RESEARCH STUDIES AND REPORTS ON EXAMINATIONS

Summary of Papers

W.B. Elley: Informal Tests for Classroom Use, emphasizes the need for the school teacher to make his classroom tests as reliable and valid as possible. To do so he must state the purpose of his test, state the objectives of his lessons and his examination and then construct questions tailored to measure these objectives. A variety of question types are listed and some hints are given for constructing improved examinations. Post-examination item analysis is recommended in the pursuit of better assessments.

L.S. Skurnik: Item Banking is put forward as a flexible, economic procedure for assessing the achievement of students in specified subject areas. An item bank is a library of examination questions or items which are found through both expert judgement and experimental evidence to be effective for measuring proficiency on a common scale of achievement although different curriculum objectives have been pursued toward that end. An item bank, it is argued, holds promise as a useful means for expediting curriculum reform, can contribute to better teaching, better examining and, by extension, better services to the community for whom the benefits of education are ultimately intended.

P.J. Hitchman: Examining Spoken English, reviews the problems associated with examining spoken language and describes how better standardization can be obtained among examiners and a more suitable form of examination developed. The central target is to assess what people actually do in the use of language and the newer tests of language include the assessment of reading, conversation, group discussion skills and public speaking abilities. Problems of reliability and validity are in need of research to ensure both accuracy and truthfulness of the examination results.

R.H. Dave and Y.B. Patwardhan: Improving Practical Examinations in Science Subjects, considers many of the shortcomings of practical examinations arising from a poor sampling of prectical work skills, the absence of reliable criteria of assessment and a lack of comparability between exercises as well as examiners. An improved procedure is recommended whereby practical skill is specified in terms of both the process and product of performance. Detailed behaviours are listed and it is suggested that both reliability and validity will be improved by: increasing the number of exercises used, making them more objective, improving the sampling of both abilities and content, and improving both the scoring procedures and the reporting procedures. The encouraging results of a pilot project are described and a number of recommendations are made to improve the effects of science education.

L.D. Mackay: A Study of Optional Questions in Examinations, reveals a number of disturbing results that arise when candidates are offered a choice of questions in an examination. The research reported that candidates can lose 3-10 per cent of marks by selecting a different set of questions to answer from those of their peers of equal ability. Markers apply different standards to different questions and 30 per cent of marks can be lost through a poor choice of questions and an unlucky draw of markers. Less able students tend to attempt the more difficult questions and there is a considerable difference in reliability with which optional questions are marked. Further work is in progress.

K.E. Sinclair: The Influence of Anxiety on Several Measures of Classroom Performance, shows that under experimental conditions anxiety is observed to interfere with factual recall but not with reasoning tasks in an examination. Open-book examinations are recommended where reasoning can be most effectively measured and the effects of anxiety can be minimized. Low anxiety students are found to profit from some tension-producing conditions but the more highly anxious student is found to suffer a reduction in observed performance.

R.E. Traub and H.A. Elliot: Development of the Canadian Scholastic Aptitude Test, describes the development and initial use of the CSAT and discusses briefly its purpose, from and results. The CSAT consists of verbal and mathematical problems and was found to be highly reliable. A parallel test of ability is being developed for Francophone students and plans were underway for equating the two tests with each other as well as with the American SAT.

Ng Fook Kah: The Development of Examination Techniques for Technical Subjects, reports the aims and objectives of trade testing in Singapore schools. With an increasing interest and participation in technical education in the secondary schools, a systematic programme has been developed for specifying both education and examination objectives as well as criteria of performance which will characterize desired, objective levels of achievement.

A.E.G. Pilliner: Testing with Educationally Disadvantaged Children, reveals the difficulties and dubious success in the development of so-called culture-free or culture-fair tests. He argues that efforts would be better devoted to the development of better methods of teaching and testing in language as the most effective way of reducing educational disadvantage. There are a large variety of human talents which may be developed and assessed, aside from those which may be more or less common to western cultures, and educators and examiners might pay due regard to them.

G.J. Matys: Tests and Measurement Procedures, Review and Evaluation, takes the reader on an excursion through the various functions and uses to which tests and examinations are put. He shows how education and the devices used to assess achievement are intimately related and argues that the evaluation function is too important to be postponed until a course of study has been completed. A wide variety of measurement procedures are recommended for use.

S.M.S. Chari: Public Examinations and the Curriculum, discusses the role of examinations in the education system in India and urges the integration of curriculum objectives with examination objectives. The need for continual adaptation of both the school experiences of the pupil and the assessment system used to measure proficiency in learning is affirmed and it is emphasized that the last word on reform of education and examinations has not yet been uttered.

INFORMAL TESTS FOR CLASSROOM USE

W.B. Elley,
Assistant Director
New Zealand Council for Educational Research.

WHY TEST?

In the course of his career, every teacher will have occasion to prepare, administer and mark hundreds of tests of his pupils' attainment. Sometimes these tests will consist of carefully selected formal written exercises with a rigid time limit and an elaborate marking system; sometimes they will be spontaneously constructed, orally presented and evaluated only superficially. But all will have these characteristics which distinguish them from external examinations and standardized tests:

(i) Classroom tests are prepared by the teacher or headmaster for local rather than national use. They are usually prepared at short notice, without the benefit of special expertise, panel review, or pretesting of questions.

(ii) Classroom tests are usually designed to evaluate the pupils' mastery of short units of work recently studied, or objectives which are specific to a school, a class, or a lesson. A teacher may prepare a short test on multiplication and division with decimals, or the causes of World War II, or a list of irregular French verbs. By contrast, an external examination usually evaluates the fruits of a year's study - or more, while a standardized test normally samples basic objectives which are developed over an extended period of time, and are not dependent on the teaching of a particular course.

However sophisticated his test preparation procedures, it should be apparent to all that a teacher requires the results of classroom tests to make decisions about his teaching:- whether to proceed or to back track, whether to change a teaching method, or to introduce a new topic. Judgements have to be made about classification and selection of pupils, advice must be offered about course changes and vocational plans, suitable materials and approaches must be found for children at all levels of ability, pupils must be identified for special treatment. Whether a teacher's decisions are required for groups or for particular individuals, they are more likely to be sound if they are based on accurate information about the abilities and attainments of his pupils. If this information is to be helpful, it should be obtained from tests which are both reliable and valid. Tests which are too easy or difficult, tests which are too short or too long, tests which sample only part of the course, or which weight certain parts too heavily, tests which are ambiguous in their directions, or which leave too much to chance, tests which cannot be marked with reasonable objectivity - such tests may mislead both teacher and pupil, confirm erroneously-held prejudices, and occasionally lead to injustices with far-reaching effects. How can teachers prepare classroom tests which will produce results in which they can place confidence? What test construction methods are likely to produce tests of adequate reliability and validity? First we must examine these criteria of a good test. What does it mean to say that a test is reliable and valid?

RELIABILITY

Tests are reliable if they produce consistent results, if they produce similar marks on different occasions. If a pupil gains 100% in a foreign language dictation test today, and only 50% tomorrow, then the results are not consistent, the tests are not sufficiently reliable to base judgements on. If a pupil is placed first in his class in a test of multiplication and division of decimals on one occasion and is 20th in a subsequent test of the same skills, we can conclude that the tests are not reliable indicators of his ability.

To be reliable a test must normally be long enough to minimize the effects of chance factors in the content and skills included in the test. With a short test, a pupil may be lucky, because he happened to know or guess correctly the few questions that were asked, whereas he knew very little about the areas untouched by the test. A standardized test of reading, mathematics or language, normally requires at least 40 sound objective-marked questions to reach a satisfactory level of reliability. To make decisions about individual pupils, a teacher-made test will probably require more questions than this. For judgements about groups, a teacher may get by with fewer. Just how long a particular test should be depends on the type of material tested, the amount of supplementary information available, and the importance of the decisions being made. Thus a test of a highly specific skill, such as arithmetical computation, or typing, may produce reliable results within ten minutes. If however, we wish to examine a pupil's grasp of mathematical relationships, or his understanding of a period of history and to make decisions about future schooling on the basis of the results, we may wish to extend the test over two hours to gain maximum reliability. For such skills as essay-writing ability, or oral expression, it is commonly found that pupils vary so much in their performance from day to day that the only way to gain adequate reliability is to test the pupils on several topics (over several occasions), and to combine the marks given by two or three independent markers.

Other requirements of a reliable test are clear, precise directions and reasonable time limits. The questions should be unambiguous, neither too easy nor too difficult; they should discriminate well between good and poor pupils, and they should be capable of reasonably objective scoring.

VALIDITY

A good test must be valid. This means that, in addition to measuring a pupil's attainments reliably, it should be relevant to the needs of the tester. It should cover the unit or course adequately, sampling each content area and skill in appropriate proportions. If a teacher knows precisely what his objectives are, he can usually tell, by analysing the questions of a test, whether they conform closely to the objectives he has adopted i.e. whether the test is valid for his purposes.

To illustrate, a 100-item test of mathematical computation may be highly reliable, and yet be quite invalid for measuring achievement in a course of modern mathematics which emphasizes concepts, relationships and reasoning. The objectives of the test do not match the teaching objectives. Again, a test of geography which focusses on isolated details about populations, areas, climate, exports, capital cities and the like, would produce irrelevant results for a teacher who stressed broad concepts, generalized skills and underlying relationships. A valid test of such

objectives may require novel or fictitious situations on which to base questions so that a pupil can demonstrate that he has attained these objectives, regardless of the particular factual details he has acquired.

To ensure maximum validity for his tests, then, it is important for a teacher to spell out, as clearly as possible, precisely what his objectives are, and to build his questions around these, in the appropriate proportions. Tests which develop without such planning often degenerate into factual quizzes of the low-level, isolated, easily testable fragments of the course.

DEVELOPING THE TEST

(i) Once a teacher has decided on the purpose for his test, he should consider the various objectives he has in mind, and how he might best classify them. For a content-oriented course, such as science or history, he might first divide the course into the main content areas, and ensure that each receives a fair ratio of questions. A general science course may be classified into three main areas - say chemistry, physics and biology. A more specific classification, for a biology course, might be living organisms, life processes, conservation, heredity and reproduction, and evolution. In addition a teacher should ensure that questions test different levels of understanding. Some tests concentrate on examining for recall of specific information, some for understanding of important ideas, some for application to new situations, and so on. In a science test, a useful classification system for the objectives might be
   (a) Knowledge of facts and conventions
   (b) Understanding of concepts and principles
   (c) Ability to apply the scientific method to new problems
   (d) Knowledge of industrial applications.

   In language subjects the content areas are less easily defined and it may be more appropriate to classify the objectives of the course according to the skills to be tested - reading, writing, translation, dictation, etc.

(ii) Once objectives are classified, a blueprint or table of specifications can be drawn up which sets out the content areas and the objectives, and allows them to be weighted on some rational basis, before the test questions are prepared. For maximum validity, a test will normally weight most heavily those topics or objectives which have been given most emphasis in the course or unit taught. But all areas should be tested where possible.

An example of a fictitious test blueprint is set out below.

Sample Blueprint for a Mathematics Examination

Objectives	Nos.& Numerals	Measurement	Fractions	Geometry	Total
Knowledge of terms, facts	10	5	5	5	25
Understanding of concepts	10	5	5	10	30
Routine calculations ..	5	5	5	0	15
Application to new problems ..	15	5	5	5	30
Total emphasis ..	40	20	20	20	100

(iii) The third stage in developing the test requires a decision on the form of the questions to be asked. There is no question type ideally suited for all purposes. For instance, short-answer questions which require pupils to fill in the blank or complete a sentence are useful for covering a wide range of facts in a short time. Outside of mathematical subjects they are less useful for estimating depth of understanding without introducing some ambiguity in the question or subjectivity in the marking. Multiple-choice questions are widely used in standardized tests and external examinations because they can sample the whole course widely and efficiently, and test higher objectives, but such questions are not easy for classroom teachers to prepare and they do not examine the ability of the pupil to generate and organize his own ideas. Matching questions are best suited to measuring knowledge of homogeneous sets of facts or conventions. Pupils may be asked to match books with their authors, chemical compounds with formulae, countries with exports, etc. They should not be used however, unless the contents of each list form a homogeneous group, so that each item on one list is a plausible match for each item in the other list. True-false questions may have some value in classroom tests since they enable the teacher to sample widely in a short time, but they are frequently superficial, they are unsuitable if the truth of each statement is not absolute, and they are prone to be unreliable due to guessing on the part of the pupils. They can be modified of course, by requiring pupils to correct false statements, or to classify a statement as "sometimes true", depending on other factors. Perhaps their greatest value is as a starting point for classroom discussion. Essay questions compensate for some of the deficiencies of other question types in that they do require the pupil to express his own ideas, and to demonstrate fluency and organization. However, they cannot measure as many aspects of a course as do short-answer questions, and they are difficult to mark reliably.

Before deciding on the kinds of questions to use then, a teacher should consider the various pros and cons outlined above, in relation to his own expertise in item-writing, the number of pupils involved, the time available for setting and marking, the degree of reliability required and the kinds of decisions to be made with the results.

(iv) Preparing the questions to fit the blueprint is the fourth stage. Here there are many pitfalls, and no short-cuts to success. So often when questions are hurriedly prepared they turn out to be ambiguous, too easy, too difficult, or unsuitable for some other reason. The following checklist may alert teachers to the kinds of weaknesses likely to be found in their questions.

(a) - General

- Keep questions brief, simple and free from complex verbal instructions, double negatives etc.

- Test only important facts and skills; avoid trivia, catch questions, and irrelevant material.

(b) Completion Questions

- Use a single blank in each question.

- Place the blanks near the end of the sentence.

- Ensure that there are a finite number of correct answers.

- Make all blanks approximately the same length.

(c) Multiple-Choice Questions:

- Use only plausible distractors.

- Ensure that there is only one acceptable answer.

- Avoid the stereotyped language of textbooks in the correct answer.

- Beware of grammatical clues and verbal associations which help the uninformed.

- Make the correct option the same length as the distracting option.

- Avoid overlap in the options.

- Avoid any discernible pattern in the correct answers.

(d) Matching Questions:

- Clarify the instructions so that pupils know the basis for matching.
- Use only homogeneous sets of items in each list.
- Make an unequal number of items in each list.
- Use fewer than ten items in each list.

(e) Essay Questions:

- Ensure that the question as it is asked cannot be more adequately measured by another approach.
- Structure the question in such a way that pupils know what to include, what to omit, and how much to write.
- Ask several short questions of different types rather than one long question.
- Avoid giving pupils a choice of questions unless it is absolutely necessary.
- Prepare a model answer before the test, but be prepared to revise it in the light of pupils' answers.

These principles may not always be applicable or even justifiable, but they do point to frequent sources of weaknesses in classroom tests. Such weaknesses can often be overcome by working with a colleague or panel of teachers, by requiring somebody to answer the questions while the test is being prepared, or by pre-testing the questions on a sample of pupils similar in ability to those for whom the test is designed.

There are no perfect paper-and-pencil tests. All are somewhat artificial; all are subject to pupil fluctuation in concentration; all provide only a sample of pupil knowledge; all therefore are to some extent unreliable. A close observance of the principles outlined above, however, should help teachers to polish up their testing procedures. Further improvements can be effected by studying textbooks on the subject, by examining well-constructed standardized tests, by item analyzing one's own tests, and by discussing one's efforts with other teachers.

ITEM BANKING

Larry S. Skurnik
West African Examinations Council

An item bank is a library of examination questions or items which are found through both expert judgment and experimental evidence to be effective for measuring achievement at specified levels of proficiency. An item bank in any one subject would enable examination boards to overcome a number of difficulties in the assessment of the results of their schools. An item bank can serve its patrons by:

(a) storing a collection of high quality questions from which examiners can draw materials for individualized assessment, continuous assessment or terminal evaluation;

(b) supplying evidence (through item and test statistics) upon which standards of achievement can be established and maintained between schools in one country and even between schools in different countries which issue certificates of achievement that are intended to have international acceptance;

(c) promoting curriculum reform, since the patrons are required to carefully review their course objectives and specify, through an examination blueprint, the abilities they wish to develop and assess and the items they judge to be relevant to their course;

(d) trial testing examination materials among a wider field of candidates than is generally available to existing examination boards, giving greater stability and meaning to the items and their operating characteristics;

(e) scoring examinations, storing results, supplying computor processing services and other support to examiners which most boards cannot easily afford on an individual basis;

(f) accumulate a wealth of evidence on the reliability and validity of examinations, on the efficiency and effectiveness of various teaching and examining procedures, and on the associated characteristics of candidates and teachers.

II. BACKGROUND

The idea of developing item banks in one or more subjects is not entirely new, although it has only been explored during the past decade. A pilot study of the full process has been investigated in England through the development of an item bank in mathematics (Wood and Skurnik, 1969). Scriven (1967) discussed the possibilities offered by item banking. Education Testing Service has been storing, retrieving and printing tests through the use of a computor since 1965 (Rock, Epstein and Melton, 1967; Epstein, 1968) and the U.S. Navy has been employing computors to control both the teaching and the examining process in certain technical programs (Katzenmeyer and Swanson, 1968).

The state of Oregon has embarked upon building a Computor-Based Test Development Centre (COMBAT) which aims to produce a number of item banking services (Gage, 1968).

> "The purpose of COMBAT is to establish a computor centre responsive to educator demand in which a large pool of items designed to measure how far students with particular characteristics achieve educational objectives in all curricular areas and grade levels can be stored. Thus, once the information is stored, a teacher will only have to inform the computor of the instructional objectives and the characteristics of the students or student to be tested and an appropriate educational device will be prepared. Hopefully educators will have on demand the best of two testing worlds, the tests will not be subject to the technical weaknesses of most teacher-made tests nor to the content and normative weaknesses of standardized tests. They will be of professional technical quality and based on local objectives and local student characteristics. They might also contain items in a form not now available in any test."

Item banking, although relatively new, is not untried. It holds promise for pooling resources and encouraging progress in examining and education at a level which is seldom considered possible through most other educational innovations. However, all of the work to date has been narrowly focused. The beneficiaries of the systems developed have been limited to some geographical areas, to the subjects chosen for banking such as U.S. Navy technical training or secondary school mathematics in the U.K., and have a limited effect upon national or international development. Item banks in GCE subjects can hasten the development of secondary education and examinations, improve the quality of training offered and maximize the efficiency of examination boards.

III. FEASIBILITY

What types of problems are likely to be encountered in establishing and operating an item bank? Will teachers/examiners be able to produce useful blueprints? Will a sufficient number of high quality items be obtainable? How technically efficient can the bank items be? Can a bank accommodate international differences in educational programs and examinations?

1. Planning the Bank

The work of Wood and Skurnik (1969) and others has shown that item banking is highly feasible if adequate planning is carried out beforehand. One of the first priorities is to meet with representatives of the profession for whom the bank is being designed, to explain what the program is all about and to solicit suggestions for implementation. When sufficient people have been enlisted to cooperate and funds have been obtained from independent sources then working parties can be organized to represent specialist areas where advice and assistance will be required in the development of blueprints and item specifications, item writing and editing and moderation of examination results. The work of these representatives will lead to the production of the bank "deposits".

Technical measurement specialists will need to be engaged in work on both the hardware and software problems of system design to make the bank ready for "deposits, withdrawals and preparation of financial statements."

Meetings should also be convened with established examining boards to work out mutually convenient procedures for the pre-testing of items and administration of examinations. Since an item bank functions as a clearing house of information and not usually as an examining body, the formal registration of candidates, administration of papers, distribution of results, etc. would remain in the hands of the existing bodies, although efforts may be made to simplify and rationalize procedures and forms so that the system operates smoothly. The bank could be able to supply questions, mark papers, help examiners to produce blueprints and content specifications as well as interpret statistics produced by the examinations. The statistics may help to decide questions about pass/fail criteria, accuracy of assessment, etc. and other related problems.

2. Preparing Blueprints

Although the task of specifying subject details and specific terminal behaviours which constitute an examination blueprint is a difficult one, it is not beyond the competence of many teachers. Experience has shown that teachers have problems articulating their own blueprints, but they would have minimum difficulty making choices from a comprehensive blueprint. This blueprint would approximate an exhaustive map of behaviours and subject matter associated with the area of interest and if presented in the form of a menu card, it would allow patrons to check-off their choices without having to labour through the chore of initial preparation of a basic blueprint. The response patterns of the examiners who complete these check-blueprints would effectively determine the type and quantity of items required for the bank. Since examining boards would always be free to add and items, questions or other assessment procedures to any examina assembled with bank material, there should be no fear that the bank will exert an inhibiting influence upon those who are currently engaged in teaching and examining.

3. Producing Items

Existing resources of test and examination items would undoubtedly be of help in the early stages of development of an item bank. However, if a first-class library is to be assembled then serious efforts will need to be devoted to the construction of many new items which are carefully matched to the terminal behaviours expected by the teachers. Extensive workshops will need to be held since:

> "Item writing is an art. It requires an uncommon combination of special abilities. It is mastered only through extensive and critically supervised practice. It demands and tends to develop high standards of quality and a sense of pride in craftmanship." (Ebel, 1951).

4. Pre-testing Items

The pre-testing of items which are general or common to most syllabuses are readily field tested as long as due regard is given to the basic canons of sampling. Items which are relatively uncommon, perhaps measuring esoteric knowledge which is taught in only a few places, will be a bit more difficult to evaluate, since it is essential that an adequate sample be obtained for all tryouts. It is however possible to 'deposit' items in the bank which are so rare that perhaps only one school or center is interested in using them. These items can be embedded in the body of the examination of the one school needing them, but not scored in the first cycle. Subsequent

item analysis and calibration of achievement against other items in this examination or items which may be used as bench marks will help to reveal the quality of even the most rare questions or items in an examination.

5. Technical efficiency of the items

Although the published literature has clearly demonstrated that individual examination items are notoriously unreliable as isolated measures of achievement, it also shows that a carefully chosen set of items, in concert as an examination, can yield very precise estimates of achievement. Even a test of very modest length can produce an accurate assessment of the achievements of groups of candidates although there may have been large differences in syllabus, method of study and methods of assessment. Although it will be essential to evaluate item operating efficiency as parts of a coherent examination, the final aim of the exercise will be to calibrate items against independent criteria which define the terminal behaviours expected from a course of study. These items statistics can be a key to international comparability of standards and are of fundamental importance to the interpretation of the meaning of the test scores.

6. Technical efficiency of bank examinations

Perhaps the most useful results to come out of the few item banking studies that have been conducted is the observation that custom-built tests are at least as efficient as the usual examinations and are often better. They can provide a more comprehensive coverage of a subject in a limited period of examination time, provide very high reliability and validity (when correlated against the marks on traditional but parallel examinations) and, perhaps most importantly, yield additional information about the absolute accomplishments of the candidate as well as the rank order among the candidates entered.

7. International problems

Despite differences in culture, custom and language, tests can be readily translated from one language to another with only a limited change in content. The vital factors are the degree of care taken in the translation and retranslation, and careful review by experts in the subject who are conversant in the two languages, and the accuracy of checks carried out to verify that all the alternatives as well as the questions are relevant to the syllabus.

IV. SUMMARY

An item bank is a flexible, economic procedure for assessing the achievement of students in specified subject areas. It also holds promise as a useful means for expediting curriculum reform. It can serve to certify accomplishments of greater importance than the operation of a motor vehicle, which is virtually the only skill that has a universally recognized licence. It can contribute to better teaching, better examining, and by extension, better services to the community for whom the benefits of education are ultimately intended.

References

Ebel, R.L. 'Writing the test item.' In Lindquist, E.F. ed., Educational Measurement. Washington, D.C., American Council on Education; 1951.

Epstein, M. Computor assembly of tests from an item bank. Paper delivered at the annual convention of the American Psychological Association, San Francisco; 1968.

Gage, G.E. Teaching teachers to write objectives. Unpublished manuscript; 1968.

Katzenmeyer C.G. and Swanson, C.L. An evaluation of SEQUIN, a computorized test construction procedure. Paper delivered at the annual convention of the American Psychological Association, San Francisco; 1968.

Rock, D., Epstein, M. and Melton, R. An exploratory study of computor assembly of a mathematics test. R.M. 67-4. Princeton, N.J.: Educational Testing Service; 1967.

Scriven, M. 'The methodology of evaluation.' In Tyler, R.W. et al. Curriculum Evaluation, Chicago: Rand McNally; 1967.

Wood, R. and Skurnik, L.S. Item Banking. London: National Foundation for Educational Research; 1969.

EXAMINING SPOKEN ENGLISH

P.J. Hitchman,
Department of Education,
University of Nottingham, England.

For centuries the teaching of English in the United Kingdom has been very largely confined to teaching children to write their language and to study its literature from the printed text. Learning to speak their own language has been largely left to chance - the chance of social background. A middle-class house has provided much richer linguistic opportunities than a working-class environment. This has tended to a rigid stratification of class structure, with educational opportunity and career achievement unfairly tilted in favour of the English middle-class.

Now there is a belated recognition of the importance of spoken English in the world of today; and for the first time in our educational history its backing is everywhere being taken seriously.

To test a subject in our schools is to give it importance and to fix its status. Written examinations are part of our educational tradition. Now, in the last few years, we have the novelty of tests in spoken English for our school children. This is giving status to the oral language in our schools.

Official tests are in operation at three levels: General Certificate of Education Advanced Level (within the area of the Joint Matriculation Board), Ordinary Level (organised by the London University School Examinations Department), and at the level of the Certificate of Secondary Education. (C.S.E. examinations are taken in most schools in England that do not take G.C.E. O Level examinations). The J.M.B. 'A' level test in Spoken English can be taken only by those candidates taking General Studies, the London '0' level test by those taking the English Language paper; in neither case is it compulsory. Candidates taking C.S.E. English must take a test in oral English unless excused by reason of speech defect.*

The aims of these various oral examinations can be summed up in the words "communicate, communication". This is implicit in the syllabuses of the 'A' and '0' level authorities and is made explicit in those of most of the 13 C.S.E. boards. Probably all intend "Communication" to be understood as two-way. The North Regional Board (covering schools in the northeast of England) states: "The English Language examination will attempt to ascertain the candidate's ability (i) to communicate clearly with other people and (ii) to understand other people when they attempt to communicate with him, both orally and in writing".

The most important use of speech deliberately made audible is as a means of communication between human beings. The aims of the various

---

* In 1969 approximately 2,000 candidates took the 'A' level test in spoken English, and approximately 20,000 the '0' level test. It is not possible to give the total for the C.S.E. test in spoken English, but in one region alone - the Metropolitan - the figure was 11,000.

examinations clearly recognise this; and their forms of examination are coming to be - as they should - a reflection of these aims.

At first the examining boards played safe. They had to their hand a well-tried test-item in reading aloud. The reading by pupil to class has been a popular form of educational activity in schools from ancient times, and teachers have always felt impelled at intervals to give a mark for attainment. From the inception of the teacher-training system in the 1840s students had been tested in prose reading by Her Majesty's Inspectors on their annual visitations and marks had been awarded. Reading aloud is still a popular classroom activity.

It is thus easy to understand why the Examining Boards included a test of prose reading in their new spoken English examinations. Reading aloud is a good test of a candidate's ability to interpret and present the ideas and words of others. What was also needed was a test of his ability to present his own ideas and words. Thus Conversation came to be chosen as the second item of a ten-minute test. The J.M.B. developed these forms of tests at 'A' level in the early 1950s, the University of Durham Examinations Board used them for several years until its demise in 1964, London has used them for its 'O' level examinations since their inception in 1964, and they are by far the most popular forms of spoken English test with the new C.S.E. authorities, which instituted examinations in 1965 and 1966.

A test composed of these two items has certain advantages. They are nicely balanced with a contrast of oral interpretation and oral composition. They make for a pleasant variety, and together they test a candidate's ability to communicate ideas and feelings to others. They are easy to administer and need not take more than ten to twelve minutes per candidate. Research has shown that in the hands of competent examiners they have a reasonable statistical validity and reliability.

These tests in Reading and Conversation are, in general, private affairs between the single candidate and the examiner, with no other people present. The Boards have been experimenting for some time with examinations in a group situation and, at the same time, with items other than Reading and Conversation. Conversation is, in its nature, talk between two persons. If three or more take part its form and nature are subtly changed; it becomes Discussion. For the last two years Conversation in the J.M.B. tests has been a three-handed affair - it has become a discussion (on any subject chosen by the candidates) involving two candidates and the examiner. The London Board has also been experimenting and proposes to introduce in 1971 (in addition to its single-candidate Conversation) a Discussion involving three candidates and an examiner. In these two examinations all candidates in any one discussion are being tested. In the 13 regional C.S.E. areas Group Discussion is a compulsory part of the oral English test in two areas and optional in five (which means that candidates can take some other option if they wish). Of the seven syllabuses involved two have group discussion in which all candidates are being tested; in the others each candidate is tested separately, talking with the group.

The London Board is retaining Reading in its new 1971 examination, but as a group activity, each candidate being required to read aloud to a group comprised of the examiner and other candidates. In the J.M.B. tests Reading is now also a group activity, but optional to the giving of a Talk. A new feature is that the candidate sits after his reading and answers questions from members of the group on any matter arising. Reading aloud

is still a very important test item in the C.S.E. examinations, appearing in 12 of the 13 syllabuses, sometimes compulsory, sometimes optional; sometimes a private affair between candidate and examiner, sometimes in the presence of a group of other candidates. In three regional areas the reading tests are conducted with the candidate sitting at a table close to the examiner, in others he stands and speaks at a distance; in one or two areas sitting or standing is at the choice of the candidate. (The South East region also tape-records its candidates for purposes of moderation.)

The size of group for the 'A' level J.M.B. tests is six or seven, for the 'O' level London three. This means that each member of a group spends much more time in his examination than if he were tested solo - in the former test one and a half hours (instead of 12-15 minutes), in the latter half an hour (instead of ten minutes). This is excellent for the generation of a group rapport. The size of the C.S.E. groups varies. In the Metropolitan region Group Test the size is twelve. (This test is unique. The candidate introduces the passage to the other eleven in the group, answers questions on it, and then reads it to the group a second time, two marks being given, one for either reading, and one for the quality of his answers.)

The last important development is the institution of the Talk as a test-item. In the J.M.B. tests it is optional to Reading and is delivered to an audience of six or seven (the rest of the group and the examiner). Questions and answers from the candidate follow. The candidate is handed a printed card containing three topics, from which he chooses one. He is allowed a few minutes to prepare his talk and five minutes to deliver it. He can speak from notes. His audience then questions him on matters arising. The Talk is a compulsory element in the C.S.E. test in four regions and optional in four. Where it is compulsory the syllabus states that the candidate will talk on a topic of his own choice. In all areas but one the talk is delivered to an audience. In five of the eight areas, Question and Answer form part of the test.

Thus the tests now in existence comprise Reading, Conversation, Group Discussion, the Talk to an audience. We have moved from the private to the group situation. Tests are coming to be more realistically based; that is to say, they are concerned more with what people actually do in real life situations - they talk to each other singly or in groups, they lead a discussion (or are led), and they stand up and address an audience.

The tests in spoken English most usually taken by candidates for whom English is not the mother-tongue are those for the Cambridge Proficiency Certificates - comprising Prose Reading and Conversation with the examiner. Among the speech elements the examiner is concerned with in those tests are those that help him to answer the very important question - how English does this candidate sound? They are his pronunciation (his use of vowels and consonants), his intonation (or tune-patterns), his articulation, and, perhaps most important of all, his pattern of stressing, a complicated alternation of stressed and unstressed syllables. These are the major elements in the English speech rhythm. In Conversation the examiner is also looking for the use of an acceptable vocabulary, acceptable grammar, word-order and idiom.

Now the English candidate, however poor his speaking in other respects, at least sounds English - he can't help it. So the examiner in the tests in England under discussion is not primarily listening for the way

these speech elements are used. In Reading the examiner is looking for the candidate's ability to interpret the page before him, to communicate to the audience the author's meaning and mood. He is asking the candidate to exercise his imagination as well as his communication skills of voice and speech. In the other test elements - conversation, discussion, the talk - the examiner is considering the candidate's ability to use the English Language efficiently in face-to-face communication - that is, to make a statement clearly, to develop a theme, to rebut an argument, to inform, to persuade. He is, of course, also considering the various aspects of delivery (the use of the voice, dictions, bodily stance and gesture).

It will be seen that judgments made by an assessor about a candidate's speaking are necessarily highly subjective - he has to make the decision as to whether a speaking performance is a good one or a poor one and thus whether to award a high or a low mark. This means that examiners may disagree sharply about the performance of particular candidates. In fact the quality of the examiner is of crucial importance. As much as possible is done by test-designers, by the examining body and by chief examiners to minimise the possibility of disagreement. This is done by "standardising". A rating scale is prepared which shows what qualities are being looked for in a candidate's speech and what mark is to be awarded for the strongly positive presence of a quality and what for the almost total absence of this quality (e.g. CONTENT and ORGANISATION OF TALK: Main points made clearly in a logically developing argument. Content clearly organised to show an introduction, a middle and a conclusion. Material interesting, relevant, sufficient, of good quality - AWARD 7 to 10 marks. Talk badly arranged. No logical development or argument: main points do not stand out clearly. Material of poor quality, uninteresting, irrelevant, insufficient - AWARD 0 to 3 marks). Then there will be a briefing meeting at which the Chief Examiner will take the assistant examiners carefully through the rating scale so as to establish an identity of understanding as to what is intended. It is probable that a few "guinea-pig" candidates will be examined by the Chief and the other examiners so that the latter can have a preliminary experience of both examining and marking, and so that standards shall be set and absorbed. When the examiners are in the field examining the Chief will pay each a visit for a day or half a day and make his own assessments of the candidates. Later these will all be scrutinised and the assistants' assessments raised or lowered or left as they are. It is a chancy business. But research has suggested that the judgments of experienced examiners in spoken English (at least in Reading, Conversation and the Talk) are at least as valid and reliable as those of written essay-type examinations. (There is considerable doubt about the reliability of group discussion assessments).

If examinations (over the whole range of educational activities in school) are to remain as a vital element in assessment it seems certain that the assessment of spoken English will become a "growth" industry. Perhaps the most important problem will concern the calibre and training of potential assessors. (In the C.S.E. testing of spoken English almost all teachers of English are likely to be drawn into assessment.) There will also be the search for new and appropriate test-elements that test achievement in school courses in spoken English, which must themselves be geared to the needs of human beings in adult society. Finally, there will be continuing research into methods of assessment and their attendant problems of validity and reliability.

IMPROVING PRACTICAL EXAMINATIONS IN SCIENCE SUBJECTS*

R.H. Dave and Y.B. Patwardhan

National Institute of Education,
National Centre for Educational Research and Training, New Delhi, India.

Examination reform has now been accepted in our country as a very powerful instrument to improve quality in education. During the implementation of the reform programme particularly in science subjects it was felt that the work in these subjects would be incomplete unless the practical examinations are also reformed. In fact, the sixth conference of chairmen and secretaries of the boards of secondary education in the country held in November 1964 discussed at length the necessity of improving practical examinations and passed a resolution that the boards may take up this work in collaboration with the Central Examination Unit of the NCERT.

It was a source of great satisfaction that soon after this conference, the Board of Secondary Education, Rajasthan, came forward to take up the new venture in the field of examination reform. A series of experimental tryouts were carried out to evolve an improved system of practical examination and then the board implemented the new plan in its higher secondary examination of 1968 in the subjects of physics, chemistry and biology after making adequate preparation in collaboration with the NCERT. We give a brief résumé of the experimental studies conducted for the development of an improved pattern of practical examinations in science subjects and its large scale implementation.

SHORTCOMINGS OF THE PRESENT PATTERN

A qualitative study based on verbal reports by a number of experienced examiners in Rajasthan revealed that the then existing pattern of examination suffered from the following major shortcomings:

1. Poor sampling: Each experiment given being very complex, comprehensive, and time consuming, only a few experiments (e.g., two in physics) could be set in the limited time available. As such it could only measure a small fragment of the content and a few of the many aspects of skill that practical work is expected to develop. This reduced both the validity and reliability of the practical examination. This was very discouraging to the pupil as well as to the teacher, especially on account of a high degree of chance factor operating in such a system.

2. Absence of reliable criteria of assessment: The criteria of assessment were very general and examiners were inclined to assess the performance of students according to varied standards leading to loss of consistency and uniformity.

---

* This article is reprinted, with permission, from the NIE Journal, September 1969.
(Publication Unit, NIE Campus, Sri Aurobindo Marg, New Delhi 16, India).

3. Non-comparability of exercises: The few exercises (e.g., two in physics) each pupil would get, varied very much in complexity and nature of skill involved. It was not justifiable to compare the performance of different candidates as obtained on these differing instruments.

BASES FOR A NEW PATTERN

It was felt that practical tests are more costly and time-consuming, and so they should be used only when other more convenient techniques such as written tests cannot be used. It is in the realm of practical skills in these subjects that written tests are not usable and hence practical tests should essentially be used to measure practical skill although other objectives such as knowledge, understanding or application need not be entirely eliminated. For this purpose practical skill was defined under the heads:

(a) process of performance, and

(b) product of performance.

They were further clarified in each subject to delimit their scope. In physics, for example, they were delimited as follows:

Process of Performance

The pupil

1. selects appropriate apparatus, tools, etc.
2. checks apparatus, tools, regarding their working.
3. detects errors and limitations in the fitting up of apparatus.
4. rectifies errors, if possible, under laboratory situation.
5. cleans apparatus, tools, etc.
6. sets up apparatus, tools, etc.
7. sketches arrangement of apparatus (if necessary, at the outset).
8. prepares and follows a systematic and sequential plan for taking observations.
9. states the principle, formula (explaining the symbols, etc., useful in the experiment).
10. manipulates apparatus, tools, etc., while performing the experiment.
11. measures quantities and reads instruments, apparatus, etc., accurately.
12. takes precautions in handling instruments, substances, etc.
13. makes accurate observations of parts, specimens, processes, etc.
14. records observations and makes calculations where necessary.
15. verifies observations.
16. performs experiments with reasonable speed.
17. performs experiments with reasonable accuracy.
18. performs experiments with neatness.
19. adapts himself with somewhat new and different apparatus in setting novel experiments.
20. explains orally the procedures, principles, etc., involved in the experiments.

Product of Performance

The pupil

1. summarizes observations.
2. calculates and finalizes the results.
3. interprets data and draws conclusions.
4. records experimental procedure and conclusions.
5. dismantles and cleans the apparatus, where necessary.
6. arranges the apparatus, substances, etc., at their appropriate places at the end of the work.

Sessional Practical Work

The practical exercises performed by pupils in the higher secondary classes are recorded in specially developed record books. The skills and traits attained while performing these may also be evaluated in board examinations. It may not be possible and also not desirable to evaluate all the traits developed, but a few like completeness, neatness and regularity may be evaluated with the assistance of the subject teacher.

Some of the skills that may be appraised from this aspect of practical work are:

1. Drawing diagrams and sketches from observed facts.
2. Collecting specimens like that of ores, minerals, crystals, etc.
3. Displaying material collected.
4. Improvising simple apparatus.
5. Constructing models.

DEVELOPMENT OF A NEW PATTERN

For the purpose of improving validity and reliability the practical examinations were modified in the following respects:

1. Increasing the number of exercises: Instead of giving few long exercises, many short exercises are introduced, e.g., in physics one major comprehensive experiment is retained and the other is replaced by four or five short exercises. The maximum marks and the time, however, are kept the same.
2. Making the exercises objective-based: Exercises are to be set to test predetermined specific aspects of skill (or understanding) as laid down in the objectives. As mentioned earlier they include within the process and the product of performance. This tends to improve the validity of the practical examination.
3. Improving the sampling of abilities and content: Increase in the number of exercises enables the test to cover many different abilities as specified under the specifications of the skill objective and also to cover a variety of content areas. This helps in improving the coverage and consequently tends to improve reliability and validity.
4. Improving scoring procedures: Very detailed marking schemes giving minute analytical details of assessment of pupil performance are developed not only for major and short

exercises but also for sessional work and viva. Detailed instructions are developed for the use of examiners and candidates for this purpose. This helps in improving objectivity of scoring and controlling inter-rater reliability by minimizing the variability in scoring by examiners emerging from extraneous factors like personal likes and dislikes.

5. Improving reporting and interpreting procedures: Detailed proformas and instructions for their use are developed for the use of examiners. When these reports would be properly used by schools, they will be able to improve science teaching in many respects.

TRY-OUTS

Four examiners in each subjects of physics, chemistry and biology who were involved in the development of the new pattern tried out these procedures three times in actual situation specially arranged for this purpose. In the first try-out 10 candidates and in the second and third try-outs 15 to 20 candidates were involved in all the three subjects. In all the try-outs the four examiners observed simultaneously and marked independently. The experience of earlier try-out was always invariably used to improve the exercises and refine the scoring schemes of the subsequent try-outs. The results of assessment were then studied and correlations found. The findings in the try-outs of biology are given here. In other subjects similar findings are available.

Comparative Study of Three Sets of Examiner Inter-correlations in Biology Practical Examination in Three Try-outs

A. Averaged across questions

Examiner/Try-out	AB	AC	AD	BC	BD	CD	N
I.	.78	.65	.67	.83	.74	.68	12
II.	.94	.95	.95	.92	.94	.93	15
III.	1.00	1.00	1.00	.99	.99	.99	15

B. Based on total test scores

Examiner/Try-out	AB	AC	AD	BC	BD	CD	N
I.	.89	.88	.59	.78	.82	.63	12
II.	.87	.45	.84	.55	.90	.45	15
III.	.97	.99	.98	.98	.98	.97	15

C. Based on ranking of difficulty indices of questions

Examiner/Try-out	AB	AC	AD	BC	BD	CD	N
I.	.88	.90	.90	.98	.78	.75	9
II.	.93	1.00	.98	.93	.97	.98	9
III.	.98	1.00	1.00	.98	.98	1.00	9

Significance levels of Rhos:

When N = 9: .05 = .600; .01 = .783
When N = 12: .05 = .506; .01 = .712
and When N = 15: .05 = .439; .01 = .623

In this subject, during the first try-out three out of six examiner inter-correlations are not significant at the .01 level but are significant at the .05 level. The agreement among the examiners in Try-out II is at or above .92. All the inter-correlations are substantially increased in Try-out II. Again, in Try-out III, the examiner inter-correlations averaged across the questions have reached unity in 3 out of 6 cases, and in the rest they are at .99. Thus the inter-examiner agreement in biology practical tests reached almost to the optimum as a result of intensive training, practice and development of well-designed scoring procedures.

IMPLEMENTATION

1. Preparation: Encouraged by the above findings, the Rajasthan Board decided to launch the reform programme on a large scale throughout the state. It, therefore, developed brochures in each subject entitled "Improved Pattern of Practical Examination" with the help of the NCERT and the examiners who participated in the try-outs, and circulated them to schools. The board also trained all examiners in new pattern of practical examination in the three science subjects. They were given training in the theory of conducting the examinations through four-day workshops organized for the purpose. They acquired practical experience in conducting the new type practical tests in actual examination, which were specially arranged at various places as a part of the training programme.
2. Implementation: With these and other preparatory steps carefully executed, the board introduced the new pattern in the higher secondary examination of 1968. Its impact on school practices is being closely watched. The preliminary review of the impact has been found to be quite encouraging.

SOME PROBLEMS

During implementation some problems were faced which were already envisaged.

1. Lack of equipped laboratories: Many schools do not have good laboratories. For want of such laboratories it becomes difficult to conduct the examination effectively. This applies to the old pattern of examination also.

   Some laboratories do not have trained assistants. Services of trained assistants are essential.
2. Number of candidates per examiner: This pattern envisages close observation of pupil performance during the period of examination. One examiner cannot obviously cope up with 20 candidates at a time as is the practice in vogue. Perhaps, 10 may be a manageable number.
3. Trained examiners: For some years, till the examiners are acquainted with the new pattern, it will have to be seen that every examiner is fully acquainted with the spirit and technique of the new pattern of examination before he is entrusted with the job.

IMPLICATION

For the efficacy of this new examination co-operation from different agencies will be needed. Some implications to such agencies are indicated below:

Departments of Education and Boards of Secondary Education

1. School laboratories will have to be better equipped.
2. Practical syllabus may be reviewed.
3. Flexible time-tables will have to be permitted.
4. Better inspection and guidance programme will have to be developed.
5. Only qualified and trained examiners will have to be selected.
6. Examiners' reports will have to be scrutinized and the findings reported to schools for action.

Schools

1. More initiative on the part of individual teachers and pupils will be needed.
2. Rigidity of time-tables will have to be reduced.
3. Laboratories should be better equipped.
4. The evaluation data should be used for remediation and improvement.

Teachers

1. Variety of practical activities will have to be designed and organized to develop specific skills among pupils and to discourage the tendency of mechanical repetition of standard experiments.
2. Initiative on the part of pupils should be encouraged.

CONCLUSION

Practical work in science subjects is aimed at the realization of some specific purposes which cannot be otherwise achieved. Practical examinations, therefore, have to be so planned that they measure the degree of success achieved by practical work as a contribution towards the multi-faced development of pupils' innate powers and subsequent achievements. The new pattern suggested here aims at this. It defines the outcomes of the process and the product of performance, stresses the need for developing valid and reliable tools of measurement and builds in ways to evaluate the results of measurement. It also envisages sound feedback procedures to utilize the results of evaluation in improving school practices. It is hoped that given a fair trial this pattern will work as a catalytic agent in making science education a dynamic process and a creative activity in our schools.

A STUDY OF OPTIONAL QUESTIONS IN EXAMINATIONS

Lindsay D. Mackay

Faculty of Education, Monash University, Melbourne, Australia

The case for inclusion of optional questions in an examination is frequently argued on the grounds that a particular examination is not designed to measure whether a student possesses knowledge or facts, but whether he has developed particular abilities or skills which can be assessed independently of the particular question or questions answered. If the assumption that these skills and abilities can be assessed independently of the particular question answered is accepted, the inclusion of optional questions in a public examination allows the teacher greater freedom to develop these skills and abilities in any of a number of sections of the broad subject area, and it allows greater freedom for the individual student to pursue his interests through independent study.

From a measurement point of view there are considerable difficulties in the use of optional questions in an examination. Many people would argue that the use of optional questions in an extended answer examination adds one further source of variability to the subjectivity and inaccuracy that already exist in extended answer examinations. If a student is permitted to choose which questions he wishes to answer, the basis for comparability of scores is considerably weakened, because different students will answer samples of questions which are not comparable in content, abilities or objectives. As a result, the content validity of the examination differs considerably for different students. If a choice of questions is available, the questions answered by a particular student are likely to provide a limited, and perhaps distorted, sample of that student's achievement in the course, because he will tend to choose those questions he is best prepared to answer, and because the availability of options gives him greater opportunity to use materials prepared by others to produce an answer which does not reflect his ability. Rather than being fairer to all students, as some advocates of optional questions in an examination would argue, the opportunity to choose among optional questions may help the poorly-prepared student considerably more than it helps the well-prepared student.

This report is a brief summary of some of the results of a study of the effects of optional questions in examinations. As part of this study, the marks awarded to optional questions in a number of examinations, ranging from the sciences to the humanities, have been analysed in an attempt to determine -

(i) the extent to which optional questions differ in difficulty,

(ii) the extent to which optional questions and different marker interact,

(iii) the extent to which students of different ability select optional questions of different difficulty, and

(iv) the extent to which optional questions differ in their marker reliability.

(i) Differences in difficulty of optional questions

One problem in estimating the difficulty of optional questions is that the groups of students attempting each optional question differ in ability. In an examination which consists of a compulsory section in addition to optional questions, it is possible to use a student's score on the compulsory section as an index of his ability in the subject being examined. It is then possible to calculate the average score on each optional question for candidates in each range of scores on the compulsory section (i.e. in each ability range), and to use these average scores to estimate the average score that would have been obtained on each optional question if all students sitting for the examination had attempted each optional question. If the examination does not contain a compulsory section, the average score obtained by the students on all other questions attempted on the examination can be used in a similar way, as an index of the student's ability in the subject of the examination.

An indication of the differences in difficulty of combinations of optional questions available to students can be gauged from the results in Table 1. In this table, the differences between the estimated average scores on the least difficult combination of questions available to students are given for six examinations.

Table 1: Differences in average scores between least difficult and most difficult combination of questions available to students in six examinations containing optional Questions.

Examination	No. of students used in the analysis	Choice available	Difference in estimated average mark between least and most difficult question combinations (expressed as a percentage of possible marks on optional questions)
1. Grade 11 Physics	4360	3 out of 5	9.7%
2. Grade 12 Physics	4450	5 out of 7	4.7%
3. A Grade 12 Humanity	2230	5 out of 14	3.5%
4. A Grade 12 Humanity	2150	5 out of 13	3.3%
5. Grade 12 History	7410	3 out of 13	6.5%
6. Under-graduate Physics	190	5 out of 11	10.2%

The entries in the table can be regarded as estimates of the differences in average marks that would have been obtained had the same candidate attempted two different combinations of questions, or had two candidates of "equal ability" attempted different combinations of questions. It is apparent that considerable differences in the results on an examination containing optional questions can arise from differences in the difficulty of the options offered, and that these can produce a difference in average marks obtained by candidates of equal ability of up to 10% of possible marks.

(ii) Interaction of examiner and optional questions

It has long been recognized that different markers mark essay questions to different standards. In the analyses of examination papers in this study it is clear that markers do not mark all questions to a consistently hard or easy standard, and there is evidence of interaction between marker and optional question which results in a marker marking different questions to different standards. An idea of the magnitude of the effect of this can be gauged from the results in Table 2, in which are given the maximum observed differences in the average marks that would be obtained by candidates of "equal ability" who attempted different question combinations, and whose scripts were marked by different markers. Results in the table are based on analyses of Examinations 3, 4 and 5 in Table 1. Examination 3 was marked by 15 markers, Examination 4 by 14 markers and Examination 5 by 31 markers. The system of allocating scripts to markers was such that each marker marked every possible question.

Table 2: Maximum observed difference between average scores obtained by students of "equal ability" as a result of a number of sources of variation.

Source of variation	Maximum observed difference between average scores obtained by students of "equal ability" as a result of this source of variation.		
	Examination 3	Examination 4	Examination 5
	%	%	%
Different questions (irrespective of marker)	3.5	3.3	6.5
Different markers (irrespective of question)	6.8	5.1	9.5
Same questions, marked by different markers	7.8	10.0	23.1
Same markers, marking different questions	10.8	9.6	18.8
Different questions, different markers	13.0	14.1	29.9

It is apparent that there is considerable interaction of question and marker and that two students of "equal ability" would be expected to obtain average marks which differed by up to 30% of possible marks, depending on the optional questions they selected and the particular marker to whom their scripts were assigned. The entries in Table 2 are the observed maximum differences in average scores between students of equal ability due to various sources of systematic variation in the marks awarded to students. Some students would experience differences considerably greater than these average values, and other would experience differences considerably less than the average values.

If a statistical correction had been applied to the marks awarded to different questions by different markers, so that students of "equal ability" would be expected to obtain the same average mark on each question irrespective of the marker who marked the question, then a considerable percentage of the students in the above examinations would have received a different final grading.

(iii) Ability of students attempting different question combinations

In an examination which consists of a compulsory section and optional questions, it is possible to use the compulsory section mark as an index of the ability of students who attempt different combinations of optional questions. As described earlier, a measure of the difficulty of the question can be obtained by estimating the average mark that would have been obtained had all students attempted th question. There is evidence that the most difficult question combinations were chosen by a group of students of significantly lower average score on the compulsory section than the group of students who chose the least difficult question combinations. Some of the results obtained are summarized in Table 3 for Examinations 1 and 6 in Table 1.

Table 3: Differences between average scores on the compulsory sections of two examinations for students attempting the least difficult and most difficult combinations of optional questions.

Examination	Difference in estimated average score of least and most difficult question combinations (as percentage of total score on all optional questions).	Average compulsory section score of students attempting least difficult questions minus average compulsory section score of students attempting most difficult questions (expressed as a percentage of possible marks).
	%	%
1	9.7	+21.5
6	10.2	+ 8.6

(iv) Differences in marker reliability of optional questions

Each of the scripts completed by the 7410 students who sat for a Grade 12 History examination (Examination 5 in Table 1) were independently marked by two of the 31 markers who marked scripts for this examination. The consistency with which the same answer was independently marked by pairs of markers for each of the 13 optional questions on this examination is expressed in three ways in Table 4. Firstly, for each question, the correlation between the marks awarded by two markers to the same answer is given. The mark/remark correlation values in the Table vary from .53 to .67 for the 13 questions on this examination. Secondly, the percentage of variance in the marks awarded to an answer by one marker which is common to the mark awarded by the second marker is given. This percentage represents the percentage of common variance in the marks awarded to the same answer by two markers. The percentage of the variance of scores which is unreliable variance or error variance can be obtained by subtracting the percentage of common variance from 100%. In the Table, the percentage of variance which is common to the two marks awarded to answers on each question varies from 29% to 46%; that is, the percentage of variance of marks which is error variance varies between 71% and 54%. The third measure of consistency in the Table is the average size of the difference of marks awarded to the same answer by two markers. The values obtained range from an average difference of 9.9% of possible marks to an average difference of 12.5% of possible marks.

Table 4: Correlation between marks awarded independently by 2 markers to answers on each of 13 questions on a Grade 12 History Examination.

Question	Correlation between marks awarded independently by 2 markers.	Percentage of variance common to the marks awarded by 2 markers.	Average difference between marks awarded to the same answer by 2 markers (as a percentage of possible marks.)
		%	%
1	.56	31	11.6
2	.57	33	12.5
3	.53	29	10.5
4	.60	36	10.5
5	.67	45	12.3
6	.55	31	10.0
7	.67	46	10.9
8	.65	43	10.5
9	.65	42	11.4
10	.63	40	9.9
11	.65	43	11.3
12	.65	43	10.4
13	.66	44	12.4

Summary

This report briefly summarises some results of a study of the effects of optional questions in six examinations. The results indicate that

(i) Differences in difficulty of different combinations of optional questions available to students can result in average differences of between 3% and 10% of possible marks between students of equal ability who select different question combinations.

(ii) Different markers mark different optional questions to different standards, and, as a result, students of equal ability who answer different questions and whose answers are marked by different markers could be expected to obtain average marks which differ by up to 30% of possible marks.

(iii) The group of students who select the most difficult combinations of optional questions on an examination are of significantly lower ability than students who answer the least difficult question combinations.

(iv) There are considerable differences in the reliability with which different optional questions are marked.

The report has indicated a number of difficulties associated with the use of optional questions in examinations. Further work is currently in progress to investigate other effects of optional questions in examinations.

THE INFLUENCE OF ANXIETY ON SEVERAL MEASURES OF CLASSROOM PERFORMANCE*

Kenneth E. Sinclair,
University of Sydney,
New South Wales, Australia.

The area of research concerned with the influence of anxiety on human learning and performance has significance for both educational practice and psychological theory. Within an educational context it has particular relevance for procedures used in student evaluation and testing. We live today in a highly test conscious culture. Decisions of major consequence to the individual are increasingly being made on the basis of his performance in tests. It is important, therefore, that the various factors that influence test performance be identified and the nature of their influence determined. There is growing evidence that anxiety is a factor of considerable importance in influencing test performance.

Beyond its relevance for educational measurement, research in this area is also contributing directly to a more precise understanding of human learning and performance. Investigators from quite varied backgrounds have carried out research in this area. Behaviorists (Spence and Spence, 1966), neuro-psychologists (Hebb, 1955; Malmo, 1959) and psychologists adopting a more psychoanalytic position (Sarason et al., 1960) have all developed rival theories designed to explain the influence of anxiety on learning and performance. Within an educational context, Sarason's psychoanalytic position has been found to have greatest relevance.

The influence of anxiety on performance in a variety of laboratory tasks is now quite well documented. Laboratory studies have established that the complexity of the task to be performed and the level of stress (usually defined in terms of level of ego-involvement) inhering in the task are two factors, in particular, which must be considered in explaining the influence of anxiety. Anxiety appears to facilitate performance on simple, straightforward tasks where there is little response competition and to interfere with performance on more complex tasks where response competition is likely (Taylor, 1951; Spence and Taylor, 1951; Taylor and Spence, 1952; Montague, 1953; Standish and Champion, 1960). In conditions where ego-involvement is low, a number of studies have found anxiety to be unrelated to performance (Lucas, 1952; Deese, Lazarus and Keenan, 1953; I.G. Sarason, 1957b; Kalish et al., 1958; Nicholson, 1958; Feshbach and Loeb, 1959), although some studies have found that anxiety facilitated performance (I.G. Sarason, 1956, 1957a; Longnecker, 1962). In conditions of high ego-involvement, anxiety has typically been found to interfere with performance (I.G. Sarason, 1956, 1957a; Nicholson, 1958; Harleston, 1962).

---

* This research was supported by a University of Sydney Research Grant. The cooperation of the N.S.W. Department of Education and the principals, staff and students of the Canterbury, Crows Nest and Drummoyne Boys' High Schools is gratefully acknowledged.

This article is reproduced, with permission, from The Australian Journal of Education, Volume 13, Number 3, October 1969.

While these relationships have frequently been demonstrated in relation to laboratory tasks, rather fewer studies have dealt with the question of the relationship between anxiety, task complexity, level of stress, and performance in more naturally occurring situations such as the classroom. Wrightsman (1962), however, in one study, varied level of stress in relation to aptitude test performance. He found no relationship ($r = -.06$) between anxiety and performance in the low ego-involvement condition and a significant negative relationship ($r = -.37$) in the condition of high ego-involvement. While there had been little change in the performance of low anxious (LA) subjects in the two conditions, the performance of high anxious (HA) subjects was reduced by almost one standard deviation by the stress of the instructions.

In a study with college students as subjects, Paul and Eriksen (1964) carried out a similar analysis using a classroom achievement test. A regular psychology class examination was administered on the morning of the experiment (the high stress condition) and a parallel form of the test was administered to the same individuals at night under conditions designed to minimise anxiety (the low stress condition). When their data were analysed using only subjects from the middle range of intelligence, a significant interaction was found between level of stress and level of anxiety. In the high stress condition, LA subjects were superior to HA subjects, while, in the relaxed condition, the HA subjects were superior.

The absence of experimental control over the learning materials and process may be a limiting factor in this study. Wide variation would be expected among the subjects as to the notes and texts used in studying for the examination, as well as for the time spent in studying for the examination.

These difficulties were substantially overcome in a study carried out by Caron (1963). He presented high school students with a 1700 word passage (consisting of an explanation of Atkinson's motive-expectancy-incentive model) to be studied in the experimental situation and, following the study period, obtained measures of rote learning and comprehension. The rote learning questions involved the reproduction of formulae and the definition of symbols contained in the passage, while the comprehension questions required the subjects to apply principles concerning risk preference that were presented in the passage. One half of his subjects studied the passage and were tested under examination conditions while the other half did so under conditions designed to induce curiosity. The condition was established by informing the subjects that the purpose of studying the passage was to enable them to interpret their own personality profiles which had been obtained in a previous testing session. For the rote learning task, there were no differences between HA and LA subjects in either treatment condition. For the comprehension task, there was no difference between HA and LA subjects in the curiosity condition. In the examination condition, however, LA subjects were superior to HA subjects. Caron (1963, p. 537) interpreted these findings as supporting the conclusion ". . . that the performance of anxious subjects on 'simple' tasks does not deteriorate under stress . . . whereas on 'complex' tasks their output suffers markedly."

While Caron's study contains many attractive features, a problem in interpreting some of his results arises because of the shortness of his measuring instruments. Only six rote learning questions and four comprehension questions were used (personal communication) and this may have operated to reduce reliability and, through this, the possibility of obtaining

significant differences between the LA and HA subjects. With respect to the rote learning task, a significant difference in favour of the LA subjects might well have been expected in the examination condition. The subjects were given only fifteen minutes to study the 1700 word passage so that learning that took place might be expected to be rather unstable and unorganized, resulting in considerable response competition in the performance situation. As has already been noted, in these circumstances anxiety may be expected to disrupt performance.

In the present study, the influence of anxiety on the performance of typical classroom tasks was again studied. As in Caron's investigation, the subjects were required to study a prose passage in the experimental situation and were then tested on several performance measures. In the present investigation, the measures obtained were of factual learning and reasoning and by increasing the number of questions asked, an attempt was made to ensure that a satisfactory level of reliability was reached for each measure. On the basis of scores on the High School Form of the Test Anxiety Scale, groups of LA, MA (moderately anxious) and HA high school students were obtained who completed the performance tasks in conditions of either high or low ego-involvement.

Hypotheses

Anxiety is conceived of as a hypothetical construct mediating between certain situational stimuli and various specifiable responses. The stimulus situation which evokes the anxiety reaction is assumed to be such that the individual anticipates a strong threat to his self-esteem. In classroom test situations, the anticipated threat to self-esteem is, most often, failure on the test.

In learning and performance situations, it is the view of Sarason and his colleagues (Mandler and Sarason, 1952; Sarason et al., 1960), that anxiety acts as a cue to elicit both responses that are relevant to the learning or performance task, and responses which are irrelevant. Task-relevant responses are observed in an increase in effort, concentration, and in procedural strategies previously found to facilitate learning and reduce anxiety. Task-irrelevant responses may be observed in the intrusion of thoughts concerning the consequences of failure, of self-depreciating ruminations and by ego-defensive avoidant responses designed to protect the individual from loss of self-esteem. These task-irrelevant responses compete with responses relevant to the task and typically have an interfering effect on learning and performance.

The extent to which interference to performance is caused by anxiety will depend upon level of ego-involvement and task complexity. When ego-involvement is low and performance is not perceived as having important ego-related consequences, little anxiety and few associated task-irrelevant responses will be elicited. In such a situation, therefore, performance for all individuals would be expected to be relatively free of the influence of anxiety. As ego-involvement increases, however, so will the tendency to react with anxiety increase and with this the tendency for interfering task-irrelevant responses to be elicited. When ego-involvement is high, individuals reacting with high levels of anxiety will respond with many more task-irrelevant responses than individuals who react to the same conditions with lower levels of anxiety. When the task is complex requiring concentration and careful processing of information, the intrusion of these task-irrelevant responses would be expected greatly to disrupt performance, so that level of anxiety would be inversely related to performance.

In the present study, the complexity of both performance tasks was such that anxiety, when elicited, was expected to have a debilitating effect on performance. On the factual learning task, the intrusion of task-irrelevant responses was expected to interfere with both the learning and the recall of the material studied. Because of the limited exposure to the study passage, overlearning would be unlikely so that what was learned would be relatively unstable and unorganized and, as such, highly susceptible to interference resulting from anxiety. Even greater interference was expected on the reasoning task. The presence of task-irrelevant responses was expected to have a particularly disruptive effect on the application of the complex cognitive processes required for performance on this task as generalizations were made, inferences drawn and hypotheses formulated and tested.

On the basis of these considerations two hypotheses were examined:

Hypothesis 1. In low ego-involvement conditions, anxiety has no influence on performance. With both tasks, there will be no difference in the performance of LA, MA and HA groups of subjects.

Hypothesis 2. In high ego-involvement conditions, anxiety acts to disrupt performance in complex tasks. In performing both tasks, LA subjects will be superior to MA subjects and MA subjects will be superior to HA subjects.

Differences in performance for the various anxiety groups were also expected under the two ego-involvement conditions. For the factual learning task, ego-involvement was expected to facilitate the performance of LA and MA subjects. For these subjects the enhancing effects of the increased motivation induced by the high ego-involvement instructions were expected to outweigh any negative effects due to the intrusion of task-irrelevant responses associated with anxiety. Thus it was expected that their performance would be superior in the high ego-involvement condition. For the HA subjects, however, the facilitating effects of the increased motivation were expected to be completely counteracted by the interfering effects of anxiety.

With the more complex reasoning task, the interfering effects of anxiety were expected to be greater than for the factual learning task. Because of this, only the performance of LA subjects was expected to be superior in the high ego-involvement condition. For MA subjects similar levels of performance were expected for the two ego-involvement conditions. For HA subjects the interfering effects of anxiety in the high ego-involvement condition were expected to be substantially greater than any facilitating effects that might occur, so that their performance was predicted to be superior in the low ego-involvement condition.

On the basis of these expectations, two further hypotheses, concerned with difference in performance under the two ego-involvement conditions, were examined.

Hypothesis 3. With the factual learning task, the performance of the LA and MA subjects will be superior when ego-involvement is high. However, HA subjects are expected to perform no better when ego-involvement is high than when it is low.

Hypothesis 4. With the reasoning task, the performance of LA subjects will be superior when ego-involvement is high, the performance of MA subjects will be similar in the two conditions of ego-involvement and the performance of HA subjects will be superior when ego-involvement is low.

Method

The subjects of the study were 173 sixth form male high school students attending three metropolitan boys' high schools in Sydney.

The content of the study passage consisted of a description of life among the Trobriand Islanders of the South Pacific. (1) This content appeared to be particularly suitable, since it was closely related to content typically taught at the high school level and yet there was little chance of the subjects having had any prior experience with it. To control the difficulty level of the vocabulary used in the passage, only words from the Thorndike-Lorge lists (1944) which occur in reading materials with a frequency of six or more times per million words were included. Thorndike and Lorge state that words appearing with this frequency are suitable for use with students in 3rd form and above. The passage contained 1332 words and one illustration, and filled almost six quarto pages of typescript.

Two performance tests were constructed. One measure, the factual learning measure, consisted of 20 multiple-choice questions for which the correct answer was explicitly stated in the study passage. The second measure, the reasoning measure, contained 12 multiple-choice questions for which the correct answer was not explicitly stated in the study passage. In answering these questions the subject was required to make deductions, and to draw inferences and implications from the given information.

Three weeks prior to the test administration, the High School Form of the Test Anxiety Scale (Mandler and Cowen, 1958), specially adapted for Australian conditions, was administered. A split-half reliability coefficient of 86 was obtained for this measure. Subjects scoring in the lower, middle and upper thirds of the anxiety distribution were designated as LA, MA and HA respectively. For each level of anxiety, the subjects were divided into two groups by use of a table of random numbers, one group being allocated randomly to the high ego-involvement condition and the other to the low ego-involvement condition.

To establish conditions of high ego-involvement (2), the subjects were informed that the test was one of scholastic aptitude and that their results would be made available to their headmaster. When the testing was completed, they were informed as to the actual purpose of the test. To establish conditions of low ego-involvement the subjects were informed that

(1) An earlier version of the study passage and performance measures was used in a previous study (Sinclair, 1965).

(2) The administration of the instruments in the high ego-involvement condition was carried out by the author in each school. The administration of the instruments in the low ego-involvement condition was carried out by T. Heys and W.J. Fenley whose assistance is gratefully acknowledged.

the purpose of the test was to establish whether the study passage was a good one for sixth form students or whether the questions were too easy or too difficult.

Twenty-five minutes were allowed for study of the passage. Twenty minutes were provided in which to answer the twenty factual learning questions and a further twenty minutes were provided in which to answer the twelve reasoning questions. These time limits were sufficient to enable all subjects to complete both tests. So that performance on the reasoning measure would not be influenced by the subjects' ability to recall information from the passage necessary for answering the questions asked, they were instructed that they could use the study passage in answering these questions.

Results

The design of the study was a 2 x 3 factorial, involving 2 levels of ego-involvement (high and low) and 3 levels of anxiety (high, moderate and low). This design was used for each of the two performance measures (factual learning and reasoning) with unequal numbers of subjects in each cell.

For the factual learning measure, the means of scores of the different anxiety groups are presented in Table 1. A reliability coefficient (K.R.20) of ·59 was obtained for this measure.

TABLE 1
Mean Factual Learning Scores for LA, MA and HA Groups of Subjects in Two Conditions of Ego-involvement

Anxiety Level	Low Ego-involvement			High Ego-involvement		
	N	$\bar{X}$	sd	N	$\bar{X}$	sd
LA	28	13·82	2·20	31	16·16	1·81
MA	28	14·32	1·94	29	14·62	2·58
HA	24	13·71	2·74	33	14·03	2·26

TABLE 2
Summary of the Analysis of Variance for the Factual Learning Measure

Source	Sum of Squares	df	Mean Square	F
Ego-involvement	41·73	1	41·73	8·17**
Anxiety	36·02	2	18·01	3·53*
Interaction ..	39·20	2	19·60	3·84*
Error	853·16	167	5·11	

** $p < ·01$.
* $p < ·05$.

A summary of the results of the analysis of variance carried out on these data (Winer, 1962, pp. 241-244) is presented in Table 2. Both main effects and the interaction were found to be significant. When individual group mean scores were examined by the Newman-Keuls procedure, it was observed that the performance of the LA group in the high ego-involvement condition had largely accounted for the significant results. As predicted, there were no significant differences found between the anxiety groups in the condition of low ego-involvement. In the high ego-involvement condition, as predicted, the performance of the LA subjects was superior to that of MA and HA subjects. The expected significant difference between the MA and HA groups did not emerge. Finally, again as hypothesized, the performance of the LA group in high ego-involvement conditions was superior to that of the LA group in low ego-involvement conditions while for the HA groups performance was similar in these two conditions. The expected superiority of the MA group in the high ego-involvement condition was not found.

TABLE 3

Mean Reasoning Scores for HA, MA and LA Groups of Subjects in Two Conditions of Ego-involvement

Anxiety Level	Low Ego-involvement			High Ego-involvement		
	N	$\bar{X}$	sd	N	$\bar{X}$	sd
LA	28	8·00	1·89	31	8·48	2·06
MA	28	7·36	2·08	29	8·38	1·82
HA	24	6·88	2·58	33	7·76	2·05

TABLE 4

Summary of the Analysis of Variance for the Reasoning Measure

	Sum of Squares	df	Mean Square	F
Ego-involvement	27·16	1	27·16	6·28*
Anxiety	24·77	2	12·39	2·87
Interaction	2·23	2	1·11	
Error	721·68	167	4·32	–

* $p < .05$.

In sum, the hypothesized relationships for the LA and HA groups in the two conditions of ego-involvement were all confirmed. Those for the MA group were not confirmed, the performance of that group being no different from that of the HA group.

For the reasoning measure, the mean scores of the different anxiety groups are presented in Table 3. A reliability coefficient (K.R.20) of ·68 was obtained for this measure.

A summary of the results of the analysis of variance carried out on these data is presented in Table 4. In this analysis only the mean square for level of ego-involvement was significant, indicating a general superiority in the high ego-involvement conditions. When pairs of means were analysed, again using the Newman-Keuls procedure, it was found that there were no differences between the anxiety groups in either ego-involvement condition. This was predicted for the low ego-involvement condition but for the high ego-involvement condition an inverse relationship between level of anxiety and performance had been predicted. All anxiety groups performed better in the high ego-involvement condition (although in no case did the difference reach an acceptable level of significance). This was predicted for the LA subjects but not for the MA and HA groups. In fact, for HA subjects superior performance had been predicted for the low ego-involvement condition.

Discussion

With respect to the factual learning task, the results obtained confirmed, in large measure, the hypotheses that were developed for testing, In test-like conditions, anxiety was observed to debilitate performance on that task. With respect to the reasoning task, however, few predicted relationships were supported. Despite the complexity of the task, anxiety did not appear to influence performance in the test-like condition. A possible reason for this latter result is to be found in the manner in which the reasoning test was administered. So that all subjects would have approximately equal access to the factual information upon which the reasoning items depended, the subjects were allowed to consult the study passage while answering the questions. This would make the reasoning task rather comparable to an open-book examination in which the student is able to consult certain reference material on answering the question asked. This procedure, by providing a memory-support (Sieber, 1969) in the performance situation, may well have had a reassuring, anxiety-reducing effect on the HA subjects so that interference to performance due to anxiety may have been minimal.

The results obtained provide a number of conclusions that bear directly on classroom practice and on the different theories that have been developed to explain the influence of anxiety on learning and performance. With respect to the factual learning task, the results support the conclusion that anxiety operates to debilitate performance when a complex task is to be performed in test-like conditions. This conclusion suggests that in important examinations, the HA student will be at a considerable disadvantage. When competing with other students for scholarships, university entrance, school prizes, employment opportunities or simply place in class, anxiety will act to interfere with and reduce the level of his performance.

The results also support the conclusion that while instructions designed to increase level of ego-involvement will raise the level of performance of LA students, it will not do so for MA and HA students. Sarason's theory suggests that for the MA and HA student, the positive motivational benefits deriving from the ego-involving instructions are cancelled out by the operation of task-irrelevant responses which are also elicited.

This conclusion suggests that the widely adopted practice in education of attempting to motivate students by placing strong emphasis upon the importance of examinations and the need to do well and avoid failure will be of value only to low test anxious students. In the present study with respect to the performance of moderately and high test anxious students on the factual learning task, little was achieved by increasing level of ego-involvement and, through this, anxiety. In fact, it may be that this emphasis, from a long term view, will have quite harmful effects. Since, at high levels, anxiety is such an unpleasant and exhausting experience, this emphasis may serve to engender a strong dislike of school which may eventually lead the student to drop out of school prematurely. Some support for this possibility is provided by Spielberger (1962) who observed, in one study, that HA college students had a higher drop out rate than LA students of comparable ability.

In addition to the implications provided for education practice, the results of the present study also provide implications for theory. The conclusion that in a test-like situation, anxiety will interfere with performance on a complex task is, as we have seen, consistent with the viewpoint of Sarason and his colleagues (Mandler and Sarason, 1952; Sarason et al., 1960). It is also, however, consistent with the Spence-Taylor theory, although in this theory it is the drive function of anxiety that is emphasised rather than the cue function. Spence and Taylor (Spence and Spence, 1966), conceive of anxiety as a drive which combines multiplicatively with the habit strengths of responses present in the individual's response hierarchy. When the desired response is not clearly dominant in the response hierarchy, as tends to be the case in complex performance situations, increase in drive (anxiety) serves to heighten competition among potential responses and in so doing disrupts performance.

The conclusion reached that increase in level of ego-involvement (stress) serves to raise the performance of LA individuals but not MA and HA individuals is, again, consistent with Sarason's theory. This conclusion, however, is not easily accounted for by the Spence-Taylor theory. Although, in the most recent statement of their position (Spence and Spence, 1966), they give passing reference to the question of situational factors (such as ego-involving instructions) that serve to elicit anxiety, they have not considered this question in detail, nor attempted to manipulate such factors in their research studies.

A number of directions for future research are suggested by the results of the present study. In this study the subjects used were male and of above-average ability. There is a need, then, for research to be carried out to determine if the conclusions reached in this study also hold for females and students of average and below-average ability. It is important, too, that ways be found to control the interfering effects of anxiety in the classroom. In particular, ways need to be found by which the HA student may be challenged but his anxiety kept within non-debilitating limits. One suggestion that arises from the present study is the possibility of using open-book examinations where reasoning is the major objective of assessment. Being able to consult appropriate reference material in the examination situation reduces the strain of having to remember and recall large bodies of information and in so doing may serve to reduce anxiety and the interference to reasoning that results. Sieber (1969) in an important recent article, provides further experimental evidence that

the provision of memory supports will be a particular aid to HA students in counteracting the interfering effects of anxiety. In that article she also suggests a number of other ways by which the HA student may be helped to perform more effectively. In particular she discusses the benefits that may be derived from instruction in the use of verbal encoding skills, diagrams, mnemonic devices, notational systems and outlining systems for organizing general ideas prior to the development of detail. There is a need for these suggestions to be followed up in classroom-oriented research.

Bibliography

Caron, A.J. "Curiosity, Achievement, and Avoidant Motivation as determinants of Epistemic Behaviour" J. abnorm. soc. Psychol., 67, 1963, 535-49.

Deese, J., Lazarus, R.S., and Keenan, J. "Anxiety, Anxiety Reduction and Stress in Learning" J.exp. Psychol., 46, 1953, 55-60.

Feshbach, S, and Loeb, A. "A Further Experimental Study of a Response-interference versus a Drive-facilitation Theory of the Effect of Anxiety upon Learning" J. Personal., 27, 1959, 497-506.

Harleston, B.W. "Test Anxiety and Performance in Problem solving Situations." J. Personal, 30, 1962, 557-573.

Hebb, D.O. "Drive and the C.N.S. (Conceptual Nervous System)." Psychol. Rev., 62, 1955, 243-254.

Kalish, H.I., Garmezy, N., Rodnick, E.H., and Bleke, R.C. "The Effects of Anxiety and Experimentally induced Stress on Verbal Learning." J.gen. Psychol., 59, 1958, 87-95.

Longnecker, E.D. "Perceptual Recognition as a Function of Anxiety, Motivation, and the Testing Situation." J. abnorm. soc Pyshchol., 64, 1962, 215-221.

Lucas, J.D. "The interactive Effects of Anxiety, Failure, and Intraserial Duplication." Amer. J. Psychol., 65, 1952, 59-66.

Malmo, R.B. "Activation: a Neuropsychological Dimension." Psychol. Rev., 66, 1959, 367-86.

Mandler, G., and Cowen, J.E. "Test Anxiety Questionnaires" J. consult. Psychol., 22, 1958, 228.

Mandler, G., and Sarason, S.B. "A study of Anxiety and Learning." J. abnorm. soc Psychol., 47, 1952, 166-173

Montague, E.K. "The Role of Anxiety in Serial Rote Learning" J. exp. Psychol., 45, 1953, 91-98.

Nicholson, W.M. "The Influence of Anxiety upon Learning: Interference or Drive Increment?" J. Personal., 26, 1958, 303-319.

Paul, G.L., and Eriksen, C.W. "Effects of Test Anxiety on 'Real-Life' Examinations." J. Personal., 32, 1964, 480-494.

Sarason, I.G. "The Effect of Anxiety, Motivational Instructions and Failure on Serial Learning," J. exp. Psychol., 51, 1956, 253-259.

Sarason, I.G. "Effect of Anxiety and Two Kinds of Motivating Instructions on Verbal Learning" J. abnorm. soc. Psychol.54, 1957a, 166-171.

Sarason, I.G. "The Effect of Anxiety and Two Kinds of Failure on Serial Learning", J. Personal., 25, 1957b, 383-392

Sarason, S.B., Davidson, K.S., Lighthall, F.F. Waite, R.R., and Ruebush, B.K. Anxiety in Elementary School Children, New York, John Wiley, 1960.

Sieber, J.E. "A Paradigm for Experimental Modification of the Effects of Test Anxiety on Cognitive Processes." Amer, Educ. Res.J. 6, 1969, 46-61.

Sinclair, K.E. The influence of Anxiety and Level of Ego Involvement on Classroom Learning and Performance. Unpublished Ph.D.

dissertation, University of Illinois, 1965.
Spence, J.T., and Spence, K.W. "The Motivational Components of Manifest Anxiety: Drive and Drive Stimuli" in C.D. Spielberger (Ed.), Anxiety and Behaviour. New York: Academic Press, 1966, 191-326.
Spence, K.W. and Taylor, J.A. "Anxiety and Strength of the UCS as Determinants of the Amount of Eyelid Conditioning," J. exp. Psychol., 42, 1951, 183-188.
Spielberger, C.D. "The Effects of Manifest Anxiety on the Academic Achievement of College Students." Ment. Hyg., 46, 1962, 420-426.
Standish, R.R., and Chamption, R.A. "Task Difficulty and Drive in Verbal Learning". J. exp Psychol. 59, 1960, 361-365.
Taylor, J.A. "The Relation of Anxiety to the Conditioned Eyelid Response". J. exp. Psychol., 41, 1951, 81-92.
Taylor, J.A. and Spence, K.W. "The Relationship of Anxiety Level to Performance in Serial Learning. "J. exp. Psychol., 44, 1952, 61-64.
Thorndike, E.L., and Lorge, I. The Teacher's Word Book of 30,000 Words. New York: Bureau of Publications, Teachers College Columbia University, 1944.
Winer, B.J. Statistical Principles in Experimental Design. New York: McGraw-Hill, 1962.
Wrightsman, L. S. "The effects of Anxiety, Achievement Motivation, and Task Importance upon Performance on an Intelligence Test." J. educ. Psychol., 53, 1962, 150-156.

DEVELOPMENT OF THE CANADIAN SCHOLASTIC APTITUDE TEST

Ross E. Traub
The Ontario Institute for Studies in Education

and

H.A. Elliott
Service for Admission to College and University

The Canadian Scholastic Aptitude Test (CSAT) for English-speaking students is one of a battery of objectively scorable tests being developed by the Service for Admission to College and University (SACU). CSAT is designed to measure what may be described as the general verbal and mathematical abilities of students in their final year of secondary school. The test is divided into four sections each requiring 30 minutes of administration time. Two sections test verbal ability, one with antonym, verbal analogy and sentence completion items, the other with questions on the content of several short essays dealing with diverse topics. The remaining two sections of the test each contain a heterogeneous collection of mathematical items which assess the examinee's ability to reason numerically, algebraically, or geometrically. In addition to these four sections, CSAT contains a fifth section composed of items being pretested. Performance on this section, which also requires a half-hour of administration time, does not count towards the scores an examinee achieves on the test. The motivation of examinees to perform pretest questions is maintained by concealing their identity in the test.

CSAT was first given in February, 1969 and is presently scheduled for administration once per year. The test is new each year in the sense that it consists of a different set of items drawn from a secure item pool. It should be noted that different forms of CSAT can be made very similar in the sense that each form can be built from items drawn to match similar specifications with respect to type, difficulty and level of discrimination. The specifications are laid down by a test development committee consisting of one member from each Canadian province plus representatives of SACU and of the Ontario Institute for Studies in Education, the institution responsible for item development and test assembly under contract with SACU.

Two "raw" scores are derived for each examinee who takes CSAT, a verbal and a mathematical score. Both scores are obtained from a formula in which a fraction of the number of wrong answers is subtracted from the number of correct answers. The penalty is imposed to discourage random guessing on the test. Examinees are fully informed of the penalty and the rationale underlying its use in a handbook they are given to study several weeks before the administration date. The handbook also contains practice items to make examinees better informed of the nature of CSAT.

For the purpose of reporting on the performance of an examinee both to him and to the universities he designates, raw scores are converted to standard scores. The distribution of standard scores has a mean of 500 and a standard deviation of 100. Thus the effective range of CSAT scores is from 200 to 800.

At this point several questions that are frequently asked with reference to CSAT warrant consideration:

1. Why does Canada need a national university admissions testing program? One answer to this question begins by recognizing two facts of contemporary Canadian education, that it is a matter of provincial responsibility and that control of education is becoming decentralized in important respects. Provincial control of education has resulted in the development of interprovincial differences in secondary school programs. The differences are substantial enough to make difficult any direct and meaningful comparison of records of school achievement from different provinces. Decentralization of control has occurred in some provinces to the extent that the teaching and administrative staff of individual schools have sole responsibility for determining curriculum and evaluating student performance. Consequently, it is often difficult to compare in a meaningful way the school records of applicants from different secondary schools in the same province. National admissions tests, such as CSAT, hold forth the hope of providing universities with a valid basis for comparing applicants from different schools and different provinces.

Another reason for university admissions tests is that they help the universities reach early admission decisions. The policy of admitting some applicants several months before they finish secondary school, subject only to the proviso that they successfully complete secondary school, has been forced on Canadian universities by the practice of provincial governments to finance universities on a formula basis and by the practice of most students to apply to more than one university. Under formula financing, universities typically receive government grants in direct proportion to the number of students they have enrolled. To ensure a full first-year enrolment, thereby ensuring full enrolments in succeeding years and qualifying for maximum government grants, a university will admit students in two phases. Many of the applicants admitted in the first phase will decide ultimately to go elsewhere. When a university knows which applicants are not coming, it is able, in the second phase of admissions, to complete its first-year roster by admitting from the pool of remaining candidates. Inasmuch as the first phase of admissions is made in the absence of a completed secondary school record, universities find it advantageous to have the information provided by valid admissions tests to guide their decision making.

2. Why does CSAT attempt to measure verbal and mathematical ability, nothing more nor less? Our first response to this question is that these abilities seem to represent characteristics of considerable practical and theoretical significance. Over the past 60 years psychologists and educators have found them, either separately or in combination, to correlate moderately well with academic accomplishment of many different types. Moreover, verbal and mathematical ability appear with great consistency in factor-analytic studies of academic achievement. They are central to theories such as Vernon's (1961) on the structure of human abilities. One would expect a good test of verbal and mathematical ability to provide scores with considerable relevance for admissions work.

Another response to the second question focuses on the fact that verbal and mathematical scores are only moderately correlated. This implies that a test like CSAT should provide information about two substantially different aspects of an examinee's capabilities. Consequently, the test should enable university admissions officers to judge applicants in terms of the ability most relevant to their proposed programs. Verbal ability should probably receive more weight in comparing applicants who want to study a language or history. On the other hand, mathematical ability would be expected to receive more weight in comparing applicants for work in mathematics, physics or engineering. More than this, the availability of scores on both abilities should enable admitting institutions to counsel students about the advisability of entering one program of study as opposed to another.

3. Why does Canada require its own test? Why can it not use admissions tests prepared elsewhere? It is true that CSAT is very similar to the Scholastic Aptitude Test (SAT) of the College Entrance Examination Board (CEEB) of the United States. The reason for this is not coincidence. In fact, CEEB has been very generous in its provision of assistance to SACU in initiating CSAT. Moreover, the underlying rationale for the two tests is very similar. These facts notwithstanding, there are points that can be made in support of the development of a Canadian test.

One quite obvious point is that the population of examinees for CSAT differs in some respects from the population for SAT. For example, it appears that in order for the test to be ideally suited to Canadian examinees the mathematical items in CSAT must be somewhat more difficult on the average than the mathematical items in SAT. Also, differences between Canadians and Americans in cultural background and in the use of English means that some questions that would be appropriate for use in one country are inappropriate for use in the other.

Another reason for Canada to build its own university admissions tests is that by so doing it retains control of the specifications for the test. If CSAT should prove to be unsatisfactory in certain respects, given its present specifications, it will be possible to make revisions in an attempt to achieve a better instrument. Such revisions would probably be difficult to have incorporated in a test designed primarily for a United States population.

An additional factor which suggests the need for a Canadian test for English-speaking students is the parallel requirement in Canada for a test for Franco-phone students. Such a test, Test d'aptitude générale aux études post-secondaires (TAGEPS), has been developed by the Institut de recherche pédagogique and was also administered for the first time in 1969. The design of TAGEPS is essentially identical with that of CSAT but the items have been produced and validated independently. In the near future, an attempt will be made to equate scores on CSAT, TAGEPS and SAT.

4. What results have been observed to this point? There has been only one administration of CSAT thus far. The test performed well in the sense that satisfactory estimates of internal consistency reliability were achieved. The coefficients for both the verbal and mathematical scores exceeded .90. Moreover, the distributions of scores were as desired, being unimodal and roughly symmetrical and bell-shaped. Satisfactory discrimination among students across a broad range of ability levels appears to have been achieved in that standard scores extended the full range from 200 to 800. What is relatively unknown at the moment is the predictive validity of CSAT. This cannot be determined satisfactorily until those students who took CSAT in 1969 complete at least their first year of university in 1970. However, some indication of validity is available for a test similar to CSAT which was administered in Ontario in 1967 and 1968. For that test, validity coefficients as high as, or higher than, .60 have been observed for some programs in some Ontario universities. The median validity coefficient across all programs in all Ontario universities was unfortunately considerably smaller, about .30. Thus, it is with an air of wary optimism that we await the initial validity results for CSAT itself.

Reference:

Vernon, P.E. The Structure of Human Abilities (2nd ed.), Methuen, London, 1961.

THE DEVELOPMENT OF EXAMINATION TECHNIQUES FOR TECHNICAL SUBJECTS

Ng Fook Kah
Chief Examinations Officer
Ministry of Education, Singapore

Background

Prior to 1968 technical subjects received little attention in schools as more than ninety per cent of school leavers in Singapore schools were in the academic stream.

In 1968 the Technical Education Department was formed within the Ministry of Education to be responsible for general policy relating to technical education and industrial training and to concentrate on the problems of training to meet Singapore's anticipated shortages of craftsmen and related skilled workers for industrialisation.

By 1971, the Technical Education Department established the

(1) Successful restructuring of secondary schools system to give the required technical bias as well as a fully developed Technical Stream up to Pre University level.

(2) Channelling of secondary students into Technical Stream.

(3) System for industrial training within the Vocational Institutes.

(4) Industrial technician training programme arrived at turning out a higher calibre worker who could be further trained on the job for factory-floor supervisory functions.

At the end of 1972, the Technical Education Department had completed the groundwork for the industrial training programme of workers within industries. An Industrial Training Board would take over the responsibility for Industrial training both institution and industrial based.

On 19 February 1973, the Industrial Training Board was officially established which replaced the Technical Education Department. The reorganised technical education system within the schools under the former Technical Education Department are under the responsibility of the Ministry of Education.

Part I

Trade Testing

The Trade Testing Unit of the Examinations Division in the Ministry of Education at the outset of the reorganisation in 1968 planned a National Trade Testing Council designed to co-ordinate the work of the various Trade Testing Committees on standards for certification in the various trades.

Development of the National Trade Testing System

The Ministry is grateful for assistance given by ILO and UNESCO Advisers in the development of a system for national trade testing.

The establishment of the National Trade Testing system comprises two main areas:

(1) Preparation of Skill Analyses and National Trades Standards.

(2) Working out an Assessment System for National Trade Testing.

The Trade Testing System is the uniform national mechanism by which a candidate's skill is gauged objectively. It is, therefore, important that tests are reliable and valid. The reliability and validity constitute a sound foundation on which every test should be constructed and scored. A reliable test yields accurate and consistent results. A valid test yields accuracy with which it measures what it was designed to measure, in this case, the candidate's skill against the National Trade Standard and to distinguish between the semi skilled (Grade III), the skilled (Grade II) and the highly skilled (Grade I) candidates. In Trade Testing 2 kinds of validity are of importance: Content-Validity and Face Validity.

Content Validity

Test items must be sampled so that a test will show the actual stage of skill in the trade concerned on the basis of trade test syllabus.

Face Validity

(a) The instruction given to candidates must be adequate in order to provide them with full information and detailed job requirements.

(b) The acceptability of the marking scheme by markers in their applications.

Preparation of a Marking Scheme will consist of the following Areas:

(a) Marking criteria contain all the parameters which should be taken into account when scoring a test (e.g. trade of Turner).

---------- time taken to perform the test.
---------- quality of work.
---------- economy in the use of materials.
---------- use of materials and tools.
---------- function of work pieces.
---------- the observance of safety precautions.

(b) Basic awards and multiplication (weighting) factors

The Four-Point Scale for grading a test performance combined with a simple weighting system, has proved to be a reliable and a valid measuring instrument.

(i) Basic Award

3 points (very good performance)

2 points (good performance)

1 point (slightly below standard)

0 point (poor performance)

(ii) Weighting System will comprise the following multiplication factors

3 points (difficult test item)

2 points (test item of medium difficulty)

1 point (simple test item)

(c) Pass standard

The Pass Standard states the minimum score level which a candidate has to reach in order to pass the test.

Grade III	(semi-skilled)	60% Pass Mark
Grade II	(skilled)	60% Pass Mark
Grade I	(highly skilled)	70% Pass Mark

Objective Type Testing is far more applicable to the field of Trade Testing such as
(i) True-false,
(ii) Multiple choice,
(iii) One work answer,
(iv) Completion,
(v) Short statement (listing of facts),
(vi) Matching-type,
(vii) Trade calculation, and
(viii) Trade drawing items.

Item Bank

Constructing reliable and objective type tests are time consuming. An item bank saves time, provided the bank items are analysed and acceptable ones to the items. The card will also contain the level of difficulty and discriminate power of the items.

Practical Tests

In order to maximise the reliability of the marking procedure, markers operate in pairs. Each of the markers in a pair assesses the same candidate independently. An average mark is then taken into account for final scoring.

Preparation of National Trades Standards (Syllabuses) Testing System

The purpose of the system of National Trade Standards is to provide an objective method by which it is possible to establish measured national work performance standards and to gauge skills against these standards. The system defines each skill within the Rules and Regulations for the award of the National Trade Certificate which is nationally acceptable. The system also provides a means for discovering ineffective training areas and sets out the standards which should be achieved.

Preparation of Skill Analyses (Specification)

In the preparation of skill analyses, the Trade Testing unit has developed the following skills for

(1) Motor Vehicle Mechanic
(2) Heavy Duty Diesel Mechanic
(3) Vehicle Body Repairer
(4) Metal Fabricator
(5) Arc Welder
(6) Gas Welder
(7) Machinist Fitter
(8) Platemaker
(9) Turner
(10) Electrician
(11) Refrigeration and air conditioning mechanic
(12) Plumber/Pipe Fitter
(13) Barbender/Concretor
(14) Bricklayer/Plasterer
(15) Building Draftsman
(16) Building Carpenter
(17) Wood machinist
(18) Furniture Maker
(19) Construction Painter/Decorator
(20) Decorative Tile Setter
(21) Offset pressman
(22) Compositor
(23) Photo-engraver (Blockmaker)
(24) Letterpress Pressman
(25) Book binder.

The Trade Skill Analyses are prepared so that the National Trade Standards meet the up-to-date requirements of industry. Each skill analysis is carried out on the basis of Trade Definition and Scope of Activity. The Trade Definition is based on international classification (International Standard Classification of Occupations, published by ILO). The Trade Definition also states the technical skills that are expected. The Scope of Activity specifies the type of work expected at semi-skilled level (Grade III), Skill-level (Grade II) and highly skilled level (Grade I). The Analysis is intended to show the operations and related knowledge that a skilled worker necessarily will require when working in a specific trade for earnings.

The implementation of trade tests on a national scale from full-time Vocational Institutes students to industrial workers on the job will identify workmen of varying degree of skills. This will in turn encourage effective employment of skills which has a direct bearing on the eventual development of the national economy in sustaining high productivity. The National Trade Testing System will motivate workers to upgrade themselves. The results in trade tests will exhibit ineffective areas of training and, therefore, the gearing and adjusting of training programmes to meet national industrial requirements.

Part II

Technical Education in Schools

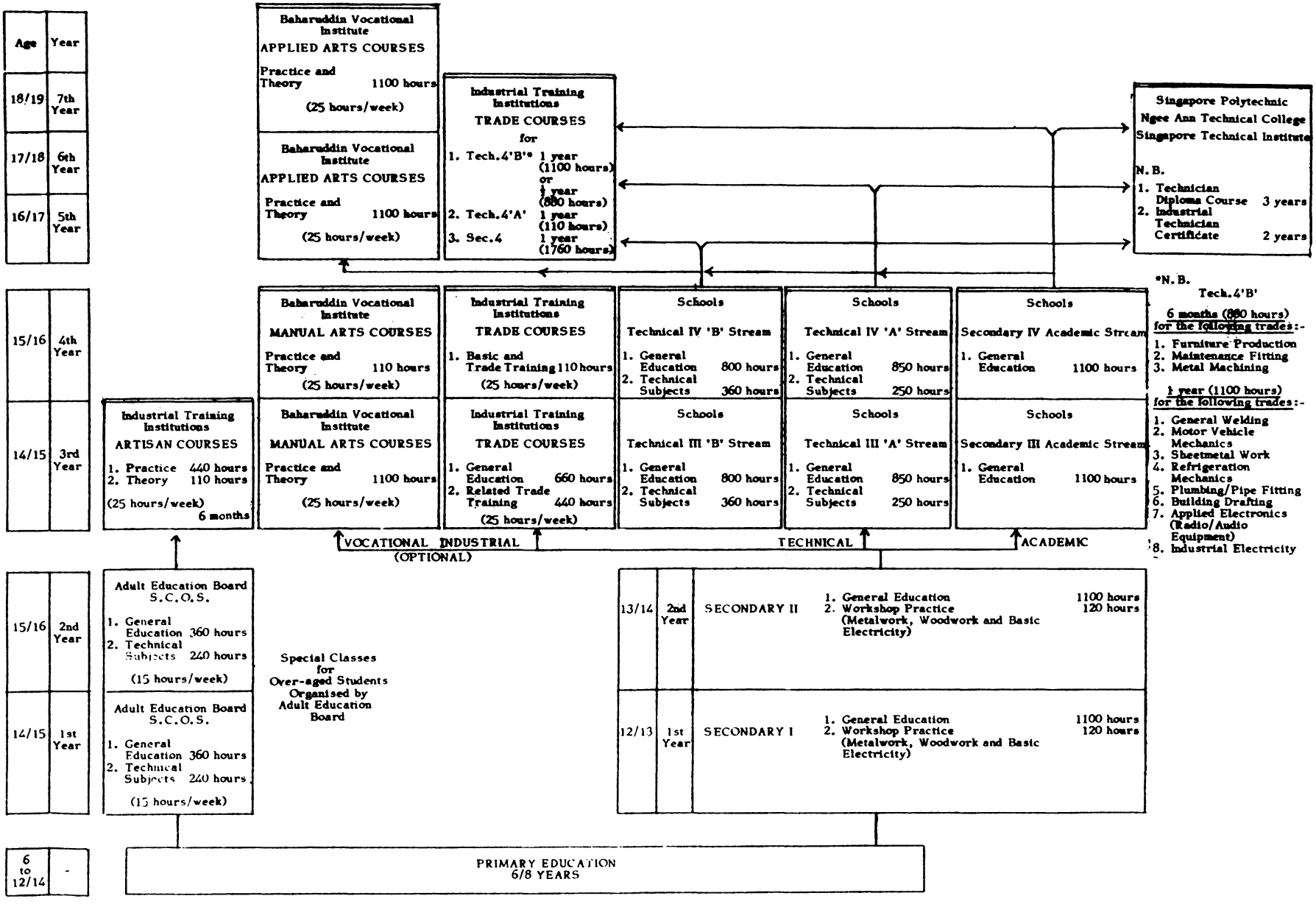
Since 1970 all male students in their first and second year of secondary education after they have completed a six year Primary Education and 50% female students are given technical workshop practice for 3 hours weekly outside school curriculum hours in Metalwork, Woodwork and Basic Electricity. All students in addition are taught Technical Drawing as a classroom subject.

A system of channelling students who completed Secondary II into the technical and academic streams for Secondary III and Secondary IV was introduced in 1970, through an aptitude test, performance in workshop subjects and the cognate group of academic subjects to assist parents in exercising their option.

In the Pre U II classes, no school candidate offered technical subjects in 1968. In 1972, 202 candidates entered technical subjects in the H.S.C. Examination in Metalwork and Geometrical and Mechanical Engineering Drawing. These subjects are given equal consideration as other subjects by the Faculty of Engineering of the University of Singapore for purposes of undergraduate admissions.

The Ministry of Education runs 8 Vocational Institutes and a hotel catering training school in the engineering, refrigeration, building, electrical and electronic, woodwork and building, metal, automotive manual and applied arts trades. Examinations are conducted at Trades and Artisan levels to produce an adequate number of skilled workers from the Vocational Institutes. Trade Tests at the Vocational Institutes are conducted to determine whether the students have attained the required level of proficiency to qualify for the award of the certification bestowed by the institutes upon completion of their courses.

EDUCATION SYSTEM OF SINGAPORE
SHOWING RELATIONSHIP BETWEEN SCHOOL EDUCATION AND INDUSTRIAL TRAINING



TESTING WITH EDUCATIONALLY DISADVANTAGED CHILDREN

A.E.G. Pilliner,
Director, Godfrey Thomson Unit for Academic Assessment,
University of Edinburgh, Scotland.

Every country has its educationally disadvantaged children, even those in which educational development is most advanced. Britain is now replacing a selective system of education by a comprehensive one in an endeavour to eliminate, or at least reduce, unequal educational opportunity. Superimposed on this task, she is now faced with the responsibility of educating an increasing number of young immigrants from other Commonwealth countries. The United States of America, despite the fact that in principle her educational system has never been other than comprehensive, has not yet achieved her avowed aim of de-segregation and the quality of education offered to some of her citizens is still inferior to that enjoyed by others.

It should not surprise us that this state of affairs exists a fortiori in countries at an earlier stage of development. In such countries, through sheer force of circumstances, education in any way comparable in quality and adequacy to that taken for granted for the majority in some advanced countries is available to only a small minority. The allotment of a large proportion of scarce resources to the more extensive education of a relatively small proportion of children is understandable. Pre-requisite to speedier progress in the future, further technical advance, increased economic development and wider educational expansion, is the production now of a necessarily small number of people possessing the knowledge, skills and dedication essential to the achievement of these aims.

It is no accident that in many of the new countries educational objectives tend increasingly to resemble those in others more fortunate in having advanced further along the path of development. According to Doob (6), the pressures forcing the new countries in the same direction are inevitable, irresistible and irreversible. This does not mean that all will arrive at the same place. A country on its way 'up' will be selective in what it absorbs and will adapt its acquisitions from elsewhere to its own traditions and needs. Nevertheless, since both less and more developed countries share a number of the same objectives, their educative processes will have much in common. At the same time, a process evolved over a lengthy period and geared to the norms of a society or culture already well developed cannot be transferred ready-made to another less so without giving rise to problems, even though the objectives are similar.

These problems are reflected in the testing procedures which are an integral part of any and every educational process. The situation previously mentioned implies that a relatively small number of pupils must be selected for secondary education from a very large primary school population. Countries where this situation exists are likely to be characterised by primary education of poor and uneven quality. Children living in towns may be more fortunate in their primary education than others living in villages.

In this case, restrictions are imposed on the interpretation of scholastic attainment test results. Though such tests may still accurately

measure a pupil's achievement in specific subjects to date, their use as prognosticators of future success is precluded or at least limited, however successful they may be in this respect with children more fortunately circumstanced. The poor performance on an arithmetic test of a pupil who has hitherto been taught arithmetic either badly or not at all is a fair index of his present ability in that subject. But as a predictor of his likely progress if this defect is remedied its value is questionable. If assessment of potential or aptitude is at issue some other means must be found.

If experience in countries more educationally advanced is anything to go by, the use of tests of verbal reasoning might seem to offer a solution. Tests in this category differ from tests of scholastic attainment in that they are less closely geared to the school curriculum; good performance on them is less dependent on exposure to the usual range of school subjects. They have been extensively used for 11+ selection in Britain, where numerous follow-up studies have consistently shown them to be among the best predictors of academic success.

However, difficulties still remain. There may be several native languages or dialects while the accepted medium of instruction in the secondary schools is a second language such as English, the pupil's acquaintance with which is limited by factors such as his primary teachers' command of it.

Bernstein's (3) work on language habits in Britain bears on this situation. He points to the relation between class structure and the varieties of English used by school children. Social stratification is related to differential availability of language codes. The lower working class child has a group-oriented 'restricted' code; the middle class child has both this and an individually-oriented 'elaborated' code. These codes differ in that the first is more fluent, repetitive and predictable, the second more hesitant, idiosyncratically planned and complex. Educationally, the child from a poor background is at a disadvantage since he finds himself having in effect to translate what he hears from his teachers. As Bernstein points out, differential difficulty in communication is likely to be reflected in differential verbal test performance.

The problem is exacerbated when the differential is not merely intra- but inter-language. It is therefore natural to consider the possibility of assessing pupils' aptitude for further stages of education by some testing procedure which avoids the use of the differentially unfamiliar second language. On the face of it, one way of doing so would be to couch the tests employed in the pupils' own native languages. This however may be difficult in practice if several languages are involved. There is the further technical difficulty of equating the performances of different children on different tests - for, let there be no mistake about it, even the same content translated into different languages produces different tests, the results of which, expressed numerically, are not necessarily comparable. Moreover, the problem of unequal primary school opportunity, and its implications for scholastic attainment, will still remain.

On all these counts, it may be thought desirable to go one step further, to eliminate the use of language so far as is practically possible, and to rely on non-verbal or non-language tests. Here, surely, it might be argued, is the way out of the difficulty. If the use of language-bound tests is seen as impracticable or leading to injustice, should not their substitution by non-language tests reduce the practical problems and promote 'fairness' for all concerned?

This is the kind of thinking behind the more general concepts of 'culture-free' and 'culture-fair' testing. The intention is wholly admirable. Any measure which will help to redress the balance in favour of children who are culturally deprived or otherwise educationally disadvantaged is surely to be encouraged. The laudable objective is to reduce these obstacles by the use of testing devices which transcend or remove cultural differences or educational inequalities.

However, the problem is by no means simple. The concept of 'culture-free' tests is highly dubious. Anastasi (1, p.256) is surely right when she says: 'No test can be truly "culture-free". Since every test measures a sample of behaviour, it will reflect factors that influence behaviour. Persons do not react in a cultural vacuum.' Wesmen (16, p.269) is even more forthright. 'I do not wish to impugn the high social motives which stimulate the search for such devices; I do wish to question that such a search, in its usual setting, is sensible. A culture-free test would presumably probe learnings which had not been affected by environment; this is sheer nonsense.' These statements represent the general view of most contemporary psychologists. Few would now regard the quest for culture-free tests as other than chimerical.

The prospect for 'culture-fair' tests is, on the face of it, less unpromising. In principle it is possible to build tests which, though not free of cultural influences, sample only behaviour common to several cultures. An alternative description of such tests is 'cross-cultural'. The amount of effort that has gone into the construction of allegedly cross-cultural tests is vast, particularly if we include also tests intended for comparisons among sub-cultures within a larger culture. Only a few can be mentioned here. In the nature of things, they are non-verbal in content. They fall into two main categories: performance tests, designed for individual administration, and in the main involving manipulation of objects; and non-verbal or non-language group tests, normally paper-and-pencil tests which do not demand of the testees the skills of reading and writing. Most such tests do however depend on oral instructions, it being assumed (perhaps too lightly) that these are of such simplicity that no semantic problems arise in their translation and that different language versions do not differ in difficulty. A few tests have been constructed in which the instructions can be mimed or demonstrated.

Examples of tests in the performance category are: Form-board (Sequin, Pintner-Paterson), Mazes (Porteous), Picture Completion (Healy), Block Manipulation (Kohs); Stencil Design (Arthur); Analogies (Leiter); and, of course, the General Performance Scale of the WISC (Wechsler). Examples from the group non-language category are the Draw-a-Man (Goodenough), Matrices (Raven), Pictorial Problems (Davies-Eells), Semantic Symbols (Rulon); and a number of tests intended to probe, using pictorial or diagrammatic material, mental functions - analogies, odd-man-out , series and the like - similar to those frequently occuring in verbal tests (Moray House Picture, Jenkins Non-Verbal, Cattell IPAT).

On closer examination, however, the prospect of producing 'culture-fair' tests is only slightly less unpromising than for tests that are 'culture-free'. By restricting test content to elements common to several cultures the relevance of the results in respect of any one of them is made questionable. To the extent that different cultures display unique features,

nurture disparate traditions and values, or foster or suppress different abilities or modes of behaviour, tests restricted in this way may miss their targets. To quote Anastasi (2, p.299) again: 'If we were to rule out cultural differentials from a test, we might thereby lower its validity against the criterion we are trying to predict'. It is as though in trying to please everybody, we succeed in pleasing nobody. Or, to change the metaphor, although the wave pattern for the fundamental tone emitted by different musical instruments is the same for all, it is the superimposed over-tones or harmonics which endow each with its peculiar timbre, its richness of quality.

The concepts of 'culture-free' and 'culture-fair' tests once received plausible support from the contemporary psychological theory. 'Native intelligence', like original sin, was reified and came to be regarded as a fixed entity rather than a developing attribute. By the exercise of sufficient inventiveness - Wesman (16) speaks of 'ingenious mining devices' - the influence of differential exposure to learning could be eliminated and the 'innate intelligence' of the individual revealed and recorded on a scale for all to see.

More recent theory is less accommodating. Hebb's (9) distinction between Intelligence A and B corresponds broadly to the geneticist's distinction between genotype and phenotype. Like the genotype, Intelligence A is not directly observable, still less measurable. Only Intelligence B, corresponding to the phenotype, can be observed; it results from the interaction of both nature and nurture. The title of a once popular song sums it up neatly:; 'It's what you do with what you've got that counts'. Vernon (14) playfully, in the first place, one suspects, but then more seriously, had added a further category. Intelligence C is what tests measure. It varies with difference in test content and is therefore not unique in the prediction it affords of Intelligence B. Hebb's theory offers but cold comfort in the search for instruments equally fair to differentially disadvantaged testees.

On the fact of it at least, the theory of 'fluid' and 'crystallised' intelligence attributable to Cattell and Horn (5) is distinctly more hopeful. They suggest that the general factor emerging from studies of batteries of disparate tests is a mixture separable into two components: G_f ('fluid' intelligence), reflecting constitutional equipment; and G_c ('crystallised' intelligence), the results of experience such as cultural and educational pressures. Unlike Intelligence A, G_f is measurable by tests tapping adaptability to situations so unfamiliar that previous learning experience is of no help. G_c, corresponding roughly to Intelligence B, is manifested in cognitive behaviour already patterned by previous experience. Even before biological maturity is reached, diversity in cultural opportunities, interests and personality traits produces substantial individual differences in G_c which, according to the theory, should not be parallelled for G_f.

This theory underlies the construction of the Cattell IPAT Culture Fair (formerly Culture Free) Intelligence Test. Predictably, the greatest success in removing 'contamination' by cultural differences is claimed for subtests involving mazes, identification of similar drawings, picture classification and symbol copying. At best, however, the success achieved is only partial. In view of the IPAT, Tannenbaum (12, p.454) concludes that 'the goal of demonstrating equality among national and international subpopulations by some measures of general ability has not been reached

by this test.' He questions whether this is a goal worth pursuing. 'Even if it were possible to devise a test so antiseptic as to clean out inequality not only among subcultures but also among other groups showing differences in test intelligence, such as those classified by sex, age, geographic origin, body type, physical health, personality structure, and family unity - what kind of instrument would we have then? Since such a test must perforce be so thoroughly doctored as to omit tasks that reveal these group differences, or substitute others that show "no difference", what could it possibly measure? What could it predict?' Vernon's (15, p.25) conclusions are equally definite. 'The main weakness in his (Cattell's) theory is the claim that fluid ability tests are largely immune to cultural influences. The skills required for reasoning with these abstract materials would appear to be built up in just the same way as those involved in verbal reasoning; and the evidence ... demonstrates at least as great variation attributable to cultural differences'.

For a very complete and up-to-date survey of this evidence, reference should be made to Vernon (15). Only some of it can be cited here. As already stated, the IPAT was found to be only partially successful in ironing out cultural differences. Although in cultures similar to that in which the test was developed the same norms were approximately applicable, this was not so for cultures more dissimilar; for these, average performance was often much lower. Bernstein (4) reports smaller differences in performance on Raven's Matrices between middle and working class groups than on tests of verbal reasoning. But in other studies, particularly in African countries, test results were positively correlated with amount of education. The Goodenough Draw-a-Man (8) test has gone through several revisions. After extensive use with a number of different cultural and ethnic groups, its authors have abandoned their original optimistic view and in their more recent reports have concluded that a culture-fair test of whatever attribute 'is illusory'.

The Davis-Eeels Games (7) were specially designed for American use to be relatively independent of social class bias. But differential educational disadvantage was still reflected in differential performance on these tests no less than on more conventional intelligence tests which were in addition more predictively valid in respect of tested achievement and teachers' assessments.

One of the most interesting and definitive studies in this area is that conducted by Ortar (10). She administered both a Hebrew version of the WISC Verbal Scale and also the Performance Scale to upwards of 1000 Israeli children. These were divided into five groups with different cultural backgrounds ranging from recently arrived Oriental immigrants to an Israel-born 'high status' group (mainly of European parentage). After re-standardising both Scales for Israeli children, she found the 'cultural distances' between the groups to be larger on the Performance than on the Verbal Scale. In a similar study conducted with Scottish children Tsakalos (13) found differences in social status to be reflected in differential performance on the Jenkins non-verbal test no less than on Moray House tests of verbal reasoning and scholastic attainment.

The conclusion is inescapable that it is fruitless to search for testing instruments that will somehow transcend cultural differences and educational inequalities. What are the implications?

In the first place, it must be recognised that belief in the essential

equality of man receives little support from the considerable research in this area which it has stimulated. It remains an act of faith. This need not deter us from acting on that belief. A warrant from psychologists qua psychologists is not essential to the maintainance of a fundamental principle on which the advance of civilisation is predicted.

Secondly, it has to be accepted that educational disadvantage is endemic and that there is no simple counter to it by way of tests purporting to reveal intelligence, talent, potential, or whatever we may choose to call it, irrespective of differences in cultural, social or educational background. Such tests are of dubious value to a primary school teacher in Britain faced with an influx into her class of immigrant children without a word of English among them. There is no simple way of helping her to differentiate among them, or between them and their native-born peers, in terms of 'basic' intelligence. Her best practical policy still is to do all she can to make them feel welcome and to teach them English. Likewise, such tests offer no panacea to a developing country where, because of scarce resources, stringent selection is necessary and too many children are chasing too few places in the educational sun. The brutal truth must be faced that there are plenty of other children whose claim for preferment is no worse than that of the fortunate few selected. The solution to the problem is economic, not psychometric.

From an educational stand-point, the best hope of advance in general and amelioration of educational disadvantage in particular, lies in the field of language-teaching. The mother-tongue may suffice if it provides for effective communication with other nationals and is suitable as a medium for advanced education. If not, a second language is necessary, taught, as Vernon points out, not peripherally, but as a central tool of comprehension and thought.

What then should be the role of the psychologist? There is no reason why it should change materially, though possibly a shift of emphasis is indicated. Any still engaged in the search for testing instruments equally 'fair' in different cultures should bear in mind the fruitless quest of the alchemists for the philosopher's stone; though they may console themselves by reflecting that (in a different sense from the original alchemists') the transmutation of metals has now been accomplished. There is a lesson here. That achievement was the outcome of 'pure' research not specifically aimed at transmutation, nor concerned with its consequences. So too with the psychologist. He should listen to Anastasi's (2, p.302) warning: 'It is not (the psychologists') role to provide ready-made solutions to insoluble problems. It might be salutory if testing gave less heed to the pull of practical needs and more to the thrust of behavioural sciences'.

But less heed is not the same as no heed at all. The psychologist, like the physicist, has responsibilities outside his laboratory. Despite all that has been said, he has yet much to give in the field of testing in the service of education. It is a truism that the best indicator of a child's learning potential is a test sampling previous learnings which are relevant to the criterion or criteria we wish to predict. For long enough this maxim has guided with reasonable success the construction of tests for educational purposes within western cultures. There is still room for further research of the kind that Schwarz (11) has engaged in, aimed at discovering the previous relevant learnings in cultures elsewhere in a stage of transition.

Let Vernon (15, p.229) have the last word. 'What is important is that in concentrating on abilities recognised by western cultures, psychologists should not neglect special talents that might be more highly developed in other countries'. To extend a metaphor employed earlier, in seeking out these special talents we may be taking a small but useful step towards the assembly of a cross-cultural orchestra.

REFERENCES

1. Anastasi, A. (1964) Psychological Testing. (2nd Edition). New York: Macmillan.

2. Anastasi, A. (1967) 'Psychology, Psychologists, and Psychological Testing'. In American Psychologist, 22,4, pp. 297-306.

3. Bernstein, B.B. (1961) 'Social Class and Linguistic Development: A Theory of Social Learning'. In Education, Economy and Society (Halsey, A.H.) Glencoe: The Free Press, pp. 288-314.

4. Bernstein, B.B., and Young D. (1966) 'Some Aspects of the Relationships between Communication and Performance in Tests'. In Genetic and Environmental Factors in Human Ability (Meads, J.E. and Parkes, A.S.). Edinburgh: Oliver and Boyd, pp. 15-23.

5. Cattell, R.B. (1963) 'Theory of Fluid and Crystallised Intelligence: A Critical Experiment'. J. Educ. Psychol. 54 pp. 1-22.

6. Doob, L.W. (1960) Becoming More Civilised. New Haven: Yale University Press.

7. Eeels, K., Davies, A., Havighurst, R.J., Herrick, V.E., and Tyler, R.W., (1951) Intelligence and Cultural Differences. Chicago: University of Chicago Press.

8. Goodenough, F.L. and Harris D.B. (1950) Studies in the Psychology of Children's Drawings': II. 1928-1948. Psychol. Bull. 47, pp. 369-433.

9. Hebb, D.O. (1949) The Organisation of Behaviour. New York: Wiley.

10. Ortar, G. (1963) 'Is a Verbal Test Cross-Cultural?' In Scripta Hierosolymitana. Jerusalem: Magnes Press, The Hebrew University, pp. 219-35.

11. Schwarz, P.A. (1961) Aptitude Tests for Use in Developing Nations. Pittsburgh: American Institute for Research.

12. Tannenbaum, A.J. (1965) Review of IPAT Culture Fair Intelligence Test. In Mental Measurements Year Book (Ed. Buros, O.K.) New Jersey: The Gryphon Press, pp. 453-4.

13. Tsakalos, P. (1966) 'Is a Non-Verbal a Culture Fair Test?' Master of Education Thesis (unpublished). University of Edinburgh.

14. Vernon, P.E. (1960) Intelligence and Attainment Tests. London: University of London Press.

15. Vernon, P.E. (1969) Intelligence and Cultural Environment. London: Methuen.

16. Wesman, A.G. (1968) 'Intelligence Testing'. In American Psychologist, 23, 4, pp. 267-274

TESTS AND MEASUREMENT PROCEDURES, REVIEW AND EVALUATION.*

G.J. Matys
Curriculum Research and Development Unit
(Measurement and Evaluation Section),
Ministry of Education, Ghana.

PART I

TESTS AND MEASUREMENT PROCEDURES:

To a considerable extent modern education is characterised by the emphasis it places on adapting the educational programme to the needs of the individual child. Since these needs are governed by the child's level of ability and by the degree to which he has mastered the educational contents to which he has previously been exposed, it is important to determine these factors as accurately as possible. Once this information has been obtained, much can be done to individualize the educational process for each child, in part by grouping children into homogeneous instructional groups and in part by differentiating instruction within the classroom. It follows then that accurate educational measurement is a prime and key factor in modern educational trends.(1)

So if in modern education the emphasis is on teaching the individual it follows that we must have knowledge of, and differentiate between, individuals. Knowing the individual requires evaluating and testing. At the same time everywhere in the world the role of testing and evaluation in education is being questioned, criticized and scrutinized as never before.

Now before one goes further into the subject a definition or two should be made. The first is that Educational Evaluation is much broader than "testing". Educational Evaluation uses a variety of methods to measure and assess. These include questionnaires, surveys, cumulative records, projects, class work, oral answers, role playing and so on. Secondly the tools of educational evaluation are not a precise measure as are the tools of the engineer and physical scientist. In education one is trying to measure aptitudes, or intelligence, or content and processes - all very intangible and very difficult to assess. So it must be taken as a premise that the best of tests, under the best of conditions, provides large factors of error. When conditions (the training of personnel, administration and the tests themselves) are not ideal, results are even less reliable and less meaningful.

Testing had a traditional and fixed role for many years. It marked the end of one phase and the beginning of another. It meant passing out of one grade and into another. Successful completion of examinations allowed one to enter a profession, such as medicine. This type of testing has something in common with the initiation ceremonies characteristic of many non-Western societies and of various secret or exclusive groups within Western society. The examination or the initiation ceremony is a more or less

---

(1) Test Service Notebook - Test Bulletin, Harcourt Brace and World Inc.

* Originally included in the documentation for the Commonwealth Conference on Education in Rural Areas, held at the University of Ghana, Legon, Accra, Ghana, 23 March to 2 April 1970.

difficult procedure; if the examinee reaches a certain level of performance, a level agreed upon by the elders of the society, then his status becomes altered and he becomes permitted to practice as a doctor: the high school leaver is enabled to seek employment. The characteristic of this type of examination is that it marks the end of one phase in the person's life, and it demonstrates that he is competent to enter a new phase. It is perhaps worth making another distinction between (a) the terminal examination in a particular vocation, which is intended as one indicator of the individual's competence (and the examination should not be the only indication which is used), and (b) the final examination at the end of a non-vocational course. The former should ensure that society is not plagued with incompetent doctors and other professionals, but the latter has a less clear social raison-d'etre, and would appear to be a potent agent in the development of a society in which every adult's status is determined by his scores on tests taken during his adolescence.

The second reason advanced for examining is as part of a continuous process of education, a method by which the teacher assesses what each student has and has not learned. This use of examinations has been going on for many years and it might seem hardly to deserve comment, - but the present renewed interest in "Measurement and Evaluation" seems to stem from a more precise analysis of the assessment process than used to be practised.

In modern education this first type of testing plays a less and less important role. Testing now is seen as one type of evaluative procedure and only a part of the overall educational process. There was a period recently when some advanced countries had an almost religious faith in tests, whether they were tests of aptitude, ability, achievement or intelligence. Test results were felt to provide final, definite, and reliable answers to many questions. In another, earlier, period educators ignored and discounted tests and their results as useless. It was felt that one could not measure "intelligence" and other "intangible" processes of man.

Today a middle-of-the-road approach which avoids either of the above extremes is gradually emerging. Educators are realising that tests are far from useless, and yet far from providing all absolute answers. It is realised that tests are only one factor, one piece of information which becomes valuable when combined with school marks, common-sense evaluations of teachers and a multitude of other data, some scientific, some less so. Also it must be kept in mind that measurements are only tools, a means to an end, and not an end in itself. We do not weigh or measure an article just for the sake of knowing how heavy or how long it is. We use this knowledge in some way. So it is with testing - there must be some definite purpose in the testing we do. Are the tests simply to measure achievement? Are they diagnostic tests only, to be used to diagnose teaching weaknesses and learning difficulties? Are the tests to measure a person's potential for academic or other fields, his capacity or aptitude? Do the tests try to measure interest, attitude, intelligence or personality? There are instruments available today which attempt each of these, or some combinations of these, and so the purpose must be clear from the start.

Generally the types of tests listed above are Standardized Tests - tests built by experts over long periods of time and with carefully selected norms and so on. However, both these and teacher-constructed tests have a role or purpose in the classroom. What are some of the purposes of classroom testing?

Let us examine some of the purposes and uses of such testing (this applies whether the tests are "teacher constructed" or "standardized"):

(1) To test pupils' achievement. This is probably the most common purpose of testing. The teacher should have constant feed-back on how well the skills taught have been mastered and how well the concepts and understandings can be applied. However, the diagnostic aspects of achievement tests should not be overlooked at any time.

(2) To assess the effectiveness of instruction. Educators are so prone to say, when looking at the results of a test, that the pupils have done "well" or "poorly". Frequently the results of a test are more an indication of how well the teaching has been done. If the results of a test show weaknesses, the teacher has the opportunity to re-teach, change the method of approach, or seek for other methods of increasing the effectiveness of instruction.

(3) To motivate pupils to improve in their work. Test results will encourage most pupils to put forth their best efforts. A word of caution is in order here. Every class has pupils of varying abilities. We do not expect all of them to run at the same pace when they are racing. In the same way it would be wrong to expect all pupils, the bright and slow ones, to achieve the same standards on a test. It would be wrong, therefore, to compare the mark of a pupil with lower ability with that of a pupil with higher ability. It is sound, however, to stimulate pupils to improve their own marks on successive tests rather than comparing them with the brightest pupils.

(4) To discover individual problems and weaknesses. The test results will identify pupils who have particular problems and the teacher then has the opportunity to provide individual help and instruction to such pupils.

(5) To provide a sound basis for keeping parents informed regarding pupils' progress. Parents are usually interested in knowing how their children are performing. If full records of test results are kept by the teacher, these form a good basis of communicating pupil progress to the parents.

(6) To locate or identify weak areas in the teaching-learning situation. If test results are analysed carefully "gaps" or weaknesses may be discovered and necessary steps taken to deal with them. Methods of doing this are mentioned later.

(7) To gain information for grouping pupils for instructional purposes. It has already been mentioned that pupils within a class vary greatly in their innate ability to learn. The slower pupils require simpler explanations and more instruction and drill. This frequently becomes boring to the brighter ones and causes them to lose interest. Test results would indicate which pupils might be grouped to provide the most suitable instruction.

(8) To gain knowledge about individual pupils for guidance purposes. If full records of test results are kept pupils' strengths and weaknesses as well as their special interests will be discovered. This information can be used in guiding pupils to make proper choices when they go on to further education or when choosing a vocation.

The above suggestions apply particularly to teacher-made classroom tests. Standardized Tests that are prepared for a wider use, such as throughout a school system or country, would serve other purposes as well. Such tests should provide even more help and information to teachers, parents, administrators, curriculum makers and educational planners and policy makers. This additional information would help:

(1) to evaluate courses or syllabuses for the purposes of revision, etc.;

(2) to compare different methods of instruction and assess teaching methods;

(3) to ascertain standards of classes within a school, a district, a region or the country as a whole;

(4) to assess the work of individual teachers;

(5) to assess pupils and recognise the individual characteristics of each pupil (we should know: (i) his difficulties and weaknesses
(ii) his strength and present knowledge);

(6) to provide a means for pupil and teacher review, an integral aspect of learning;

(7) to provide information for educational planning and policy making.

SOME OTHER GUIDELINES

It is worth emphasizing that testing should be done only when we know how and by whom the results will be used. Of all the functions for which tests may be used, the least valuable function educationally a test can perform is when it is used only by administrators and only in passing, failing, admitting or screening students. Yet examinations with this function alone are still quite common. The most valuable function of tests is in helping the pupil and teacher communicate, and derive benefits from the learning process. In addition tests should be constructed with a clear knowledge of all their educational objectives and should be critically evaluated to see whether they are valid (measure what they are supposed to measure) and are reliable (consistently measure the same thing in the same way). The interpretation of course must be logical, attributing no more, or less, to a particular result than it deserves.

It is worth noting also that if testing and evaluation are to become an integral part of the educational system, teachers must know something about the field. Generally today teachers learn how to demonstrate, explain and put a point across. But little or nothing is given them on how to evaluate

get feedback and measure what changes have taken place in the pupil. And yet to be truly effective a teacher must know (a) what the pupil has already; (b) what he has failed to learn and, if possible, why; (c) what the pupil is capable of learning.

There are, then, certain problems and pre-conditions to good testing that must be attacked simultaneously with any attempt to enlarge the role of evaluation in schools. These include:

(1) A FOUNDATION OF TEACHER TRAINING in the understanding, interpretation and use of tests. Programmes must be developed to improve this in the colleges and through in-service training. A basic record-keeping system (cumulative records) is needed together with teachers who can use it properly.

(2) DEFINED OBJECTIVES. Good evaluation programmes can help to show how far the school programme meets the objectives of education. This presupposes that there are measureable objectives set forth for general education as a whole and also detailed objectives for each subject. It is only against some objectives, however simple, that one can evaluate.

(3) OTHER GENERAL PROBLEMS TO TEST DEVELOPMENT IN A DEVELOPING COUNTRY.

(a) Difficulty with control groups due to seemingly high turnover among pupils and teachers, very different levels of teacher training, lack of records of ages and other data etc.

(b) Lack of pupil and teacher familiarity with the notion of (i) carefully timed tests (ii) objective tests.

(c) Greater differences than in developed countries between urban and rural cultural factors.

(d) Administration problems (developing the effective machinery necessary).

(e) Language problems. Literature and other evidence suggests that any standardized test meant to measure "anything". In countries where English is not the first language it will, in fact, measure largely facility with the English language.

THE INFLUENCE OF TESTS

In a modern technological world there is bound to be a great concern with accurate measurement. Scientists can calculate an exact point and time for a moon landing 240,000 miles away. It is inevitable, then, that this desire to evaluate accurately should spill over into education. And, as is pointed out elsewhere in this paper, the two extremes (a) of attributing too much, and (b) too little, significance to the role and value of testing, both exist.

One author says "Measurement touches upon and influences every phase of education. Whether it is marking, promotion, guidance and

counselling, curriculum development, instruction or some other aspect of the work, measurement plays an important part."(2) Examinations and marks can be called the currency of education. By these marks, or value assigned, people are passed, granted certificates, promoted, given degrees and so on. We often judge a man's worth by his academic percentages!

There is general agreement then that testing can and does have a profound effect on the educational system of a country. The methods of teaching, the emphasis in the curriculum, the attitude of teachers and students, are all affected or sometimes dominated by the examinations. The types of things stressed in examinations largely determine what happens in the classroom. It matters little what teaching notes or syllabuses are prepared unless the examinations reflect the same spirit and aims. This is especially true where there are large scale and important external examinations.

EXAMINATIONS

Although examinations should measure what is being taught in the classroom, it is very easy for the situation to develop where we teach what is tested rather than test what is taught. Curriculum development and examinations cannot be separated and should be developed in close harmony at all points. Persons sitting on Curriculum Panels or Examination Panels should both be familiar with the general national aims and objectives of education as well as the specific spirit and aims of a given sullabus.

All this does not mean that examinations are the only determining factor in education nor is this a criticism of external examinations. The important thing is that these things should be in the right order and priority; tests should serve the educational goals and needs, not determine them.

The author recently sat on a committee the members of which were drawn from the Ghana Ministry of Education and the West African Examinations Council. A paper produced as a result of these meetings had in part this comment on examinations:

"In spite of inherent weaknesses external examinations are useful and necessary in many situations. In Ghana, for instance, some common measure is needed to provide objective norms and maintain a common standard owing to great disparities in:

i) staff;

ii) training facilities;

iii) libraries and supply of textbooks, etc."

The point here, then, is not to weigh the advantages and disadvantages of external examinations not to debate how much external examinations can affect classroom practice, for such argument or debate has limited value.

---

(2) V.H. Holl, Introduction to Educational Measurement, Haughton Mifflin Co., Boston, 1965.

INTEGRATING EXAMINATIONS AND CURRICULUM

The important thing, then, is to recognise

(1) that there is interaction between examinations and curriculum;

(2) that examinations are not simply passive instruments of assessment but an integral and vital part of the educational process;

(3) that both examinations and curriculum are important and powerful forces for change;

(4) that both form part and parcel of the educational process; and

(5) that both should be under constant review in terms of relevance to changing needs.

The central problem - which must be true of every educational system - is to find the most effective ways of ensuring that curriculum planning and examinations complement each other and work towards the same end. In other words, what should be done is to make sure that the examinations used (a) reflect the same goals (b) promote the same spirit, objectives, emphasis and priorities that the curriculum planners had in mind. Without the proper integration with curriculum, a tester starting from the same written syllabus could build several examinations, each one providing a different emphasis and different educational objectives and goals. It is for these reasons that the contacts between curriculum builders and the examiners must be continuous, and at all stages of development.

DEFINING OBJECTIVES

To establish contact between curriculum planners and examiners it is essential that the objectives and goals of education, both general and specific, should be clearly defined and clearly set out. Without clear direction as to the goals and objectives in education it can follow that there can be the situation where the main emphasis will be teaching what is tested rather than testing what is taught. Examinations can either lead or follow in education. When examinations become the key determiners of curriculum and education, it is usually by default, because the curriculum planners and syllabus writers have not been clear enough in their directions and objectives. Similarly if objectives are clearly defined but examiners are not properly informed about these and cannot translate them into the examination material the same unhappy situation may occur.

In order to meet today's needs, curriculum panels or testing panels must be conversant with:

(1) modern testing ideas;

(2) general aims and objectives of education for the country;

(3) the desirable objectives, spirit, and emphasis for that particular syllabus or subject.

CONCLUSIONS

Curriculum and examinations are two sides of the same coin and it is only when they operate together that the goals and objectives of education can be adequately reached.

The pre-requisites, then, are:

(1) that curriculum makers must build into the original curriculum evaluation and testing goals and objectives, and

(2) see to it that they reach the examiners who actually make up the tests;

(3) that the examiners (i.e. the examining body) must keep themselves fully informed at all stages of curriculum planning and defining objectives, and

(4) become conscious of the spirit, aims and implications of the written syllabus;

(5) that both curriculum planners and examiners, working as a team, appreciate where they are leading and heading from the earliest stages in terms of what will be measured and how.

PART II

REVIEW AND EVALUATION:

As indicated in Part 1, the focus today in progressive education is on individualizing education, focussing on the individual. This means getting to know the pupil. Knowing the pupil in turn requires a number of practices including the necessity of measuring and evaluating each pupil in order to recognize individual differences. Hopefully the day of considering examinations as something separate and apart in education is over. The curriculum, the teaching and learning process and test and evaluation procedures should all be part and parcel of one complete and integrated process. Evaluation procedures turn education from a teacher-to-pupil 'monologue' into an effective 'dialogue' and communication. Continuous evaluation makes the process truly fruitful and meaningful.

Two main reasons have been given earlier for having examinations: the first is the need to provide evidence of an individual's competence to move from one social status to another; the second is to provide, as part of a continuous process of education, a method by which the teacher can assess what each student has and has not learned.

An analogy to educational assessment could be the study of cybernetics, a comparatively new science which is concerned with the behaviour of control systems in the physical and biological worlds. Perhaps the basic law of this science is that goal-seeking systems are error-actuated. What does this mean? Here is an example: a missile which "homes" on a target does not, in fact, go straight to it. Its course is constantly changing and its direction is modified in accordance with "feedback" information about its errors. The missile receives feedback which tells it how it is off target, and it then makes appropriate compensatory movements, though it can never be said to be exactly "on target". The same process can be recognised in

many human goal-seeking activities, and it seems directly applicable to the educational process. The teacher can apply this process consciously if he draws up a list of objectives which he hopes to achieve (he hopes to effect certain changes in his students as a result of the course he is teaching) and then by frequently evaluating his pupils' progress. He then has the feedback information which is necessary to reduce the errors inherent in progress towards any goal.

Most control systems have an optimum frequency for receiving feedback, they will swerve this way and that in their progress towards their target, but if they receive feedback too frequently, they may be unable to process the data at sufficient speed. (e.g. the feeling when one set of essays is due for collection before you have finished marking the previous set?) So the first question is: How frequently should one obtain such information? As we shall see later the answer is, as often as possible, in fact, continuously (but systematically).

It is a common occurrence to find the classroom teacher most surprised at the results of a testing programme. Assuming the test is a good and valid one, it shows how often teachers know little or nothing about the progress and capabilities of individuals, or indeed of even a whole class. Perhaps the most vital point in sound educational evaluation is the fact that to be effective evaluation must be both diverse and continuous. Part 1 of this paper, on the role of testing, emphasizes that tests are not precise instruments and that even the best of tests under the best conditions leaves large margins for error. For the sake of accuracy and reliability alone, evaluation must be continuous and not just a periodic event for selection or admission. However, as has been pointed out already, testing should be the other side of the educational coin, the means whereby communication and feedback to the educators is established. So, in order that evaluation procedures may provide accurate information to (1) pupil, (2) teacher, (3) parent and (4) administrator, and in order that evaluation may become an integral and useful part of the educational process, it must be continuous.

As a periodic event applied at certain times for screening and selecting pupils, testing has a limited educational function. And this is mainly an administrative function since it allows those in charge to pick people, pass and fail people, and assign a certificate or value to the person in question. It has limited educational value because it does not necessarily help the teacher to teach better, or the pupil to learn more effectively, which is, after all, the core of the educational process. Evaluation taken in the many forms discussed later, and used for diagnostic and remedial purposes on a day-to-day basis is the type of evaluation which really is a valuable part of the teaching-learning process. In this sense, written tests, oral questioning, quizzes, projects, reports, classwork, homework, etc., etc., are all considered as evaluative measures. Records are kept and results are analysed first in the classroom to help the teacher teach better and pupil to learn better, and secondly in the larger area of planning, curriculum, etc.

This is not really something dramatically new. It simply involves a consciousness on the part of teacher and other educators of the need for constant feedback and communication from the pupil. It shows a mature realization of the weaknesses and deficiencies of any one test or group of tests. It simply means a greater emphasis on gaining more and better information about the pupil and what is happening in the classroom and then using this in the next stages. It also involves a recognition of the importance of individual differences and the realization that to know pupils well and

accurately, we must assess (1) very frequently and (2) in as many ways as possible. It means that the focus is taken off the class as a single entity which must absorb a certain amount of material and be able to regurgitate this in an examination. Instead, education with evaluation as an integral part focusses on:

(1) developing the full potential of each individual to his or her capacity;

(2) seeing education as a dialogue between the teacher and the learner, where both, communicating effectively and constantly, 'grow' together;

(3) providing more accurate information as a result of continuous, diverse, and multi-faceted evaluation to all the people who need it - first of all the teacher and pupil, secondly parents, thirdly curriculum builders, planners and administrators.

But to be successful, a wide-scale use of testing and evaluation as a continuous element in the schools requires certain basics and pre-requisites. The main one is in the training of the teachers both in training institutions and through in-service work. Before testing can play the role described above, an understanding of some testing theory along with enough technical knowledge to understand, use and interpret tests and evaluation procedures must be basic to all education officers and teachers. They must be able to make effective use of objective as well as essay questions and to be able to use all the other evaluation techniques. In the field of objective tests they should have a practical classroom knowledge about the construction and interpretations of the various types of objective tests such as multiple choice, fill-in blanks, true-false, matching and so on. In addition a "guidance approach" to the child and test results and cumulative records should be a part of the teacher's equipment.

It should be emphasized as well that testing and evaluation does not mean only large scale sophisticated standardized tests. Widespread use of classroom testing combined with effective records and use of results, could provide much of the evaluation data now lacking. "Measurement devices and techniques prepared by the teacher are often the best and sometimes the only means of determining how well a class or individual pupils are progressing towards the objectives of instruction."(3)

But in such continuous classroom testing the teacher must develop a certain level of sophistication. For instance, the ability to analyse test results (do a simple item analysis) can add a great deal to the teacher's knowledge of the effectiveness of his teaching and the extent of the learning. It is also a powerful tool for test improvement. Item analysis also indicates which items are too easy or too difficult to discriminate between better and poorer examinees and it can be done simply and with little loss of time in the classroom. Then, too, teachers must understand validity in testing - that is, that the test measures what it is supposed to measure.

---

(3) Victor H. Holl, Introduction to Educational Measurement, Haughton Mifflin & Co. Boston, 1965.

The classroom teacher who has a sound knowledge of testing and who has available cumulative test records is in a far better position to understand the learning problems and difficulties of individual children. He is able to identify the most capable youngsters, who need enriched learning experiences, as well as the slow learners who may need special help and modified assignments. The slow learner who achieves less because he is slow mentally is a perfectly normal child; he should in no sense be considered a failure simply because he does not reach the average level of achievement of children of his own age or grade. On the other hand, the child with high ability who does mediocre work, is, in a truer sense of the word, a school learning problem. The concept of failure in school is one with which we could very easily dispense since it is never possible to determine with certainty who is failing - it may be the school quite as much as the child.

lt should be stated emphatically that standardized testing here is no complete substitute for an effective evaluation programme on the part of the classroom teacher. Such an evaluation programme includes the teacher's own locally constructed tests as well as ratings on specially assigned projects and daily classroom recitations. Nevertheless, the professional technicians who develop standardized tests can offer the classroom teacher many suggestions for evaluating the results of classroom instruction. Indeed, much in-service training is needed in this important area.

At several points it has been mentioned that evaluation must (1) be continuous (2) employ a wide variety of evaluative procedures. What are some of these ways of evaluating? Here is a list of some of the many common-sense methods for diverse and multi-faceted evaluation:

(1) Tests - there are many types including:

  (a) Achievement:

    (i) informal teacher-made

    (ii) standardized

  (b) Mental ability

  (c) Personality

  (d) Aptitude

  (e) Interest;

(2) Rating scales;

(3) Checklists, surveys, inventories and questionnaires;

(4) Observation;

(5) Records and reports:

  (a) cumulative folders,

  (b) anecdotal reports,

  (c) diaries and logs;

(6) Interview;

(7) Sociometry;

(8) Role-playing:

(a) sociodrama,

(b) psychodrama;

(9) Situational or performance tests;

(10) Student papers and projects:

(a) papers,

(b) notebooks,

(c) reports,

(d) autobiographies,

(e) personal data sheets;

(11) Case studies;

(12) Case conferences.

Tests must measure all the important outcomes of instruction such as course objectives, factual knowledge, understanding of human nature, the proper weight for each topic and so on. Benjamin Bloom(4) in his taxonomy lists six main objectives or outcomes of learning in the cognitive domain that should be measured. These include:

(a) Knowledge of specifics, ways and means of dealing with specific universals, abstraction from specifics;

(b) Comprehension, involving abstraction, interpretation, extrapolation of communication;

(c) Application of knowledge;

(d) Analysis of elements, relative principles;

(e) Synthesis;

(f) Evaluation.

Before continuous evaluation becomes a full and integral element in education there must be clearly defined educational objectives. Good evaluation shows how far school progress meets the objectives set out. These must be clear both to the policy makers, examination bodies and teachers. Unless these are clarified, testing cannot play its proper role.

---

(4) Bloom, B.S. (ed) - Taxonomy of Educational Objectives, Handbook 1: Cognitive Domain, David McKay & Co., Inc., New York, 1956.

Then too in newly developing countries there are other problems which often require close attention before testing can be effective; these include:

(a) Keeping records. Records of pupils' ages are basic to many kinds of testing. Continuous evaluation is only useful when a cumulative record of each individual is kept and used.

(b) Both teachers and pupils must become familiar and at ease with such things as objective tests, the concept of "time tests" where every minute counts, and so on.

(c) Often in such countries there is a less settled population of teachers and pupils, both of whom move about and leave schools frequently. This makes for difficulties in establishing norms, control groups and experimental groups.

In newly developing countries differences between urban and rural groups tend to be bigger. Urban groups quickly become sophisticated in a variety of aspects and ideas while rural pupils remain almost totally unaffected. Language problems, administrative problems and others must all be tackled in a special way for rural areas.

It was mentioned earlier that in addition to evaluation being continuous it must include a wide variety of techniques and methods. In testing, we are measuring people and their responses and knowledge and not bricks or bridges. Physical things can be measured accurately and completely with a ruler or scale. Because of the complexity of man and the complexity of the facets we wish to measure, there is more chance of accurate assessment if a variety of techniques are used and used often. And by accumulating and combining results, we are more likely to measure accurately the many processes and facets that we see as the goals of education.

In summary, then, worthwhile testing should meet a number of criteria and requirements. An attempt is made overleaf to include the key elements in a graphic form:

GOOD WORTHWHILE TESTING

Proper Administration

Timing, testing conditions, uniformity, etc., are all properly done.

Good Sampling, Clear Objectives

Questions measure different aspects and levels (Bloom) - valid and reliable tests - the tests measure all the objectives of education and measure it always in the same way.

Continuous Evaluation

A variety of types of ways of testing are used continually as an integral part of the teaching and learning process. Testing must be continuous:

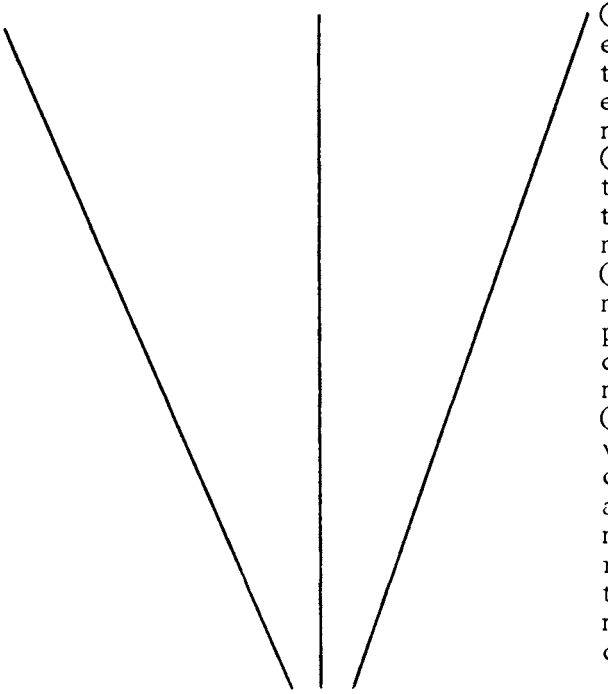
(1) to provide communication and dialogue between teacher and pupil;

(2) to make evaluation an effective teaching and learning tool by providing pupils and educators with feedback information on what is happening;

(3) to allow pupil and teacher to assess progress, assess their work and make adjustments;

(4) to provide up-to-date information to teachers, pupils, parents, administrators, curriculum builders and policy makers;

(5) to compensate for the inherent weaknesses in the results of any one test. The average results of a great many (continuous) assignments are much more valid, reliable, accurate and meaningful than the results of any one or two major tests however carefully constructed.



PROPER USE OF TESTS

Interpretation

Test results are interpreted for what they are. No more or less evidence or value is attached to them than they deserve. Test results are not regarded as "god-like" nor as "useless". They are given their proper due and right and used as one piece of information along with all other information available.

Well used information

A truly useful test provides information to all of:

1) pupils
2) teachers
3) administrators and inspectors
4) parents
5) school curriculum builders and policy makers.

Variety of Technique

Evaluation must take a variety of forms and include as many different kinds of evaluative techniques as possible. These should range from carefully constructed standardized terms to common-sense evaluative observations and ratings of teachers. This is because

1) we are assessing complex human beings;

2) the qualities being measured are tangible, abstract and difficult to measure;

3) more kinds of measurements taken more often ensure more accuracy, reliability and validity.

Part I of this paper presented some points to stimulate discussion on the topic of the nature, place and influence of tests and measurement procedures including examinations; Part II has focussed on continuous review and evaluation. As was pointed out in the opening statements of Part I, the author feels that the two topics are indeed only one. Evaluation is a continuous and integral part of education. The frequent repetition in Part II of points from Part I are meant to emphasize this point.

However, for organizational purposes, the paper was divided into two parts. It is hoped, however, that the overlapping and repetitions of similar points in the two parts may lead to a line of thinking that combines the two ideas.

This marks the end then of a few brief ideas in the field of measurement. Not all points are covered, nor are those that are mentioned covered adequately or completely. However, hopefully, enough has been said to provide the raw material so that discussion can whittle away the rough edges and produce a refined and finished product on this important and contraversial subject. Perhaps because of its deficiencies, this paper will serve the better to stimulate discussion, which is, after all, its purpose.

PUBLIC EXAMINATIONS AND THE CURRICULUM

Dr. S.M. Chari
Joint Educational Adviser
and Chairman, Central Board of Secondary Education, India

Public examinations play a vital role in the educational systems of many countries. The degree of their dominance over the educational system may vary as between advanced and developing countries, but the fact remains that they have come to stay in many systems. In India, an examination of this kind is usually taken at the end of ten or eleven years of schooling. Some States provide for a common examination at the end of seven or eight years of schooling. The examinations are conducted by State Boards of Education for each State in India and by the Central Board of Secondary Education for such schools as follow all-India course of studies. Sometimes more than a lakh of students take an examination. It is altogether an elaborate and involved process.

Owing to historical reasons, the public examinations in India came to dominate the educational system. In the middle of the nineteenth century, universities were established in India 'to ascertain by means of examinations the proficiency acquired by candidates and to provide a test of eligibility for Government service'. Thus, to succeed in examinations meant job opportunities. A thirst for education on a mass scale was evident in the initial phases of our educational development arising from the close connection between obtaining a certain standard in examination and getting good job opportunities. Examinations obtained a unique responsibility in society. They denoted excellence and were a passport for entry into the realm of lucrative careers. Examinations thus became the all-important motivating factor in education and relegated the true purpose of education to the background. As time passed by, not only was the true aim of education lost sight of, but also the true purpose of examinations. The purpose of examinations is, without a doubt, to reveal what progress the pupils have made, whether a desire has been instilled into them to continue to learn, and whether they have developed the capacity to reflect, enquire, investigate and draw conclusions. Indeed, examinations should serve the purpose of education. To do so, their techniques should be subject to constant evaluation and reform. Rigid, formal assessments should give place to internal assessment.

During the last three decades there has been a growing concern all over the world about the dominance of public examinations over the educational system. The question whether the entire system of examinations should not be abolished has also engaged the attention of the educationalists. In India, several Commissions expressed dissatisfaction with the examinations system. They found that the public examinations merely sought to measure the level of achievement of a pupil at a given moment of time. Their evaluation did not embrace the whole personality of the pupil. Even within their limited sphere, they were unreliable owing to their defective mechanism. But academically speaking, some sort of evaluation is necessary to ascertain whether the growth sought through formal education has been in the right direction and if so, to what extent. It was this view which was at the back of the 'National Policy on Education' issued by the Government of India in 1968, which laid down 'A major goal of examination reform should be to improve the reliability and validity of examinations and to make evaluation a continuous process aimed at helping the student to improve his level of achievement rather than at certifying the quality of his performance at a given moment of time'. It follows from this that in order to rid the traditional system of too much dependence on public examinations alone, it is vital that a two pronged attack is made. The first is changes in the evaluation techniques and reforms in the curriculum.

Taking note of the limitations of the public examination system, the Central Advisory Board of Education in India has recently set up an Examination Reforms Committee, which has produced a well-considered report. Among the reforms it has suggested is permitting a system of 'autonomous schools'. It recommended that some of our best schools which have a good tradition should be identified and freed from the restrictions of the public examination and allowed to examine their own pupils. If this recommendation is carried out, it will open the way for experimentation not only in the system of examination, but also in the methodology of teaching and learning in curricular construction and evaluation. It will help the effort to ensure that (a) examinations do not become an end in themselves but serve the purpose of education; and (b) that purpose of education is clearly reflected in the syllabus, textbooks and gradually in teaching practice. It will lead to improved techniques of assessment such as oral tests, questionnaires, anecdotal records, checklists, project reports, rating scales and so on, with the aim of testing the student's ability to comprehend clearly to apply his knowledge in new situations.

Public examinations are necessarily written examinations. They test one type of ability, the ability to answer questions within a limited period of one to to three hours. Pupil evaluation has, however, to be a continual process and should go side by side with learning. If education is to build up and foster the integrated and harmonious development of the human personality, which indeed, is its aim, the curriculum must provide for such growth. The techniques of teaching should be directed towards the discovery of truth. Examinations are only a part of the over-all education to promote this discovery. External examinations suffer from being too remote from the individual classroom and the teacher does not always have the opportunity to test. The teacher's fitness should not be confined to teaching but should extend to testing as well. And this is where internal assessment comes in. In India, we are now moving towards the gradual introduction of this type of assessment. This will supplement the annual examination and help the examiners to obtain a broader student profile. The weightage given to both ranges from 25% to 50%.

If examinations should effectively serve the purpose of education, both examiners and teachers should have a new orientation. They should be exposed to the several facets which are inherent in the educational system. It should be their endeavour to delineate curriculum objectives in terms of pupil behaviour, to frame objective based test items specifically to test certain behaviour, to validate those test items, to draw a balanced blue-print for a question paper keeping the various objectives in view.

All this has assumed great significance in this era of change and challenge and in the context of the upsurge of knowledge, the early obsolescence of this knowledge and its rapid replacement of more up-to-date knowledge. It necessitates a corresponding training of teachers on a massive scale at all levels, the proper orientation of paper setters and the strengthening of the dialogue between the examining bodies and the framers of curriculum. India has awakened to the need for these urgent explorations. For, the quality of education depends, to a large extent, on the quality of evaluation.

In recent years, much has been taken in hand, both at the Centre and in the various States of India, to promote an awareness of the need for continual examination reform and for the development of proficiency in the various areas of educational growth. There is no doubt that there is a great deal more which requires to be accomplished, to enable us to move from a situation where there is equal measurement of unequal opportunities towards a more rational system, which encompasses not only educational goals but also national needs. The point to bear in mind is that there should be a continuous quest, and efforts towards reform have to be periodical. As in all fields of human endeavour, in this field also, there is no room for thinking that the last word has been said.