for x : $2^{2x} + 6.2^{x} = 7$
3. Find the maximum value of y in:
$$(x^2 - 9)\,(3^{y+2} - 18) \geq 0 \quad \text{if} \quad |x| < 3$$

The examiner must also decide into what taxonomic level the question should be placed. The subjectivety involved in the latter decision should not be minimised. Guidelines are given with each taxonomy, but even after considering these carefully there are still grey areas which lead to differences in interpretation. The examples given above should now be considered.

Eg 1. When met for the first time this question definitely involves algorithmic thinking, but thereafter knowledge of the basics of exponential equations is the important feature. The student makes use of the

Table 2

Process Objectives / Content Objectives	A	B		C
	Knowledge	Comprehension	Application	Analysis
	Recall, Recognition	Algorithmic thinking and generalisation		Synthesis / Open Search
Relations and Functions				
Quadratic Equations and Inequalities				
Indices Logarithms and Surds				
Sequences and Series				
Mathematical Induction				

Blueprint applying Avital and Shettleworth's Taxonomy to the Higher Grade Algebra Syllabus for Senior Certificate

following recall even though he may not have worked with the exact numbers given here:

(i) To solve an exponential equation the bases must be the same. Therefore convert 27 to 3^3

(ii) Equate indices $x + 1 = 3$

$x = 2$

The method would definitely have been memorised as far as most students who are writing the end of year examination are concerned. As Avital and Shettleworth state when considering the examples chosen to illustrate the assessment of Comprehension objectives : "for any of the following items to test performance in the category of Comprehension rather than in that of Knowledge, the student must not have been exposed to them before." (p.11). In considering whether questions belong to the Knowledge or Comprehension category one must clarify what "being exposed to them before" means.

If it means a passing reference only, then one could hardly regard this as knowledge, but if in contrast it means having studied the section of work thoroughly then, at the Standard Ten level, both the algorithmic needs and the simple method used in this problem must result in its being placed in the category Knowledge.

Eg 2. $2^{2x} + 6.2^x = 7$

$2^{2x} + 6.2^x - 7 = 0$

$(2^x + 7)(2^x - 1) = 0$

$2^x = -7$ or $2^x = 1$

(inadmissable) $x = 0$

> algorithmic rule for exponential equations follows with a final complication created by the introduction of logarithmic theory as well. According to Avital and Shettleworth the essential characteristic of open search is the non-routine manipulation of previously learned material and at a higher level, the discovery of relationships among previously unrelated concepts and propositions (p.19). Thus this example can be placed in the category Analysis and Synthesis.

It is obvious that the placing of questions into hierarchical levels is subjective and the question arises whether the use of objective-type questions would not be an advantage. Firstly, use of a blueprint is aided if the questions are objective in form eg of the multiple-choice type, because the process response with this kind of question can be isolated more easily than with the more traditional questions set in South Africa and Europe, which demand the setting-out of a solution or a mathematical proof. The reason for this is that the questions are more specific and usually test one aspect of a solution or proof at a time. The setting of objective questions (objective from the marking point of view) in mathematics is fraught with difficulties, however, one of which is that many questions can be answered by substituting the solutions given into the original question, and thus deciding, which is the correct answer by avoiding the intent of the question and in this way obtaining answers to many problems. Much of mathematical benefit is also lost when pupils are not required to set out answers in a logical way. It is also difficult to set objective-type questions

which require analysis and synthesis. Because much reasoning is involved, questions at this level should be allocated more marks, but if this is done then the weaker candidate will be doubly penalised as he will not be able to obtain recognition for those aspects of the problem he does understand. In addition unless all public examination papers are handed in with the scripts of the candidates, an Item Bank on which the examiners concerned can draw will soon be exhausted with the more than eighty public question papers that are now set in Mathematics each year in South Africa. If public examination papers are taken in with the students' scripts the principle of open access to these papers will be violated. It is questionable whether the minimal advantage gained using a blueprint on objective-type questions rather than traditional outweighs the disadvantages outlined above because placement within the blueprint is still subjectively based. Also subjective marking in Mathematics is not as serious a problem as in many of the other subjects.

In any examination one of the factors affecting subjectivity in the placing of questions into the different categories is the educational history of the student. In a system which is largely textbook based one can surmise the type of problems most students would have been taught. However, there will be situations where certain teachers have enriched the curriculum in certain fields of study thus helping the student, and other situations, by contrast, where a number of students, especially in the country areas suffer, because either the teacher is unqualified or there is no mathematics teacher at all. Hence a hard and fast classification for every example is impossible and the examiner in drawing up a public examination paper

will have to base his decisions on "Mr Average Pupil" - not that he exists. The teacher at a school is in a better position to assess into which taxonomic level each question falls as he will know what has been taught and how it has been taught, as far as his classes are concerned. In a large school this knowledge will not be as sure, but with the necessary staff consultation the examiner will have information at his disposal that will enable him to identify the taxonomic level with fair accuracy.

Willmot and Hall (1973 p.117) asked a number of teachers to classify the questions of a certain examination into taxonomic levels. They found much lack of agreement although there was considerably greater concensus of opinion on questions which required a high knowledge factor. Fairbrother (1975 p.208) in this connection notes that he questions the value of publishing detailed examination specifications if teachers disagree about the category of individual items, but at the same time he states that without some monitoring procedure, there would be the danger of just doing that which was easiest and assessing mainly knowledge. The examiner is helped by the monitoring procedure.

Having decided on the weighting and also on the taxonomic level, the examiner should enter the number of the question and the marks allocated to that question in the relevant part of the blue-print. The examiner will then be able to ascertain how closely the examination paper he has drawn up approaches a norm previously decided upon. For example, using Wood's Taxonomy, Scopes (1973 p.164) suggests that according to the amount of

time spent on the various areas of teaching a weighting can be drawn up as in Table 3.

Scopes' suggested proportioning of marks using Wood's taxonomy

	OBJECTIVES OF INSTRUCTION	PERCENTAGE OF MARK
A	Knowledge	20
B	Skills	25
C	Comprehension	25
D	Application	20
E	Inventiveness	10

Table 3

In the case of a Higher Grade paper in South Africa the author is not aware of any such criteria being laid down, although there is a sub-division of marks according to content. This means that all teachers virtually work towards their own ideas as to what constitutes a satisfactory assessment of learning on the higher grade. Without being prescriptive, it appears to be advisable that some guidelines should be given as teachers can be helped to consider their papers more carefully. The purpose of the examination is continually in mind when a blueprint is used, and as a result more thought is perforce given to the balance of the paper as more attention is paid to the various levels of ability. Use of a blueprint, however, does not mean that the paper set will have objectively selected questions as a result. As stressed before each selection is subjective.

Remembering that a Senior Certificate Paper must be set within the bounds of the syllabus and with summative evaluation being the prime aim, it is difficult to set questions which will lead to new

knowledge so that questions involving inventiveness will normally not be part of such an examination. But in school assessments where these restrictions should not apply as regidly, questions should be set with this in mind.

Using a taxonomy carefully should result in a greater evenness of standard in the papers set over the years. Teachers will be aware of all the process levels in their teaching. It could also mean less teaching to the examination in a literal sense in that it is easy to teach towards knowledge, and essentially when certain method types predominate in an examination paper that is what is being striven for. It is counter-productive to teach towards examination question types when there is greater emphasis on Application, and Analysis and Synthesis type questions. This last consideration leads us to review what effect Bloom's Taxonomy and Behavioural Objectives have had on instruction.

Chapter 3

Application of Educational Objectives and Bloom's Taxonomy to Instruction

Examinations have a great influence on what is taught, and this is especially true in South Africa, where the public examination at the end of standard ten with all its influence as far as employment opportunities is concerned plays an even greater role than in most countries, even to the extent that one of the fastest ways to change instruction is to change certain aspects of the examination paper. In the same way, with the view that an assessment should test different abilities, has come the idea that instruction should be more varied in intent. The suggested variety in assessment was initiated by those who believed that observation of the students' behaviour indicated whether higher-order abilities or knowledge only had been used. Thus it was logical that when instruction seeking greater variety in processes in education was considered, instructional objectives were seen in behavioural terms and Bloom's taxonomy was applied to instruction and behavioural objectives were espoused by many. Examples of instructional (behavioural) objectives were given in chapter 1.

Proponents of instructional objectives state that one of the main reasons for using these objectives is that it is consistent with the concept of accountability. This accountability despite being anathema to many educationists is a force in society, especially when one considers the growing

economic problems and the growing involvement of parent groups. Educationists, increasingly, may not live to themselves in a modern society. If behavioural objectives are accepted then standards of performance can be measured by society, and also by the teacher in order to test the adequacy of his instructional program. If students do not master the objectives then various changes may be considered in the instructional program. But this accountability can only be demanded of a teacher if the assessment is a reflection of the instruction. It is not right to set a test or examination in which analysis is required to answer the questions, if all the instruction has been at the level of knowledge alone, unless that instruction has ignored the objectives agreed upon. Thus the assessment is controlled by the instruction if there are instructional objectives, and accountability is then meaningful.

Proponents also argue that students are spared the frustration and time-consuming effort of trying to guess what the teachers expect of them, when they are told beforehand what they are expected to learn and what level of ability is required. As a result of clearly specified objectives, curriculum planners are better able to arrange sequences of courses of instruction, and should be able to prevent overlap as a result. Teachers can also discuss what is being taught meaningfully, if there is a clear idea of what is to be learned, and can also design the instructional means to greater purpose. In addition it is possible for the teacher to determine the student's present level of mastery for any prescribed objective both at the start of the process and during it, as well as at the end.

Tests can be constructed along the way so that evaluation can be formative aiding the instruction as it is seen how the pupil's behaviour compares with the objectives laid down. Because the teacher can determine the level which each student has reached at various stages, individualisation of instruction becomes possible, and the meeting of the needs of each child can be considered seriously. (Kibler pp.4-5). Also one can consider under what circumstances every child will be required to reach the same goals, and if this proves to be necessary one can then provide the extra time needed in order for all pupils to master what is regarded as essential.

In considering instructional objectives it is not true that the same kind of objective is used for each educational level. Instructional objectives depend on the level of education reached, the individual's capacity to perform different skills and his behavioural pattern. It frequently becomes more difficult to formulate objectives and measure the outcomes of instruction as the educational level increases, because they both become more complex. With this more complex situation at the higher level it is necessary to develop objectives that apply to each individual, and this requires great thought and analysis on the part of the teacher both of the subject-matter and of the pupil. With this analysis the teacher can only improve his instruction.

In order to assist teachers to use behavioural objectives Popham and Baker (1973) have written programmed tests to develop teacher competency in the Curricular Instructional and Evaluative spheres. Most classical behavioural objectives are limiting

in that they are mainly concerned with acquiring knowledge as seen in Mager's book (1962). Popham and Baker suggest how to express objectives at a higher cognitive level eg.:

"The learner will show that he perceives the meaning of Shakespeare by writing an essay that describes the purpose of any given subplot."

In each case the result is measured in terms of the intended behaviour and this is one of the main bones of contention about the use of educational objectives - their behavioural basis and the control of the education process that is exerted by the teacher.

Popham here and elsewhere (eg Popham 1968) evinces the attitude that the objectives model should be adopted as a basis for large scale action by every - one, although according to Davies (1976 p.70) he has now retreated from this extreme position. It is important to bear in mind that a model of objectives is not a thing, not a fact we cannot do without, but is a conceptual scheme. "We do not have objectives : we choose to conceptualise our behaviour in terms of objectives." It is important to realise that there may be other ways of organising one's teaching and organising it successfully although it must be stated that the number of ways almost varies with the number of teachers, and there is also the dangerous attitude that planning is not necessary at all which is held by some who are opposed to educational objectives. What leads many people to reject behavioural, or more euphemistically, instructional objectives?

Chapter 4

Criticism of Behavioural Objectives and the Taxonomic Approach

Criticism of Behavioural Objectives

Kneller in his article "Behavioural Objectives? NO!" (1975 p.34) writes as follows:

"The use of behavioural objectives in instruction is characteristic of a culture which sets a high value on efficiency and productivity. Such a culture seeks to measure accomplishment in standard units. Theoretical justification for behavioural objectives comes from behavioural psychology. This type of psychology defines learning as behaviour that is changed in conformity with predicted, measurable outcomes, and with little or no measurable 'waste'".

Ormell writes (1974 pp.5-6)

"Is it not absurd to try to classify educational objectives in terms of behaviours when the primary objective (on most interpretations of education) is the mind of the child? ..... the taxonomy reflects an attitude to education which may be described as 'materialist'..... One's doubts about this perhaps reduce to the feeling that it pre-supposes that no one is asking what education is, after all, for ..... One gets little sense of the appeal of education for education's sake, and a lot of the feeling of the training of children to be conspicuously effective in exercising skills on tasks of various levels of complexity..... One might describe the taxonomy therefore as being predisposed to fit a pragmatic, materialistic, meritocratic view of society, in which few profound questions of value arise and in which explicit

evaluation is always expected and preferred."

These two comments which evince a completely opposite philosophical viewpoint to that of those who support educational objectives, present an oversimplified logical sequence - almost a labelling. Both comments are a reaction to the use of behavioural objectives to the exclusion of all else and as such perform a very real service. Taking such an extreme position on either side, however, is not helpful at all. Opponents of behavioural objectives should rather see how the work done by the behaviourists can be used, and when this happens then course objectives designed to foster educational experience for students can be developed (Giroux 1979 p.418). Incidentally both Kneller and Ormell acknowledge that much of the work done by the behaviourists has merit, but they do not wish the behavioural models to be accepted uncritically.

Freudenthal in his criticism of educational objectives attacks the notion that there can be such a thing as general instructional objectives arguing that this leads to superficiality, and does not take the realities of the classroom situation into account. He states that the quest for instructional objectives is legitimate "but why if they are so important do people do so little about them," (1978 p.105) and this after sixteen years of instructional objectives being in vogue. He continues: "Formulating objectives should be preceded by profoundly scrutinising analysis of the subject matter. There is no cheaper way; an educationist who does not know enough mathematics is better advised to keep off objectives of mathematical learning (p.116)..... Of the general ones (objectives of instruction) I do not know as yet

whether they can involve more than the expression of a background philosophy of some use as such if they are cautiously watched. Of the operational ones I lack the proof of existence since they have not been represented by convincing examples. It will take time and trouble, if it ever succeeds to pull the objectives of instruction out of the swamp of slogans. It is a pity that show words as taxonomy and model more often than not radiate smattering rather than learning." (p.184).

The above comment is needless to say a little polenical, but does point to a very real problem that of triviality and superficiality as well as the problem of empirical support to back the philosophical viewpoint. It is important to note that none of the writers quoted above are against objectives as such. To be against all objectives leads to an attitude that all planned education is wrong - this will soon lead to chaos. What is being attacked is the view that only those objectives which lead to an observable and measurable change in behaviour are to be formulated. At present those who support behavioural objectives (now preferably called educational objectives by them) no longer take the extreme position that they did in the early sixties. When asked now "Can I use non-behavioural objectives in my teaching?" the reply would now probably be: "Certainly but make certain that stating your intentions in a non-behavioural way is appropriate to what you have in mind." (Davies 1976 p.62). This modified position is more defensible and is useful - it is, however, countered by the unfortunately more entrenched position of those that are opposed to behavioural objectives take instead of using what is of benefit from the approach.

In mentioning what objectives are considered to be useful one immediately meets the question: "Where do these so-called objectives come from and how are they derived?" In other words just because an objective is stated clearly it does not mean to say that it is valid. Responsibility for content cannot be evaded by concentrating on the issue of clarity alone, the objectives must have content validity. Associated with this is an assumption that must be questioned and that is, is it possible to know and readily identify the educational objectives for which one strives? We may have many objectives which are achieved on the way, but in doing so we may be completely unaware of a final though very worthy goal. One must recognize that unanticipated outcomes do occur and must allow for them as they are often very important. They occur because our memories constantly remember incidentals often at the expense of those items to which they have been directed.

In considering unintended outcomes one must realise that innovation in the curriculum can be hampered by too early a demand for isolating and identifying objectives. Also with the vast number of outcomes that are worth striving for some must, of necessity, be missed, and there will be a great deal of time wastage identifying the required objectives. This time could more profitably be spent in preparing the actual instruction. Against this one would use one's professional judgement in deciding how far one's subdivision of an area of study into objectives should go. This subdivision has in it the seeds of fragmentation, fragmentation in which one can lose the overall educational purpose, and can indeed even change the norms by which that purpose is judged.

"Creative curriculum development and good teaching do not necessarily begin with a blueprint. The constantly evolving richness, the intricate subtleties, the never-ending succession of refinements - all of which lie at the very heart of the creative process can sometimes be lost in the apparent materialism of objectives." (Davies 1976 p.66). To many, objectives portray little heaps of knowledge, rather than an integrating structure, especially if all that the objectives are is in reality a list of contents prefixed by words such as "to list" and "to analyse."

In a real sense there is a danger that if curriculum planning is confined to predetermined objectives then content in education cannot be seen as a whole, and is reduced to an instrumental role and all we have done is to explain what is in mind. A poem for example is more than a list of objectives to be achieved in studying that poem. A geometrical proof is more than just the analysis and synthesis required to write down that proof. The content of a whole is more than the sum of its parts, and this aspect cannot be measured by subdivision into parts. It must be seen in its totality rather than simply in a series of performance outcomes.

In the practical sphere there are a number of problems associated with the use of objectives. When teachers concur with a list of educational objectives they have to agree to the values that these objectives represent, and once having done so the question is whether they can make use of them in the intended way. Teachers differing operationalisation of objectives in the classroom makes it very difficult to justify the same objectives for

a whole group of classes unless the aim is to present stereotyped knowledge objectives. As soon as higher level abilities are considered the personality differences between teachers must result in widely differing interpretations of the objectives. Subjectivity is at the root of all educational objectives. In addition with the use of objectives there is the danger of uniformity of response being prized above variety, because the objective approach rests upon an assumption that it is possible to predict what the final results of teaching will be, without the realisation that what is predicted is only the minimal result and that one cannot set a standard against which the total achievement can be measured. As a criterion of accomplishment objectives are useless beyond the original specification which the objectives seek to represent.

When considering the arguments for and against educational objectives one must concede that both sides express very important principles. The only problem is to make sure that these principles are not distorted by over-emphasis and the total rejection of the opposing viewpoint. Thus Popham (1968) sought to refute all argument against behavioural objectives. In contrast those who are against educational objectives all too often fail to realise that goal-free education is impossible. "Creation is an organic development, but some sort of management is nearly always required otherwise agonising interaction is inevitable." (Davies 1976 p.71). To imagine that objectives must necessarily limit creation is to deny something of the very act itself. As Henri Poincare points out in describing mathematical creation, "the process may be unconscious, the

solution may appear in a sudden flash of understanding, but our will pursued a perfectly determined aim. To invent is to choose, and choice can only be made in terms of some ultimate yardstick or goal." (Davies 1976 p.72). Objectives, being commitments, help with the allocation of priorities - one is brought back to reality and objectives help to clear the way ahead. Objectives must,, however, never be regarded as the be-all and end-all of any creative or learning process, but as guides because the means can be individualised in order to help with the attainment of the ends. Objectives help to promote clarity of thought for the teacher and the pupil to show where they are going even if it is only because they force curriculum developers and teachers to be rather more precise about their intentions. Care must, however, be taken that the valid criticism of only being concerned with what has been verbalised in objective form should be watched carefully. Finally in view of the widespread endorsement of behavioural objectives one may expect to find many examples of their effective use. That this is not the case suggests that practical application of the concept may involve some difficulties. (Ebel 1970 p.172).

Criticism of the Taxonomic Approach

Freudenthal states (1978 p.81)

"this taxonomy can only be understood with the background knowledge of a homogeneous instruction and strict instructional norms created by a strong communis opinio; whoever applies these patterns of norms is thoroughly acquainted with what students know, the kind of courses they attend, the instructional methods which are the general custom, but he is also indoctrinated with a sharply

defined educational philosophy strongly depending on culture, time and country. Only with such a background are the valuations of the Taxonomy meaningful."

Because the taxonomy is based on a behavioural approach its basic philosophy and the classification cannot be regarded as independent of questions of value : particular classifications, like particular institutions tend to be suited to the expression of certain values and unsuited to the expression of others. The business of classifying educational objectives is not a purely technical matter. The problem lies in that if we see the order of objectives as self-evident we are presupposing the kind of value-system which fits the taxonomy most neatly. To use the taxonomy we must agree that behaviouristic change is measureable and that there should be broad concensus on the means of evaluation. In this connection it must be recognised that not all behaviouristic change is measurable. The teacher when teaching can know if his class is understanding what he is saying by their attitude, especially when he has taught the pupils for some time and knows them. One cannot use the taxonomy for this intuitive assessment. One must always remember the original reason for the establishment of the taxonomy and that was to classify examination questions. This point is amplified by Freudenthal when he states:

"The original target of the taxonomy was to examine an extensive but nevertheless well-defined sector of instruction,..... it should facilitate the grading of examination results. Applying this pattern to curriculum development and the preparation of classroom teaching is a dangerous transgression. It reinforces the tendency to identify

the objectives of instruction with examinations and to teach only what can be examined; if finally the contents of examinations are also determined by the pattern of norms, the vicious circle is firmly closed." (p.83).

If one accepts the above criticism then in an education system like that of South Africa's where the final examination carries so much weight, the taxonomy should be a useful tool, because there is strong common opinion on what is to be learnt and how it is to be learnt, because of the control exerted by the examination through the textbooks which dominate most teachers. Freudenthal does, however, overstate his case when he asserts that homogeneous instruction and strict instructional norms are required as background before the taxonomy can be used. All that is needed is a general agreement on what instruction should be carried out, thus allowing for individual means of presentation within the broad framework. This is especially true if the taxonomy chosen is not too detailed.

Strict application of Bloom's taxonomy to objectives of instruction must be avoided, because of the need to allow for inventiveness and imagination in the classroom situation, two very important activities that must be encouraged. However, the reality of the examination is with us, and the taxonomy will of necessity have some influence on instruction, even if it is only to indicate that teachers must not concentrate only on items that require recall or recognition. Another reason for avoiding strict application of Bloom's taxonomy is that a taxonomy should be a guide to pedagogical action and should be dynamic and readily adaptable to new develop-

ment something which Bloom's is not, it is merely a classifying device for pigeon-holing objectives that already exist. (Stones 1979 pp.155 and 158). But it is doubtful whether any classification can ever really be a guide to action, although Stones does try with his generating objectives.

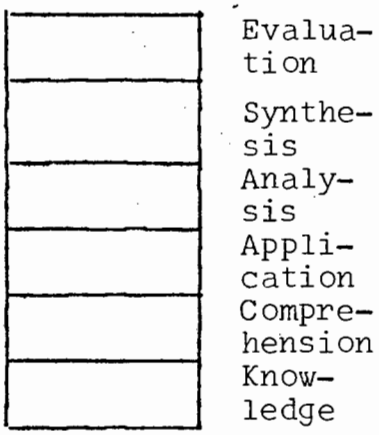
The danger of circularity which Freudenthal points out if the taxonomy is used as a guideline for instructional objectives, must be watched carefully. Circularity of reasoning also occurs within the taxonomy itself in that propositions about the mind are described by what can be observed. Propositions about the mind should have been examined by philosophical analysis, but instead Bloom placed much of the burden of defining educational goals and cognitive levels on test items, the correct response to which was taken as the necessary evidence of the attainment at issue. "Thus the authors took as the only viable alternatives the operational development in which the intended student behaviour was implicit." (Furst 1981 p.442).

Another problem that faces Bloom's Taxonomy is whether it actually does represent a hierarchy. There are a number of researchers who would prefer to see a horizontal classification with each area of knowledge the same as a hierarchical level of Bloom's taxonomy and in each category various levels of difficulty called by Ormell "levels of visibility" eg the factorisation of $x^4 - 1$ despite needing difference of two squares in the process is not as "visible" as the factorisation of $a^2 - b^2$. Ormell's ideas are illustrated in Figure 1.

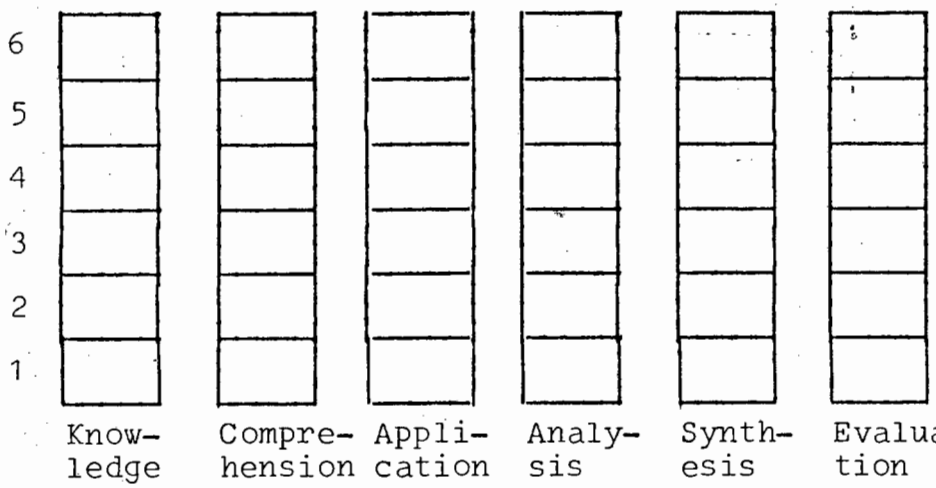
Figure 1

Classification to show Ormell's levels of visibility approach.

Instead of a pattern like this:



there would be a pattern as follows:



The scale from 1 to 6 represents levels of visibility in each category. There need not be six levels.

The difficulty of placing questions into the various categories even with a handbook for constant reference has been viewed with concern especially in the field of mathematics, because of its precision which contrasts markedly with the imprecise nature of the taxonomy. This conflict increases with the increasing difficulty of higher mathematics (Ormell 1975 p.3). This imprecision is more marked at the higher ability levels and this compounds the problem even further.

There is a view that there should be a non-hierarchical classification as well which should be aimed at describing the range of mathematical abilities thus complementing the hierarchical taxonomy. P. Human at Stellenbosch University suggests the following non-hierarchical classification of learning outcomes for Junior Secondary Mathematics

1. Execution - Use of Algorithms
2. Evaluation - True/False questions, spot the mistake
3. Manipulation - Changing from one form into another eg factorisation
4. Translation - Changing from one notational symbol system to another eg algebraic to graphical systems
5. Problem-solving - Student must decide what to do
6. Reproduction of knowledge
7. Terminology - use of terms

At Senior Secondary level more outcomes would be added to the list. One would be able to place any mathematical question into this horizontal classification. The only additional aspect that would have to be considered would be the level of difficulty required in each outcome so that a balanced paper can result.

Even amongst those that accept the hierarchical nature of Bloom's taxonomy there are many who question the order. Some have suggested a different structure. One of these is the Y-shaped structure of Madaus et al (cited by Seddon 1978 p.311) as illustrated in Figure 2.

Others are very critical of the position of Evaluation as the pinnacle of the hierarchy. In

Figure 2

Madaus et al's Structural Adaptation of Bloom's Taxonomy

Measuring achievement depending on learning ability	(	Analysis	Evaluation Synthesis	) Measuring general ability
	(	Application		
	(	Comprehension		
	(	Knowledge		

research mathematics, analysis is as important as synthesis. In Bloomian terms the word synthesis implies more than the mere assembly of disparate bits to form a whole : it implies that the bits fit together with a kind of authority, and clearly this cannot be achieved without an acute awareness of what will fit well together ie using evaluation. In this sense evaluation is bound up with, and presupposed by, anything deserving the title synthesis. But in today's pluralistic society there is no one pre-eminent system of values within which evaluation can be conducted, and for this reason evaluation must not appear as the final educational objective eg contrast an American business school students' attitude and a Maoists' attitude to economic theory - they will not be doing comparable thinking or producing comparable measurable outputs. (Ormell 1974 p.4). This philosophical argument leads back to the view that the taxonomy is pre-disposed to fit a pragmatic, materialistic and meritocratic view of society. But even in American society which is supposed to be the personification of this view the position of evaluation is challenged as will be seen when Kunen et al's experiment is considered in the next chapter.

Thus far in this survey all that has been considered can be summed up by the words of Eliza Doolittle:

"Words, words, words", and one's response like hers is "I am sick of words show me what you blighters can do." Arguments presented thus far have been based on philosophical views - the empirical factor must now be considered.

Chapter 5

Empirical Findings on Behavioural Objectives and Bloom's Taxonomy

Twelve years after Bloom's taxonomy first appeared, Wood commented as follows:

"Perhaps the messianic fervour with which Bloom has been hailed in some quarters has diverted attention from the fact that apart from one study (Kropp and Stoker 1966) little attempt has been made to validate the assumptions on which it rests, but the fact that it has spawned numerous inquiries into the nature of achievement in fields as diverse as mathematics and social hygiene indicates how suggestive the efforts of Professor Bloom and his associates have proved to be." (1968b p.85).

Kibler another proponent of Bloom and of educational objectives writes in 1974:

"there have been only fifty or so experimental studies focused on instructional objectives. Unfortunately the results of these studies are inconsistent and provide no conclusive evidence about the effect of instructional objectives on learning." (p.5).

Davies (1976 p.83) echoes Kibler's statement, but adds the following significant comment:

"It is important for teachers to bear in mind that while objectives have sometimes failed to be helpful, they have never been shown to be harmful."

Melton echoes this last comment by stating more emphatically when commenting on certain studies about behavioural objectives "in none of these instances did the availability of behavioural

objectives depress such learning." (1978 p.292).

The question arises as to the degree of discrepancy in the results. Kibler reports as follows (1974 pp. 6-7):

Of thirty-three studies that compared the effects of student possession of instructional objectives only eleven reported that instructional objectives improved learning while the other twenty-two found no difference. Of seven studies that investigated whether student learning was improved by the use of specific objectives as compared to more general objectives only two showed that possessing specific objectives helped the student - in the other five no difference was reported.

Of eight studies that investigated the effects of teacher possession and use of instructional objectives only three reported significant positive effects - in the other five no difference was reported.

Of seven studies that investigated the effect of the possession of instructional objectives on the use of student time only two found that study time was effectively reduced.

Kibler does suggest that there were a number of methodological weaknesses which may explain the inconsistent findings - one must, however, view this suggestion with scepticism in the light of Winne's review which concerns itself with an important aspect connected with Bloom's taxonomy.

Winne (1979) considered the effect of teachers classroom use of higher cognitive questions as opposed to lower cognitive questions on student learning and in doing so summarises eighteen experiments which have to do with this effect. In nine of these experiments called by the author "training"

experiments, the teachers were trained in the use of higher cognitive questioning, but were not monitored as to how they used this type of questioning in the classroom. In the other nine experiments the teachers were trained, and their classroom use of questioning was monitored and these Winne called "skills" experiments. Of the training experiments only one of the nine were judged to be sufficiently valid to permit relatively accurate inferences about the effect of higher cognitive versus factual questions on student achievement, while only six out of nine skills experiments were regarded as sufficiently valid. Taking these valid studies - and tallying the results on the basis of each time a pair of treatment groups were compared - in 64% of the cases no difference was recorded between the effects of asking higher cognitive questions as opposed to factual questions on pupils' achievement. When all the studies were taken into account the figure was 60%. Thus, Winne concluded "regardless of whether studies are methodologically sound the studies reviewed here indicate that whether teachers use predominantly higher cognitive questions or predominantly fact questions makes little difference in student achievement." (p.43).

Andre (1979) arrived at the opposite conclusion to Winne giving support to the facilitating of learning as a result of the use of higher level questions, but he added that they showed limited transfer and generalisation.

As far as Bloom's taxonomy is actually concerned, Seddon comments (1978 p.321) that there is no evidence that the properties do not exist nor is

there evidence that the properties do. The strongest supportive evidence concerns the cumulative hierarchical relationship between the categories Knowledge, Comprehension, Application and Analysis, but the evidence is by no means conclusive even though it should spur on the carrying out of further experiments.

Recently a study on Bloom's Taxonomy (1981 Kunen et al) has been published which deserves close attention. This study uses Levels-of-Processing Analysis which is a Framework for Memory Research first suggested by Craik and Lockhart (1972) which is based on the premise that retention is a function of depth, and thus that memory performance is a positive function of the level of processing required by the orienting task (p.678). In an experiment using this analysis subjects would be stimulated at various levels while being exposed to the same content. After the process is completed a surprise memory test is given, and the scores obtained are analysed. Craik and Tulving (1975) conducted a number of experiments and came to the following conclusion (p.290):

"It is abundantly clear that what determines the level of recall or recognition of a word event is not intention to learn, the amount of effort involved, the difficulty of the orienting task, the amount of time spent making judgments about the items or even the amount of rehearsal the items receive; rather it is the qualitative nature of the task, the kind of operations carried out on the items that determine retention."

A number of important points are raised in Kunen et al's study:

1. Central to the question of whether the Taxonomy represents a cumulative hierarchy is the

issue of whether the amount of learning increases or decreases as students engage in thinking at successively higher taxonomic levels. There is nothing in the assumptions underlying the taxonomy that leads to a clear prediction of whether scores should increase or decrease. Kropp et al (1966) were the first to formulate that mean scores should decrease as the level of category increases.

2. Many previous attempts to validate the hierarchical structure of the Taxonomy have been hampered by the absence of an appropriate technique that permits qualitative comparison of learning across levels. In order to do so identical content must be used at each level so that information common to all taxonomic levels can be tested.
3. The critical test of the cumulative hierarchical assumption depends on whether there is a trend across taxonomic levels. This means that taxonomic questions must be administered on a pre-study basis to orient subjects to acquire information compatible with each level. If given afterwards the subjects will orient themselves to the material predominantly at the lower level. In this connection Gagne and Britton (1982) suggest that the best time to reveal objectives is after the subjects have gone through the material, and before it is reviewed - performance is significantly better under such circumstances.
4. It is suggested that Bloom's Taxonomy fits appropriately in a Levels-of-Processing framework, since the taxonomy is assumed to represent levels of cognition that become successively more extensive and complex. The theory is that

> if students can be oriented to study materials at either a high or a low taxonomic level then the resultant memory by-products should last longer following high rather than low level questions.

In the experiment various groups of subjects in both America and Australia were tested at the following levels : Knowledge, Application, Synthesis and Evaluation. The questions set to orientate the students to the different levels were checked by a number of academies to obtain agreement as to whether they fell within the level suggested. This is remarkable when one considers Fairbrother's comment discussed already - it would appear as if there was much consultation amongst the people concerned. At each level the attention of the student was drawn to the same thirty principles which were drawn from college textbooks. In the assessment because the Knowledge and Application tasks took half the time that the other two did when pilot tests were carried out, it was decided to intersperse the questions in these categories with some simple mathematics problems to make sure that the students had to use their memories for the same length of time as those doing tests at the Synthesis and Evaluation levels. Thus all four groups completed their assessment tasks at the same time. Then to the students complete surprise they were given a memory test at the end - this test was the same for everyone, testing information common to all levels. The reasoning behind the formulation of the experiment was that if the Taxonomy truly represented a cumulative hierarchy, then presenting questions on a prestudy basis should result in subsequent memory performance that increases as taxonomic level increases.

The results on the memory test are given in Table 4.

Hierarchical Level	Memory test score
Knowledge	8,58
Application	11,58
Synthesis	15,09
Evaluation	11,24

Table 4
Average Memory Scores of students oriented to different process objectives
(from Kunen et al 1981 p. 207)

With the exception of Evaluation where the results of both the American and Australian students indicated low performance, there is reason to assume confirmation of the hierarchy. In the case of the American students the Synthesis group provided the highest memory score, while with the Australian students the Application and Synthesis groups provided similar scores.

It would appear that the criticism that Evaluation should not be regarded as the highest level in the hierarchy is supported in this study. The authors state that the results "provide moderately strong support for the assumption that the Taxonomy represents a cumulative hierarchy of categories of cognitive operations." (p. 208). Another aspect of the results that is important is that the recall produced by teaching focused at the Knowledge level was poor so teachers should try to prepare their questions to cover all taxonomic levels.

To sum up the empirical findings, it would firstly appear as if there have been far too few experiments carried out on Bloom's taxonomy. Because of the confusing results obtained by

researchers over the past ten years it would seem as if there has been a more careful consideration of method lately. It would definitely be of advantage for there to be follow-up of Kunen et al's study. It would also appear as if researchers avoid mathematics when considering either Bloom's taxonomy or the effect of higher order objectives on achievement, as nearly all the empirical findings concern themselves with the behavioural and natural sciences despite Wood's confident assertion in 1968:

"Although Bloom's taxonomy is intended to have universal application it is particularly relevant to mathematics where most significant behaviours appear to have cognitive origins." (1968b p. 86).

Ormell's comment on the imprecision of the taxonomic instrument as opposed to the precision of mathematics cited on p. 32 seems to be justified.

It would also seem that while teaching by objectives is important, teaching using behavioural objectives is not necessary although there is no harm done if one uses them.

Chapter 6

Applications of Educational Objectives and Taxonomics to Evaluation in Mathematics in South Africa

It is the author's belief that less-stereotyped questions in examinations, especially those on the higher grade, can help in the development of improved teaching practice in the classroom. This does not mean that all the questions must be different in type to those set in the past; this would be self-defeating in that the skills required in answering questions of higher ability depend on a thorough knowledge of the content. The reproduction of knowledge must, therefore, be an important component of any examination. Many educationists quite rightly point out that one can increase one's knowledge by understanding, but the converse is also true : one's understanding is increased by knowledge gained, for example pupils often obtain insight into the meaning of certain algorithmic processes once they know how to use the processes thoroughly. Thus knowledge must never be underestimated. In addition no examination should consist of questions on higher abilities alone because of the, sometimes unconscious, attitude of learning not being worthwhile that tends to be generated in the pupil, if he sees no reward for his labours. This attitude then leads to the inability to handle higher level questions. Process objectives must be seen in terms of content, and it is significant that Bloom's taxonomy has been adapted to be of use in the mathematical field. There must be a balance, and if the taxonomic approach can help in creating this balance then it deserves serious consideration.

At this stage it would be of use to consider a taxonomic breakdown of the Cape Senior Certificate Higher Grade First Papers in Mathematics for the period 1979 to 1981. In table 5 a blueprint has been drawn up for the three years.

Because the questions are non-multiple-choice it is quite possible that part of a question will be in the knowledge category while other parts will belong to higher abilities. The highest ability required by the question will indicate the level at which the questions will be categorised. One other aspect must be borne in mind constantly, that of what is to be regarded as pure recall or algorithmic thinking in contrast to the higher levels of ability. Without a knowledge of how the students were taught this decision becomes difficult, and the only means by which an examiner can decide is to use his classroom experience and to consider past examination papers. This is very subjective. Unfortunately in a public examination an examiner is not in a position to consult others.

Analysing table 5 it is clear that there was little variation in the number of marks allocated to the various content subdivisions. The syllabus was therefore covered in a consistent manner from year to year. Where there is inconsistency, however, is in the marks allocated per taxonomic level in each year. There was a steady decline in the influence of pure recognition and recall which is in line with the view that Higher Grade papers should not contain many questions of this nature. In addition it is to be noted that the total number of marks at the Analysis and Synthesis level remained reasonably constant. In a three-hour

Appendix B

Homework Exercise to Illustrate the Format of the Test Paper

Part A (10 marks)	Answer	Confidence
1. Solve for x: $\frac{3x}{8} = 6$	1.	
2. Select the best response ie A, B or C and circle it in the answer column if: A is a Rational Number B is an Irrational Number C is an Imaginary Number What kind of number is		
(a) $\log_6 7$	A B C	
(b) $\frac{22}{7}$	A B C	
(c) $(-3)^{\frac{1}{2}}$	A B C	
(d) $\sqrt[3]{81}$	A B C	

Part B (10 marks)

Do your working in the space provided.

3. Solve for x: $-\sqrt{x + 2} = x$

Answer	Confidence

Induction	-	15	-	-	15	1980
	-	8	7	-	15	1981
Non-Algebraic	-	12	-	-	12	1979
Questions	-	-	11	-	11	1980
	-	-	-	-	-	1981
Totals	23	89	56	32	200	1979
	17	125	36	22	200	1980
	8	82	83	27	200	1981

Table 5

Completed blueprint for Cape Senior Certificate Higher Grade Mathematics First Papers for November 1979, 1980 and 1981.

Process / Thought Process / Content	Knowledge	Comprehension	Application	Analysis and Synthesis	Totals	Year
	Recall, Recognition	Algorithmic Thinking and Generalisation		Open Search		
Relations	6	32	10	11	59	1979
Functions and	3	34	6	5	48	1980
Graphs	-	38	26	-	64	1981
Equations	6	13	20	-	39	1979
and	6	26	-	17	49	1980
Inequalities	-	6	30	7	43	1981
Indices	9	14	9	16	48	1979
Logarithms	8	42	5	-	55	1980
and Surds	8	30	11	5	54	1981
Sequences	2	10	15	-	27	1979
and	-	8	14	-	22	1980
Series	-	-	9	15	24	1981
Mathematical	-	8	2	5	15	1979

paper it is considered to be detrimental to the candidates to have too many questions which require detailed analysis and synthesis because of the time factor.

The first main difference that should be noted is in the number of marks allocated to Comprehension in 1980 as compared to the other two years. This made the 1980 paper easier for the Higher Grade candidate who does not really experience any sort of challenge when presented with questions at this level. Thus in 1980 about 70% of the marks were allocated to questions which either required actual recall or recall of well-practised methods. The second main difference is in the number of marks allocated to the Application level where the questions could be answered reasonably quickly, but required the candidate to consider them carefully and apply his knowledge, because they were different from the more usual type of question given in textbooks. The greater number of marks allocated to Application in 1981 meant that only 45% of the marks were allocated to actual recall or to recall of well-practised methods.

From this analysis it can be seen that the character of the three papers was different, this difference being most marked between the 1980 and the 1981 papers. The character of these last two papers was markedly different, and it would seem as if a consistent policy was not in force. Even though the category choice for each question is subjective, the differences shown here indicate that it would be advantageous to suggest that approximate percentage totals for each hierarchical level on some simplified taxonomic scale be implemented. As long as there is no attempt to be prescriptive this can

be of benefit to all concerned, because such variability in the character of papers in successive years is detrimental.

This will not make the setting of papers any more scientific because of the subjectivity involved, but it will cause the examiner to consider his questions and the balance of the paper more carefully. One can easily criticize the taxonomic approach, but it is evidently in this sense that Francis (1981 p.19) writes the following: "Despite these criticisms it has proved to be of real practical use in the sphere of attainment testing." The word 'proved' as used here must surely be in its loose conversational sense - not in a mathematical or statistical sense.

How Higher Grade and Standard Grade papers should compare is of great concern. To many the difference between the two in Mathematics lies in the greater content that the Higher Grade candidate has to learn, whereas to others the difference must lie both in the content and in the process ability level of the questions asked. Joubert (1980) has highlighted this difficulty when he analysed five Higher Grade and five Standard Grade 1978 Senior Certificate Geometry papers from various examining bodies in South Africa. There is not much difference in the average number of marks allocated at the various hierarchical levels as can be seen from table 6. Joubert has used Avital and Shettleworth's taxonomy in his comparison.

Thus even at the Standard Grade level over one-third of the marks were allocated to questions considered to require Analysis and Synthesis. This would seem to indicate that the difference between the two levels was regarded as solely being

	KNOWLEDGE	COMPREHENSION/ APPLICATION	ANALYSIS AND SYNTHESIS
HG	30,9%	25,5%	43,6%
SG	33,9%	30,6%	35,5%

Table 6

An analysis of Geometry questions in 1978 Senior Certificate papers
(adapted from Joubert 1980 p.54)

that of the quantity of material that had to be learnt. If one now considers the Geometry questions in the 1981 Cape Senior Certificate Standard Grade paper there appears to be a marked difference in character in the paper as the paper contains no difficult Geometry problems. Again the variety from year to year in level of questions should be less so that those who enter for the examination may know what is expected of them. This is especially true at the Standard Grade level where the pupils writing are generally not mathematically inclined, and also especially while the syllabus is basically a watered-down higher grade syllabus being more academic than is necessary. Thus a guide to the standard of the questions that will be set is of great importance. Again this must not be prescriptive - the examiner must have latitude and again it must be realised that subjectivity plays a large role in deciding the hierarchical level of questions.

A study of examination papers using a taxonomy can draw attention to apparent shortcomings in instruction in the subject as has been shown by Smith (1980). He investigated examination papers set at a number of schools at the Standard 5 level to Standard 7 level. His results on a hierarchical scale are given in Table 7:

Process Level	Standard 5	6	7
Knowledge	4,2 - 4,4	12,7	9,3
Comprehension	72,1-72,3	74,1	78,2
Application	23,5	13,1	12,1
Higher Abilities	0	0,27	0,5 - 1,8

Table 7

Average percentages of marks allocated at various taxonomic levels for Standard 5 to Standard 7 Mathematics papers translated from Smith (1980 p.86)

Smith states that the comprehension level includes the use of algorithms, the illustration of definitions and translation, ie the student has learnt the content, but the situation differs from the original. In other words this level consists mainly of well-practised algorithmic methods. It would appear as if there is an over-concentration of questions at the knowledge and knowledge of methods level, and too few questions at the Application and Higher Ability levels. What is of concern is the decline in the marks allocated at the higher ability levels from Standard 5 to Standard 6. The reason for this would seem to be that in Standard 5 the syllabus states that there must be a paper in which problems are to be set, whereas this does not apply in the Standard 6 and Standard 7 syllabus. This decline is most unfortunate especially in view of the demands of the Higher Grade syllabus in Standards 8 to 10. A possible future research area could be the effect of an over-emphasis on algorithmic processes in Standard 7 on ability to do well in Higher Grade Mathematics.

Because one can ascertain shortcomings in instruction teachers should consider analysing tests written during the term in order to assess the understanding by the pupils of the subject matter, not setting tests only to make sure that the pupils work. These tests should be carried out frequently so that feedback on instruction can be obtained quickly. The tests should be so designed as to ascertain problem areas so that formative evaluation followed by remedial instruction can be carried out. The teacher will then no doubt have to make time for the greater individualisation of the teaching program, and will also have to prepare to a greater extent. Class size becomes a definite factor in this process.

It is notable as far as research in South Africa on taxonomies applied to mathematics is concerned, that the worldwide trend of acceptance without empirical testing has been followed here - this acceptance no doubt being encouraged by the dominance of the examination in this country. When this attitude of acceptance is combined with a worldwide lack of empirical findings involving mathematics, then a wide area is open for investigation. A number of points about this wide area need to be made:

1. It would be profitable to ascertain pupil opinion on the questions in an examination paper, especially with regard to whether they have answered that type of question before. Establishment of this is a prerequisite in any study using a taxonomy.
2. It would be of use to compare on what type of questions pupils who obtained high, average or low marks gained their marks.
3. A question by question analysis of marks gained

at the various hierarchical levels to test whether the hierarchy applies would be worthwhile, although one must bear in mind Kunen et al's criticism that in order to test the hierarchy the same material should be tested at the various levels.

4. There needs to be an agreement as to the proportion of the marks that should be allocated at each hierarchical level. In order to do this the opinion of teachers and tertiary education lecturers should be sought. The latter should be consulted in that they are aware of the mathematical thought processes that are required in tertiary education. Especially would this be applicable to Higher Grade Mathematics as a pass on higher grade is a virtual necessity for entrance to university mathematics.

As far as instructional objectives are concerned it would appear that one does not have to prepare detailed specified behavioural objectives as generalised objectives seem to be as effective in improving pupil's achievement according to empirical investigations (Davies 1976 p.86). But teaching by means of objectives is definitely superior to the apparently casual non-objective approach so often used by many South African teachers. Careful use of higher level cognitive questions seems to help pupils' understanding and achievement to a greater extent. More confirmatory research on objectives of instruction is, as indicated before, needed before wide acceptance of this method can be propagated.

Research into what constitutes good assessment is absolutely essential especially with the changes that are taking place and will continue to take

place in South Africa with the coming educational population explosion. If the concensus is that certain well-defined needs must be fulfilled by education then the assessment must meet these needs. If the concensus is, in contrast, that education is there for education's sake then the assessment should reflect this. The assessor may not divorce himself from the direction of the educational process as a whole - he must respond to it, and in order to do so research on the effects of the changes on assessment must be carried out as with other aspects of education at this time.

Conclusion

Bloom's taxonomy with necessary adaptation is a useful tool for examinations, but it must never be viewed as a scientific tool nor must it be regarded as the only means of classification. A classification such as that by Human can be combined with Bloom to give a more complete picture of the mathematical abilities being tested. The main use of a taxonomy lies in the way that it can help the user think about the questions set in the paper concerned, thus helping to maintain a more even standard from one year to the next. Applying Bloom's taxonomy to the instructional sphere in education must be done carefully as the taxonomy is not geared for generating action, because the taxonomy is not dynamic and thus is not readily adaptable to new developments, although it can be used to point out weaknesses in instruction.

References

Ammons, M. : "Objectives and Outcomes" in Ebel, R.L. (ed) Encyclopedia of Educational Research 4th ed Toronto, Macmillan, 1969.

Andre, T. : "Does Answering Higher Level Questions while Reading Facilitate Productive Learning" Review of Educational Research Vol 49 No 2 Spring 1979 pp 280 - 318.

Avital, S.M. and S.J. Shettleworth : Objectives for Mathematics Learning Ontario, Ontario Institute for Studies in Education, Bulletin No 3, 1968.

Barker, D. and W.G. Hapkiewicz : "The Effects of Behavioural Objectives on Relevant and Incidental Learning at Two Levels of Bloom's Taxonomy" Journal of Educational Research, Vol 72, No 6 Jul/Aug 1979 pp 334 - 338.

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

Bloom, B.S., J.T. Hastings and G.F. Madaus : Handbook on Formative and Summative Evaluation of Student Learning. New York, Mc Graw-Hill, 1971.

Craik, F.I.M. and R.S. Lockhart : "Levels of Processing : a Framework for Memory Research." Journal of Verbal Learning and Verbal Behaviour. Vol 11 1972 pp 671 - 684.

Craik, F.I.M. and E. Tulving : "Depth of Processing and the Retention of Words in Episodic Memory." Journal of Experimental Psychology : General. Vol 104 No 3 1975 pp 268 - 294.

Davies, I.K. : Objectives in Curriculum Design London, Mc Graw-Hill, 1976.

De Landsheere, V. : "On Defining Educational Objectives" Evaluation in Education Vol 1 pp 73 - 190 1977.

Ebel, R.L. : "Behavioural Objectives : A Close Look" Phi Delta Kappan Vol 52 No 3 pp 171 - 173 1970

Fairbrother, R.W. : "The Reliability of Teachers Judgements of the Abilities being Tested by the Multiple-Choice Items." Educational Research Vol 17 No 3 June 1975 pp 202 - 210.

Francis, J.C. : "Writing Education Aims and Objectives for Examination Syllabuses." Curriculum Vol 2 No 2 Autumn 1981 pp 15 - 20.

Freudenthal, H. : Weeding and Sowing : Preface to a Science of Mathematical Education. Dordrecht, D. Riedel, 1978.

Furst, E.J. : "Bloom's Taxonomy of Educational Objectives for the Cognitive Domain : Philosophic and Educational Issues." Review of Educational Research. Vol 51 No 4 Winter 1981 pp 441 - 454.

Gagne, E.M. and B.K. Britton : "The Role of Objectives in finding the Organisation of Information Learned from Text." Contemporary Educational Psychology Vol 7 No 1 Jan 1982 pp 15 - 25.

Giroux, H.A. : "Overcoming Behavioural and Humanistic Objectives" Educational Forum. Vol 43 No 4 May 1979 pp 409 - 419.

Halls, W.D. and D. Humphreys : European Cirriculum Studies No 1 Mathematics (In the Academic Secondary School). Strasbourg, Council for Cultural Co-operation 1968.

Husen, T. (Ed) : International Study of Achievement in Mathematics Vol 1 New York, Wiley, 1967.

Joubert, G.J. : Enkele Vernuwingsmoontlikhede in die Evaluering van Meetkunde in die Senior Sekondêre Skoolfase. M.Ed. Dissertation, University of Stellenbosch, Stellenbosch 1980.

Kapfer, M.B. (ed) : Behavioural Objectives : The Position of the Pendulum. Englewood Cliffs, Educational Technology Publications, 1978.

Kibler, R.J., D.J. Gegala, L.L. Barker, D.T. Miles : Objectives for Instruction and Evaluation. Poston, Allyn and Bacon, 1974.

Kneller, G.F. : "Behavioural Objectives? NO!" South African Journal of Pedagogy. Vol 7 No 1 June-July 1973 pp 34 - 37.

Krathwohl, D.R., Bloom, D.S., and B.M. Masia : Taxonomy of Educational Objectives Handbook 2 : Affective Domain. London, Longmans, 1964.

Kropp, R.P., W.M. Stoker and W.L. Bashew : "The Validation of the Taxonomy of Educational Objectives" Journal of Experimental Education Vol 34 1966 pp 69 - 76 (cited in Kunen et al).

Kunen, S., R. Cohen and R. Solman : "A Levels-of-Processing Analysis of Bloom's Taxonomy." Journal of Educational Psychology Vol 73 no 2 April 1981 pp 202 - 211.

Mac Intosch, H.G. and D.E. Hale : Assessment and the Secondary School Teacher. London, Routledge and Kegan Paul, 1976.

Mager, R.F. : Preparing Instructional Objectives. California, Fearon Publishers, 1962.

Melton, R.F. : "Resolution of Conflicting Claims Concerning the Effect of Behavioural Objectives on Student Learning". Review of Educational Research Vol 48 No 2 Spring 1978 pp 291 - 302.

Ormell, C.P. : "Bloom's Taxonomy and the Objectives of Education". Educational Research. Vol 17 No 1 Nov 1974 pp 3 - 18.

Popham, W.J. : "Probing the Validity of Arguments Against Behavioural Goals'. in Kibler, R.J. et al : Objectives for Instrucion and Evaluation. Boston, Allyn and Bacon, 1974 pp 9 - 17.

Popham, W.J. and E.L. Baker : Systematic Instruction Englewood Cliffs, Prentice-Hall, 1970.

Provincial Administration of the Cape of Good Hope : Junior Secondary Couse Syllabus for Mathematics. Cape Town, 1973.

Scopes, P.G. : Mathematics in Secondary Schools - a Teaching Approach. Cambridge, Cambridge University Press, 1973.

Seddon, G.M. : "The Properties of Bloom's Taxonomy of Educational Objectives for the Cognitive Domain" Review of Educational Research, Vol 48 No 2 Spring 1978 pp 303 - 323.

Smith, J.C. : Doelstellings en Doelwitte in Wiskunde Onderwys in die Junior Sekondêre Skoolfase. M.Ed Dissertation, University of Stellenbosch, Stellenbosch, 1980.

Stenhouse, L. : An Introduction to Curriculum Research and Development. London, Heinemann, 1975.

Stones, E. : Psychopedagogy : Psychological Theory and the Practice of Teaching. London, Methuen, 1979.

Taba, H. : Curriculum Development : Theory and Practice. New York, Harcourt, Brace and World Inc. 1962.

UNESCO : New Trends in Mathematics Teaching 1972. Paris, UNESCO, 1973.

Willmot, Alan S. and G.G.W. Hall : O Level Examined : the Effect of Question Choice. London, Schools Council Research Studies, Macmillan Education Ltd 1973.

Wilson, J.W. : "Evaluation of Learning in Secondary School Mathematics" in Bloom et al Handbook in Formative and Summative Evaluation of Student Learning 1971 pp 643 - 696.

Winne, P.H. : "Experiments Relating Teachers' Use of Higher Cognitive Questions to Student Achievements." Review of Educational Research Vol 49 No 1 Winter 1979 pp 13 - 50.

Wiseman, S. and D. Pidgeon : Curriculum Evaluation. London, National Foundation for Educational Research, 1970.

Wood, R. : "Exploring Achievement" in Examinations and Assessment. Nelson, Association of Teachers of Mathematics MT Teaching Pamphlet No 14 1968a p 13 - 34.

Wood R. : "Objectives in the Teaching of Mathematics" Educational Research, Vol 10, 1968b pp 83 -98.

Wood, R. : "The Item Bank Project", in Examinations and Assessment. Nelson, Association of Teachers of Mathematics. MT. Teaching Pamphlet No 14 1968c pp 34 - 44.

CONFIDENCE SCORING METHODS

A CRITICAL REVIEW OF THE METHODS

WITH AN EXTENSION OF THE APPLICATION

OF CONFIDENCE-WEIGHTING TO TRADITIONAL

MATHEMATICS TESTING

David A. Norton

A paper submitted to the Faculty of Education,
University of Cape Town, in partial fulfilment of
the requirements for the degree of Master of Education

1982

Abstract

This paper seeks to introduce the three main methods of confidence scoring namely Confidence-Weighting, Distractor Marking and Probabilistic Scoring, and to examine the strengths and weaknesses of each, as well as the approach as a whole. The possible influence of confidence-weighting on teaching is considered, while consideration is also given to the extension of confidence-weighted scoring procedures to the more traditional kind of mathematics evaluation used in South Africa, where the effect of confidence-weighting on partial knowledge can be considered more fully. A group of thirty-eight Standard Ten pupils were given a traditional mathematics test which was scored by confidence methods. The experiment confirmed the hypothesis that confidence-weighting has a greater effect on part scores than on full and zero scores. It was also ascertained that there was an increase in reliability with the use of confidence-weighted scoring despite there being little increase in the range of scores. It was found that the effect of confidence-weighting on the score was the same regardless of whether the error made was careless or an error of misunderstanding.

Contents

Introduction		p.1
Chapter 1	Confidence-weighted scoring	p.3
2	Marking of distractors	p.17
3	Probabilistic Testing Methods	p.20
4	Evaluation of Confidence Testing	p.27
5	Confidence Testing as an aid to the Learning Process	p.31
6	Confidence-Weighted Scoring of a Mathematics Test of a Traditional Type	p.35
Conclusion		p.51
Appendices		p.54
References		p.77

Tables

1. Confidence-Weighted Scoring in True-False tests p.4
2. Scoring systems for Dressel and Schmid's Free Choice and Degree of Certainty Tests p.7
3. Confidence-weighted scales using three levels of confidence p.9
4. Theoretical consequences of adopting different strategies on various three-levels of confidence scoring systems p.11
5. Coombs, Mulholland and Womer's Distractor Marking and conclusions therefrom applied to four-answer multiple-choice items p.17
6. Personal Probabilistic Confidence Tests using simplified response methods p.24
7. Reliability and Validity Coefficients of certain probabilistic tests p.25
8. Analysis of the Effect of Types of Errors on Scores in a Traditional Mathematics Test p.46

Appendices

A. Instructions for preliminary homework exercise for test on Sequences and Series p.54

B. Homework exercise to illustrate the format of the test paper p.56

C. Test on Sequences and Series p.57

D. Memorandum of marking for test on Sequences and Series p.60

E. Item analysis on multiple-choice test p.62

F. Raw scores obtained from test on Sequences and Series after rejection of certain multiple-choice items p.64

G. Graphs to illustrate the spread of scores of Conventional compared with Confidence-weighted marking of the Mathematics test p.66

H. Reliability Coefficients for each part of the test and Correlation Coefficients between parts p.67

I. Traditional Questions : Analysis of marks by question
Effect of Confidence-weighting p.68

J. Statistical analysis of marks obtained for Questions 5, 6 and 7 p.69

K. Statistical test of the effect of Confidence-weighting on marks obtained in traditional questions p.71

L. Error analysis of traditional questions p.72

M. Statistical tests on error analysis data p.74

Introduction

Ever since multiple-choice questions were first introduced as a means of testing there have always been those who have criticised their use because of their concentration on marks given for totally right answers only, and because of the encouragement given to guessing. On the first count it is often stated that partial knowledge needs to be given credit as it usually is in more traditional forms of essay type testing, and for this reason it is suggested that ways and means should be found that evaluate partial knowledge and/or test the degree of confidence that students have in their answers. This thinking has led to confidence scoring methods of various kinds. Echternacht in his 1972 review of the various forms of confidence testing defines it as "a method of testing where weights are assigned directly or indirectly to item responses in such a way as to reflect the examinee's belief in the correctness of the alternative or alternatives so marked". (p.217)

Many view these methods as being of great benefit, especially theoretically, because more information is ascertained concerning the students' ability - information which otherwise would be unknown to the evaluator. In contrast Wood (1977 p.239) writes as follows:

> "Despite all the ingenuity and effort which has gone into developing methods for rewarding partial information.... there is little evidence that any one method provides measurable gains. Elimination scoring has possibilities, as does self-scoring, except that it is presently too limited in scope. Confidence weighting is, I think too elaborate and beyond the average

> candidate. One is left with the conclusion that if the items in a test are well constructed, if candidates are advised to go over their answers since changing answers seems to pay, and if the testing conditions are such as to inhibit blind guessing with candidates being encouraged to attempt all items, number right suffices most needs."

These are highly contrasting views and neither touch on a further necessary consideration. What effect does a method of testing have on the learning situation? For example does the necessity of expressing a certain degree of confidence in one's answer make one more aware of the knowledge that one is required to learn, and thus enhance the learning process or does it have no effect? Little has been done experimentally on this although recent studies conducted by Sieber (1978 and 1979) throw some light on this aspect.

Before considering Sieber's work, however, the various types of tests that have been developed with their strengths and weaknesses should be reviewed. The tests can be divided into three categories:

Confidence-weighting of one answer considered to be correct;
Elimination of distractors; and
Probabilistic testing.

Chapter 1
Confidence-Weighted Scoring : A Critique

In this type of testing the examinee is asked to indicate what he believes the correct answer to be, and how certain he is of the correctness of his answer.

Confidence Weighting applied to True-False tests.

This type of testing was first employed to adjust true-false scores by Hevner in 1932 (Echternacht 1972 p. 218). It was later modified by Soderquist in 1936, and Ebel (1965a) suggested a different system of scoring. The scoring systems used by these researchers are given in Table 1.

In her work, Hevner compared four different scoring techniques: (a) the number right, (b) the number right minus the number wrong, (c) the weighted correct score and (d) the weighted correct minus the weighted wrong score; testing each for reliability according to Spearman-Brown's formula for a test of double length. This resulted in the weighted right score giving the highest reliability coefficient of 0,76 compared to 0,67 for conventional scoring. Soderquist found that the weighted-right minus the weighted-wrong score gave the highest reliability according to the Spearman-Brown Split-half method- 0,85 compared to 0,72 for conventional scoring. According to Soderquist the difference between his results and Hevner's was probably due to his telling the examinees how the test would be scored, whereas Hevner did not inform her music students. Ebel recorded significant increases in reliability on tests for students on Educational Measurement,

Tester / Answer	Hevner	Soderquist	Ebel
	Confidence Rating	Confidence Rating	Confidence Rating
Correct	High confidence +3	High confidence +4	Probably +2
	Middle confidence +2	Fair confidence +3	Possibly +1
	Low confidence +1	Low confidence +2	
		No real confidence +1	
Omission	0	0	$\frac{1}{2}$
Wrong	Low confidence -1	No real confidence -1	Possibly -1
	Middle confidence -2	Low confidence -2	Probably -2
	High confidence -3	Fair confidence -3	
		High confidence -4	

Table 1

Confidence-Weighted Scoring in True-False tests

these being 0,71 compared with 0,57; 0,83 compared with 0,77 and 0,82 compared with 0,73. (Hopkins et al 1973 p.137). In addition to this Ebel pointed out how a student could maximise his score by stating that if a student answered more than two-thirds of the questions correctly, then he would score more highly at the higher confidence response, whereas he would score more highly at the lower level of confidence if he answered less than two-thirds correctly (Ebel 1965b p.132). This indicates that knowledge about the quality of his own knowledge is necessary otherwise a student's mark will be affected

adversely. Thus an evaluator can obtain increased information about a student by the use of this method of scoring if he is prepared to analyse the results.

True-false items are, however, often regarded as unsatisfactory tests because of the limited choice that is involved with the consequent influence of guessing on the results, and because of their unreliability. Ebel does suggest, that because true-false tests are so seriously affected by guessing, confidence-weighting should always be applied to them. The question now arises as to how effective confidence-weighting would be in multiple-choice testing.

Confidence Scoring applied to Multiple-Choice tests.

1. Dressel and Schmid (1953) were the first to apply confidence-weighting to a multiple-choice test. They scored the tests in four different ways, each of which was known to the particular examinees concerned, and two of these involved confidence:
   a) a Free Choice test where examinees were asked to mark as many answers as they thought were correct; and
   b) a Degree of Certainty test where examinees were asked to indicate on a four-point scale their certainty in a single answer selected.

The methods of scoring for these two tests are given in table 2.

Dressel and Schmid found that there was only a marginal increase in reliability between the test marked conventionally and the degree of certainty test, from 0,70 to 0,73 - an increase which is definitely not significant. What was shown was that the performance of students seemed to be related

positively to expressed certainty, and also that students show decreasing confidence with increasing difficulty of items. The free-choice items seemed to differentiate the good student from the average and poor student, but failed to distinguish between the average and poor student. The degree of certainty test on the other hand differentiated between all three levels of students about equally well. (Dressel 1953 p.592). This is possibly due to the effect of the increased range of marks created by the scoring system. Thus the ability of degree of certainty testing to aid in the differential grading of students was demonstrated for the first time.

As far as the scale of confidence is concerned Average and Low Performers were positive and fairly certain about their results approximately equally while more High Performers were positive about their results. Another significant difference was that low performers tended to guess more wildly than the others. These last observations confirm what would naturally be expected i.e. that the person with the higher ability would be more confident of his answers while a person with low performance will tend to guess more. Even though this is not a striking result it does confirm what was believed. If all that can be gained is this result and a greater spread of marks then one must question whether the increased testing and marking time required is justified.

One further disadvantage of Dressel and Schmid's work which they acknowledge themselves was that the students were not familiar with this type of testing. This lack of familiarity no doubt affected the result, and also meant that the method could not exercise any influence on the examinee in his learning situation.

Free Choice Test

Number of alternatives marked	Item score	
	Correct alternative marked	Correct alternative not marked
1	4	-1
2	3	-2
3	2	-3
4	1	-4
5	0	

Degree of Certainty Test

Certainty value	Item score	
	Correct alternative marked	Correct alternative not marked
Positive	4	-4
Fairly sure	3	-3
Rational guess	2	-2
No defensible choice	1	-1

Table 2

Scoring systems for Dressel and Schmid's Free Choice and Degree of Certainty tests. (Adapted from Echternacht 1972 pp. 220-221)

2. Three levels of Confidence Scale Tests

The work of Ahlgren, Rothman (1969), Paton (1971), Hopkins, Hakstian and Hopkins (1973) and Fredman (1977), should now be considered. Each worked on a three levels confidence approach using different scales for scoring which are given in table 3.

Consideration of the Scales

In the scale used by Ahlgren and Rothman what is of immediate concern is that guessing is rewarded even when the answer is wrong, with the result that the person who honestly admits that he does not know an answer to a particular question is at a disadvantage. Thus one of the originally stated aims of confidence testing ie counteracting the effects of guessing is itself counteracted by this scale.

Paton (1971 p.53) adjusted the scale so that there would be greater countermarking for incorrect answers, thus reducing greatly a student's chances of obtaining a significant score when he knows little or nothing about a particular question, and also eliminating any reward for incorrect guessing (1971 p.55) Paton's non-linear scale was drawn up to take student strategies into account. It seems, however, that the person who is very sure is not sufficiently rewarded compared to the person who obtains the correct answer by guessing, because the difference between a correct guess and an omission is too great when compared with the credit obtained by an answer given in the very sure category. Thus in trying to cater for theoretical consequences of student strategies other disadvantages result. In addition if the scores of a correct and an incorrect guess are added, the result is the same as that of correct and incorrect answers at the very sure level, but less than those at the fairly sure level, which means that using the fairly sure level is a good risk. The main disadvantage of the scale is that the reward for being very sure is too small.

Hopkins et al's and Fredman's A scale returns to the linear format and again the scale has the advantage that credit is not given for wrong guessing. Fredman's B scale is interesting in that it provides the

Table 3

Confidence-weighted scales using three levels of confidence

Researchers / Levels of Confidence	Ahlgren and Rothman	Paton	Hopkins et al	Fredman		
				A	B	C
Correct very sure	+ $\frac{4}{3}$	$\frac{5}{3}$	3	3	5	10
fairly sure/sure	+ 1	$\frac{4}{3}$	2	2	4	8
guess	+ $\frac{2}{3}$	1	1	1	1	3
Omission	0	0	0	0	0	0
Incorrect guess	+ $\frac{1}{3}$	- $\frac{1}{3}$	-1	-1	-1	-3
fairly sure/sure	0	- $\frac{1}{2}$	-2	-2	-3	-8
very sure	- $\frac{1}{3}$	-1	-3	-3	-5	-10

greatest reward for those who are sure compared to those who are guessing. Fredman stated that the reason for drawing up this kind of scale was that he and his head of department concluded that the various grades of confidence were not equal steps, and that the step between sure and very sure is less than the step between guess and sure. (Fredman 1977 p.100). It is also important to note that Fredman regards the step between omission and guess to be the smallest of all, equal to that between sure and very sure, which is in direct contrast to Paton's scale where this is the largest step. A quirk in the scale is that at the sure level a correct answer earns a score of 4 while an incorrect answer obtains a score of -3. This has an effect on strategies students may employ.

The theoretical consequences of the strategies that students may employ using the various scales are presented in table 4. Fredman's C scale is not considered as it gives similar results to that of the

A scale.

Strategies using the Different Scoring Systems

Paton (1971 p.54) points out that when using Ahlgren's and Rothman's scales far too much credit is given for guessing. A significant positive result of 40% can be obtained even when one has no knowledge about what is being tested. It pays to guess even when one knows three choices to be incorrect, unless one has done poorly in the rest of the test when one's score should be maximised by choosing the very sure category. Using this category in this way is a misnomer as one cannot be very sure if one is uncertain as to which of two or more alternatives should be taken. The fairly sure category is meaningless in that it does not give any advantage to the student. Paton's modification does allow for the meaningful use of all the levels of confidence. Thus strategies devised to maximise the examinee's score will include the obvious very sure if all the incorrect distractors are known, while if three incorrect choices are known then it is worthwhile to answer at the fairly sure level of confidence. It is evident that if only one or two distractors are known to be incorrect then the student should either omit the answer or guess.

The advantage to the teacher of using Paton's scale is that he can gain an idea of the student's partial knowledge if he does use a deliberate strategy, whereas in the case of Ahlgren's scale very little can be ascertained concerning partial knowledge. The kind of partial knowledge obtained is limited to the student's confidence in the answer he has chosen, and certainly does not make us aware of his actual knowledge as he has not indicated which distractor(s) he recognises to be such. We are also really only aware of his partial knowledge if he follows the maximising strategy strictly, so that use of this scale does not provide

Student's knowledge		Knows 4 choices incorrect	Knows 3 choices incorrect	Knows 2 choices incorrect	Knows 1 choices incorrect	Nil
% correct by chance (Scoring 1 or 0)		100	50	33,3	25	20
Percentage score obtained when confidence in answers on various scales marked as	Ahlgren & Rothman					
	Very sure	133,3	50	22,2	8,3	0
	Fairly sure	100	50	33,3	25	20
	Guess	66,7	50	44,4	41,7	40
	Paton					
	Very sure	166,7	33,3	-11,1	-33,3	-46,7
	Fairly sure	133,3	42,7	+11,1	- 4,2	-13,3
	Guess	100	33,3	+11,1	0	- 6,7
	Hopkins et al & Fredman A					
	Very sure	300	0	-100	-150	-180
	Sure	200	0	- 66,7	-100	-120
	Guess	100	0	- 33,3	- 50	- 60
	Fredman B					
	Very sure	500	0	-166,7	-250	-300
	Sure	400	50	- 66,7	-125	-160
	Guess	100	0	- 33,3	- 50	- 60

Table 4

Theoretical consequences of adopting different strategies on various three levels of confidence scoring systems

useful understanding of a pupil's partial knowledge.

When considering Hopkins et al's and Fredman's A scale it is evident that the student who wishes to maximise his marks must know all four distractors in which case he should choose the very sure category. If he does not know all four distractors then he should omit the answer, or at most answer at the lowest level of confidence in order to be penalised least. As Fredman states (1977 p.101) although he is mistakingly commenting on his B scale "The theoretical consequences of using this scale mean that the student will add to his score only when he knows the correct answer. The opportunity to enhance his marks by the adoption of any strategy (apart from his true knowledge) is reduced to the absolute minimum." Fredman's conclusion is no doubt very important for medical testing, but it does eliminate all hope of ascertaining the partial knowledge of the student. It is noteworthy, however, that in the Fredman B scale if one knows three of the distractors then one can obtain a positive score using the fairly sure level of confidence, so that some measure of the student's partial knowledge can be learnt despite Fredman's comment on this scale.

The choice of scale used will depend on the aim of the person setting the examination. Clearly if the aim is to ascertain the student's whole and partial knowledge then the best scale to use is Paton's. If the aim is to give maximum credit to sure knowledge combined with the elimination of student strategies then Hopkins et al's scale should be chosen. Incidentally the sure category on this scale is meaningless. In this connection it is worthy of note that Fredman now uses a scale with only two levels of confidence, sure and guess, with scoring +2, +1, 0, -1, -2 (Fredman 1979 p.416). He finds that satisfactory results are

obtained with this simpler scale which eliminates the almost meaningless middle level of confidence.

Influence of Confidence Weighting on Reliability

Ahlgren reported an increase in reliability in 20 out of the 25 cases in which he used confidence-weighted scoring using the K-R20 formula for the conventional test and the Cronbach α test for the weighted scores. Hopkins et al (1973 p.139) reported that the typical pattern of a slightly higher reliability estimate for confidence scoring was observed although the difference in reliability estimates did not reach the criterion of statistical significance. Their confidence-weighted score was 0,92 compared with 0,88 whereas the criterion test reliability was 0,93. Hopkins et al conclude that these findings suggest that the added reliable variance may be an irrelevant response style variance (p.140).

Unfortunately Fredman's reliabilities are not acceptable seeing that he used the K-R20 reliability formula for the confidence-weighted scores when it may only be used if the scoring is either 1 or 0. This applies both to his 1977 and 1979 papers. Ebel, to whom Fredman refers, states categorically: "If the scores of the tests are corrected for guessing or if other forms of weighted scoring are used, more complex variations of the formula must be employed"(1979 p.279). It is further strange that he should have accepted his 1977 results when he obtained reliability coefficients greater than one.

Thus a general increase in reliability is noted, but this must not be regarded as the all important factor on deciding whether to use a test or not. Too many researchers are concerned with reliability to the exclusion of all else.

Generally there is very little difference in the validity coefficients of confidence-weighted as opposed to conventional scores when compared with a criterion. Most view that because of the marginal increase in reliability the validity of the test cannot really be affected (Hopkins et al, 1973 p.136). Hopkins et al found that the validity coefficient for his confidence-weighted score of 0,667 was slightly lower than that for the conventional test which was 0,701, and this despite the tendency for higher ability students to respond with greater confidence. The authors concluded that this would seem to indicate that the greater confidence was not allocated wisely among the items (p.139). This thinking accords with a comment by Ebel on the confidence-weighted True-False test, where he pointed out that the advantages to be gained from confidence-weighting depend on how rationally it can be applied ending with "the results of recent experimental studies suggest that sometimes the more capable students are not much more successful than their less capable classmates in deciding when to answer confidently and when to answer cautiously." (1965a p.56)

3. Usefulness of Confidence-Weighted Scores

There are two main uses to which Confidence-Weighted scoring can be put. Because of the generally increased reliability there is a measurable increase in effective test length, for example Rothman (1969 p.238) quoted an increased test length of 1,46. This automatically means that a shorter test can be set using confidence-weighted scoring, which will be as reliable as a longer test using conventional scoring. This increase in test length was very important to Fredman in that it enabled him to set shorter tests during the year for the guidance of his students, and still know that the results obtained would be as reliable as longer conventionally marked tests (1977 p.107). This would also mean that

his students would obtain practice in the techniques involved in answering confidence-weighted tests, as well as having a good guide as to their ability in the subject. Thus a series of short tests designed to consolidate the knowledge learnt could be organised for the benefit of all the students involved. In the past the increased time required to mark the test compared to normal marking time was a problem, but with computerised marking becoming more sophisticated this is no longer the case, and so the series of short tests can be marked quickly providing the extra reinforcement so necessary for such a program.

The second main use is in the differential grading of students. Instead of all the marks ranging between 0 and 100 with most scores falling in the range 40 to 60, the range is increased in the case of Fredman's A scale from -300 to +300. This marked increase results in an increased spread of scores, and with this increased spread one can grade students more easily. Using this scoring system a large number in the class will score less than zero and the majority will obtain less than 50%. This is no problem if the confidence-weighted test is used for grading purposes only, and is combined with other methods of testing designed for certification in order to obtain an overall assessment of the candidate. Thus instead of one type of test attempting to fulfil the functions of certification and grading, it is suggested that two different tests be set specifically to fulfil these functions separately. Fredman advocates this use strongly. Two tests are set. One is a conventional multiple-choice paper in which no question will score less than 70% of the student responses correct. If the student passes this test then he will not fail the examination as a whole, and thus this test acts in a certification capacity. A grading paper is then set with more

difficult questions and is confidence marked with the marks being adjusted by means of standard scores so that a mean of 60% and a standard deviation of 12% is obtained. Confidence-weighted scoring differentiates more readily between candidates when the questions themselves are more difficult, so that the candidate is made to decide whether he is sure about an answer or not. Confidence-weighting tends to be an academic exercise when the questions are easy.

Whether the two main uses justify the extra marking time is a matter of opinion. Palva and Korkunen regard such very strict ranking as usually not being meaningful and definitely not a justification for the extra marking load (1973 p.179). It must, however, be remembered that computers have developed markedly since 1973 so that the marking load is not that much greater. In addition, because of the increase in effective test length, tests can be made shorter and will be still as reliable as longer conventionally marked tests thus easing the marking load.

It would now be of benefit to look at other schemes of scoring that will enable one to obtain more information about the partial knowledge that examiners have. Thus far only Paton's system has been able to provide this and only when a particular strategy is followed. As to the effect that confidence-weighted scoring can have on learning and retention, this will be considered in chapter five.

Chapter 2
Marking of Distractors

The marking of distractors was first suggested by Coombs, Mulholland and Womer (1956), because they were interested in ascertaining information about the examinees' partial knowledge. Lord and Novick (1968 p.315) comment that, in addition to this, the scheme often tends to discourage guessing. The system of marking used in the scheme and the conclusions drawn from it on a student's partial knowledge are given in table 5.

Number of Distractors Marked	Conclusion on Student's Knowledge	Scoring
3	Complete information	+1 for each distractor marked correctly -3 for a correct answer marked as a distractor
1 or 2	Partial information	
Correct answer only marked	Complete misinformation	
Correct answer marked amongst distractors	Partial misinformation	

Table 5

Coombs' Mulholland and Womer's Distractor Marking and conclusions therefrom applied to four-answer multiple-choice items

It is evident that in order to use the information obtained from this type of test properly, each candidate's answers will have to be analysed carefully and the results commented upon, so that the examinee can obtain feedback. It is only when the information that can be

obtained from a test is used that a testing process can be regarded as being worthwhile. The strength of this testing lies in its ability to isolate some of the misinformation and to highlight the missing gaps that form part of a candidate's knowledge. The confidence-weighted test is unable to do this.

Considering the test further Coombs et al reported that little additional testing time was required (1956 p.15), and that with the scoring from -3 to +3 there was an increased variance and increased discrimination between the variables (p.35). This last advantage is the same as that obtained by confidence-weighted scoring. As far as reliability is concerned there was an increase to a degree equivalent to a 20% increase in effective test length. The effect on reliability is, however, dependent on the difficulty of the test with an increase in reliability for more difficult tests. The validity of the test is about the same as for a test using conventional scoring (p .36). The examinees preferred this type of test to the conventional type as they regarded that it gave a truer representation of their knowledge.

Archer in 1962 (Echternacht 1972 p.222) found that this testing method was only slightly more reliable than conventional testing, so slight that the effective test length was reduced, but there was a definite increase in reliability on the more difficult items. The test did not correlate as highly as the conventional test did with a criterion.

Except for Wood's comment (1977 p.239) that this method seems too limited in scope, the method appears to have been ignored recently, and it would appear as if the low increase in reliability and the increased marking time have discouraged its use even though it is simple to

administer. Perhaps educators are not really interested in information on partial knowledge. Such information to be of benefit must be analysed and this means increased work! Another possible reason is that theoretically at least it is possible to obtain 66,7% without knowing the correct answer to a single question. Tests marked in this way would, however, be very useful as mid-course tests when the information gained on the partial knowledge can be used to greatest advantage.

Chapter 3
Probabilistic Testing Methods

These testing methods were first introduced in the 1960s and were the first to have a theoretical basis founded on a high degree of mathematical sophistication. This method was first proposed by de Finetti in 1965 when he introduced decision theory and personal probability into confidence testing basing the method on assumptions of examinee behaviour. (Echternacht 1972 p.225). There are drawbacks to probabilistic methods and amongst these are the difficulties that examinees have to estimate their own personal probabilities which they are to attach to the alternatives in a question, so that it is extremely difficult to develop these probabilities on a continuous scale. In addition to this as soon as opinions of this nature are sought then the administration time for the test is far longer. The scoring table is of necessity far more complex, and this can lead to students not understanding the processes by which the scoring is calculated, which is a most negative factor. Examinees must be prepared for this type of test through an easing-in process so that each person can gradually grasp the notion of attaching a quantity to his belief in the truth of the alternatives.

Because of the difficulty of a student suggesting his own probabilities de Finetti suggested a five-star system (Echternacht 1972 p.224). The examinee is given five stars and is asked to place them on the alternatives in such a way that the weights indicate his relative strength of about the alternatives. If on five alternatives the stars are allocated as follows:

three on one, two on another, and nil on the other three; then the resulting probabilities will be:

0,6 0,4 and 0,0;

and the scores given would be

21 16 and 6

If one star were to be allocated to each answer, 15 would be scored which seems rather high when if all five stars are given to one alternative and none on the rest only 25 marks are awarded. The scores allocated for each distribution of stars are listed in a table, and all that is required is that they be read off. No experiments have been conducted using this method as far as the author is aware however, probably because of the complicated nature of the scoring mechanism, and the high result of 60% that could be obtained even if nothing is known by the candidate. But de Finetti's work did spawn a number of experiments.

Shuford, Albert and Massengill in 1966 developed a theory of valid confidence testing using Logarithmic, Spherical and Euclidean functions. The last type of function is suitable for scoring items using a distribution of confidence (Rippey 1968 p.211). The scoring weights are obtained from the confidence that each pupil has in the answer he has chosen. The relation between confidence and the scoring weights is not linear, but is graduated approximately as indicated (note that there are 26 different degrees of confidence which the examinee is helped to find by using a Scorule):

Confidence	100%	80%	60%	40%	20%	0%
Weight	1,00	0,90	0,78	0,60	0,30	-1,00

The effect of this scale is to penalise dishonest reports of degrees of confidence severely (Ebel 1968 p.353).

The whole process appears to be very complex and costly (20 cents per examinee for the kit alone in 1968). There is also rather incomplete evidence for increases in

reliability and validity mainly because of a lack of control and small sample size, and with marking time twice as long as conventional marking few experiments other than those conducted by Shuford et al have been reported on. Shuford has stated that people who take tests must be trained in the use of the Scorule because of the complexity of the task when met for the first time; but after they have been trained they can use it reasonably quickly.

Rippey (1968 p.211) made use of Shuford's functions in testing in order to test the assertion that increases in reliability as a result of probabilistic administration and scoring could be anticipated. He did not use the Scorule, but had developed a series of computer programs which he used to mark the tests. It is interesting to note that he carried out four tests on different subjects using graduate students, high school freshmen, under-privileged fourth grade children and sophisticated second year medical students. No previous training on how to answer the test was given to any of the groups with the result that the response from the fourth grade pupils was a complete disaster - so poor that he did not bother to tabulate the results. It would appear from this that without careful training only more mature students are able to respond to this type of test. Of the other groups it was only the results of the graduate students that recorded any significant increase in reliability viz 0,79 to 0,85, this being the equivalent of a 50% increase in test length, but even this increase only matched the increased time taken by the candidates to do the test, so that the assertion of increased reliability was thus not proved for this style of probabilistic testing. The candidates did, however, respond favourably to the test.

On Shuford et al's assertion that increased practice in the use of the test would improve the accuracy of the scoring Hansen in 1971 found that there was "a characteristic tendency to be either certain or uncertain which was relatively stable from one exam to the next, and which could not be fully accounted for on the basis of the stability of knowledge. This tendency further appeared to be only slightly related to the knowledge possessed by the subjects. Hansen finally concluded that training with the confidence system did not improve the accuracy of the scoring system." (Echternacht 1972 p.231).

Further probabilistic forms of confidence testing were developed which were easier to administer and to score. These are summarised in table 6 and the results of reliability and validity coefficients carried out on these tests are given in table 7.

The increase in reliability is greatest in Pugh and Brunza's research. It should be noted that the subjects in this empirical study were very mature having an average age of 25,5 years. It would seem as if Rippey's results combined with these would indicate that using mature subjects leads to more effective testing when the personal probabilistic approach is used. In addition one must not view reliability as the most important feature of a test because it alone is of little value in describing the effectiveness of a test. "The reason for this....is that a large reliability coefficient can be obtained by administering the test to a sufficiently heterogeneous group of examinees." (Echternacht 1972 p.233).

What is important is to ascertain whether a more complicated response and scoring procedure provides

Formulator of test	Response Method	Scoring Method
Michael (1968)	Allocate 10 points to various alternatives	Score = proportion of points allocated to correct answer
Hambleton, Roberts and Traub (1970)	As for Michael	Similar to Shuford and Massengill
Boldt (Pick-one confidence)	Choose alternative believed to be correct. Indicate sureness on 5-point scale	Scoring uses a table
Krauft and Beggs (1973)	Assign 4 points among 4 alternatives to correspond to belief in correctness of alternatives	Number of points allocated to the correct alternative
Pugh and Brunze (1975)	Assign a number from 0 to 5 to each of the 5 options	Score = number allocated to correct answer

Table 6

Personal Probabilistic Confidence Tests using simplified response methods

significantly greater information than a simple conventional procedure does; or whether there is contamination of the scores as a result of a personality factor. Whether probabilistic methods provide more information than can be ascertained from other simpler methods is questionable - it certainly does not provide more

Formulator of Test	Reliability	Validity
Michael	Conventional 0,764 Probabilistic 0,84	No information
Hambleton et al	Conventional 0,71 } insignifi- Probabilistic 0,66 } cantly down	Conventional 0,62 } insignifi- Probabilistic 0,72 } cantly up
Boldt	no information	no information
Krauft and Beggs	no statistical difference	no information
Pugh and Brumza	Conventional 0,57 } signifi- Probabilistic 0,85 } cantly up	Regarded as improved because: (a) increased reliability; (b) no increase in relative difficulty of items; (c) no significant personality bias.

Table 7

Reliability and Validity Coefficients of Certain Probablistic tests

information on partial knowledge than Coombs et al's method, but further consideration will be given to this in the next chapter.

Other aspects that must be considered are firstly Ahlgren's assertion that "those students whose confidence was more appropriately founded retained the content of the course better in the long run" (Rothman 1969 p.237); and secondly whether confidence testing can be applied in some form to a more traditional style of examination. These aspects will be considered in the last two chapters.

Chapter 4

Evaluation of Confidence Testing

One of the reasons for the introduction of confidence testing was the desire to reduce the influence of guessing in multiple-choice examinations. One popular way to do this was to make use of a correction for chance formula, but this type of formula ignores a number of very important considerations. Firstly those students who choose an incorrect response as a result of misinformation are penalised twice. Secondly poor guessers are penalised more than good guessers, and thirdly an assumption of the correction for chance formula is that the answer is either known or it is not known - partial knowledge is not considered (Marshall and Hales 1971 p.225). It is generally held that confidence-weighted scoring definitely reduces guessing depending on the scale used, because of the negative marking that takes place. Against this Krauft and Beggs report that their examinees guessed more on the confidence test than on the conventional test, but this appears to be an isolated instance (1973 p.75).

A more vexed problem is whether subjects respond to training in realism of confidence judgements. Adams and Adams in 1961 offered some evidence that this did occur especially with intentional learners. Ebel, however, felt that personality variables or response styles could contaminate the student scores on confidence-weighted tests (1965a p.51) and Jacobs added his voice to this claim (Echternacht 1972 p.220). Opposing viewpoints are very evident, and it therefore behoves us to consider the views of authors who have studied

Risk Taking per se.

Kogan and Wallach in their book published in 1964 state: "It should be emphasised that confidence is evidently not a 'strategy' variable in either an explicit or implicit sense. Rather, it is an index of a subject's introspective conviction regarding the correctness or appropriateness of his judgement or decision....the variable has salient personality overtones." (p.125).

This raises the question as to whether one's attitude can be changed towards confidence in judgement with training, seeing that confidence is an index of one's introspective conviction. Hansen (1971) noted that there was a characteristic tendency to be certain or uncertain which was relatively stable from one examination to the next, and it could not be accounted for on the basis of stability of knowledge. He therefore definitely concluded that training with a confidence system did not improve the accuracy of the scoring system (Echternacht 1972 p.230). Koehler in 1974 added to this when he stated that confidence response methods produce variability in scores that cannot be attributed to knowledge of subject matter (Wood 1977 p 235). This all tends to confirm an early experiment by Swineford in 1938 who identified a personality variable. She derived a "gambling" score which was sufficiently reliable and yet independent of the right-wrong achievement score in the true-false test she was undertaking. She concluded with subsequent tests that boys tended to gamble more frequently than girls, and that both had a greater tendency to gamble more on unfamiliar material than on familiar (Echternacht 1972 p.219). As stated before, Shuford does not accept that practice cannot change the way students treat confidence weighting - he holds that it will remove undesirable personality effects. In summing-up, Echternacht

feels that an open mind should be exercised in this matter. The whole question is most involved because Krauft and Beggs (1973 p.76) state that the findings of their investigation "tentatively suggest that there is no relationship between subject-weighted test taking procedures and risk taking as measured by the Kogan and Wallach instrument. Furthermore risk taking may be unrelated to obtained score on a statistics achievement test." However, they do acknowledge that more work needs to be carried out before a final decision can be made (if ever).

The next query is whether it is psychometrically worthwhile to switch to confidence-weighting? Koehler (Wood 1977 p.235) states that conventional testing is preferable because it is easier to administer, takes less testing time and does not require the training of the candidates. Hanna and Owens in 1973 (Wood p.235) concluded that greater validity would have been obtained by using the available time to lengthen the Multiple-Choice test rather than to confidence-mark the items. Krauft and Beggs comment that "it appears as if the particular weighting procedure employed in this study did not encourage examinees to respond any differently than they would have responded on a conventional multiple-choice test, thereby, it added no additional measurement information for the practitioner" (1973 p.76). This conclusion goes directly against one of de Finetti's original reasons for introducing personal probabilistic methods. He wished to ascertain partial information something which cannot be found using the conventional multiple-choice test:

"If partial information exists, there can surely be no doubt that it is important that it should be revealed. It is interesting to detect and measure it, because it is a component of the mental processes being studied. The importance of this component per se is also high, both

Chapter 5
Confidence Testing as an Aid to the Learning Process

In this chapter the sources of reference are exclusively Sieber et al 1978 and Sieber 1979.

"Despite the practical and theoretical interest in confidence estimation as a component of problem solving, there has been little research to examine its effectiveness. Among the questions that need to be answered are the following:
Is performance improved when problem solvers are required to make confidence rating on their solutions?
What problem conditions and individual variables affect the accuracy of confidence estimates and the bias of estimates towards over- and under- confidence?
How are over- and under-confidence related to problem-solving performance?
Does confidence estimation lead to error diagnosis and subsequent acquisition of relevant information?" (1979 p.273).

Sieber in her 1979 study tested 4th, 5th and 6th grade students. This is in strong contrast to other tests using confidence techniques as they have generally concentrated on university or high school students with emphasis on the former. Indeed on one of the two occasions when younger children were used Rippey reported that the whole process was disastrous. If one's aim is to see whether confidence estimation creates improvement in diagnostic and learning skills, then this should be applied to subjects at as an early age as possible. In Sieber's 1978 study the relation between the age at which warranted - uncertainty skills were generated and the effect on the retention of those skills was considered.

It was found that the ability to generate warranted uncertainty was retained best by students who had learned it earliest (in the fourth grade in this case). It was also learnt that there are "stable individual differences in the ability of elementary school students to diagnose when they have given a poor answer (and hence have grounds for uncertainty) in solving math, spelling and logic problems. The ability to generate warranted uncertainty, in turn, was related to student's ability to improve their performance subsequently and overall level of ability. Thus it seems to be an indication of intelligent openness to problems." (p.263). It is also interesting to note that once the intellectual skill of warranted uncertainty had been learnt it was retained and applied to new content areas. Thus it would seem to be important that elementary school children should be the subjects of any investigation which concerns improvement in knowledge as a result of confidence testing.

Approximately 1 200 students were included in the testing program, which consisted of a Verbal Skill test, two spelling tests, a mathematics test, a logic test and a Certainty Estimation Test. The students were divided into three groups. The first group consisted of students who used a confidence rating procedure in connection with the spelling tests, and in addition their teachers attended a four-week workshop where they received instruction in methods of teaching students to generate warranted uncertainty. In the second group the students were taught to use a confidence scoring procedure in connection with the spelling tests. The third group acted as a control group receiving neither of the treatments given to the first two groups.

Four main conclusions were drawn from the experiment:

1. Giving confidence estimates with one's responses

produced significantly more correct responses then responding without giving confidence estimates, but only in the case of those students who had been trained to value warranted uncertainty and to encourage its development in themselves. It must be noted here that the children were not trained to answer the test, they were encouraged to estimate their ability in all spheres, and to value their judgement in the estimate of their ability. This was a generalised program; this was not the practice in testing that Shuford had in mind, but something more encompassing, and this more encompassing program had the effect of significantly increasing the number of correct responses.

2. Students whose teachers had been trained to value warranted uncertainty and to nurture its development in students gave more accurate confidence estimates than did other students, and erred more in the direction of underconfidence.
3. Greater accuracy and underconfidence were also characteristic of girls. This accords with Swineford's "gambling"score results measured earlier. It is also reflected in Kogan and Wallach's conclusion that males are prepared to take more risks than females (1964).
4. Those students who estimated their confidence accurately or were underconfident in their estimation tended to give more accurate answers to problems. This is an important result in that it shows that confidence estimation serves as a component of rational decision-making. It seems that when one can estimate one's confidence in handling a task then problem-solving ability is increased, but only if introduced in a setting where warranted uncertainty is generally valued and nutured. When this is combined with the findings of the 1978 study, that firstly the ability to recognize when it is warranted

> to be uncertain and secondly that one's bias in confidence estimation are both seemingly fairly stable characteristics over time and over various types of problem material, then one has the situation where more effective teaching can be nutured.

It is evident from the foregoing that teacher training in the ability to encourage warranted uncertainty would be most helpful, so that the approach can be integrated maximally into classroom teaching and in this way improve the child's ability to solve problems. Anything that can help with this very important aspect deserves, perhaps even demands, further consideration and investigation.

Chapter 6

Confidence-Weighted Scoring of a Mathematics test of a Traditional Type

The question now arises as to whether Confidence Testing can be usefully employed in the conventional testing of mathematics at the secondary level. A number of empirical studies have made use of university students, but very few have been conducted at the high school level. In addition few studies have used Mathematics, although a number have used Statistics as the test subject matter. Confidence testing has also been limited to true-false and multiple-choice questions. In South Africa questions of these types are not normally set in mathematics papers, and, indeed, as far as the Senior Certificate is concerned, their use is discouraged, as it has often been found that they waste examination time, especially when questions are set that require the candidates to think carefully about their answers, although this waste of time may result from the questions being included with more traditional questions in the same paper.

Because of the nature of the answers to Mathematics questions at school level, i.e. because there is a definite solution or proof to each question, even though the actual solution or proof may vary from candidate to candidate, the question arises as to whether a declaration of confidence in the solution to the problem asked would be of benefit and use. The questions that need to be asked are the following:

1. Is there an increase in reliability as a result of confidence marking?
2. Is there increased correlation between the various sets of marks as a result of confidence- weighting?

3. Is grading made easier because of a significant increase in the spread of the marks?
4. What type of mistake is penalised by confidence-weighting and to what extent is the mistake penalised?
5. What effect does confidence-weighting have on the memorandum of marking?
6. Will the type of error be reflected in the confidence level chosen by the candidates?

In order to investigate these questions a short pilot study was undertaken. In this study a short Mathematics Higher Grade test was given to 38 students in Standard Ten at a high school in Cape Town. It must be emphasised that this study cannot be regarded as definitive by any means, because the test involved was short, and also concerned itself with only one aspect of the Mathematics syllabus in Standard Ten ie. Sequences and Series.

The test was given on the Friday before the pupils began their midyear examinations. The pupils were informed about the test nine days beforehand. The justification for this pre-examination test was that the pupils had not written any tests on Sequences and Series, and would have to answer examination questions on this aspect of the work without having any experience in answering questions on it, which would be to their detriment in the examination. The reason for the author's setting of the test ie. that he was investigating confidence-weighting marking procedures was then explained to the pupils. The way in which the test was to be set and also the way in which it was to be marked were then discussed making use of the information sheet that appears in Appendix A. The pupils were also informed that they were to complete a homework exercise which

illustrated the format of the test paper. They were to mark this exercise (Appendix B) themselves and a responsible pupil was appointed to check that the exercise had been completed. The pupils were also informed that if they omitted to answer a question they would receive nil for it. Time was allowed for questions after the explanation. The pupils all did the homework exercise and were thus considered to be familiar with the method of answering the paper. From the way in which the pupils answered their papers it was evident that this conclusion was valid.

The reason for including multiple-choice questions in the test paper was to provide some contact with the way in which confidence weighting of responses has been used in published empirical research, and also despite the few questions actually involved, to provide some means of validating the traditional type of question against those of multiple-choice if both aspects of the test were found to be reliable. It would probably have been better to validate the test by comparing the scores against another criterion.

The test was administered during school hours in a class teacher's period. The instructions as in Appendix A were again given to the pupils at the beginning of the test. The test itself is given in Appendix C. The pupils found that they had sufficient time in which to answer it - no one was unable to finish for lack of time. The tests were marked according to the memorandum given in Appendix D. When the tests were marked each was photocopied, the original was given back to the pupil, and the photocopy was used for statistical purposes. This was unfortunate as two scripts were so faint that they could not be used for the error analysis.

An item analysis (Appendix E) was carried out on the multiple-choice test, and as the P-values on questions 2(a) and 2(b) were both higher than 0,90 it was decided to reject these two items, and thus in the statistical tests the other seven items only were used. It was also decided not to include question one in any of the statistics as too few short questions were asked.

Effect of Confidence-Weighting on the Memorandum of the Traditional Questions

In setting a traditional test where the pupils have to show how they obtained their answers most of the marks are given for the actual working-out of the problem, and few are given for the answer alone. In this connection it would, therefore, defeat the basic purpose of giving credit where it is due to award -6 if the answer is wrong, when a total of 6 marks is awarded for a fully correct answer. This would magnify the effect of one mistake and would also not differentiate between a candidate who makes one mistake compared to one who makes two or even three errors in his work. In multiple-choice tests the amount of calculation or reasoning is limited because each question generally tests one aspect only of the work. It was, therefore, considered reasonable to make the confidence-weighting affect the answer only, and this to a maximum of 2 marks.

On examining the memorandum (Appendix D) it will be noted that each answer has been awarded two marks so that the answer would be marked on the following scale:

Correct and sure	+2
Correct but not sure	+1
Answer omitted	0
Incorrect and not sure	-1
Incorrect but sure	-2

To some people this places undue emphasis on the answer. In question four for example two out of the four marks are given for adding three numbers together while in question five, two marks out of six are given for the simple addition of 1 to both sides of an equation. In the above two cases a half and a third of the marks respectively are allocated to the answer, whereas one would normally only award one mark to each of these steps. In questions six and seven where the answers depend on understanding what sort of result is acceptable, it is more usual to award two marks, and thus confidence-weighting has no effect on the distribution of marks as such. The first deleterious effect, therefore, that weighting of the answer has is that sometimes too much credit is given to an answer. A second possibly unreasonable effect is that instead of losing one mark for a careless error, because of positive marking after the error has been made, a candidate will lose three marks - one for the error and two for the answer being wrong. This will be discussed at greater length when the error analysis is considered.

Effect of Confidence-Weighting on Scores

In the Multiple-Choice test the expected effects of the lowering of the mean and an increase in the standard deviation are evident in the statistics given in Appendix F. As shown by the top graph in Appendix G the usual increase in the spread of marks is evident and this allows for more refined grading.

The effects on the traditional test are, however, different. There is firstly a smaller decrease in the mean when the test is marked by means of confidence-weighting although the decrease is still significant, because the marks either decrease or remain equal to the original mark, and this consistent lowering is the main reason

for the significant difference noted. The spread of marks as indicated both by the smaller difference in the standard deviation between the two sets of data and also by the lower graph in Appendix G is nowhere near as wide, and this means that the advantage of being able to use confidence-weighting for more refined grading is lost. The spread of marks is now from 3 to 26 instead of 5 to 26 which is hardly any difference at all. Indeed, whereas in the conventional marking there was only one four-way tie on a particular mark, in the confidence-weighted marking there were two four-way ties, which is hardly conducive to more refined grading.

The main reason for the lack of spread in the scores is probably the limited influence that the confidence-weighting has on the total score when only two marks out of a possible six or eight are involved. This is in strong contrast to multiple-choice confidence testing on a 2; 1; 0; -1; -2 scale where the difference between a correct and wrong answer at the sure level is double the mark that can be obtained. This must lead to a greater spread in marks. It is also true that with 58% of the solutions being completely correct as in this case, confidence-weighting of answers will not play as large a role as when there is a lower mean score.

Reliability of the Tests (Appendix H)

None of the tests can be regarded as being reliable despite Guilford's comment (p.388 1954):
"As to how high reliability coefficients should be, no hard-and-fast rules can be stated. For research purposes one can tolerate much lower reliabilities than one can for practical purposes of diagnosis and prediction. We are frequently faced with the choice of making the best of what reliability we can get,

even though it may be of the order of only 0,50, or of going without the use of the test at all. For some purposes even a test of low reliability adds enough to prediction to justify its use particularly when used in a battery along with other tests." The highest reliability of 0,45 is lower than even Guilford's generous low acceptance level. If question four is rejected because of its low discriminating power then the reliability for the conventionally marked traditional test becomes 0,41 while that for confidence-weighted scoring of that test is 0,59. These higher reliabilities are probably a reflection of the greater heterogeneity in the marks because question four was omitted. Whitta, however, states that acceptable reliabilities are those in the 70s or low 80s for most purposes that involve using summaries of test scores as information about groups (1968 p.272); the obtained reliabilities fall far short of this standard.

The main reason for the low reliability coefficients is no doubt the shortness of the test. Confidence-weighted scoring did, however, increase the reliability in both cases with the increase being very dramatic on the multiple-choice test probably because of the increased spread in the scores. Because of these reliability results it will not be wise to accept unconditionally the significance or otherwise of the correlation coefficients calculated.

The Confidence Traditional scores have lower correlation coefficients with the two multiple-choice tests than do the Conventional Traditional scores, whereas the Confidence Multiple-choice scores indicate a higher correlation with the traditional scores than the Conventional Multiple-choice scores do. The question that arises from this is whether the same aspect of confidence is being tested in the traditional as opposed to the

multiple-choice tests or are the differences due to the unreliability of the tests in question? Thus further investigation is absolutely necessary before any conclusions can be drawn.

Effect of Confidence-Weighting on the Mark Obtained in the Traditional Test.

Because there is a fully written solution to each question it is possible to study whether weighting affects different questions in different ways by means of error analysis. Before this is considered, however, attention must be drawn to the effects of confidence-weighting on the score of each question separately, and also to a possible relationship between the choice of confidence level and the mark obtained.

It was decided to ignore data from question four in this section of the study as the mean was 98,7% showing that the question did not discriminate between poor and good students. The data that was used in the various statistical tests is given in Appendix I. Using a Chi-Squared test it was found that there was no significant difference between questions as far as marks being affected by confidence-weighting were concerned. Nor was any significant difference found between the effects of the choice of confidence level on the scores obtained for each question (in this case Fischer's Exact Probabil - ity test was used). As in all this work it would be advisable to draw definite conclusions only after further investigation has taken place, and after pupils have been made familiar with the weighting process over a period of time.

In Appendix J an analysis is made to ascertain whether the questions differ in their effect on the examinees obtaining full, part or zero marks. The distribution of these categories of marks is significantly different

between questions 6 and 7 as far as full and zero marks are concerned. There is no significant difference between any other conbinations. The difference would seem to be the result of the ease with which the rxaminees were able to answer question 6. It is also notable that question 7 would appear to have been that type of question in which it was difficult to obtain part marks. It would, however, appear to be true that as long as the questions discriminate between poor and good students then there is little consistent effect of confidence-weighting on differences in the scores obtained between questions.

In Appendix K extreme scores (i.e. either full marks or a zero mark) were contrasted with part marks. Using a chi-squared test it was found that part marks were more affected by confidence-weighting than extreme marks. Thus it would indicate that where there is partial knowledge the candidate's mark is negatively affected by confidence-weighting whereas full knowledge or full misinformation is generally not negatively affected by such weighting. This result was highly significant.

In addition tests were conducted to see if the results obtained by upper, middle and lower groups arranged in order of merit were differentially affected by confidence-weighting. The conclusion drawn was that this was not the case.

Discussion on Error Analysis

It is important to note that in Appendix L, and hence in the data used in the statistical tests in Appendix M, only those errors are considered which resulted in a change of score when confidence-weighting was used.

This was to consider the effects of confidence-weighting on the scores as a result of careless errors as opposed to errors of misunderstanding. It was felt that all errors could be divided into these two basic types. Where the decision was not clear-cut the author used his own judgement to decide into which category the error should be placed. Another question of import was whether candidates would attach a different confidence level to an answer following a mistake as a result of misunderstanding compared with a wrong answer as a result of carelessness.

One of the problems concerning confidence-weighting is that the theoretical effect on a score of a minor careless error can be the same or higher than the effect of a mistake as a result of misunderstanding. For example if a question asked for the number of terms required in a particular Arithmetic Series given the sum to n terms and the resulting equation was correctly simplified to:

$n^2 - 24n - 640 = 0$,

which the candidate wrote down as

$(n - 16)(n + 40) = 0$ instead of $(n + 16)(n - 40)=0$,

because of careless factorisation and so obtained answers of

$n = 16$ or $n = -40$ instead of $n = -16$ or $n = 40$;

if he further proceeded to reject correctly one of the answers he would be penalised by 3 marks in the case of confidence-weighting instead of the more normal 1 mark. This despite the fact that there was no cause for him to reject the answer as unreasonable. Confidence-weighting under such circumstances must be regarded as a suspect means for accurately assessing a student's knowledge of mathematics. In this case a careless error would result in a greater loss of marks than if a candidate had worked correctly, and not rejected the negative answer, which is an error involving misunderstanding of the meaning of what is being asked.

Thus the vexed question arises as to how one should allocate marks for confidence-weighting in a traditional style mathematics paper? In view of the argument above is it correct to mark the answers only by means of confidence or should there be differential weighting depending on the type of error? This would introduce a high measure of subjectivity into the marking process especially when a careless error leads to an answer that is mathematically unacceptable. Marking in mathematics should be as objective as possible. This will also result in an increase in marking time, and any such increase is justifiable only if there is a corresponding increase in information concerning the student's ability. The only increase in information being obtained from confidence-weighting in the traditional style examination is in the sphere of the candidate's awareness of his own ability. Unlike multiple-choice testing all other information can be gleaned from an analysis of what has been written down. To what purpose does one increase a candidate's awareness of his own ability? Is it to make him more critical of himself and thus to analyse where he goes wrong? If this is so then confidence-weighting must be used in a formative evaluation setting - it has limited use in summative evaluation. The kernel question is, therefore, where does one place the stress in confidence-weighting? On the answer, as in this test, or on the type of error; or does it really matter? Much research is necesaary to ascertain whether it does matter and from Sieber's findings it looks as if the research will be worthwhile.

What can be considered immediately is whether careless errors are measurably affected by confidence-weighting. In table 8 the number of errors which resulted in a change of score through confidence-weighting and the total number of errors made are given. The difference in the

number was caused by certain students being unable to finish the question and thus were unable to give an answer to which confidence-weighting could be applied. It can be seen that there is virtually no difference between the ratio of careless errors to errors as a result of misunderstanding as far as those errors that influence confidence are concerned and the total number of errors. It is also noteworthy that there is a high percentage of careless errors, errors which under normal circumstances would only merit the loss of one mark (unless they resulted in a situation in which the answer was obviously wrong when matters of understanding are introduced), but which under a procedure of confidence-weighting carry a higher penalty.

Type of error	Number of errors having effect on confidence-weighted scores	Number of errors having effect on all scores
Careless	18	22
Misunderstanding	21	27

Table 8

Analysis of the Effect of Types of Errors on Scores in a traditional Mathematics Test

An analysis of the data given in Appendix L shows that the total difference in score between conventional and confidence-weighted scoring as a result of careless errors is virtually the same as that for errors in misunderstanding, which shows that confidence-weighting does not appear to differentiate between the two types of error. Further to this in Appendix M the errors due to carelessness and those due to misunderstanding which influence the confidence-weighted score are compared between the upper, middle and lower groups arranged according to their conventional scores. With only three errors of each type in the Upper 12 it is not surprising that no significant difference results between that group and the other two groups occurred. However, there is a highly

significant difference between the middle and lower groups results which seems to indicate that carelessness affects the middle group to a greater extent, while misunderstanding as a cause of error is a characteristic of the lower group. When it is further realised that omitting to reject an answer could be as a result of carelessness, then the recorded difference could have been even greater. It is, however, necessary to point out again at this stage that this test is by no means definitive, and before definite conclusions may be drawn much additional research needs to be undertaken.

Comparing the various groups again, but this time considering whether there is any difference between the Middle and Lower groups in choice of confidence level when there is cause to be uncertain ie. when a mistake has been made, it is noteworthy that there is no significant difference as indicated in the second set of statistical tests in Appendix M. However, when the errors due to carelessness are isolated from the errors due to misunderstanding and each group is considered separately the result is different with the errors due to misunderstanding. Here there is a significant difference in the choice of confidence level with the lower group of 12 choosing confidence level 1 in problems where errors in misunderstanding occurred, significantly more frequently than the middle group. This again calls to mind Ebel's wry comment (1965a p.56) "the results of recent experimental studies suggest that sometimes the more capable students are not much more successful than their less capable classmates in deciding when to answer confidently and when to answer cautiously." In mitigation it should be noted that three of the misunderstanding errors in the middle group were again those of non-rejection of an answer. Thus again no definite conclusion can be drawn without further investigation.

Lastly it is essential to look within each grouping to see whether there is any difference between the type of mistake made and the confidence level chosen. The results are given as the third set of tests in Appendix M. Firstly it is shown that there is no significant overall difference between levels of confidence chosen and type of mistake made, but when the lower group is considered then there is a definite difference. It would seem that when a question is misunderstood then that error is recognised by the candidate and in consequence he is not confident of his answer. There is, however, often no such realisation when the mistake is careless. Again it must be emphasised that this investigation is only raising issues because the evaluation exercise set was of such short duration. The evaluation also resulted in an average of 68,5% before confidence was taken into account, and this could be the reason why the middle group has not used confidence level 1 much - there were few occasions when complete misunderstanding occurred. Indeed the middle group tended to use confidence level 1 for careless errors, these being mainly in question 5 where a careless error can easily lead to a situation where the answer is hopelessly wrong.

Thus results on this short traditional test are tentative, but the test has opened up a possible new line of investigation as far as mathematics and related subjects are concerned. One can investigate the types of error that cause the loss of marks as a result of confidence-weighting; something which cannot be done by confidence-weighted multiple-choice questions. By and large multiple-choice questions cannot be set to test careless errors especially if the question tests only one process, because carelessness generally arises in a calculation when the calculation is part of an overall

presentation of a problem. When questions are asked on each part of a problem and the steps are highlighted, then errors due to carelessness are minimised. If more than one process is tested in a multiple-choice question, then one can certainly not ascertain which type of error causes a candidate to be incorrect no matter how one chooses the distractors. Thus one's knowledge of the candidate's ability is reduced.

For all its shortcomings the confidence-weighted marking of this traditional test has highlighted two aspects which need further investigation. The first is that confidence-weighting does not affect complete information nor complete misinformation to a negative degree. Its main negative effect is concentrated on candidates who possess partial knowledge or whose performances are affected by partial misinformation. A question mark must hang over a scoring procedure which penalises partial knowledge more heavily than complete misinformation. There are spheres in which the correct answer only must be rewarded, but there are far more spheres of learning in which partial knowledge is an asset. Knowing a pupil's partial knowledge can also help a teacher to decide how best he can help that pupil turn that partial knowledge into full understanding.

The second aspect that has been highlighted, and this needs much further investigation, is the type of error that confidence-weighting penalises. One must decide whether a method that penalises a minor careless error to the same extent as an error of misunderstanding should be continued with. One last comment on this, however, is necessary. When a pupil who wrote this test was asked by the author what he thought about giving confidence ratings he replied, "It made me check my answer." Confidence-weighting can introduce a more critical attitude to one's own work - a point that

Conclusion

Confidence scoring methods have been used now for fifty years, but there is still no unanimity as to their effectiveness, even though many researchers report that the examinees prefer scoring methods based on some type of confidence scoring eg Coombs (1956), Hevner (1932), while others thought that they were fairer even if they preferred the conventional test (Krauft and Beggs 1973 p.76). This lack of unanimity would seem to indicate that their usefulness is limited, but what must be asked is whether the tests have been used properly, and whether the additional information supplied has been found to be useful in the teacher-learner sphere.

When the testing situation alone is considered it would seem as if priorities must be sorted out before any form of confidence-testing is considered. The scale chosen in confidence-weighting of answers should depend on what one wishes to gain from the student. If, as in medical testing, one is concerned that the student should only gain marks if he is certain of an answer then a scale such as Fredman uses should be used. If the major concern is ascertaining how much knowledge a student has then Paton's scale is more useful especially if the students are informed about the advantages of the use of different strategies depending on the state of their knowledge. From the point of view of grading Fredman's scale would appear to be the best, but then as he points out a grading test must be considered to be only a part of the entire testing program.

From the point of view of ascertaining partial knowledge

Paton's system of confidence-weighting only informs one of how much partial knowledge a student has and gives no clue as to the areas in which he is lacking. Here Coombs method of marking distractors would appear to give the greatest amount of information. The other advantage of Coombs et al's method over the probabilistic methods is that it is far simpler to administer, and requires no subjective judgement on the part of the student. It is surprising that others have not used his method to a greater extent - perhaps because it takes longer to mark and teachers do not make use of the additional information obtained. Probabilistic methods in the opinion of the author create too much uncertainty in the mind of the candidate, and would certainly be too complicated for use in schools especially in the lower standards. There is enough stress in an examination as it is, without adding to the uncertainty by forcing the candidate to make up his own mind as to how he rates the different distractors.

Should confidence testing be extended to more traditional styles of questioning? This really depends on further investigation and also on how much useful information can be ascertained from the process especially as to how much light the process will throw on the effects that testing for confidence has.

No matter what method of confidence testing is used training in its use is absolutely essential - all too often testing has been carried out without the candidates having been thoroughly prepared for its use. Training in testing also has a side-effect on teaching, and if the decision of how much confidence a student must express in his answer helps firstly in the understanding of the work and secondly in its retention as Sieber suggests, then much that can be of benefit will be

achieved. The ultimate test of any examining method within the school is whether it adds to the knowledge gained and retained by the child. The results of further research on this line would thus be of great interest and importance.

Appendix A

Instructions for Preliminary Homework Exercise for Test on Sequences and Series

The test will be divided into two parts A and B. In part A the questions will be short answer questions or multiple-choice. In part B the questions are of the kind that are usually asked.

In both parts A and B you will be required firstly to enter all your answers in the space shown, and secondly to say how confiden you are of your answers by writing either "1" or "2" next to each answer. Write "1" if you are guessing at the answer or if you are not quite sure of it. Write "2" if you are confident of your answer.

Each short or multiple-choice question will be worth 2 marks so a pupil who writes every answer with a confidence rating of 2, and who gets every answer correct will score +20 marks for 10 questions. However, if a pupil marks every answer with a confidence rating of 2 and gets every answer wrong he will score -20 marks.

Eg 1

Solve $\frac{x}{3} - 15 = 0$

	Question Number	Answer	Confidence	
If you wrote (a)	1	45	2	you would be awarded +2 marks
(b)	1	45	1	you would be awarded +1 mark
(c)	1	5	2	you would be awarded -2 marks
(d)	1	5	1	you would be awarded -1 mark

In each long question the confidence with which the pupil treats his answer will be limited to the range -2 to +2 as above, even if the number of marks allocated to the question is 10.

Process / Thought Process / Content	Knowledge	Comprehension	Application	Analysis and Synthesis	Totals	Year
	Recall, Recognition	Algorithmic Thinking and Generalisation		Open Search		
Relations	6	32	10	11	59	1979
Functions and	3	34	6	5	48	1980
Graphs	-	38	26	-	64	1981
Equations	6	13	20	-	39	1979
and	6	26	-	17	49	1980
Inequalities	-	6	30	7	43	1981
Indices	9	14	9	16	48	1979
Logarithms	8	42	5	-	55	1980
and Surds	8	30	11	5	54	1981
Sequences	2	10	15	-	27	1979
and	-	8	14	-	22	1980
Series	-	-	9	15	24	1981
Mathematical	-	8	2	5	15	1979

Appendix C

Test on Sequences and Series

Part A (24 marks)	Answer	Confidence
1(a) If $Tn = 3n^2$, calculate T2	(a)	
(b) Calculate the next term in the following sequences:		
(i) 11; 8; 5; 2	(b)(i)	
(ii) 27; 9; 3; 1	(ii)	
2 Select the best response ie A, B, C or D and <u>circle</u> it in the answer column. For each of the following series decide whether it is: A. An Arithmetic Series B. A Geometric Series C. Another type of series D. No series at all		
(a) $\frac{1}{2} - 1 + 2 - 4 + ...$	A B C D	
(b) $6 + 3 + 0 - 3 - .....$	A B C D	
(c) $1 + 4 + 9 + 16 + ...$	A B C D	
(d) $(2a - b) + (a) + (b) ..$	A B C D	
(e) $322 - 3 + 86 + 1064 - 21-$	A B C D	
(f) $3^5 + 3^6 + 3^7 + 3^8 + ..$	A B C D	
3. Circle the best response in the answer column if: A is $Tn = a + (n - 1)d$ B is $S_n = \frac{n}{2}[2a + (n - 1)d]$ C is $Tn = ar^{n-1}$ D is $S_n = \frac{a(1 - r^n)}{1 - r}$ E is none of these		
(a) You are asked to calculate a man's salary in his tenth year of work if his starting salary was R5000 p.a. and his yearly increase R500 p.a. Which of the above should be used to find the correct amount?	A B C D E	
(b) If the sum to 12 terms of a Geometric Series is 4095, and the first term is 1 which of the above would you use to find the common ratio?	A B C D E	

(c) Which of the above would you use to calculate: 1 + 3 + 6 + 10 + ...to 15 terms

Answer	Confidence
A B C D E	

Part B (26 marks)

Do your working in the space provided.

4. Calculate $\sum_{n=3}^{5} (n^2 - 5)$

(4)

Answer	Confidence

5. Find the number of terms in the following sequence $\frac{1}{9}; \frac{1}{3}; 1 \ldots\ldots; 729$

(6)

Answer	Confidence

6. If the sum of 18 + 13 + 8 + ... is 21, find the number of terms

(8)

Answer	Confidence

7. If $(3x - 3);(2x + 2)$ and $2x + 8$ form a Geometric Sequence, calculate x

(8)

Answer	Confidence

Appendix D

Memorandum of Marking for Test on Sequences and Series

1. (a) 12 (2)

   (b) (i) -1 (ii) $\frac{1}{3}$ (4)

2. (a) B (b) A (c) C (d) A (e) D (f) B (12)

3. (a) A (b) D (c) E (6)

4. $\sum_{n=3}^{5} (n^2 - 5) = 4 + 11 + 20$ (2)

   $= \underline{35}$ (2) (4)

5. $T_n = ar^{n-1}$

   $729 = (\frac{1}{9})3^{n-1}$ (2)

   $6\,561 = 3^{n-1}$ (1)

   $3^8 = 3^{n-1}$ (1)

   $\underline{9 = n}$ (2) (6)

6. $S_n = \frac{n}{2}\left[2a + (n-1)d\right]$

   $21 = \frac{n}{2}\left[36 - 5(n-1)\right]$ (2)

   $42 = n\,(36 - 5n + 5)$

   $42 = 35n - 5n^2 + 5n$ (1)

   $5n^2 - 41n + 42 = 0$

   $(5n-6)(n-7) = 0$ (2)

   $n = \frac{6}{5}$ or $n = 7$ (1)

   $\underline{n = 7}$ (2) (8)

7. $\frac{2x + 2}{3x - 3} = \frac{2x + 8}{2x + 2}$ (2)

$4x^2 + 8x + 4 = 6x^2 + 18x - 24$ (1)

$0 = 2x^2 + 10x - 28$ (1)

$0 = x^2 + 5x - 14$

$0 = (x + 7)(x - 2)$ (2)

$\underline{x = -7 \text{ or } x = 2}$ (2) (8)

TOTAL 50

Appendix E

Item Analysis in Multiple-Choice Test (N = 38)

Q2(a)

	a	b	c	d
Upper 19	0	19	0	0
Lower 19	0	17	1	0

P = 0,97 Item rejected

Q2(b)

	a	b	c	d
Upper 19	19	0	0	0
Lower 19	19	0	0	0

P = 1,00 Item rejected

Q2(c)

	a	b	c	d
Upper 19	0	0	19	0
Lower 19	2	4	13	0

P = 0,84 Item retained. Check distractor d

Q2(d)

	a	b	c	d
Upper 19	16	1	1	1
Lower 19	1	2	2	14

P = 0,45 Item retained

Q2(e)

	a	b	c	d
Upper 19	0	0	1	18
Lower 19	0	0	4	15

P = 0,87 Item retained, but check distractors a and b

Q2(f)

	a	b	c	d
Upper 19	1	16	2	0
Lower 19	2	10	7	0

P = 0,68 Item retained. Check distractor d

Q3(a)

	a	b	c	d	e
Upper 19	16	3	0	0	0
Lower 19	13	6	0	0	0

P = 0,76 Item retained, but distractors c, d and e must be checked

Q3(b)

	a	b	c	d	e
Upper 19	0	1	1	15	2
Lower 19	0	1	4	13	1

P = 0,74 Item retained but distractor a must be checked

Q3(c)

	a	b	c	d	e
Upper 19	1	1	0	0	17
Lower 19	1	5	1	1	11

P = 0,74 Item retained

Appendix F

Raw Scores obtained from test on Sequences and Series after rejection of certain Multiple-Choice Items

Part of Test / Pupil	Multiple-choice Conventional Marking	Multiple-choice Confidence weighted Marking	Traditional part Conventional Marking	Traditional part Confidence weighted Marking
A	12	9	24	21
B	14	14	10	6
C	12	8	20	18
D	12	8	22	20
E	12	10	23	18
F	12	10	13	10
G	12	10	26	26
H	12	10	21	19
I	12	10	26	26
J	12	10	26	26
K	14	14	18	18
L	12	10	26	26
M	14	14	12	8
N	12	10	26	26
O	14	11	16	15
P	12	8	14	12
Q	14	11	12	12
R	10	6	26	26
S	10	7	25	23
T	10	7	26	26
U	10	6	9	7
V	6	-2	12	12
W	8	2	11	9
X	6	-3	18	15
Y	10	6	7	3
Z	8	3	15	15
AA	8	1	26	26
BB	8	5	25	24
CC	6	0	17	15
DD	12	7	24	21
EE	12	7	16	13
FF	6	-2	14	14
GG	10	4	11	9
HH	6	0	13	12
II	6	-2	5	4
JJ	10	6	13	13
KK	8	1	12	11
LL	6	-1	17	14
ΣX	390	235	677	619

Part of Test / Pupil	Multiple-choice Conventional Marking	Multiple-choice Confidence weighted Marking	Traditional part Conventional Marking	Traditional part Confidence weighted Marking
$\bar{x}$	10,26 = 73,3%	6,18 = 44,2%	17,82 = 68,5%	16,29 = 62,7%
SD	2,68	4,84	6,45	6,99

Difference between the means was assessed by the matched-groups t-method.

a) Conventional and Confidence-weighted marking of Multiple-Choice; and

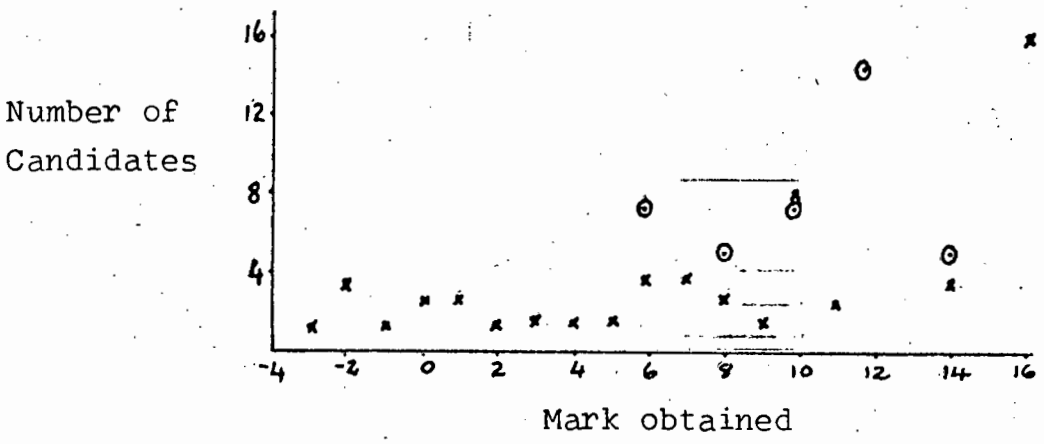
b) Conventional and Confidence-weighted marking of traditional questions;

are both significantly different at the 0,01 level.

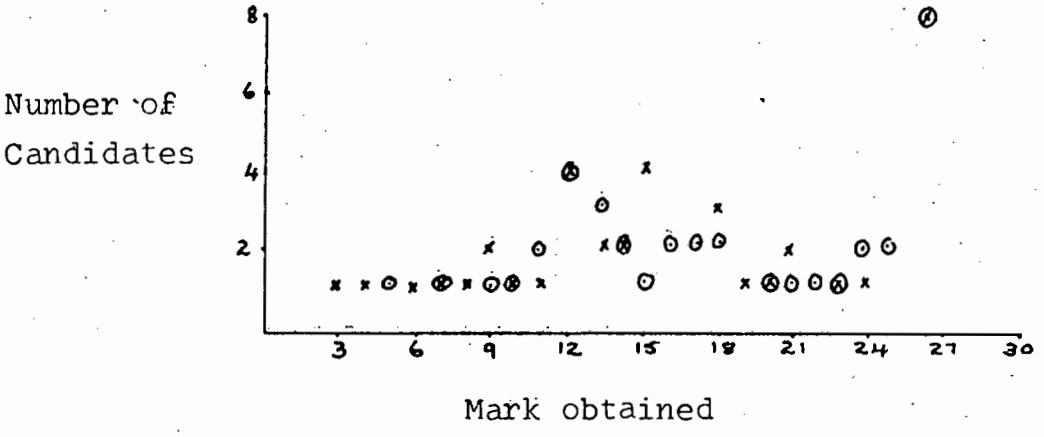
Appendix G

Graphs to Illustrate the spread of scores of Conventional compared with Confidence-Weighted Marking of the Mathematics tests

Multiple-Choice Test



Traditional Test



⊙ Conventional marking

× Confidence marking

Appendix I

Traditional Questions : Analysis of Marks by Question
Effect of Confidence-Weighting

Question	4	5	6	7
Median Score	4	3	8	7
Mean Score	3,95	3,29	5,97	4,63
Mean %	98,7	54,8	74,7	57,9
SD	0,226	2,46	3,03	3,57
Full marks:				
Affected by Confidence Level 1	-	-	2	-
Not affected by confidence marking	36	14	21	14
Part marks:				
Affected by Confidence Level 1	-	5	5	4
Affected by Confidence Level 2	2	10	3	5
Not affected by confidence marking	-	3	4	5
Zero mark:				
Affected by Confidence Level 1	-	1	-	3
Affected by Confidence Level 2	-	1	-	1
Not affected by confidence marking	-	4	3	6
Total	38	38	38	38

Appendix J

Statistical Analysis of Scores Obtained by Pupils for Questions 5, 6 and 7 (N = 38)

	Full Score	Part Score	Zero Score
Question 5	14	18	6
6	23	12	3
7	15	13	10

Ho: There is no difference between questions 5 and 7 as far as full, part or zero scores obtained are concerned.

Chi-squared test used $\chi^2 = 1,078$

This is not significant

Ho is not rejected.

Ho: There is no difference between full, part and zero scores as far as questions 5 and 6 are concerned.

Comparison of full and part scores :

Chi-squared test used $\chi^2 = 2,434$

This is not significant

Comparison of full and zero scores.

Fischer's Exact Probability test used $p = 0,09$

Comparison of part and zero scores

Fischer's Exact Probability test used $p = 0,29$

Ho is not rejected

Ho: There is no difference between questions 6 and 7 as far as full, zero and part scores obtained are concerned.

Fischer's Exact Probability test used

Comparison of (a) full and part scores $p = 0,13$

(b) full and zero scores $p = 0,02$

(c) part and zero scores $p = 0,10$

Ho is rejected for the comparison between the full and zero scores at the 5% level of significance.

It is therefore likely that the questions are significantly different as far as the scoring of full and zero marks is concerned.

Appendix K

Statistical tests on effect of Confidence-Weighting on marks obtained on traditional questions

H_o : There is no significant difference in the effect of Confidence-Weighting on full and zero marks as opposed to part marks.

	Full and Zero	Part	Totals
Score lowered	8	32	40
Score not lowered	62	12	74
	70	44	114

f_o	f_t	f_o-f_t (corrected)	$(f_o-f_t)^2$	$\frac{(f_o-f_t)^2}{f_t}$
8	24,561	-16,061	257,956	10,503
32	15,439	16,061	257,956	16,708
62	45,439	16,061	257,956	5,677
12	28,561	-16,061	257,956	9,032

$\chi^2 = 41,920$

$P = 0,001$

H_o is rejected at the P = 0,001 level

It is therefore highly likely that part marks are lowered more often by Confidence-Weighted Scoring than Full or Zero marks.

Appendix L

Error Analysis of Traditional Questions

Analysis of errors which cause a difference in score between Confidence-Weighted and Conventional Marking. The 39 errors have been divided into two main groups which are further subdivided. The error analysis could not be made on two scripts because the photocopies were too faint. The 36 remaining scripts were arranged in rank order according to conventional marking. Score difference column indicates penalty caused by confidence-weighting.

Group / Errors	Upper 12		Middle 12		Lower 12		Totals	
	Number	Score diff	Number	Score diff	Number	Score diff	Number	Score diff
Careless								
a. Numeracy	2	-3	6	-8	2	-4	10	-15
b. Copying	1	-2	1	-2	1	-2	3	-6
c. Learning	0		4	-8	1	-2	5	10
Sub-totals	3	-5	11	-18	4	-8	18	-31
Misunderstanding								
a. Index Laws	0		1	-2	4	-5	5	-7
b. Meaning of S_n and T_n in AP and GP	0		0		2	-3	2	-3
c. Equations	0		1	-2	0		1	-2

Group / Errors	Upper 12		Middle 12		Lower 12		Totals	
	Number	Score diff	Number	Score diff	Number	Score diff	Number	Score diff
d. Meaning of answer	3	-6	3	-6	2	-3	8	-15
e. Guess or complete misunderstanding	0		0		5	-6	5	-6
Sub-Totals	3	-6	5	-10	13	-1	21	-33
Totals	6	-11	16	-28	17	-25	39	-64

Appendix M

Statistical Tests on Error Analysis Data

Fischer's Exact Probability Test used

First Set of Tests

H_o : There is no difference between Upper, Middle and Lower Groups as far as careless errors and misunderstanding are concerned.

	Upper 12	Middle 12	Middle 12	Lower 12
Careless	3	11	11	4
Misunder-standing	3	5	5	13

	Upper 12	Lower 12
Careless	3	4
Misunder-standing	3	13

p = 0,273 p = 0,011 p = 0,194

H_o is rejected for the Middle and Lower Groups at the 0,01 level.

It thus seems that it is very likely that there is a difference between the Middle Group and Lower Group as far as errors as a result of carelessness, and errors as a result of misunderstanding are concerned.

Second Set of Tests

1. H_o : There is no difference between the Middle and Lower Groups as far as the choice of confidence level by an examinee is concerned when a mistake is made.

	Middle 12	Lower 12
Confidence 1	4	9
Confidence 2	12	8

$$P = \frac{16!17!13!20!}{4!12!9!8!33!}$$

$= 0,077$

H_o is not rejected

2. H_o : There is no difference between groups as far as the choice of Confidence Level 1 and Confidence Level 2 are concerned when the error is due to carelessness or to misunderstanding.

Statistical tests involving the upper group were not considered because there are three items only of each error type.

Errors due to carelessness

	Middle 12	Lower 12
Confidence 1	4	0
Confidence 2	7	4

$$P = \frac{11!4!4!11!}{7!4!4!0!15!}$$

$= 0,242$

Not significant

Errors due to misunderstanding

	Middle 12	Lower 12
	0	9
	5	4

$$P = \frac{5!13!9!9!}{0!5!9!4!18!}$$

$= 0,015$

Significant

H_o is rejected for errors of misunderstanding between Middle and Lower Groups.

It is therefore likely that the middle and lower groups differ as far as choice of confidence level is concerned in errors of misunderstanding, the lower group tending to choose Confidence Level 1.

Third Set of Tests

H_o : There is no difference within groups between examinees' choice of Confidence Level 1 and Confidence Level 2 as far as errors due to carelessness and misunderstanding are concerned.
Upper 12 not considered

All 36 students:

	Confidence 2	Confidence 1
Careless	13	5
Misunderstanding	12	9

$$P = \frac{25!14!18!21!}{13!12!5!9!39!}$$

$= 0,166$

Not significant

Middle 12 students:

	Confidence 2	Confidence 1
Careless	7	4
Misunderstanding	5	0

$$P = \frac{12!4!11!5!}{7!5!4!0!16!}$$

$= 0,181$

Not significant

Lower 12 students:

	Confidence 2	Confidence 1
Careless	4	0
Misunderstanding	4	9

$$P = \frac{8!9!4!13!}{4!4!0!9!17!}$$

$= 0,029$

Significant

Ho is rejected for the Lower Group only.
It is therefore likely that there is a difference between the examinees' choice between Confidence Level 1 and Confidence Level 2 as far as errors due to carelessness as opposed to those due to misunderstanding are concerned in the Lower Group, the difference being in favour of Confidence Level 2.

REFERENCES

Adams, J.K. and P.A. Adams: "Realism of Confidence Judgements". Psychological Review Vol 68 pp 33-45, 1961.

Coombs, C.H., J.E. Mulholland and F.B. Womer: "The Assessment of Partial Knowledge". Educational and Psychological Measurement Vol 16 pp13 - 37, 1956.

de Finetti, B.: "Method for Discriminating Levels of Partial Knowledge concerning a test item". British Journal of Mathematical and Statistical Psychology Vol 18 Part 1 pp 87 - 123, 1965.

Dressel, P.L. and J. Schmid: "Some Modifications of the Multiple-Choice Item". Educational and Psychological Measurement Vol 13 pp 574-595, 1953.

Ebel, R.L.:"Confidence Weighting and Test Reliability." Journal of Educational Measurement Vol 2 pp 49 - 57, 1965a.

Ebel, R.L.:Measuring Educational Achievement. Englewood Cliffs, Prentice-Hall, 1965b.

Ebel, R.L.: "Valid Confidence Testing- Demonstration Unit". Journal of Educational Measurement. Vol 54 No 4 Winter 1968 pp 353 -4

Ebel,R.L.: Essentials of Educational Measurement (3rd ed). Englewood Cliffs, Prentice-Hall, 1979.

Echternacht, G.J.: "The Use of Confidence Testing in Objective Tests". Review of Educational Research, Vol 42 pp 217 - 236, 1972.

Fredman, M.: An Examination of the Objective Evaluation of Student Achievement in Anatomy, with an Enquiry into the Results of Cycling Marking Programs and Confidence Weighting of Responses. Cape Town, University of Cape Town, Ph.D. Dissertation, 1977.

Fredman, F.: "Towards a Rational Approach to Student Evaluation by Examinations". Medical Education, Vol 13 pp 414 - 419, 1979.

Guilford, J.P.: Psychometric Methods. New York, McGraw Hill, 1954.

Hopkins, K.D., A.R. Hakstian and B.R. Hopkins: "Validity and Reliability Consequences of Confidence Weighting". Educational and Psychological Measurement Vol 33, pp 135 - 141, 1973.

Kogan, N. and M.A. Wallach: Risk Taking : A Study in Cognition and Personality. New York, Holt, Rinehart and Winston. 1964.

Krauft, C.C. and D.L. Beggs: "Test-taking Procedure, Risk-taking and Multiple-Choice test Scores." The Journal of Experimental Education. Vol 41 No 4 pp 74 - 77, Summer 1973.

Lord, F.M. and M. R. Novick: Statistical Theories of Mental Test Scores. Reading, Addison-Wesley, 1968.

Marshall, J.C. and L.W. Hales: Classroom Test Construction. Reading, Addison-Wesley, 1968.

Palva, I.P. and V. Korkunen: "Confidence-Testing as an Improvement of Multiple-Choice Examinations." British Journal of Medical Education Vol 7 pp 179 - 181, 1973.

Paton, D.M.: "An Examination of Confidence Testing in Multiple-Choice Examinations." British Journal of Medical Education Vol 5 pp 53 - 55, 1971.

Pugh, R.C. and J.J. Brunza: "Effects of a Confidence-Weighted Scoring System on Measures of Test Reliability and Validity." Educational and Psychological Measurement Vol 35 pp 73 - 78, 1975.

Rippey, R.: "Probabilistic Testing." Journal of Educational Measurement Vol 5 No 3 pp 211 - 215 Fall 1968.

Rothman, A.I.: "Confidence Testing an Extension of Multiple-Choice Testing." British Journal of Medical Education Vol 3 pp 237 - 239, 1969.

Sieber, J.E., R.E. Clark, H.H. Smith and N. Sanders: "Warranted Uncertainty and Students' Knowledge and use of Drugs." Contemporary Educational Psychology Vol 3 pp 246 - 264, 1978.

Sieber, J.E.: "Confidence Estimates on the Correctness of Constructed and Multiple-Choice Responses." Contemporary Educational Psychology Vol 4 pp 272 - 287, July 1979.

Whitla, D.K.: Handbook of Measurement and Assessment in the Behavioural Sciences Reading, Addison-Wesley, 1968.

Wood, R.: "Multiple Choice : A State of the Art Report." Evaluation in Education Vol 1 pp 191 - 280, 1977.

THE SCREENING AND SELECTION OF PUPILS FOR A FAST MATHEMATICS CLASS IN STANDARD SEVEN BY USING TESTS

David A. Norton

A paper submitted to the Faculty of Education, University of Cape Town, in partial fulfilment of the requirements for the degree of Master of Education

1982

Abstract

This paper discusses means of selecting pupils for a fast mathematics class of 30 pupils out of 170 at a local high school using IQ, achievement and test scores. It is argued that both Non-Verbal and Verbal IQ scores should be considered separately, as the Total IQ score alone is not sufficient, especially with bright children. Mathematics schievement scores are also used in order to provide information specifically on the pupil's ability in the subject. The use of a problem-solving test is advocated in an attempt to measure divergent thinking abilities to some degree, as is also the use of a test with a high ability ceiling. The results of the use of these tests are discussed, and various methods of using all the scores obtained in the selection of the pupils are considered.

Contents

Introduction		p. 1
Chapter 1	The Initial Screening Process	p. 4
Chapter 2	The Testing Process	p. 13
Chapter 3	Selection of Pupils for a Fast Mathematics Class	p. 24
Conclusion		p. 33
Appendices		p. 35
References		p. 89

Table

Use of IQ scores and Achievement Scores in Initial Identification of Pupils for a Gifted Program - p. 25

APPENDICES

A.	Pupils Accepted for Further Testing Using IQ and Achievement Scores	p. 35
B.	IQ Scores of the 30 Pupils Accepted for Testing on the Basis of IQ Scores	p. 37
C.	Differences in Influence of Non-Verbal, Verbal and Total Scores on Choice of Pupils for Acceptance into the Test Group for a Fast Mathematics Class	p. 38
D.	Correlation Coefficients between IQ Scores	p. 40
E.	Comparison between Pupils Selected by means of IQ Scores and those chosen by Achievement Scores	p. 41
F.	IQ and Achievement Scores of 58 Pupils Accepted by the Initial Screening	p. 42
G.	IQ and Achievement Scores of 112 Pupils Not Accepted by the Initial Screening	p. 44
H.	Statistics on IQ and Achievement Scores for the Initially Screened Group of 58 Pupils	p. 47
I.	Statistics on IQ and Achievement Scores for the Other 112 Pupils in Standard Six	p. 48
J.	Various Correlations on Pupils Scoring Highly on IQ or in Mathematics Achievement	p. 49
K.	Different Style Questions Test and Multiple-Choice Mathematics Test for Standard Six : Instructions given to Invigilators	p. 50
L.	Standard 6 Different Style Questions Test	p. 51
M.	Standard 6 Multiple-Choice Mathematics Test	p. 56
N.	Summary of Results Obtained on Multiple-Choice Test	p. 63
O.	Item Analysis on Multiple-Choice Test	p. 65
P.	Scores Obtained by Pupils on Problem-Solving and M-C tests	p. 69

Q.	Correlations of IQ Scores, Achievement Scores and Test Scores	p. 71
R.	Ranking of IQ, Achievement and Test Scores	p. 72
S.	Number of Common Rankings of Pupils According to Two Different Sets of Scores	p. 75
T.	Best Rankings of IQ/M-C test, Achievement and Problems Test Scores	p. 78
U.	Pupils Selected According to Methods Discussed in Text	p. 82
V.	Statistics Based on Data in Appendix U	p. 85
W.	Contribution of the Rankings of Each of the 8 Sets of Scores to the Final Selection of 29 Pupils Using this Method	p. 87
X.	Contribution of the Rankings of Each of the 3 Groupings to the Final Choice of 29 Pupils Using this Method	p. 88

Introduction

Much has been made of the needs to establish programes for the gifted and talented in South African schools, since it has been realised that very little thought has been given to these pupils. As in America the gifted have been generally neglected. We, along with the Americans, have the lock-step system of promotion from year to year, and this has resulted in teaching to a particular syllabus only. In order to cater for the brighter pupils in the Senior Secondary phase higher grade syllabuses have been formulated, but these have proved to be insufficient as far as the intellectually gifted are concerned or even for those who are talented in a certain area only. They need even greater stimulation. Higher Grade purports to cater for the upper third of the intellectual spectrum in a particular subject, but at many schools pupils of lower ability are allowed to study subjects at this level, and this affects the teaching of the subject since teachers generally gear their lessons to the weakest in the group. Thus the brighter pupil tends to be adversely affected. When it is considered further that there is generally a greater difference in ability between the brightest pupil and the tenth in order of merit than there is between any other succeeding group of ten pupils, it is evident that the brightest pupils are not extended sufficiently by higher grade material.

In the Junior Secondary phase the position may be even worse, as all pupils follow the same syllabus, and it is especially true in Mathematics that the syllabus for Standard Seven contains little that is

new - there is far too much consolidation of ideas introduced in Standard Six. This consolidation is necessary for the average pupil, but does stultify the intellectual growth of the brighter pupil.

Because of this lack of challenge in the Standard Seven Mathematics syllabus it was decided at a school in the Southern Suburbs of Cape Town that the best group in Mathematics should proceed at a faster rate than the other classes. Accordingly it was decided that the mathematics pupils who had the highest achievement scores in the 7A and 7B classes would come together to form this group. This fast mathematics class was formed on a voluntary basis in that the pupils concerned in 7B were invited to join the 7A group. This identification procedure does result in one essential aspect of the make-up of a fast mathematics class being highlighted, and that is motivation. In the Study of Mathematically Precocious Youth (SMPY) conducted at Johns Hopkins University in Baltimore, it is stated categorically that "students should never be pushed into attending a fast-paced class. This type of class is intended for the highly motivated child who is eager to learn at a fast pace." (Fox 1976 p.47). Secondly the importance of achievement scores must never be underestimated in that it has been shown by many investigations "that previous mathematics grades are the best predictor of later mathematics achievement." (Aiken 1973 p.408). In this connection it must, however, be pointed out that these achievement scores covered the full range of pupil ability and were not scores from tests designed for the talented or gifted group only. Unfortunately, because of time-tabling difficulties, high achievers in the other classes could not be admitted to the group.

Chapter 1

The Initial Screening Process

In identifying talented pupil for a fast mathematics group one must be aware of three dangers. The first is that of regarding the identification procedure as foolproof, discovering all who have the ability to profit from being members of such a group and omitting none. The second danger is highlighted by Renzulli in introducing his "Revolving Door Identification Model" when he states:

> "..... we continue to view giftedness as an absolute concept - something that exists in and by itself without relation to anything else. For this reason most of our identification efforts are directed towards uncovering the magic piece of evidence that will tell us if a child is 'really gifted'.... Any mistakes in the selection (or rejection) process according to the absolutist are attributed to deficiencies in identification instruments rather than the fact that giftedness could be a relative or situational concept." (1981 p. ix).

For giftedness one can read "mathematical talent". All mathematics teachers are aware how pupils respond differently to different aspects of mathematics. In this connection Renzulli's stress on "gifted behaviour" in a given set of circumstances merits attention. Thus the search must not be focused on an elusive concept called "mathematical talent", but on finding those pupils who will benefit from an envisaged mathematical learning situation.

This leads on to the third danger, and that is of identifying pupils for the sake of using identification procedures. Identification must always have the

envisaged group in mind - there must be a purpose for it. Hence justification for each procedure must be stated in terms of the requirements of the envisaged program.

In the initial screening that took place on the Standard Six pupils all the IQ scores, i.e. the Non-Verbal, Verbal and Total scores, were considered for each pupil. The reason for taking the IQ scores is that "intelligence test scores.... are valid indices of scholastic aptitude and that they usually are superior to teachers' judgements." (Daurio 1979 p.15). The scores used were those of the New South African Group Test (NSAGT) which are available to members of staff at a particular school. It must be remembered that group tests have one major disadvantage when using them for screening for highly talented children, and that is that they have definite upper limits as compared to individual tests and the NSAGT is no exception, but, as a starting point, the test is a valuable aid. The other initial screening device was the mathematics achievement scores that each pupil obtained in the second and third terms. The use of these scores has already been justified on p.2.

Use of the IQ Scores in the Initial Screening

Appendix A lists all those pupils who obtained an IQ score of at least 130 on any one or more of the Non-Verbal, Verbal or Total sections of the test as well as those pupils who achieved a score greater than 1 Standard Deviation above the mean on the June and September Mathematics examinations. Thirty pupils out of the 170 were found to have such an IQ score, and the scores for these pupils are given in Appendix B.

Appendix B shows that there is a difference of 5,9 between the mean of the Non-Verbal Scores and that of the

Verbal Scores. It was decided to undertake a matched-pairs t-test to see whether this is a significant difference. The difference is significant at the $p = 0,02$ level. It will be interesting in future to compare whether those pupils who have higher verbal scores receive more benefit from a fast mathematics program than that with higher non-verbal scores or vice-versa. Stanley (1977 p.79) notes, in giving reasons for the tests used in order to choose mathematically precocious youth, the following :"SAT-V (Special Aptitude Test - Verbal components) could be used with the higher scorers on SAT-M (SAT - Mathematical components) to assess verbal reasoning ability, which seemed likely to be more closely related to speed of thinking and of taking tests than is SAT-M". The comparison suggested above can only be accomplished using a longitudinal study, and it is only when such a study has been undertaken that the effect on a fast mathematics program of having a group of pupils with significantly higher non-verbal scores than verbal can be assessed.

What is remarkable is that only two pupils were represented in all of the Non-Verbal, Verbal and Total IQ scores. Because of this it was decided to check whether these scores chose significantly different pupils for initial screening by using a chi-squared test, the results of which are given in Appendix C, and interpreted below.

Differences in Pupils Accepted by Differing IQ Category Scores

From the chi-squared tests recorded in Appendix C it is evident that the difference between the pupils accepted as a result of the non-verbal scores and

the number of pupils accepted using the verbal scores is highly significant. From this result it would appear essential to take cognizance of both the verbal and non-verbal scores separately when choosing pupils for any type of gifted program. To many, an IQ score of 130 or more, indicating as it does the upper two per cent of all pupils, is an essential component in any initial screening for the gifted. If the tester is searching for "gifted behaviours", however, it is apparent that to consider the total IQ score alone is not a very productive exercise; both the verbal and the non-verbal scores must be considered separately in order to widen the net of knowledge on the potential behaviours that the pupils exemplify.

The difference between the Verbal and Non-Verbal scores in those pupils chosen under these categories is confirmed by the correlation coefficients given in Appendix D. There is a surprisingly high negative correlation of -0,52, N = 29; (p = 0,01). Even when only those pupils with a total IQ score of 130 or more are considered, the negative verbal/non-verbal correlation is evident (r = -0,68, N = 17; p = 0,01). It can be argued that gifted pupils should be regarded as those who obtain IQ scores of at least 130 in both the verbal and non-verbal categories if IQ is to be used as a criterion, but whatever cut-off score is chosen the two different aspects should be considered separately and should not form part of a combined score when "giftedness" is being considered, as any difference measured will then be ignored, and this difference may be highly significant, as in this sample. Incidentally the sample is supposedly typical of the kind of pupil for whom the test was originally intended i.e. white South African middle-class pupils.

Despite the marked difference between the non-verbal and verbal scores as shown by the correlation coefficient neither set of scores shows a negative correlation with the total IQ scores. This result accords with the results obtained from the chi-squared tests given in Appendix C, where the numbers of pupils initially selected according to either Non-Verbal or Verbal scores is not significantly different from those selected as a result of Total Scores. It is to be noted, though, that the non-verbal has a low positive correlation, while the verbal has a higher positive correlation which again reflects the difference between the non-verbal and verbal scores in this group of pupils.

When the achievement scores were taken into account 58 pupils formed an initially screened group. There was the same significant difference between the means of the non-verbal and verbal scores as given by the incorporated group of 30 which were identified by IQ scores alone. In contrast to the high negative correlation between the verbal and non-verbal scores, there is now a fairly low but significant ($p = 0,05$) positive correlation of 0,26 (Appendix H) suggesting that the 28 other pupils who obtained mathematics achievement scores more than 1 Standard Deviation above the mean which are now introduced into the group show positive correlation between their verbal and non-verbal scores. This correlation is, however, low ($r = 0,15$), but the combined effect of the two groups results in a significant change in the correlation. When, further, the other 112 pupils in Standard Six are considered then the correlation between the verbal and non-verbal scores is higher still at 0,60 with a highly significant 0,01 P-value. (Appendix I). It must be noted that while the difference between the means of the verbal and non-

verbal scores is lower, it is still as significant as before. These results confirm the well known statistical fact that there is greater correlation between verbal and non-verbal scores with pupils whose scores cover a wide range on IQ tests than with pupils who have high scores only. Thus the total IQ score would appear to be a better indicator of potential ability with scores less than say 120, than with those above this score, since there is then greater correlation between the verbal and non-verbal scores. This conclusion, however, must be verified with other groups.

Use of Achievement Scores in the Initial Screening

It was decided that all students who obtained a higher mark than 1 Standard Deviation above the mean in the second term examination, the third term tests or in the average of the two results would be selected. As shown by Appendix A 30, 30 and 31 pupils respectively were selected on the basis of these achievement tests. In strong contrast to the IQ scores there is high internal consistency between these results in that only 40 different pupils were selected, and this despite a 30 mark difference between the means of the second and third term results.

Only twelve students were chosen independently by both achievement scores and IQ scores. In Appendix E the ideal distribution of pupils chosen by achievement, including those chosen by IQ scores, is given - this would be the case if IQ and mathematics achievement scores measured the same aspects of talent. It is clear from the actual distribution that this is far from being the case - indeed the opposite type of distribution would appear to be closer. A chi-squared test conducted on these data indicated that the difference

was highly significant ($p = 0,001$). This could be an indication that the IQ and achievement scores measure different abilities as far as this top group of 58 pupils is concerned; and because of this, when screening for a fast mathematics class, it is evident that IQ alone must not be the sole consideration - mathematics achievement must also be considered.

The question now arises as to whether correlations of IQ and achievement scores will confirm the above result. It would seem necessary to compare the various IQ scores with the achievement scores because they differ from each other to such a marked degree.

Comparison of IQ and Achievement Scores

Appendix F lists the IQ and achievement scores of the initially screened group of 58 pupils.
Appendix G lists the IQ and achievement scores of the other 112 pupils in Standard 6.
Appendix H gives the correlation coefficients between the various IQ and achievement scores for the group of 58 pupils.
Appendix I gives the correlation coefficients between the various IQ and achievement scores for the other 112 pupils.
In considering the statistics in Appendices H and I it is firstly obvious that the Term 2 examination was far easier than the Term 3 test series. Despite this, the correlation of 0,66 between the scores of the two examinations is highly significant indicating that the achievement tests probably tested the same abilities.

What is very noticeable is the contrast between the two groups in the correlations between the various IQ scores and the achievement scores. In the group of 112 all the correlations are positive, and what is more, highly significant (all $p = 0,01$), which seems to

indicate that for this group the IQ and the achievement tests appear to be testing related or similar aspects as far as Mathematics is concerned. This is in strong but expected contrast to the screened group of 58 with a narrow range of talent where every IQ/achievement correlation is negative, although some of the results are not significant. However, the significantly different correlations do not occur when the verbal IQ is taken into account. The non-verbal and total scores produce higher negative correlations with mathematics achievement than the verbal scores, and this is confirmed by the reslts of the other 112 pupils where the non-verbal scores give the lowest correlations with achievement although the difference between these correlations is nowhere significant using Fisher's transformation. The negative correlations again suggest that in the search for those who can form a fast mathematics group it would be mistaken to ignore the achievement scores even though these negative correlations may only be a reflection of the narrow ability range of the upper group of 58. It should not be taken as read, that negative correlations will always occur when the scores of high IQ pupils are compared with the achievement scores, or indeed that there will be an increased negative correlation when only the highest IQ or achievement scores are considered. This is shown by the correlations given in Appendix J where those pupils with the highest IQ scores and highest average achievement scores have been isolated and correlation coefficients between IQ scores and achievement have been calculated. Nearly all the correlations are closer to zero than they are to any significant result. The results do, however, confirm that one may not only consider IQ scores in choosing gifted pupils for a fast mathematics class, and assume that these will give a clear indication of high IQ pupils' achievement, even

in a subject like mathematics where achievement is supposed to correlate well with IQ scores. Nor may one consider achievement only.

Thus the initial screening based on statistics at hand was completed by choosing the group of 58 pupils considering each aspect separately. Further testing was now required, but testing that conceptually should not repeat what IQ and achievement scores measure.

Chapter 2

The Testing Process

Because questions in an IQ test usually require one correct answer from the testee, convergent thinking abilities are to the fore in most parts of such a test, except where the questions demand the use of problem-solving techniques. Knowledge concerning a pupil's divergent thinking abilities should be ascertained, especially with more talented pupils, who are more likely to possess such abilities to a greater extent. One of the means of ascertaining information about such abilities is by using creativity tests, and here those of Torrance readily come to mind. However, Aiken (1973 p.412) states that "correlations between scores on Torrance's tests and mathematics achievement are usually rather low although those prepared by Guilford are significantly related to achievement." It would appear as if creativity tests are more worthwhile when seeking giftedness in Mathematics if they are specifically geared "mathematical creativity tests." General creativity tests do not seem to provide much knowledge that is helpful in identifying giftedness in Mathematics. Indeed Torrance and Gowan report that there are low correlations between verbal and non-verbal creative abilities and they appear to be largely independent (Wallach and Kogan 1965 p.7).

Further to the above, creativity tests generally involve fluency of ideas as one aspect of their assessment, and because of this they tend to become tests of how many ideas can be listed in a particular period of time. Wallach and Kogan comment that:

"Such a feature seems not to fit one's intuitions concerning the type of situation within which creativity may manifest itself most naturally." (1965 p.12). In confirmation of this the present writer has often noted that Standard Ten Art pupils state that they cannot do their best work in an Art practical examination as the idea of being tested, of its own accord, immediately inhibits their creative thinking processes, and this despite their being allowed a full six hours in which to produce their work. The inhibitions must surely be greater when the time limits are even more constraining as they are in creativity tests.

For these reasons the author decided not to use a general creativity test, and, because he could not obtain a psychologically validated mathematical creativity test, he turned to another type of test in which he had more experience, and which would involve divergent thinking processes. This is the problem-solving test where non-stereotyped mathematical problems are presented to the pupil which he is called upon to solve. Despite the testee having to work towards one correct answer, which would seem to suggest that convergent thinking processes must be dominant, in order to get to that one answer divergent thinking abilities have to be brought into play. This is because the questions asked are different from the usual style of question so that the testee is required to "doodle." Thus the same thinking abilities required in creativity testing are also necessarily used in problem-solving. It was, therefore, decided to draw up a problem-solving test where the pupils would be required to solve up to seven problems in half an hour. The reason for such a large number of problems was to provide a high enough ability

"ceiling" so that pupils with more advanced problem solving abilities would not complete the test in too short a time, and could thus be separated from those less able to do problems.

In identifying pupils for their fast mathematics classes in the United States Stanley and his associates in the SMPY project already referred to believe in the efficacy of tests which have a high ceiling. They test seventh and eighth grade volunteers from various schools in Maryland, by arranging that they sit the Special Aptitudes Test organised by the College Board. These tests are intended for twelfth grade students who are seeking entrance into colleges in the USA. The test has two parts, namely the SAT - Mathematical and the SAT - Verbal tests, both of which are regarded as being of benefit by the SMPY directors as has been mentioned already.

In testing eighth grade pupils using the SAT-M test it must be remembered that many questions contain Algebra and Geometry and in the USA neither of these are usually taught until the ninth grade has been reached. Hence, it is only those who are genuinely interested in mathematics and who are able to reason well who score highly in these tests. It must be remembered that SMPY is only searching for the top half per cent of all pupils in mathematics. In addition these students would have to be well-motivated, because they would be invited to join a special class which would not be held during school time nor in the school environment. Hence the group being looked for is very select - far more select than any school, other than one which selects top pupils on academic merit only, can provide. The SAT-M test can, however, be used with profit in South Africa for a group like the initially screened group already

selected, because Standard 6 pupils have already learnt some basic Algebra and Geometry, which means that they are better equipped than their American counterparts. There would still, however, be sufficient ceiling to the test because of the age for which the test was geared. Because the testing time available was only one hour it was decided to use only one of the battery of three mathematics tests which make up the SAT-M. The test was found to be suitable as none of the Mathematics covered in it is foreign to that taught in South African Schools. It was also decided to use the test because it was fully standardised and multiple-choice in nature. Any test constructed by the author would have had to be standardised first. In addition the College Board provides statistics on how well each question is answered, and it was felt that this would provide a good means of comparing the ability of these pupils with what is the norm for twelfth graders in the USA. The test used was administered by the College Board on April 4 1981.

Administering the Tests

On the day before the tests were written the initially screened group of 58 pupils was called together, and it was explained to them that further information was sought on their ability in Mathematics. This would be ascertained by their writing two tests which were different from what they would normally be asked. It was explained that the first test would be called "Different Style Questions", and would consist of problems which they would normally not encounter. In order to stress the difference in style the following broadly stated example was given:
"If you were given a certain number of matches that make up a square, move say three matches to make two squares."
It was stressed that they would not be able to learn for

this kind of test. They were also informed that neatness was definitely not a requirement as they would have to doodle.

All that was stated about the second test was that it would be a multiple-choice test, and that they would have to mark the correct answer.

It was stressed that the results from these tests would not be counted in their examination mark at the end of the year. But, in order to provide some motivation, the pupils were informed that the scores of these tests would be used along with their achievement scores and other knowledge that the school possessed concerning their potential, to decide who would be chosen for a fast mathematics class for the next year in Standard 7. Time was then given for questions.

The pupils were divided into two equal groups. Each invigilator had the instructions given in Appendix K. The question papers and the marking memoranda appear in Appendices L and M. It will be noted that part marks are obtainable in the problem-solving test. This is to counteract in some way the criticism by Stroker (1980 p.6): "One of the reasons why scores obtained by pupils in standardised tests need to be treated with some reservation is that the tests themselves contain "answer only" items, which militate against detection of the process of solution and depth of thought used by the pupil." It will be noted, however, that in certain questions part marks were not given because of the nature of the answer to the question where a part of an answer would give no indication as to how the problem should be solved.

A cursory examination of questions 4 and 7 would seem

to indicate that similar skills would be required for their solution. This, however, is apparently not the case because 38 correct answers were obtained for question 4 whereas only three pupils had any idea as to how to answer question 7. Perhaps the difference was a result of the pupils having been asked to count triangles in a class exercise so that it formed part of their experience, or perhaps it was because the triangles were contained within a rectangle and so more easily recognizable than finding the number of squares within a large square. Perhaps the reason was simply that many students were running out of time when they began question seven.

In the instructions to the invigilators in Appendix K the invigilator was told to ask all pupils to write down the numbers of the questions they had not answered in the multiple-choice test. The purpose was to obtain an idea as to the part played by time in the taking of the test. The reason for then asking the students to mark an answer to each question at random was to be able to do an Item Analysis on the questions.

Analysis of the Tests' Results Data

In considering the data given in Appendix N it is readily deduced from the number of random guesses recorded from question 16 onwards that time played a major role in the Multiple-Choice test. From the average results per question it would also appear that time was effective in reducing the candidates scores from question 18 onwards. The mean for questions 1 to 15 and 17, i.e. those questions in which random guessing had a less than 30% effect, was 59% while the mean for the other questions was 20%. The effect of the random guessing was also shown in the Item Analysis (Appendix O) where the questions that have to be rejected are all

those in which random guessing had a greater than 30% effect. Thus it would appear that the increased ceiling in this type of test is not a result of the quality of the individual questions asked, but of the role played by time. That there is an increased ceiling is shown by the mean of the testees being 39% when random guessing is excluded whereas their means on the achievement scores were 74% for term 2 and 66% for term 3. (It was decided, even though the students were given a score in which random guesses which were correct were included, that for statistical purposes the results they obtained without correct random guesses would be used). But the shortage of time meant that the range of scores on this test is 52 percentage points, whereas the range on the average achievement scores was 55 percentage points. Thus the test does not necessarily help with the differentiation between the better and poorer pupils.

In contrast to these results Stanley and his associates in the SMPY have found that certain of the examinees have obtained 100%. This would seem to indicate that the test as it stands is probably more useful when entrance is limited to pupils with very high ability in the subject only. In the problem test where time was not as great a factor apparently, although the results of question seven may have been affected by it, the average is again low, being 45%, and what is of significance is that the range is wider being 69 percentage points, which is more satisfactory from the point of view of differentiating between pupils. It would be interesting to see what effect an increase to forty minutes for answering the paper would have on the results seeing that nine out of the twenty-five questions, about one-third, were seriously affected by the time constraint. This, it is felt, would enable the brighter

pupils to finish the paper, thus increasing its power to differentiate between students.

The Reliability Coefficient on the problems test is very low at 0,19 (N = 58), but this can be ascribed to the high percentage of full and zero marks obtained (75%), making this the equivalent of a test where 1 or 0 could have been the score awarded to each question, and with only seven questions the reliability will be low.

In the multiple-choice test it is evident that the reliability has been greatly affected by random guessing so that the better figure occurs when only 16 questions are taken into account in which case a value of r = 0,50 is reasonable. Again one would like to test this if more time were allowed for the paper.

Correlation of the test scores with the other sets of scores

In Appendix Q correlations are given between the problems test and the multiple-choice (M-C) test and all the other scores used thus far; the necessary scores being listed in Appendix P. It is notable that the problems test does not correlate significantly with any of the IQ or achievement scores. It would therefore seem that the problems test does not test the same aspects as either the IQ or the achievement scores. Thus it may be held that it possibly tests divergent thinking abilities whereas the other types test convergent thinking abilities. It must be noted that, although weak, the correlation is positive so that this conclusion may not be pressed as one must bear in mind that, theoretically, convergent abilities are also tested by means of a problem-solving test.

The problems test does, however, correlate significantly with the M-C test which correlates significantly with the IQ scores. It would seem as if the M-C test, while generally testing convergent abilities, possibly also tests divergent abilities in some measure. As stated originally when discussing the problem-solving test, both divergent and convergent thinking abilities are required for the solution of problems which are unfamiliar. A number of the items in the M-C test were unfamiliar to the pupils and hence in their solution the students may have had to use divergent thinking abilities. However, these suggestions may be mere speculation as the test has been set only once to fewer than 60 pupils who all live in one type of environment.

What is of great importance is to note that all three aspects of the intelligence test correlate significantly positively with the multiple-choice test, and correlate negatively, although not significantly, with the achievement scores, which also occurs in the correlations between the IQ and achievement scores. Thus it can be concluded that the M-C test may test the same or similar abilities as the IQ tests do. Thus as a predictor for a fast mathematics class the M-C test, despite its high ceiling, seems to confirm what can be learnt from the IQ scores. Before a definite conclusion can be reached, however, one should take the earlier criticism on time into account. There is little spread in scores in the same way as the spread of group IQ scores is also not strongly marked. Statistics obtained using the scores given a forty minute time allowance may give an entirely different conclusion.

Ranking of Pupils According to Eight Sets of Scores

It was decided to rank the first 29 pupils on the basis of each of their IQ, achievement and test scores. The reason for this number was that this formed one-sixth of the total number of pupils in the six stream Standard 6 class and also a half of the test group. In addition because of the classroom situation at the school a class of approximately thirty pupils would have to form the fast mathematics group. The rankings are given in Appendix R. Where more than 29 pupils were ranked as a result of ties, random-number tables were used to exclude the extra ranked pupils. All the pupils have been ranked in at least one set of scores, with only four pupils ranked by all eight sets. When this diversity is considered it is important to ascertain whether there is any significant difference between the rankings according to the various chosen methods. Accordingly Appendix S lists the number of common rankings of pupils according to two different methods, which results in 28 different combinations.

From each of these combinations a 2 x 2 table can be set up in which the criteria can be compared using a chi-squared test. Theoretically, if there is no common or opposing influence, then there should be no difference between the ranked cells and unranked cells on either of the methods with 14,5 in each. If the rankings according to the two methods were exactly the same or completely different then the distribution would be as follows:

Exactly the same

	Test A: Ranked	Test A: Not Ranked
Test B Ranked	29	0
Not Ranked	0	29

Completely different

	Test A: Ranked	Test A: Not Ranked
Test B Ranked	0	29
Not Ranked	29	0

For there to be a significant result either way the cells should have 19 and 10 in them while highly significant results occur when a cell has at least 20 in it as shown in Appendix S.

The rankings are found to be significantly the same between all the achievement scores. With the exception of the Verbal/Non-Verbal comparison the IQ scores create rankings which have high significance ($p = 0,01$). The Non-Verbal/M-C test rankings are significantly the same as well. When compared with the correlation coefficients in Appendices H and Q it is noticeable that the significant ranked pairings are the same combinations as those with correlation coefficients greater than 0,4 except for the M-C/Total IQ score comparison.

The rankings are significantly different between the IQ and achievement scores with the exception of the Verbal Scores. This again accords well with the results of the correlation coefficients. On this occasion all significant negative correlations are included amongst the pairings, and the Term 2/Total IQ comparison has been added.

Thus it would seem as if there is some connection between the non-verbal scores and the M-C test. It would appear that they test the same abilities.

Chapter 3

Selection of Pupils for a Fast Mathematics Class

Thus far the various tests and scores have been compared from a statistical point of view in order to see where the common factors lie and what each aspect tests. It now becomes necessary to decide which pupils should be selected for a Standard 7 fast mathematics class. Value judgement in such a decision is always present, and indeed is more often than not the central deciding factor. The reason for this centrality is that the method that is chosen to decide who is a member of any group, even when based on statistical tests, is governed by the value judgement of deciding which tests should be used. There can be no single objective scientific means of choosing a group for even the use of a battery of tests is a subjective decision. Statistical analysis must be used and the results therefrom may not be ignored, but ultimately it will always be subject to the interpretation of the one who conducts the experiment. Depending on the method chosen, different pupils will be admitted to the fast mathematics class.

As the situation being faced in this case is a school situation, methods applicable to a regional approach cannot be used without careful modification. For example Terman's original gifted program included only those pupils with an IQ greater than 140 - merely a half per cent of the population. At the school under consideration no one would qualify under this condition since two have non-verbal scores and one a verbal score higher than 140, but in all three cases there is a marked difference between the verbal and non-verbal

scores which generally indicates that the student will not perform to his higher capability. Bearing the school situation in mind three different methods of choosing pupils making use of tests will be considered.

Use of a Combination of IQ and Achievement Scores in Pupil Selection

The Cape Education Department has suggested a simple means of initial identification of those pupils who should form part of a gifted mathematics program. This involves using IQ and Achievement Scores in a points scheme delineated in Table 1. As these two factors may test somewhat different aspects, it is useful to consider both in combination.

IQ Scores		Achievement Scores	
Score	Points Awarded	Score	Points Awarded
145+	25	100%	25
140 - 144	23	90 - 99%	23
135 - 139	21	85 - 89%	21
130 - 134	19	80 - 84%	19
125 - 129	17	70 - 79%	17
125	16	70%	16

Table 1

Use of IQ Scores and Achievement scores in the Initial Identification of Pupils for a Gifted Program

Any pupil with a combined total of 35 points will be admitted to a gifted mathematics program.

Eg. 1 Pupil with Non-Verbal IQ of 127 and achievement score of 73% will obtain 17 + 17 = 34 points and will no t be accepted.

2 Pupil with Verbal IQ of 131 and achievement score of 69% will obtain 19 + 16 = 35 points and will be accepted.

A cursory examination of the procedures outlined in Table 1 shows that anyone with an IQ score of 130 or more or an achievement score of at least 80% will be chosen for a gifted program. This could result in a very widespread net especially at a school in the middle to upper socio-economic bracket. This is of great benefit if there are limited aims for the group. A smaller more manageable group must be chosen if an ambitious program has been arranged. This is especially true if the program is being followed for the first time.

In criticizing this method of indentification, it must be emphasised that it is only intended to be an initial identification procedure and nothing more. There are three definite weaknesses. The first is that no consideration has been given to divergent thinking abilities i.e. creativity. The education program in South Africa, like most, is more concerned with convergent thinking processes, but especially when one is trying to develop all the abilities of more advanced pupils intellectually there is a need for some sort of score which reflects to some measure a pupil's divergent thinking abilities. The second weakness is that all achievement scores should first be converted to a scale of a fixed mean and standard deviation before any points are calculated, because of the great variation is scores from one test series to another. For this reason the only achievement scores used were those of the average score as the term 2 mean for all pupils was 60% while that in term 3 was 50%. Hence it is probably preferable in the case of achievement scores to consider rankings if the marks have not been adjusted. The third weakness is that no score which reflects the results of a test with a high ability ceiling has been used - thus differences between

pupils of high ability have not been ascertained.

Use of Many Different Sets of Scores

The scores that have been obtained are the following:

a. IQ Scores : Non-Verbal, Verbal and Total
b. Achievement scores : Term 2, Term 3 and Average Mathematics scores
c. Test scores : Problem-solving and M-C test

The first 29 pupils in each set of scores were ranked (See Appendix R).

In choosing pupils using these scores it is suggested that the following who show far higher than average ability in any one sphere should be included:

a. all pupils with any IQ score of 140 or more,
b. all pupils with an achievement or test score of at least two Standard Deviations above the mean

Following this the choice should be made on the number of rankings a student achieves. Those who are ranked in say six different sets of scores being chosen before those ranked in five different sets and so on. If those ranked using, say, four scores result in the chosen group being too large then it is suggested that the average rankings be calculated and that the best of these should be selected. The reason for this last procedure is that approximately thirty pupils would have to be included in this group because of the classroom situation at the school during a rebuilding program.

It would seem that this selection procedure is far more broadly based than the first method, but this is not so. Three out of the eight scores all test achievement where the scores show high correlation with each other. Another four sets of scores also correlate significantly with each other, that is the

various IQ scores and the M-C test scores. Using the information obtained from the chi-squared tests on the ranking comparisons, and the information from the correlation coefficients, it can be argued that only three different criteria are actually being tested, and that convergent abilities are still being favoured to an inordinate extent comprising seven out of the eight sets of scores. Because of this it was decided to combine certain data and to consider the effect of the combination on the selection of pupils.

Use of Scores According to Three Different Groupings

The IQ scores and the M-C test scores were considered to form one group because they correlated highly with each other. The second group sonsists of the three sets of achievement scores, while the scores obtained from the problem-solving test were treated as a group on their own because, hypothetically, the test caused the pupil to use divergent thinking abilities, and also because there was no correlation between it and any other set of scores except for the M-C test.

In each of the first two groupings the highest ranking obtained by a pupil in any of the scores in that grouping was recorded - not the average rank, because a pupil's potential ability is what is being looked for. The first 29 of these rankings were then ranked as shown in Appendix T. The selection of pupils involved firstly taking those who were ranked in all three groupings, then those in two groupings, and finally the best of those ranked in only one set of rankings.

A definite weakness of this method with this set of data is that only one problem-solving test was used, and it was nowhere near reliable as shown in Appendix N. Either a number of scores should be obtained or the test

used should have a high reliability index. This is an important consideration, especially when the results of the chi-squared tests recorded in Appendix T are considered. There is a significant difference between those pupils ranked by the IQ/M-C test scores and those by achievement scores which confirms the difference between the aspects tested. The results of the Problems test would seem to bridge the gap between the two extremes as in both cases there are more common acceptances between the problems scores and the other variable than differences in pupils accepted, this being significantly the same in the comparison with the IQ and M-C test scores grouping. Would this be true with a reliable problem-solving test? This question calls for investigation.

The Pupils Selected by the Different Means

Appendix U shows the pupils who have been selected from the initially screened group of 58 according to the methods described earlier in the present chapter. As stated in Appendix V only 8 pupils appear in all five lists and only ten pupils are not chosen by any of the five. This means that there is very little consistency of choice amongst the five methods in total even though between any two methods the choice of pupils is generally regarded as being significantly the same. But, even when the results are regarded as being significantly the same, a fair number of pupils are identified by one method and not by another. For example 20 out of 29 pupils selected by means of the eight sets of scores were also selected by the three groupings method, which means that another 18 pupils are also involved either chosen by one means or by the other - certainly no indicator of unanimity of choice.

The lack of significance between the choice of pupils using three groupings and that of the choice by using firstly the Non-Verbal scores and Average achievement and secondly the Verbal scores and Average achievement can be regarded as expected because of the predominant role (see Appendix X) played by the problem-solving test scores in the selection of pupils, these scores not being represented in the latter two methods. It must also be borne in mind that these scores did not correlate significantly with either the IQ or the achievement scores.

The results obtained from the above methods confirm the assumption that one will not obtain a definitive list of all the pupils who would benefit from a fast mathematics class by the use of one test or even a batting of tests. The variation in the pupils selected is far too wide for this. Hence it must always be remembered that other pupils who have not been chosen may benefit, and also some who form part of this group may wish to leave or even need to leave because they are not benefiting from working at a faster pace. This leads to the feeling with certain educationists that any program for a more talented group should have no built-in inflexibility at all, so that other pupils who were not chosen originally can be absorbed into it easily. Hence enrichment should dominate in order that pupils may enter or leave the program at any time.

One must, however, not miss the significance of the statistical results that are given in Appendix V. It is evident that even though only 8 are identified by all five methods, there are 13 that are identified by four methods and a further 7 by three methods including the use of the problem-solving test. This means that there is a real measure of agreement between

the methods. It can possibly be argued that the way in which the pupils should be selected is by selecting those who are accepted by the greatest number of selection procedures as in this paragraph, but this would accentuate the influence of the IQ and achievement scores to an even greater extent.

Yet another important consideration is that the pupils who should be catered for in a fast mathematics program are those who would most likely benefit from the increased pace. The most motivating way of catering for them is to give them the sense that they are being enriched by advancing at a faster rate compared to their peers. It is important that the program should not be geared to those who can possibly be part of it, but should be geared to those who are definite members, and who will therefore benefit from a fast pace even though an element of inflexibility is thus introduced. To meet the needs of the brightest in mathematics the needs of those on the fringes must not dominate. They can be catered for in another group.

In the present writer's opinion the method using the three groupings of scores is the best, because distinct areas are separate and the problem-solving score, which is deemed to be very important, especially for pupils belonging to a fast mathematics class, is neither ignored nor swamped by all the other data. In addition, however, creativity tests in mathematics should be added so as to test for divergent thinking. Yet another field that should be tapped is the opinions of the teachers who teach the pupil, and also information should be ascertained from the parents by means of questionnaires. These opinions will, however, only be useful if firstly the teachers are trained in what they are to look for so that their questionnaire can be

answered properly. Here one thinks of the oft-quoted statement in the Terman results that teacher identification as a means of ascertaining who could form part of a gifted group was less successful than if one had considered the chronological ages of the pupils and chosen the youngest. Secondly the parents must be encouraged to study their questionnaire and be given the opportunity to ask questions of the school authorities concerning the information required. In identifying pupils there must not be the familiar pandering to the notion that a safe decision lies in the quantity of data that can be accumulated. The quantity of data must be analysed and sorted according to the different aspects tested, and these aspects should be treated equally in any final decision.

The present writer used the word "opinion" at the beginning of the last paragraph. In 1983 he will be using the results obtained from the eight sets of scores because of the low reliability coefficient of the problem-solving test. It is noticeable in Appendix W that all the sets of scores contribute approximately equally to the final choice of pupils which means that the problem-solving test, while not being dominant, is not ignored. This is the greatest influence that it can be allowed to exert when there is such a low reliability coefficient.

Conclusion

It is evident from the foregoing that it is essential that testing should take place in order to determine who should form part of a fast mathematics group, but that the tests used must be of such a nature that they will provide meaningful information on the abilities that are required by pupils taking part in a fast mathematics program. Hence IQ and achievement scores are a necessity, as is also knowledge of divergent thinking abilities and knowledge on pupils' abilities to do well on tasks where the ceiling is far higher than in the usual type of achievement test.

When considering the influence of the sets of scores it is argued that those scores which show high correlation should be grouped together, and that the choice of pupil should depend on how well he evinces strengths in various different thinking abilities. In this case what is his potential both verbally and non-verbally as given by the IQ scores, can he solve problems, can he do the easier work well, can he work quickly and well?

Having thus ascertained this knowledge it seems essential that any fast mathematics course should include all of the above facets. The course cannot consist of problems alone or, for that matter, deductive reasoning alone. In addition there must be place for the development of the basic skills at each level of attainment. Hence the importance of having an idea of the ability of the pupils in different areas. Thus tests are essential, but to consider that only those who do well on the tests are those pupils who are talented mathematically, and that all of those who are selected

have great mathematical ability is untenable. The tests, however, can be used to select a group the great majority of whom should benefit from a fast mathematics program.

APPENDIX A

Pupils Accepted for Further Testing Using IQ and Achievement Scores

IQ >130			Achievement + 1SD above Mean		
Non-Verbal	Verbal	Total	Term 2	Term 3	Average
				123	123
			124	124	124
201	201	201			
211		211			
222		222			
	301	301	301	301	301
			302		302
			303	303	303
	304	304	304		304
			305		
			307		307
308					
	310	310	310		310
			311	311	311
				312	312
				313	
			314	314	314
316		316	316		316
			317	317	317
	318	318			
			319	319	319
			320	320	320
			324	324	324
			325	325	325
327			327		
			328	328	328
			330	330	330
				331	
			404		
			420	420	420
421					
				503	
			506		506
512					
			513	513	513
				514	514
521					
	525				
				529	
	601	601			
605		605			
607			607	607	607

IQ >130			Achievement +1SD above Mean		
Non-Verbal	Verbal	Total	Term 2	Term 3	Average
		608	608		
609		609	609	609	609
611		611	611	611	611
613		613			
			614	614	614
				617	617
			618	618	618
	619	619			
622					
	623				
626	626	626	626	626	626
627		627			
628					
629				629	
630				630	630

Note: Each three digit number represents a pupil

APPENDIX B

IQ Scores of the 30 pupils accepted for testing on the basis of IQ scores

Pupil	Non-Verbal	Verbal	Total
201	133	139	137
211	139	121	131
222	135	124	131
230	132	120	127
301	129	137	134
304	129	145	138
308	133	121	128
310	119	138	130
316	130	127	130
318	123	134	130
327	134	106	119
421	130	121	128
512	131	112	122
521	133	120	128
525	122	132	128
601	123	134	130
605	141	123	134
607	132	122	128
608	128	127	131
609	145+	121	134
611	133	128	132
613	133	126	131
619	129	131	132
622	137	113	125
623	123	132	129
626	135	130	135
627	135	129	134
628	132	117	126
629	130	119	126
630	137	118	128

$\overline{NV} = 131,5$

$\overline{V} = 125,6$

The difference between the means is significant at the 0,02 level according to the matched-pairs t-method

APPENDIX C

Differences in Influence of Non-Verbal, Verbal and Total Scores on Choice of Pupils for Acceptance into the Test Group for a Fast Mathematics Class

1. Choice of pupils using Non-Verbal IQ Scores as basis as Opposed to Verbal IQ scores. (N = 30)

		Non-Verbal Scores	
		Accepted	Not Accepted
Verbal Scores	Accepted	2	8
	Not Accepted	19	1

$\chi^2 = 14,464$ $p = 0,001$

There is a highly significant difference between the pupils accepted as a result of their Non-Verbal Scores and those accepted on the basis of Verbal scores. Thus it would appear essential to consider the Non-Verbal and the Verbal Scores separately when screening pupils for a gifted program.

2. Choice of pupils using Non-Verbal IQ scores as basis as opposed to Total IQ scores. (N = 30)

		Non-Verbal	
		Accepted	Not Accepted
Total	Accepted	10	7
	Not Accepted	11	2

$\chi^2 = 1,266$

No significant different in choice

APPENDIX C (contd)

3. Choice of pupils using Verbal IQ scores as basis as opposed to Total IQ scores

		Verbal	
		Accepted	Not Accepted
Total	Accepted	8	9
	Not Accepted	2	11

$\chi^2 = 2{,}046$

No significant difference in choice.

APPENDIX D

Correlation Coefficients between IQ Scores

1. Pupils chosen on basis of Non-Verbal scores compared to those chosen on Verbal scores (N = 29)

   Verbal/Non-Verbal Correlation $r = -0,52$ ($p = 0,01$)

2. Pupils chosen on basis of Non-Verbal scores compared to those chosen on Total Scores (N = 28)

   Non-Verbal/Total Correlation $r = 0,08$ (Not significant)

3. Pupils chosen on basis of Verbal scores compared to those chosen on Total Scores (N = 19)

   Verbal/Total Correlation $r = 0,31$ (Not significant)

4. Pupils chosen on basis of Total score ≥ 130 (N = 17)

   Verbal/Non-Verbal Correlation $r = -0,68$ ($p = 0,01$)

APPENDIX E

Comparison between Pupils Selected by means of IQ Scores and those Chosen by Achievement Scores

$N_{IQ} = 30$ $N_{ACH} = 40$

Ideal distribution if pupils chosen by achievement incorporate those chosen by IQ scores

Achievement	IQ: Accepted	IQ: Not Accepted
Accepted	30	10
Not Accepted	0	0

Observed distribution

Achievement	IQ: Accepted	IQ: Not Accepted
Accepted	12	28
Not Accepted	18	0

Theoretical distribution under these circumstances

Achievement	IQ: Accepted	IQ: Not Accepted
Accepted	20,69	19,31
Not Accepted	9,31	8,69

$\chi^2 = 21,64$ $p = 0,001$

There is a significant difference between selecting by means of IQ and selecting by means of achievement

APPENDIX F

IQ and Achievement Scores of 58 Pupils Accepted by the Initial Screening

Pupil	IQ Scores			Achievement Scores		
	NV	V	Total	Term 2	Term 3	Average
123	117	99	107	219	214	217
124	126	115	122	255	242	249
201	133	139	137	201	145	173
211	139	121	131	195	135	165
222	135	124	131	222	181	202
230	132	120	127	213	155	184
301	129	137	134	273	252	263
302	116	130	110	252	180	216
303	119	111	117	243	216	230
304	129	145	138	231	194	213
305	124	107	114	228	156	192
307	115	127	123	240	186	213
308	133	121	128	183	152	168
310	119	138	130	261	166	214
311	117	128	124	240	202	221
312	125	121	125	219	204	212
313	112	123	119	207	210	209
314	124	127	127	240	270	255
316	130	127	130	234	196	215
317	125	106	115	243	216	230
318	123	134	130	195	152	174
319	128	110	118	249	246	248
320	126	124	127	237	226	232
324	115	117	117	255	220	238
325	117	112	115	264	256	260
327	134	106	119	231	184	208
328	119	119	121	291	257	274
330	121	118	121	237	208	223
331	116	117	118	192	208	200
404	128	98	111	243	174	209
420	123	111	118	270	206	238
421	130	121	128	129	112	121
503	123	120	123	204	178	191
506	125	116	121	243	186	215
512	131	112	122	192	198	195
513	124	106	114	241	258	250
514	125	128	128	219	224	222
521	133	120	128	180	150	165
525	122	132	128	210	154	182
529	110	105	108	174	216	195
601	123	134	130	144	177	161

Pupil	IQ Scores			Achievement Scores		
	NV	V	Total	Term 2	Term 3	Average
605	141	123	134	204	146	175
607	132	122	128	231	274	253
608	128	127	131	228	172	200
609	145+	121	134	228	210	219
611	133	128	132	249	221	235
613	133	126	131	216	194	205
614	127	114	121	261	243	252
617	109	94	101	225	232	229
618	126	120	125	270	223	247
619	129	131	132	168	126	147
622	137	113	125	159	166	163
623	123	132	129	192	124	158
626	135	130	135	285	287	286
627	135	129	134	183	188	186
628	132	117	126	192	177	185
629	130	119	126	210	210	210
630	137	118	128	213	223	218

APPENDIX G

IQ and Achievement Scores of 112 Pupils not Accepted by Initial Screening

Pupil	IQ Scores			Achievement Scores		
Number	NV	V	Total	Term 2	Term 3	Average
101	105	97	107	141	124	133
102	93	98	95	110	123	96
103	92	91	91	126	114	120
104	106	101	103	201	128	165
105	96	98	97	132	78	105
106	110	97	103	162	114	138
107	122	106	114	198	168	183
108	102	93	97	111	88	100
109	101	102	101	132	98	115
110	109	103	107	168	138	153
111	118	110	115	171	132	152
112	129	115	124	156	88	122
113	108	103	105	171	94	133
114	112	101	105	153	140	147
115	105	113	110	180	74	127
116	86	91	88	144	76	110
117	110	111	112	147	66	107
118	114	98	106	186	118	152
119	126	111	119	210	196	203
120	118	99	108	168	128	148
121	108	105	108	114	62	88
122	109	99	104	213	152	183
125	107	105	106	168	164	166
126	104	96	100	114	134	124
127	91	91	91	138	116	127
202	110	115	113	153	161	157
203	109	112	111	153	111	132
204	116	108	113	186	91	139
205	114	114	115	159	89	124
206	122	109	116	126	93	110
207	123	121	123	105	81	93
208	119	108	114	168	103	136
209	123	114	120	105	165	135
210	111	123	118	138	87	113
212	118	126	123	111	83	97
213	114	109	112	150	119	135
214	108	100	104	114	83	99
215	125	125	127	222	197	210
216	114	116	116	204	120	162
217	121	116	120	165	175	170
218	122	128	127	195	123	159
219	115	115	116	123	89	106

APPENDIX G (contd)

Pupil	IQ Scores			Achievement Scores		
Number	NV	V	Total	Term 2	Term 3	Average
220	121	110	117	132	135	134
221	100	114	108	183	145	164
223	91	96	94	81	83	82
224	125	106	116	81	93	87
225	117	113	116	145	51	98
226	120	124	124	189	131	160
227	119	102	110	156	159	158
228	115	101	109	138	111	125
229	105	110	108	141	91	116
231	114	108	111	204	137	171
306	119	124	124	210	168	189
309	114	114	116	189	130	160
315	104	109	107	213	180	197
321	104	106	105	201	178	190
322	128	106	116	186	130	158
323	115	114	117	174	198	186
326	103	106	104	174	154	164
329	121	127	128	204	162	183
401	112	103	109	159	168	164
402	101	95	98	147	124	136
403	99	92	95	123	128	126
405	96	103	100	66	a	66
406	104	88	95	120	88	104
407	94	101	97	120	50	85
408	111	99	105	120	50	85
409	118	120	120	156	60	108
410	103	94	98	156	98	127
411	126	109	118	126	114	120
412	112	98	105	131	76	104
413	106	87	87	81	74	78
414	127	110	119	144	118	131
415	105	91	98	171	148	160
416	123	99	111	183	76	130
417	119	106	113	174	118	146
418	113	91	100	147	126	137
419	96	97	96	99	66	83
501	104	100	102	156	124	145
502	114	126	121	195	202	199
504	124	113	119	207	142	175
505	126	124	128	159	150	155
507	104	98	101	156	72	114
508	101	100	101	192	120	156
509	109	103	106	210	190	200
510	120	102	112	123	82	103
511	110	103	107	153	122	138
515	115	102	109	162	148	155
516	110	107	109	198	168	183
517	127	107	113	114	114	114

APPENDIX G (contd)

Pupil	IQ Scores			Achievement Scores		
Number	NV	V	Total	Term 2	Term 3	Average
518	118	115	118	165	118	142
519	107	87	91	171	106	139
520	90	91	91	102	82	92
522	117	112	116	165	80	123
523	111	115	115	162	156	159
524	107	111	110	168	190	179
526	120	108	115	195	126	161
527	121	112	118	203	194	199
528	–	–	–	117	182	180
602	114	119	118	171	144	158
603	122	100	111	204	199	202
604	100	110	105	207	168	187
606	122	105	114	a	122	122
610	122	128	126	210	188	199
612	109	101	106	171	146	159
615	114	106	110	168	113	141
616	116	118	118	225	181	203
620	113	109	112	174	128	151
621	125	117	123	147	137	142
624	126	128	129	210	183	197
625	105	102	104	195	161	178
631	118	116	119	138	124	131

APPENDIX H

Statistics on IQ and Achievement Scores for the Initially Screened Group of 58 Pupils

	IQ Scores			Achievement Scores		
	Non-Verbal	Verbal	Total	Term 2	Term 3	Average
Mean	126,33	120,05	124,24	222,64	197,90	210,51
Std Dev	7,77	10,67	7,98	33,75	39,91	33,59

Difference between Verbal and Non-Verbal Means is significant (p = 0,01)
Matched-pairs t test used.

Correlation Coefficients between the various IQ and Achievement scores (N = 58)

	IQ Scores			Achievement Scores		
	NV	V	Total	Term 2	Term 3	Average
NV	1,00					
V	0,26 (p=0,05)	1,00				
Total	0,68 (p=0,01)	0,88 (p=0,01)	1,00			
Term 2	-0,20	-0,12	-0,19	1,00		
Term 3	-0,22 (p=0,05)	-0,21	-0,26 (p=0,05)	0,66 (p=0,01)	1,00	
Average	-0,23 (p=0,05)	-0,19	-0,25 (p=0,05)	0,89 (p=0,01)	0,93 (p=0,01)	1,00

APPENDIX I

Statistics on IQ and Achievement Scores for the Other 112 Pupils in Standard Six

	IQ Scores			Achievement Scores		
	NV	V	Total	Term 2	Term 3	Average
Mean	111,95	106,86	109,96	159,17	124,98	141,67
Std Dev	9,70	9,99	9,51	35,19	38,84	33,99

Difference between the Non-verbal and Verbal means is significant (p = 0,01)
(A matched-pairs t-test was used)

Correlation Coefficients between the various IQ and Achievement scores (N = 112)

	IQ Scores			Achievement Scores		
	NV	V	Total	Term 2	Term 3	Average
NV	1,00					
V	0,60 (p=0,01)	1,00				
Total	0,87 (p=0,01)	0,91 (p=0,01)	1,00			
Term 2	0,31 (p=0,01)	0,38 (p=0,01)	0,38 (p=0,01)	1,00		
Term 3	0,25 (p=0,01)	0,31 (p=0,01)	0,32 (p=0,01)	0,65 (p=0,01)	1,00	
Average	0,32 (p=0,01)	0,38 (p=0,01)	0,39 (p=0,01)	0,90 (p=0,01)	0,91 (p=0,01)	1,00 (p=0,01)

APPENDIX J

Various Correlations on Pupils Scoring Highly on IQ or in Mathematics Achievement

Pupils with Non-Verbal Scores $\geq$ 130 (N= 21)

Correlation : Non-Verbal/Term 2 Achievement Scores : r = 0,12
Non-Verbal/Term 3 Achievement Scores : r = -0,01
Non-Verbal/Average Achievement Scores : r = 0,05

Pupils with Verbal Scores $\geq$ 130 (N = 10)

Correlation : Verbal/Term 2 Achievement Scores : r = 0,19
Verbal/Term 3 Achievement Scores : r = 0,03
Verbal/Average Achievement Scores : r = 0,12

Pupils with Total IQ Scores $\geq$ 130 (N = 17)

Correlation : Total/Term 2 Achievement Scores : r = 0,25
Total/Term 3 Achievement Scores : r = -0,21
Total/Average Achievement Scores : r = 0,26

Pupils with Average Achievement Scores $\geq$ 1 SD above Mean (N = 31)

Correlation : Average/Non-Verbal Scores : r = 0,13
Average/Verbal Scores : r = -0,00
Average/Total Scores : r = 0,07

None of these correlations are significant.

APPENDIX K

Different Style Questions Test and Multiple-Choice Mathematics Test for Standard 6

Instructions given to Invigilators

Please note the following:

1. Order of tests:
   a) Different Style Questions : Time 30 minutes
      Please collect scripts immediately pupils have finished writing.
   b) Multiple-Choice Test : Time 30 minutes
2. Before the M-C test is written please write the following on the board:
   Helpful information : Circle of radius r : Area = πr^2
   Circumference = $2\pi r$
   Sum of the angles of a triangle = 180°
3. After the M-C test has been written please ask the pupils to note down on the front page which M-C questions they have not answered.
   Then when this has been completed give them one minute to mark any one alternative in each of the questions not answered. Tell them that any correct answer thus obtained will result in increased marks for them.

APPENDIX L

Standard 6 Different Style Questions Test

Instructions to candidates:

1. Try to answer as many of the seven questions given as you can.
2. I do not expect to see a neat answer paper, because you are going to have to doodle in order to get to the answers.
3. Please do all your working on the paper provided preferably in the space provided. (Space was left after each question).
4. Please make sure that your answer is clear.
5. Make sure your name is on the answer sheet.

QUESTION 1

Three golfers named (believe it or not) Tom, Dick and Harry are walking to the clubhouse. Tom, the best golfer of the three always tells the truth. Dick sometimes tells the truth while Harry, the worst golfer, never does.

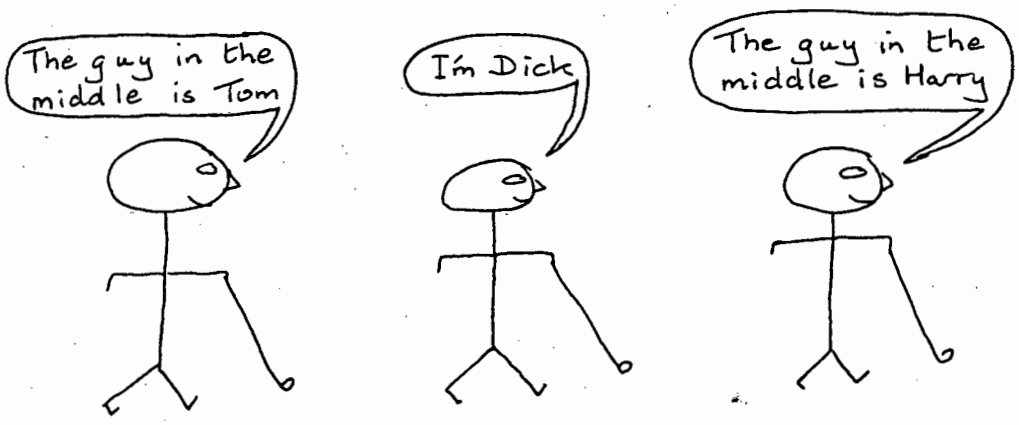


Figure out who is who and explain how you know.

APPENDIX L (contd)

Question 2

Albrecht Dürer, a German artist of the 16th century made a famous engraving entitled "Melancholy", which contains an interesting square of numbers. In this copy of the square three of the numbers have been left blank. Please calculate them.

16	3	2	13
5	10	11	8
9	6		12
	15	14	

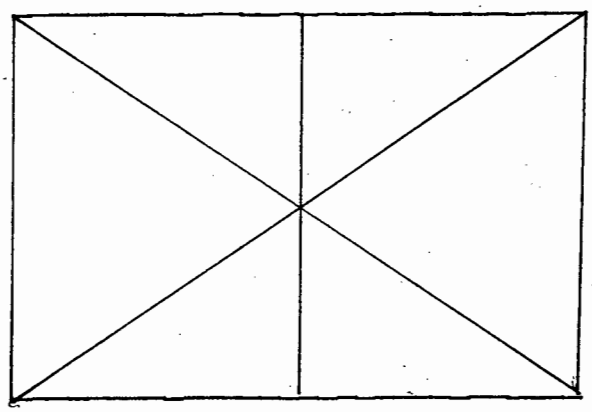
Question 3

Place one plus sign and two minus signs in any order between the numbers to make the left hand side of the equation equal to 15.

4 2 4 1 2 4 = 15

Question 4

How many triangles do you see in the following figure:



APPENDIX L (contd)

Question 5

There are 8 people in a room. Each person shakes hands with each of the other people once and only once. How many handshakes are there?

Question 6

Fill in the next three numbers in the following sequence:

1; 1; 2; 3; 5; 8; 13; 31; _; _; _

What connection do the answers to the following have with the above sequence?

$$1^2 + 1^2 =$$
$$1^2 + 1^2 + 2^2 =$$
$$1^2 + 1^2 + 2^2 + 3^2 =$$
$$1^2 + 1^2 + 2^2 + 3^2 + 5^2 =$$

Question 7

How many squares are there in a normal chess or draughts board?

Memorandum of Marking

1. Correct order : Dick Harry Tom - 3 marks
   Explanation - 2 marks (5)
2. Row 3 Column 3 7
   Row 4 Column 1 4 Row 4 Column 4 1
   One answer correct -3 marks
   Two answers correct - 4 marks
   Three answers correct - 5 marks (5)
3. 5 marks or zero (5)
4. 5 marks or zero (5)
5. 4 marks $(7 + 6 + 5 + 4 + 3 + 2 + 1 = 28)$ 1 mark
   (OR $\frac{8 \times 7}{2} = 28$) (5)
6. 34; 55; 89 2 marks
   2; 6; 15; 40 1 mark
   1 x 2; 2 x 3; 3 x 5; 5 x 8 2 marks (5)

APPENDIX L (contd)

7. Statement of what must be calculated 3 marks
calculation 2 marks (5)

TOTAL 35

Average results for each question

Q1	61%	Q5	33%
Q2	66%	Q6	46%
Q3	41%	Q7	5%
Q4	66%		

Overall average 45%

APPENDIX M

Standard 6 Multiple-Choice Mathematics Test

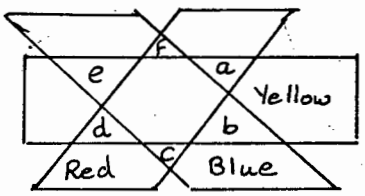
Instructions to candidates

1. Please cross the correct alternative only on the answer sheet
   eg. Question
   53 A B C D E
2. Please use the paper provided for any rough work.
3. Please hand in all the sheets including the question paper.
4. Make sure your name is on the answer-sheet.

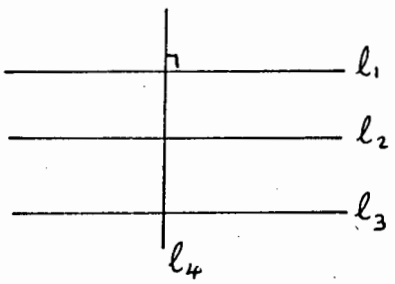
QUESTIONS

1. If $x^2 - 1 = y$ and $x = 3$, then $y^2 =$
   (A) 81 (B) 64 (C) 9 (D) 8 (E) 4

2. Red + Blue = Purple
   Red + Yellow = Orange
   Blue + Yellow = Green



   The figure shows strips of coloured glass that overlap to form other colours as shown by the colour chart. Which two labelled triangular regions would be green?
   (A) e and f (B) e and b (C) f and c (D) a and d
   (E) b and d

3. 

   In the given figure if $l_1 // l_2$, $l_2 // l_3$ and $l_1 \perp l_4$ which of the following statements must be true?
   I $l_1 // l_3$ II $l_2 \perp l_4$
   III $l_3 \perp l_4$
   (A) none (B) I only (C) I and II only
   (D) II and III only (E) I, II and III

APPENDIX M (contd)

4. If $100 \leq k \leq 400$ and k is a multiple of 5; 6; 7 and 10 then k =
(A) 105 (B) 150 (C) 210 (D) 300 (E) 350

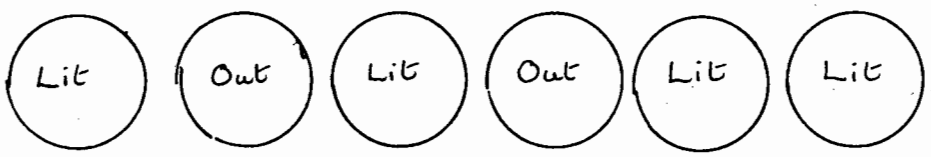
5. In a restaurant where the sales tax on a R4,00 lunch is R0,24, what will be the sales tax on a R15,00 dinner?
(A) R0,60 (B) R0,75 (C) R0,90 (D) R1,20
(E) R1,74

6. If x is a positive interger and $\sqrt{x^2 + x} = n$, which of the following could be the value of n?
(A) 14 (B) 15 (C) 18 (D) 23 (E) 30

7. The non-zero numbers shown form a triangular array. Beginning with the second row, each non-zero number in a row is the sum of the two numbers nearest to it in the row immediately above. If a sixth row is added in this fashion, what will be the sum of all the numbers on the sixth row?

1st row			0	1	0		
2nd row		0	1	1	0		
3rd row		0	1	2	1	0	
4th row		0	1	3	3	1	0
5th row	0	1	4	6	4	1	0
6th row							

(A) 8 (B) 10 (C) 16 (D) 32 (E) 64

8. 

If one more of the lit signals above were out, what percent of all the signals would be lit?
(A) 25% (B) $33\frac{1}{3}\%$ (C) 50% (D) $66\frac{2}{3}\%$ (E) 75%

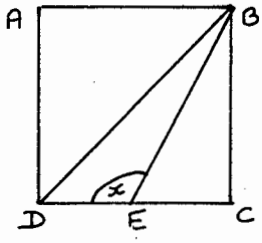
APPENDIX M (contd)

9. If 111 111 + N = 181 111 then N =
(A) 7×10^3 (B) 7×10^4 (C) 7×10^5
(D) 8×10^4 (E) 9×10^4

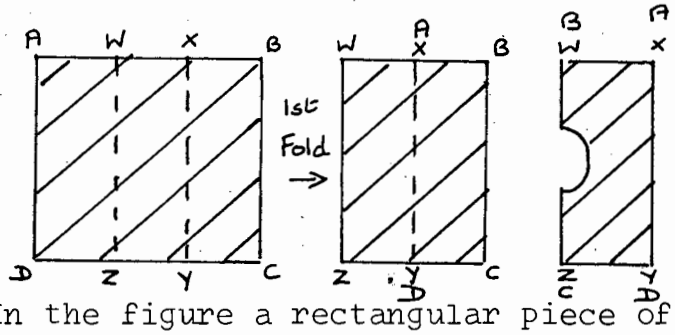
10. In the USA in a certain year, food production per person was 15% greater than food consumption per person. If the average daily consumption per person in the USA in that year was 3 000 calories, what was the average daily production (in calories). per person in that year?
(A) 3 200 (B) 3 450 (C) 3 600 (D) 3 850
(E) 4 500

11. 0,06 is the ratio of 6 to
(A) 1 000 (B) 100 (C) 10 (D) $\frac{1}{10}$ (E) $\frac{1}{100}$

12. In the figure ABCD is a square. If BE = 2EC and $B\hat{E}D = x$, then x =
(A) 100 (B) 110 (C) 120 (D) 150 (E) 160



13.

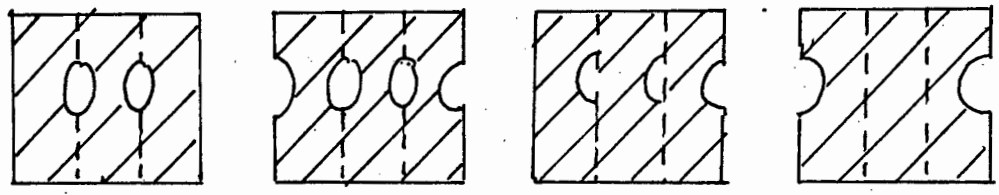


In the figure a rectangular piece of paper ABCD is folded along dotted line WZ so that A is on top of X and D is on top of Y and then folded along XY so that B is on top of W and C is on top of Z. A small semi-circle with diameter on BC is cut out of the folded paper. If the paper is unfolded

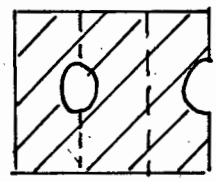
APPENDIX M (contd)

which of the following could be the result?

(A) (B) (C) (D)



(E)



14. If $5x - 3y = 8$ and $x = \frac{7y}{5}$, then $y =$

(A) $\frac{5}{4}$ (B) $\frac{8}{5}$ (C) 2 (D) $\frac{8}{3}$ (E) 4

15. Amy is twice as old as Bill. Five years ago she was three times as old as Bill was then. How old is Bill now?

(A) 20 (B) 15 (C) 10 (D) 5

(E) It cannot be determined from the information given.

16. $(3x^2 - 4x + 7) - (2x + 1)(x - 5) =$

(A) $x^2 + 5x + 12$ (B) $x^2 - 5x + 12$

(C) $x^2 - 5x + 2$ (D) $5x^2 - 13x + 2$

(E) $5x^2 + 13x + 12$

17. In $\triangle ABC$ the ratio of the length of side AB to the perimeter is 1 to 3. What is the ratio of the length of side BC to the perimeter?

(A) $\frac{1}{4}$ (B) $\frac{1}{3}$ (C) $\frac{5}{12}$ (D) $\frac{1}{2}$

(E) cannot be found from the information given

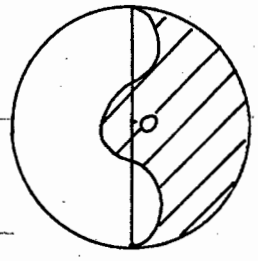
APPENDIX M (contd)

18. $\frac{(N-2)(N-4)(N-6)(N-8)-1}{2}$ is in interger

if N =

(A) 1 only (B) 2 only (C) 9 only (D) any odd integer (E) any even integer

19. Three equal semi-circles are drawn on a diameter of the circle with centre O as shown. If the area of the circle is 9π, then the area of the shaded region is:



(A) $\frac{7\pi}{2}$ (B) 4π (C) $\frac{9\pi}{2}$ (D) 5π (E) $\frac{11\pi}{2}$

20. If each angle of a quadrilateral ABCD is less than 180°, and if 3 of its angles are each x°, which of the following must be true?

(A) $x > 60$ (B) $x = 60$ (C) $x < 60$

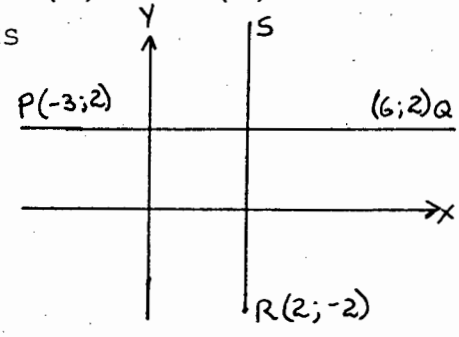
(D) ABCD is a parallelogram but not a square

(E) ABCD is a square

21. The set P consists of all numbers which are the sum of 3 consecutive prime numbers eg. the number 109 is in P because 31 + 37 + 41 = 109. The least prime number in P is:

(A) 13 (B) 17 (C) 19 (D) 23 (E) 31

22. The co-ordinates of points P, Q and R are shown in the figure. If PQ = RS and PQ ⊥ RS what are the co-ordinates of S



(A) (2;6) (B) (2;7)

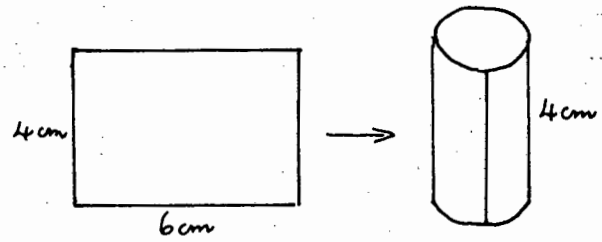
(C) (2;8) (D) (2;9)

(E) (7;2)

APPENDIX M (contd)

23. How many <u>minutes</u> will it take a rocket to travel 4 000km if its average speed is 100km every t seconds?
(A) $\frac{2t}{3}$ (B) $\frac{3t}{2}$ (C) $\frac{2}{3t}$ (D) 40t (E) 2 400t

24. If the average of x, y and 80 is 6 more than the average of y, z and 80, what is the value of x - y?
(A) 2 (B) 3 (C) 6 (D) 18 (E) cannot be determined

25. 

The figure shows a rectangular piece of paper rolled to form a cylinder. If there is no overlap what is the area of the circular base in cm?
(A) 16π (B) 9π (C) 4π (D) $\frac{9}{\pi}$ (E) $\frac{4}{\pi}$

APPENDIX M (contd)

Multiple-Choice Answer Sheet (Memorandum)

NAME: _____

Question					
1	A	(B)	C	D	E
2	A	(B)	C	D	E
3	A	B	C	D	(E)
4	A	B	(C)	D	E
5	A	B	(C)	D	E
6	A	B	C	D	(E)
7	A	B	C	(D)	E
8	A	B	(C)	D	E
9	A	(B)	C	D	E
10	A	(B)	C	D	E
11	A	(B)	C	D	E
12	A	B	(C)	D	E
13	A	B	C	D	(E)
14	A	B	(C)	D	E
15	A	B	(C)	D	E
16	(A)	B	C	D	E
17	A	B	C	D	(E)
18	A	B	C	(D)	E
19	A	(B)	C	D	E
20	(A)	B	C	D	E
21	A	B	C	(D)	E
22	A	(B)	C	D	E
23	(A)	B	C	D	E
24	A	B	C	(D)	E
25	A	B	C	(D)	E

APPENDIX N

Summary of Results Obtained on Multiple-Choice Test

QUESTION	SCREENED GROUP TESTEES (N = 58)		RANDOM SAMPLE OF COLLEGE BOARD EXAMINEES
	NO. OF RANDOM GUESSES	AVERAGE[1] %	AVERAGE %[2]
1	1	74	80
2	1	88	92
3	2	69	84
4	7	69	69
5	1	76	76
6	9	59	68
7	6	76	76
8	3	81	80
9	6	47	63
10	7	41	44
11	3	50	55
12	16	41	52
13	4	59	60
14	17	41	56
15	6	38	39
16	27	22	49
17	17	33	33
18	26	26	38
19	29	17	16
20	30	24	21
21	34	21	20
22	41	29	34
23	38	16	9
24	39	14	10
25	44	14	11

1. Includes results of Random Guessing.
2. Statistics from College Board 1982 p.45

Means of scores calculated on Questions 1 - 15 and 17

Mean of Screened Group Testees = 59%

Mean of Sample of College Board Examinees = 64%

The difference between these two means is significant ($p = 0,01$)

(A matched-pairs t-test was used)

APPENDIX N (contd)

Means of Scores Calculated on Questions 1 - 25

Screened Group Testees:

Mean including Random Guesses = 45%

Mean excluding Random Guesses = 39%

College Board Examinees (Random sample)

Mean (not adjusted for guessing) = 49%

Reliability of Tests (N = 58)

Different Style Questions Test r = 0,19

(Kuder-Richardson formula adapted for essay scores)

Total number of Answers = 406

Full score obtained in 130 answers

Zero score obtained in 176 answers

Part score obtained in 100 answers

Multiple-Choice Test (K-R 20 used)

Reliability on all 25 questions r = 0,36 (N = 58)

Reliability on questions 1 - 15 and 17 r = 0,50 (N = 58)

APPENDIX O

Item Analysis on Multiple-Choice Test

N = 58 Upper 25 pupils : Those scoring more than 11
Lower 25 pupils : Those scoring less than 11
x denotes correct answer

Q1

a	x b	c	d	e	
0	21	1	3	0	U25
1	16	0	5	3	L25

p = 0,74
Item accepted

Q2

a	x b	c	d	e	
0	24	0	1	0	U25
1	20	1	3	0	L25

p = 0,88
Item rejected because of high P-value

Q3

a	b	c	d	x e	
0	1	1	0	23	U25
3	6	0	3	13	L25

p = 0,72
Item accepted

Q4

a	b	x c	d	e	
1	1	20	2	1	U25
5	2	14	2	2	L25

p = 0,68
Item accepted

Q5

a	b	x c	d	e	
0	1	22	2	0	U25
3	3	16	2	1	L25

p = 0,76
Item accepted

Q6

a	b	c	d	x e	
0	1	5	1	18	U25
2	6	1	3	13	L25

p = 0,62
Item accepted
Alternative c should be checked

APPENDIX O (contd)

Q7

a	b	c	x d	e		
2	1	1	21	0	U25	p = 0,74
1	2	4	16	2	L25	Item accepted

Q8

a	b	x c	d	e		
0	1	24	0	0	U25	p = 0,78
1	3	15	5	1	L25	Item accepted

Q9

a	x b	c	d	e		
6	17	0	2	0	U25	p = 0,50
4	8	5	4	3	L25	Item accepted

Q10

a	x b	c	d	e		
2	13	1	1	8	U25	p = 0,40
5	7	1	1	11	L25	Item accepted

Q11

a	x b	c	d	e		
0	16	1	1	7	U25	p = 0,48
2	8	3	1	11	L25	Item accepted

Q12

a	b	x c	d	e		
3	1	17	3	1	U25	p = 0,44
5	7	5	5	3	L25	Item accepted

Q13

a	b	c	d	x e		
3	3	3	0	16	U25	p = 0,58
4	5	0	3	3	L25	Item accepted

Q14

a	b	x c	d	e		
1	7	11	5	1	U25	p = 0,42
2	5	10	4	4	L25	Item accepted

APPENDIX O (contd)

Q15

a	b	x c	d	e	
5	2	15	1	2	U25
1	4	6	6	8	L25

p = 0,42

Item accepted

Q16

x a	b	c	d	e	
6	3	10	5	1	U25
5	5	5	5	5	L25

p = 0,22

Item rejected because distribution for distractor c is better

Q17

a	b	c	d	x e	
0	13	0	1	11	U25
4	13	0	3	5	L25

p = 0,32

Item accepted because distribution for e better than for distractor b

Q18

a	b	c	x d	e	
4	0	5	7	9	U25
3	3	6	4	9	L25

p = 0,22

Item rejected. Distractor e attracts more correct answers

Q19

a	x b	c	d	e	
6	5	8	5	1	U25
4	4	10	7	0	L 25

p = 0,18

Item rejected. Correct answer does not discriminate adequately between upper and lower groups

APPENDIX O (contd)

Q20

a (x)	b	c	d	e	
6	4	9	5	1	U25
5	3	8	7	2	L25

p = 0,22

Item rejected
Correct answer does not discriminate adequately between upper and lower groups

Q21

a	b	c	d (x)	e	
8	4	3	7	3	U25
6	5	2	3	9	L25

p = 0,20

Item accepted.
Distractor a should be checked

Q22

a	b (x)	c	d	e	
3	8	6	4	4	U25
4	7	8	3	3	L25

p = 0,30

Item rejected.
Correct answer does not discriminate between upper and lower groups

Q23

a (x)	b	c	d	e	
6	0	3	11	4	U25
1	4	2	15	3	L25

p = 0,14

Item rejected.
p-value too low

Q24

a	b	c	d (x)	e	
3	6	7	3	6	U25
4	5	6	5	5	L25

p = 0,16

Item rejected.
Lower group scores better than upper group scores on distractor 'd'

APPENDIX O (contd)

Q25

a	b	c	x d	e	
2	3	9	5	6	U25
3	6	5	2	9	L25

p = 0,14
Item rejected.
P-value too
low

APPENDIX P

Scores obtained by Pupils on Problem-Solving and M-C Tests

Pupil	Problem-Solving Test	M-C Test (exc. guessing)
123	10	7
124	8	8
201	22	13
211	26	13
222	22	10
230	14	12
301	16	6
302	20	7
303	19	9
304	17	13
305	19	8
307	25	12
308	14	7
310	17	9
311	12	6
312	20	13
313	3	6
314	15	9
316	21	11
317	16	8
318	18	9
319	21	8
320	11	11
324	15	9
325	26	5
327	11	8
328	18	12
330	11	5
331	16	5
404	16	7
420	17	8
421	7	12
503	8	11
506	7	7
512	11	9
513	11	9
514	5	9
521	13	7
525	17	11
529	12	10
601	22	8
605	10	10
607	27	14
608	22	13
609	23	11

APPENDIX P (contd)

Pupil	Problem-Solving Test	M-C Test (exc. guessing)
611	25	14
613	22	12
614	11	14
617	21	9
618	12	9
619	14	14
622	10	12
623	11	9
626	20	11
627	15	9
628	13	12
629	13	4
630	23	17

APPENDIX Q

Correlations of IQ Scores, Achievement Scores and Test Scores (N = 58)

	Problems Test	M-C Test
IQ Scores	0,14	0,41
Non-Verbal		(p=0,01)
Verbal	0,19	0,29
		(p=0,05)
Total	0,16	0,43
		(p=0,01)
Achievement Scores		
Term 2	0,22	-0,12
Term 3	0,14	-0,11
Average	0,19	-0,13
Problems	1,00	0,33
		(p=0,01)
M-C Test		1,00

APPENDIX R

Ranking of IQ, Achievement and Test Scores

Pupil	Intelligence Scores			Achievement Scores			Test Results	
	NV	V	Total	Term 2	Term 3	Average	Problems	M-C
123					21	24		
124	29x			9	10	9		
201	10	2	2				8	6
211	3	25	10				2	6
222	6	20	10				8	24
230	15		26					11
301	22	4	4	3	7	3	27	
302				11		25	16	
303				14	18	16	19	27
304	22	1	1	25		29x	23	6
305				28			19	
307		15		19		29	4	11
308	10	25	19					
310		3	14	7		28	23	27x
311		12		19	29	21		
312		25			28		16	6
313		22			22			
314		15	26	19	3	5		27x
316	19	15	14	24		26	13	18
317				14	18	16	27	
318		5	14				21	27x
319	25			12	8	10	13	
320	29	20	26	22	12	15		18

Pupil	Intelligence Scores			Achievement Scores			Test Results	
	NV	V	Total	Term 2	Term 3	Average	Problems	M-C
324				9	17	12		27
325				6	6	4	2	
327	9			25				
328				1	5	2	21	11
330				22	25	19		
331					25		27x	
404	25			14			27	
420				4	27	12	23	
421	19	25	19					11
503								18
506				14		26		
512	18							27x
513				18	4	8		27x
514		12	19		13	20		27x
521	10		19					
525		7	19				23	18
529					18			24
601		5	14				8	
605	2	22	4					24
607	15	24	19	25	2	6	1	2
608	25	15	10	28x			8	6
609	1	25	4	28	22	22	6	18
611	10	12	8	12	16	14	4	2
613	10	19	10				8	11
614	28			7	9	7		2
617					11	18	13	27x
618	29x			4	14	11		27
619	22	9	8					2

Pupil	Intelligence Scores			Achievement Scores			Test Results	
	NV	V	Total	Term 2	Term 3	Average	Problems	M-C
622	4							11
623		7	18					27x
626	6	10	3	2	1	1	16	18
627	6	11	4					27x
628	15		29x					11
629	19		29		22			
630	4		19		14	23	6	1

x Excluded using Random Number Tables

APPENDIX S

Number of Common Rankings of Pupils According to Two Different Sets of Scores

	NV	V	Total IQ	Term 2	Term 3	Average	Problems	M-C Test
NV	29							
V	18	29						
Total IQ	22	25	29					
Term 2	12	12	9	29				
Term 3	10	11	10	20	29			
Average	10	12	10	24	22	29		
Problems	15	18	16	17	15	16	29	
M-C Test	20	18	17	13	13	13	17	29

From the information in each of the above cells a 2 x 2 table can be drawn up.

APPENDIX S (contd)

e.g. Non-Verbal/Verbal cell

$f_{observed}$

Verbal		Non-Verbal: Ranked	Non-Verbal: Not Ranked
	Ranked	18	11
	Not Ranked	11	18

$f_{theoretical}$

Verbal		Non-Verbal: Ranked	Non-Verbal: Not Ranked
	Ranked	14,5	14,5
	Not Ranked	14,5	14,5

χ^2 = 2,484 (not significant)

Significant distributions

	Ranked	Not Ranked
Ranked	19	10
Not Ranked	10	19

χ^2 = 4,412 p = 0,05

Significantly the same

	Ranked	Not Ranked
Ranked	10	19
Not Ranked	19	10

Significantly different

	Ranked	Not Ranked
Ranked	20	9
Not Ranked	9	20

χ^2 = 6,896 p = 0,01

Significantly the same

	Ranked	Not Ranked
Ranked	9	20
Not Ranked	20	9

Significantly different

APPENDIX S (contd)

Pairings where pupils ranked are:

Significantly the same		Significantly different	
Non-Verbal/ Total IQ	p = 0,01	Non-Verbal/ Term 3	p = 0,05
Non-Verbal/ M-C Test	p = 0,01	Non-Verbal/ Average	p = 0,05
Verbal/ Total IQ	p = 0,01	Total IQ/ Term 2	p = 0,01
Term 2/Term 3 Achievement	p = 0,01	Total IQ/ Term 3	p = 0,05
Term 2/Average Achievement	p = 0,01	Total IQ/ Average	p = 0,05
Term 3/Average Achievement	p = 0,01		

The only significantly the same result which has any educational significance is the Non-Verbal/M-C test which agrees with the correlation r = 0,41 in Appendix Q.

APPENDIX T

Best Rankings of IQ/M-C test, Achievement and Problems Test Scores (N = 58)

Pupil	IQ/M-C Scores		Achievement Scores		Problems
	Best Ranking	Ranks Ranked	Best Rankings	Ranks Ranked	Test
123			21	29	
124			9	13	
201	2	4			8
211	3	10			2
222	6	18			8
230	11	27x			
301	4	13	3	4	27
302			11	15	16
303			14	20	19
304	1	1	25		23
305			28		19
307	11	27	19	26	4
308	10	24			
310	3	10	7	10	23
311	12		19	26	
312	6	18	28		16
313	22		22		
314	15		3	4	
316	14		24		13
317			14	20	27
318	5	16			21
319	25		8	12	13

APPENDIX T (contd)

Pupil	IQ/M-C Scores		Achievement Scores		Problems
	Best Ranking	Ranks Ranked	Best Rankings	Ranks Ranked	Test
320	18		12	17	
324			9	13	
325			4	6	2
327	9	23	25		
328	11	27	1	1	21
330			19	26	
331			25		
404	25		14	20	27
420			4	6	23
421	11	27x			
503	18				
506			14	20	
512	18				
513			4	6	
514	12		13	19	
521	10	24			
525	7	21			23
529	24		18	25	
601	5	16			8
605	2	4			
607	2	4	2	3	1
608	6	18			8
609	1	1	22		6
611	2	4	12	17	4
613	10	24			8
614	2	4	7	10	

APPENDIX T (contd)

Pupil	IQ/M-C Scores		Achievement Scores		Problems
	Best Ranking	Ranks Ranked	Best Ranked	Ranks Ranked	Test
617			11	15	13
618	27		4	6	
619	2	4			
622	4	13			
623	7	21			
626	3	10	1	1	16
627	4	13			
628	11	27			
629	19		22		
630	1	1	14	20	6

x Excluded by the use of Random Number Tables

Number of Common Rankings of Pupils According to two different Groupings

Test Used : χ^2

IQ/M-C Test/Achievement

	Ranked	Not Ranked
Ranked	9	20
Not Ranked	20	9

Significantly different p = 0,01

APPENDIX T (contd)

IQ-M-C test/Problems Test

	Ranked	Not Ranked
Ranked	19	10
Not Ranked	10	19

Significantly the same $p = 0,05$

Achievement/Problems Test

	Ranked	Not Ranked
Ranked	16	13
Not Ranked	13	16

No significant result

APPENDIX U

Pupils Selected from Initially Screened Group of 58 According to Methods Discussed in Text

NV x Aver (N = 30)	V x Aver (N = 19)	Tot x Aver (N = 26)	8 Sets of Scores (N = 29)	3 Groupings (N = 29)
–	–	–	–	–
124	124	124	124	
201	201	201	201	201
211	–	211	211	211
222	–	222	222	222
230	–	–	–	–
301	301	301	301	301
–	–	–	–	302
–	–	–	303	303
–	304	304	304	304
–	–	–	–	–
–	–	–	307	307
308	–	–	–	–
–	310	310	310	310
–	–	–	–	–
–	–	–	–	312
–	–	–	–	–
314	314	314	314	–
316	–	316	316	–
–	–	–	–	317
–	318	318	–	318

APPENDIX U (contd)

NV x Aver (N = 30)	V x Aver (N = 19)	Tot x Aver (N = 26)	8 Sets of Scores (N = 29)	3 Groupings (N = 29)
319	319	319	319	319
–	–	–	320	–
–	–	–	–	–
325	325	325	325	325
327	–	–	–	–
328	328	328	328	328
–	–	–	–	–
–	–	–	–	–
–	–	–	–	404
–	–	–	–	420
421	–	–	–	–
–	–	–	–	–
–	–	–	–	–
512	–	–	–	–
513	513	513	513	–
–	–	–	514	–
521	–	–	–	–
–	525	–	–	525
–	–	–	–	–
–	601	601	–	601
605	–	605	605	605
607	607	607	607	607
–	–	608	608	608
609	–	609	609	609
611	–	611	611	611

APPENDIX U (contd)

NV x Aver (N = 30)	V x Aver (N = 19)	Tot x Aver (N = 26)	8 Sets of Scores (N = 29)	3 Groupings (N = 29)
613	-	613	613	613
614	614	614	614	614
-	-	-	-	617
618	618	618	618	-
-	619	619	619	-
622	-	-	-	-
-	623	-	-	-
626	626	626	626	626
627	-	627	627	-
628	-	-	-	-
629	-	-	-	-
630	-	-	630	630

APPENDIX V

Statistics Based on Data in Appendix U

Number of pupils chosen by all five methods = 8

Number of pupils not chosen at all = 10

Number of Choices (Common) of Pupils According to Two Different Methods

	NV x Aver N = 30	V x Aver N = 19	Tot x Aver N = 26	8 Sets of Scores N = 29	3 Groupings N = 29
NV x Aver N = 30	30				
V x Av N = 19	12 (not significant)	19			
Tot x Av N = 26	20 (p = 0,01)	17 (p = 0,01)	26		
8 Sets of Scores (N=29)	21 (p = 0,01)	15 (p = 0,01)	24 (p = 0,01)	29	
3 Groupings N = 29	15 (not significant)	13 (not signifi-cant)	19 (p = 0,01)	20 (p = 0,01)	29

From the information in each of the cells a 2 x 2 table was drawn, as in the following example:

APPENDIX V (contd)

		Verbal x Average	
		Accepted	Not Accepted
Total x Average	Accepted	17	9
	Not Accepted	2	30

From this a value for χ^2 was calculated.

The P value or the statement "not significant" in each cell follow from the value of χ^2.

APPENDIX W

Contribution of the Rankings of Each of the 8 Sets of Scores to the Final Selection of 29 Pupils Using this Method

	No. of pupils recommended
Non-Verbal IQ	19
Verbal IQ	20
Total IQ	20
Achievement Term 2	20
Term 3	18
Average	21
Problem-Solving Test	20
Multiple-Choice Test	19

It is obvious that all of the sets of scores made approximately the same contribution to the final selection of pupils.

APPENDIX X

Contribution of the Rankings of Each of the 3 Groupings to the Final Choice of 29 Pupils Using this Method

	Positive	Negative
IQ Scores and M-C test Scores	21	8
Achievement Test Scores	17	12
Problem-Solving Test Scores	27	2

Using a χ^2 test there is a significant difference ($p = 0,01$) between the positive contribution made by the problem-solving test scores and that made by the achievement test scores in the final choice of 29 pupils. There are no other significant differences.

REFERENCES

Aiken, L.R. : "Ability and Creativity in Mathematics" Review of Educational Research Vol 43 1973 pp.405-432

The College Board. Taking the SAT New York, College Entrance Examination Board, 1981

Daurio, S.P. : "Educational Enrichment versus Acceleration : A review of the literature." George, W.C., Cohn, S.J., Stanley, J.C. (eds). Educating the Gifted : Acceleration and Enrichment. Baltimore, Johns Hopkins University Press, 1979. pp.13-63

Fox, L.H. : "Identification and Program Planning : Models and Methods". Keating, D.P. (ed), Intellectual Talent : Research and Development. Baltimore, Johns Hopkins University Press, 1976 pp.32-54

Renzulli, J.S., Reis, S.M., Smith, L.H. : The Revolving Door Identification Model. Mansfield Center, Creative Learning Press Inc., 1981

Stanley, J.C. :"Rationale of the Study of Mathematically Precocious Youth (SMPY) during its First Five Years of Promoting Educational Acceleration." Stanley, J.C., George, W.C. and Solano C.H. (eds). The Gifted and the Creative : A Fifty Year Perspective. Baltimore Johns Hopkins University Press, 1977 pp.75-112

Straker, A. : "Identification of Mathematically Gifted Pupils". Mathematics in School. Vol 9 No. 4 1980. pp.4-8.

Wallach, M.A. and Kogan, N. : Modes of Thinking in Young Children. New York, Holt, Rinehart and Winston, 1965.

AN ACCELERATED PROGRAM FOR A FAST - PACED MATHEMATICS GROUP IN STANDARD SEVEN

David A. Norton

A paper submitted to the Faculty of Education, University of Cape Town, in partial fulfilment of the requirements for the degree of Master of Education

1982

Abstract

The paper briefly considers the principles governing the choice of content and the general approach to teaching of a fast-paced mathematics class. Following this a suggested course is outlined for such a class at the Standard Seven level in a South African school in which acceleration is advocated.

Contents

Four General Principles Governing the Organisation of a Fast-Paced Mathematics Group p.1

Specific Suggestions for an Accelerated Course in Standard Seven Mathematics at a Cape High School p.5

Conclusion p.8

References p.9

Four General Principles Governing the Organisation of a Fast-Paced Mathematics Group

When setting up a program for a fast mathematics group in any standard a number of general principles must receive constant consideration.

The first is that the group generally is more motivated than the average mathematics class, and for this reason it must be assumed that the pupils will interact with each other more readily in the class situation, drawing stimulation from each other. It is essential to realise that even though some clever students can work very well on their own it is a truism that "any student that autonomous and well motivated would probably have little use for school." (Stanley 1977 p.95). Stanley continues: "We have found that stimulation by one's intellectual peers within a homogeneously grouped class which is fast-paced by the teacher produces astoundingly good results for about half of the students enrolled." It must be remembered that his group of Mathematically Precocious Youth are extremely well motivated, all being volunteers, but even with only moderately higher motivation than normal it can be safely assumed that the pupils will stimulate each other. This peer stimulation must be used by the teacher, especially when dealing with new sections of the work. One of the best means is that of the students working in small groups of not more than four pupils giving each other ideas. The teacher must then make certain that all the pupils are making their contribution, and that there are no pupils who are just receiving others' ideas in a passive manner.

The second principle, which overlaps with the first, is that fast-paced groups have to be taught. It is a fallacy to think, because the students are clever, that they can usually learn on their own, receiving advice from the teacher only intermittently. Of course, individual work is very important, but the teacher ought not to have the attitude that the students can teach themselves. Indeed, for a very bright class of this nature the teacher has to be "very well trained in mathematics at a high level in order to be able to challenge and instruct the very gifted children." (Fox 1976 p.47). The teacher must act as an agent of inspiration and must be fully involved with the dynamics of the class as a group, as well as individually.

The third principle is that each pupil needs to know that he is accomplishing meaningful academic enrichment. Irrelevant enrichment, such as the study of side issues in Mathematics, will not hold the interest of the pupils for very long. It is also definite that pupils will be negatively influenced if all that they are given is "busy work." The number of examples that they should do should be sufficient only to ensure the consolidation of a principle or skill - there is no need for overlearning with a group of this nature. A far more motivating experience is the pupils' having the knowledge that they are progressing ahead of their peers. This knowledge will act as a spur, and while working ahead it is likely that the interest in mathematics will increase to the extent that the pupils will gain knowledge of other aspects of the subject. Hence any program should have acceleration as its heart, especially a mathematics program,where the subject matter builds up in such a logical fashion. Speaking for his group of mathematics educators at Johns Hopkins Univer-

sity, Stanley states: "we feel strongly that any kind of enrichment except, perhaps, the cultural sort will, without acceleration, tend to harm the brilliant student." (Stanley 1977 p.93). Khatena sums this idea up very well when he states that there is a fresh concepticalization of acceleration which includes enrichment. (1982 p.303). He continues: "We have also seen this concept directed towards a tightening of enrichment that moves away from the decorative toward the functional, that conceives of enrichment as accelerative and productive, and that requires provisions to maximise individual growth potential." Khatena is fully aware that Stanley's work concerns the upper half per cent of all pupils, but the principles involved are applicable to pupils who are not as talented. For decades Grammar Schools in England had entire classes whose learning was accelerated and who wrote their 'O-levels' a year early. Thus a program for more advanced pupils in mathematics should be an accelerated one, and this is not a fanciful idea on which there needs to be a great deal of expenditure.

The fourth principle is that mental growth and social/emotional adjustment must go hand in hand.

Acceleration raises the possibility of social maladjustment. This was not noticeable in England, but then it can be argued that, as this was accepted practice, there would be none. However, what would happen in a society where this was not the accepted practice? The United States has had a lock-step system for many years and here the attitude has been that acceleration will cause social difficulties for the students involved. Daurio (1977 p.27) in his summary on the literature points out that the warnings about social maladjustment are based more on intuitive than an empirical grounds. He states

that Bonsall in 1955 showed that although very bright accelerated children intially felt some socio-emotional handicaps, they evaluated the accelerative experience positively. This is confirmed by considerable evidence from Terman's longitudinal study wherein it is stated that mental growth and social-emotional adjustment generally went hand in hand, and even though minimal social maladjustment was reported these problems were short-lived. Daurio concludes that there has been excessive concern with this matter and "too little concern about the probability of maladjusting effects resulting from inadequate intellectual challenge." This key point is all too often overlooked.

Acceleration in South Africa's particularly rigid lock-step system must be handled very carefully, because of the ever-present requirements of the Senior Certificate Examination. This examination must, however, not be used as an excuse for a lack of experimentation in acceleration, because with pupils of greater ability a period of one month for revision purposes should be more than sufficient for them to prepare themselves for it. What would be ideal would be for the authorities to allow pupils to write the paper at the end of their Standard 9 year - one wonders whether this revolution would be allowed! Currently all examinations up to Standard 9 are under school control and as long as the syllabus for that standard has been covered at some stage the requirements have been met. The question arises as to what should be taught once the Standard Ten syllabus has been completed. The answer is simple - much university mathematics can be taught profitably to such an advanced group.

With these general principles in mind it is now essential to consider more specifically what could be taught in a fast mathematics class in a South African school.

Specific Suggestions for an Accelerated Course in Standard Seven Mathematics at a Cape High School

It is considered that it would be unwise to start accelerating pupils at the Standard Six level, this being their first year in the high school where much adjustment is necessary. It must also be noted that many new concepts are introduced at this level, specifically introductory algebra and geometry. Any enrichment required should possibly be limited to a study of a specific subject area not in the syllabus which requires a fair amount of experimentation such as Introductory Topology.

The Standard Seven Mathematics syllabus of the Cape Education Department, however, is one which is poor as far as the introduction of new ideas is concerned, basically because it is a syllabus designed for the last year in which all pupils take Mathematics as a subject. Whenever new ideas are handled the treatment is purely introductory and in geometry it is stressed that deductive proofs should not form part of the testing procedure, with the result that all the geometry has to be repeated in a more formal mathematical manner in Standard Eight. This leads to a great deal of overlap with both the Standard Six and Standard Eight syllabuses which provides an ideal situation for meaningful acceleration. Essentially the two year's work can be divided into three parts:

a. Standard 7 material not repeated in Standard 8. It is suggested that this should be completed first, especially as much of it first appears in the Standard 6 syllabus.
b. Material which is dealt with in an introductory fashion in Standard 7 and more fully in Standard 8.

c. Material which appears only in the Standard 8 syllabus which should be covered last, as much of it ties in with Standard 9 work.

Suggested Course

First Quarter

All material covered will be that which appears in the Standard 7 syllabus only. It is work that normally requires much consolidation with the average pupil, but with this type of class it is suggested that the number of examples which the students should be required to do should be limited to that which creates sufficient basic skill consolidation.

1. Sets : Basic revision, use of Venn diagrams and Set-Builder notation.
2. Ratio and Proportion
3. Simple and Compound Interest
4. Basic Algebra
   a. Monomials : Laws of Indices and Square Roots
   b. Polynomials : Definition, degree, order
      Basic Operations
   c. Removal of Brackets
   d. Simple Linear Equations

Second Quarter

In this and the last two quarters the work covered is that in which the Standard 7 and 8 syllabuses overlap. It is proposed that geometry be introduced in an informal manner making use of Standard Seven material, but that a more formal approach linked to the proofs of theorems and riders should be started as soon as possible. In algebra careful grading of the work will be essential in each subject area in order that the transition between the Standard 7 and Standard 8 work will be dealt with smoothly. Different methods of

approach must be tried at all times so that interest can be continually maintained.

Geometry :
1. Angles on a Straight Line
2. Vertically Opposite Angles
3. Parallel Lines
4. Angle Sum of Triangles and other Rectilineal Figures
5. Exterior Angles of Triangles and other Rectilineal Figures

Algebra :
1. Number System
2. Products by Inspection
3. Factorisation - all types commonly used
4. Lowest Common Multiple
5. Algebraic Fractions : a. Simplification
   b. Operations and Fractions

Third Quarter

General comment as for the second term. In the study of Quadrilaterals use will be made of the material developed by de Vries and Human (see list of references).

Geometry :
1. Congruence of Triangles
2. Isosceles Triangles

Algebra :
1. Linear Equations
2. Linear Inequalities. Use of the Number Line
3. Making Formulae and Substitution Therein
4. Simple (?) Problems

Geometry :
1. Quadrilaterals
   a. Properties
   b. Formal Parallelogram Proofs

Fourth Quarter

General comment as for second and third terms.

Geometry;1 Areas and Perimeters of Main Rectilineal Figures and Circles

2. Comparison of Areas of Rectangles, Parallelograms and Triangles
3. Theorem of Pythagoras
4. Polyhedra : Total Areas of Volumes of Pyramids, Prisms, Cylinders and Cones

Conclusion

If the above-mentioned course is followed approximately half of the Algebra and 80% of the Geometry will be completed by the end of the Standard Seven year. This will mean that these pupils should be able to finish the Standard Ten syllabus comfortably by the end of their Standard Nine year. Thus there will be meaningful academic enrichment especially if the teacher makes use of a variety of methods in the presentation of the material.

References

Cape Provincial Administration Department of Education. Syllabus for Mathematics. Junior Secondary Course and Senior Secondary Course (Higher Grade). 1973.

Daurio, S.P. "Educational Enrichment versus Acceleration : A review of the literature." George, W.C., Cohn, S.J. and Stanley, J.C. (eds). Educating the Gifted : Acceleration and Enrichment. Baltimore, Johns Hopkins, 1979 pp.13-63

de Vries, J.A. and Human, P.G. Quadrilaterals : Reconstructive Approach. Cape Town, Shell Educational Services, 1980.

Fox, L.H. "Identification and Program Planning : Models and Methods." Keating, D.P. (ed). Intellectual Talent: Research and Development. Baltimore, Johns Hopkins University Press, 1976 pp32-54

Khatena, J. Educational Psychology of the Gifted. New York, John Wiley and Sons, 1982

Stanley, J.C. : "Rationale of the Study of Mathematically Precocious Youth (SMPY) during its First Five Years of Promoting Educational Acceleration." Stanley, J.C., George, W.C. and Solano, C.H. (eds). The Gifted and the Creative : A Fifty Year Perspective. Baltimore, Johns Hopkins University Press, 1977