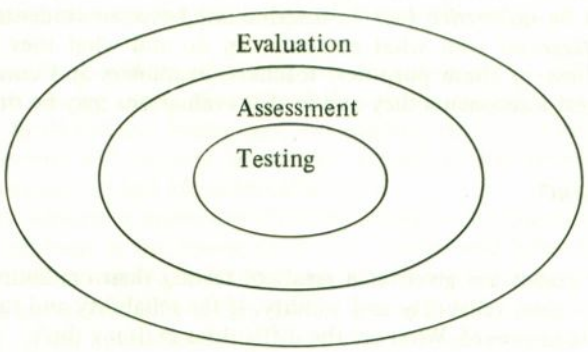
Trends in Formal Assessment

Dennis M. McGrath
RECSAM
Penang

Kebangkitan minat terhadap perkembangan kurikulum yang bermula pada tahun 1960-an di beberapa buah negara diiringi dengan pergerakan yang sama kuatnya terhadap pembaharuan penilaian dan teknik-teknik penaksiran yang sedang digunakan. Perubahan-perubahan yang berikutan ini menitikberatkan dua bidang yang utama: pertamanya, bahan-bahan kurikulum diuji di sekolah-sekolah dan penilaian formatif dilakukan semasa bahan-bahan diperkembangkan; dan, keduanya, pengembang-pengembang kurikulum, guru-guru dan pemeriksa-pemeriksa melaungkan perubahan-perubahan tentang cara-cara mengukur pencapaian pelajar di akhir kursus. Pelajaran-pelajaran yang diperolehi akibat dari perubahan-perubahan di dalam prosedur-prosedur penilaian dan penaksiran amatlah bermanfaat bukan saja untuk pengembang-pengembang kurikulum dan pemeriksa-pemeriksa tetapi juga untuk guru-guru. Kertas ini cuba menunjukkan sebahagian dari perubahan-perubahan ini terutamanya di bidang penaksiran formal, iaitu ujian-ujian dan peperiksaan-peperiksaan yang digunakan di dalam pengukuran perubahan-perubahan terhadap kemahiran-kemahiran dan sikap-sikap pelajar.

Evaluation or Assessment?

There is a very strong movement in education today towards testing and assessment. All around the world much money and effort is being put into the process of evaluation. Perhaps it is worthwhile pausing to clarify what the relationship between testing, assessment and evaluation can mean. One of the early Nuffield projects (from the United Kingdom) made the distinction with this diagram.



Evaluation in this approach is concerned with making judgements. Evaluation is often concerned with decision-making and is value-based. *Assessment* is concerned with describing a performance (which is any activity which can be observed or reacted to). This description of student attainment should be as value-free as far as possible. For example, an assessment does not say

"this student is better than this one" but simply describes what a student can do. To make evaluations we need assessments, although sometimes evaluations may be made with very little evidence from assessments. *Measurement* is the process by which assessments are made.

A *test* is any method (or device) that is used to sample students' behaviour. Tests can be of many types e.g. readiness (used at the beginning of a learning sequence), achievement (used at the end of a learning sequence) or diagnostic (used during, or at the end of, a learning sequence to help students to learn better. Students read answers and explanations of incorrect answers). Tests may be concerned with a variety of achievements too. Achievements associated with thinking (cognitive), with skills for performance (psycho-motor) and with interests, attitudes and feelings (affective) are being used more often.

Purpose of Assessment

"Good" teaching involves the setting of aims and objectives, providing students with the means of achieving the objectives (for example by lectures, discussions, textbooks, laboratories, reading lists, field trips and other aids) and finding out how well the objectives were achieved.

To find out how well students have achieved the objectives assessment procedures are designed. These can be formal examinations or class tests but their purpose is to show to what extent the students have attained the aims of courses. There is a value for both teachers and students in such assessments. For the teacher it is of value to know

- o if he is getting his message across to the students
- o whether some students have misconceptions
- o whether some students need help (by tutorials, or extra reading) and conversely if some students need extension and enrichment work, and
- o whether scholarships should be awarded to some students.

For the student there is a benefit from the knowledge of their progress and also from a recognition of their effort and achievement. Students would appear to work to achieve success in objectives which are recognised and rewarded. There is some research evidence from the United States that when a Pass/Fail grading system is used it reduces both the overall achievement and the effort that students put into their studies. Perhaps students are human after all!

Students, if they are provided with regular feedback about their progress, will discover their strengths and weaknesses and can make efforts to remedy weaknesses by further study.

Assessments can be *normative* (act to discriminate between students, such as an examination) or *criterion-referenced* (tell what students can do and what they cannot do, such as a mastery test). Regardless of these purposes, teachers, examiners and curriculum workers must attempt to make the best assessments they can so that evaluations may be the best as well.

What is Good Assessment?

When marks or grades are given as a result of testing their credibility will depend on two important properties – their reliability and validity. If the reliability and validity can be increased then assessments will be improved. What are the difficulties in doing this?

Absolute reliability is impossible to achieve. There are three factors to support this statement. Firstly, there will be variations in marking (both for one marker on different occasions and different markers on one occasion). Secondly there is always, in a test or examination, a restricted sampling of the content/or knowledge and abilities which a student might be expected to have learned or achieved. Lastly there is the matter of inconsistency of performance. A student will not perform at the same level on the same questions everytime. Factors such as motivation, tiredness and emotional state will affect his performance from one occasion to the next.

Reliability is usually referred to as consistency of measurement. When we consider the above factors it is not surprising that perfect reliability is impossible to attain.

Validity, which is really concerned with the accuracy with which we measure what we want to measure, can be increased by a number of techniques. In many of the standard texts on evaluation and assessment, for example, Theobald (1974) and Bloom et. al. (1956) different kinds of validity are referred to. One of the more recent uses of the term – curriculum validity – seems to be particularly useful for examiners and teachers. Curriculum validity refers to the correlation between the measuring instrument and the curriculum in all the three dimensions of content, objectives and teaching/learning activities. If there is a good match between the assessment procedures and what has happened during the course then a high curriculum validity would result. As an example in the early days of examining chemistry it was popular to use volumetric and qualitative analysis tasks for practical examinations. In fact, these types of tasks were only one part, and not a representative part, of the practical course as a whole. This would give a low curriculum validity to such an examination.

There is little doubt that learning is influenced by *what, how* and *why* we evaluate. Learners come to value what their teachers evaluate. If teachers tell their classes that inquiry science or practical skills are important but their tests reflect a concern for knowledge and recall of factual information, the learner will not value either inquiry or practical skills. This point seems particularly true when formal examinations are used. Examiners will often stress the more easily examined knowledge objectives and this will consequently affect the science taught in schools.

Teachers, then, need to ensure that what they evaluate does match with what they want their students to achieve. If they wish to emphasise practical skills they must give activities in the classroom that will enable students to use those skills and then evaluate (and provide feedback to the student) the extent to which these practical skills are being achieved. The evaluation can be before (through pre-tests, or entry test) during the instruction, or after the unit has concluded. If this is done then the curriculum validity of the assessment procedures will be increased.

How Can Validity and Reliability Be Improved?

In setting an examination (or test) teachers often intuitively think they have "covered the topics" they want to test. A more systematic approach is to use a *blueprint* or table of *specifications.* Often this is simply a two-dimensional grid. The first dimension is on the topic covered while the second dimension is on the outcomes of instruction stated in terms of student behaviours.

The behaviours really should be related to the objectives of the course. Some examinations formalise these behaviours by using Bloom's (1956) taxonomy (Knowledge, Comprehension, Application, Analysis, Synthesis and Evaluation) but this may be too restrictive for many teachers. A comprehensive blueprint will include behaviours related to the general areas of knowledge (cognitive), skills (psychomotor) and attitudes (affective).

Using a blueprint ensures a *balance* is obtained between the questions asked in the test or examination and the emphasis in the course (both for *content* and *behaviour*). The blueprint can also clearly indicate the weighting for each topic and behaviour as is shown in Table I.

If the weightings are indicated it is easy to check that in this example, 40% of questions are concerned with recall of information and only 20% with applying the information to new situations. It is also clear that topic 2, with 35% of the mark allocation, is regarded as the most important topic. These weightings reflect the emphasis of the course. If half of the course time was spent on topic 4 this blueprint would indicate a low curriculum validity.

There would appear to be a strong case for informing students about such a blueprint. This would clearly assist their learning of topics and help them to understand what behaviours would be expected of them.

Table 1
Table of Specification

TOPICS	BEHAVIOURS A	B	C	% Total
1	10	–	–	10
2	10	20	5	35
3	10	15	5	30
4	10	5	10	25
% Total	40	40	20	100%

In most subjects it would be a poor test or examination if the student did not have to reorganise his knowledge, to apply it to new situations or restate it in a new context, to draw contrasts and comparisons or to show independent thinking. A good examination probably requires some evidence of factual knowledge but an examination that overstresses this is probably not a good one unless the objective of the Course is concerned with factual knowledge. For example, if the objective is to master facts and regulations (anatomy, auditing) then these facts and regulations will be heavily weighted. If, on the other hand, the objective is original thinking (literature, philosophy, architecture) then a *valid* test should reward these behaviours. We may have to give weight to less reliable measures in order to increase the validity and relevance of the assessment. This means that if it is thought to assess some complex student behaviour and observational procedures are the only way to do so then they should be used. The reliability of such procedures may not be as high as one might wish but the validity (is the instrument measuring what it purports to measure) will be increased by the use of these procedures.

Reliability is a statistical concept which refers to the *results* of a test or examination. Probably the most useful notion is that of consistency of results. To achieve this consistency we need to consider the marking procedures used. To increase the reliability, and ensure greater validity too, we need to have

(1) Marking consistency – if the same series of things were measured on different occasions they should have the same measures. This brings up the ideas of marking schemes and inter-marker checks when more than one marker is used.

(2) Mark-relevance – the only performances to be measured must be relevant performances. For example if we were marking the scientific content of a sentence or paragraph it will lower the validity if marks are deducted for poor spelling or illegible writing.

If the emphasis is on knowledge of facts and principles (as in some sciences) one examination would probably be of fairly high reliability. But if essays are used as a means of assessment consistency becomes more difficult to achieve. Essay marks for the same group of students on different occasions are commonly correlated at about 0.5 (1.0 indicates perfect correspondence). This would mean that about one-third of the students who scored above 50% on one occasion would score below 50% on the second.

To improve reliability of essay marking the following points might be noted:

(1) test the topics covered (use a blueprint)

(2) take *several* estimates. One recent research (Elley, 1975) showed that the reliability of assessment decreases quickly if less than 3 essays and 2 markers are used. If one marker is used then 4 essays on different occasions should be used!

(3) the greater the choice of questions for students to answer the less the reliability of the test.

(4) it is probable that the more *structured question* is more reliable than essay questions.

(5) mark one question at a time and then reshuffle all the papers before the next question is marked.

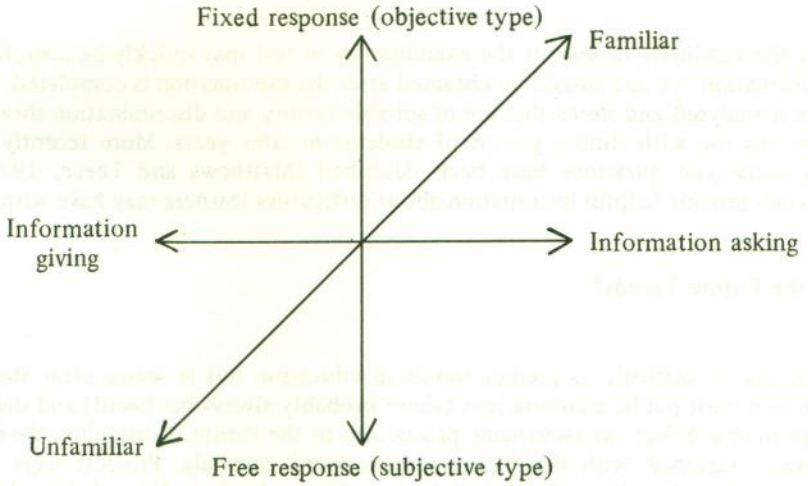
(6) the reliability seems to be increased if at least a 15 point scale (rather than 5 point) is used – provided it is clear to the marker what the points are for.

(7) impression marking, particularly if two markers are used, is much more rapid and does not seem to lower reliability.

Types of Question

If there has been a trend towards making assessment procedures more valid through some of the methods outlined above, there has been a corresponding trend towards changing the types of question asked in tests and examinations. A good illustration of this is in a comparison of two final examinations in Chemistry in Indonesia described by Ratna Wilis Dahar (1977).

One way of classifying the multitude of types of question is by the use of a three dimensional diagram as shown below.



The dimensions are concerned with the degree to which the questions contain material (whether the material is familiar or unfamiliar to the students) and whether the answer is of fixed response type (objective, or independent of the person marking) or free response (students are asked to

respond with a minimum of constraint, these are often subjective or marker-dependent answers). (Schools Council Examinations Bulletin, 1973). Between the two extremes of all these three dimensions a great deal of variety is possible. The trend for science examinations, particularly at upper secondary level seems to be towards the section of information-giving, unfamiliar and mid-way between subjective and objective type answers. This calls for skills such as investigation, problem-solving and using information in new situations.

An example from Nuffield A-level chemistry will illustrate the use of these different types of questions.

Objective test (50 items)	30%
Structured questions (50 items)	30%
Free response	25%
Practical	15%

There is a strong body of opinion that if you spell out objectives clearly the examination format should be obvious. Most objectives are probably too vague and general to be of much help in setting an examination.

If the aim is to teach your history class about political developments in Europe between World War I and World War II, you have no basis for choosing between essays, short-answers, multiple-choice ..... If you state your objectives more precisely e.g. at the end of the course students should be able to:

(1) list the immediate causes of World War II

(2) describe Hitler's rise to power

(3) explain why relations between the Great Powers deteriorated, and so on.

then it should be easier to set examination questions, and the students' learning should be more directed.

Analysis of Results

For the candidate or learner the examination or test may quickly be over; for the examiner useful information can and should be obtained after the examination is completed. Objective items can be item-analysed and items that are of suitable facility and discrimination should be kept on a card-index for use with similar groups of students in later years. More recently techniques for analysing essay-type questions have been described (Matthews and Leece, 1975). Analysis of questions can provide helpful information about difficulties learners may have with topics.

What of the Future Trends?

It is always difficult to predict trends in education but it seems clear that there is some dissatisfaction with public examinations (there probably always has been!) and that this may have a more profound effect on assessment procedures in the future. Curriculum development in the 1960's was concerned with developing teacher-proof curricula. Projects were developed that presented "packaged-deals" with student texts and worksheets and detailed teacher guides. There seems to be a growing concern that while this may be a useful procedure in the early stages of implementation of new curricula it may not be the best way of catering for the range of individual abilities and interests among teachers and students. Teachers may become more concerned with developing their own curriculum materials based on broad outlines provided by Ministries of Education.[1] If this is to happen teachers need to be trained in techniques of curriculum development, and along with these techniques procedures for assessment. Assessment may no longer mean

public examinations at the end of a course of study. Assessment may become more concerned with checking on the progress of students in a variety of objectives (the knowledge, skills and attitudes mentioned previously). If teachers develop their own (or their school) curricula based on the needs interests and abilities of the students then assessment procedures will also need to be developed to match the experiences offered to the students. Teachers will need to ensure that these procedures have high curriculum validity and will need to show concern for the points raised in this brief paper.

Note

[1]Some might argue a case for "curriculum-proof" teachers.

References

Bloom, B.Sc. (Ed.) *Taxonomy of Educational Objectives Handbook I. The Cognitive Domain.* London; Longmans, 1956.

Elley, W.B. et. al. "The role of grammar in a secondary school English curriculum". *New Zealand Journal of Educational Studies,* Vol. 10, 1975, 26-42.

Gronlund, N.E. *Measurement and Evaluation in Teaching.* 2nd. ed. New York: Macmillan Co., 1971.

Mathews, J.C. and Leece, J.R. "Nuffield Advanced Chemistry: The free response questions and assessment of practical work". *School Science Review,* Vol. 58, 1975, 362-367.

Ratna Wilis Dahar. "Evaluaton in science education at the first and second levels in Indonesia". *Bulletin of the UNESCO Regional Office for Education in Asia,* Vol. 18, 1977, 243-255.

Schools Council Examinations. *Assessment of Attainment in Sixth-Form Science.* Bulletin 27. London: Evans/Methuen Educational, 1973.

Theobald, J. *Classroom Testing: Principles and Practice.* Hawthorn, Victoria: Longmans-Australia, 1974.