



Colloquium B

Criterion-referenced Testing as High-Stakes Assessment

Panellists

Dr Thom Hudson

University of Hawaii

Ms Roxane Vigneault

International Baccalaureate Organization, Cardiff, Wales

Puan Khatija Mohd Tahir

Malaysian Examinations Council

Moderator

Dr Lee Boon Hua

Abstract

There has been a gradual and subtle shift in thinking amongst educationists as far as criterion-referenced testing is concerned, and efforts to have criterion-referenced measurement complement norm-referenced measurement are evident in the education system. However, the change has been slow, especially in the context of exams involving high stakes. This colloquium aims to provide a forum for the discussion of, and make a case for, greater use of CRT in high-stakes assessment contexts.

Summary of Panellists' Presentation

Panellist 1: Dr Thom Hudson

Role and challenges of Criterion referenced testing (CRT)

Generally, the tests constructed by testing company are norm-referenced tests (NRT) which appeals to a larger group of stakeholders

High-stake tests include CRT in 1) the way the test is constructed and 2) content in the test

CRT identifies comparison between instructed and uninstructed group. The items discriminate people who score well and those who do poorly. It measures potential achievement. It covers a well-defined domain or skill. In CRT, there's communication with all the stakeholders such as teachers, test-developers and parents.

Problems in developing CRT:

- Type of statistical analysis for CRT is not well-known as compared to NRT
- Some of the basic tools are less clear.
- Issue of cut score or standard score that is indicated as mastery or non-mastery

Panellist 2. Ms. Roxanne Vigneault

The Diploma Programme in the International Baccalaureate Organisation (IBO) is a demanding two-year pre-university course leading to examinations for highly motivated students for ages 16-19.

What makes the Diploma Programme different?

This programme combines breadth with depth and it emphasises critical, compassionate thinking. It also promotes global vision. Below are the details of the programme.

Groups 1 and 2 courses

- Language A1: Literature course for native or near-native speakers
- Language A2: Language and literature course for highly competent speakers of the target language

Language A2 Syllabus Outline

- At higher level students study 4 options
- At standard level students study 3 options
- At least one must be literary, and at least one cultural

- Cultural Options
 - Language and culture
 - Media and culture
 - Future issues
 - Global issues
 - Social issues
- Literary Options
 - Each option consists of the study of 3 works

Language A2 Assessment Outline

- Paper 1 – Comparative Commentary 25%
- Paper 2 – Essay 25%
- Written Tasks 20%
- Internal Assessment – Oral Component 30%

Language B Assessment Outline

- Paper 1 – Text Handling 40%
- Paper 2 – Written Production 30%
- Internal Assessment – Oral Component 30%

Why Criterion-Referencing?

- Population size
- Parity across subjects
- Transparency

Formulating the criteria

- Within the context of curriculum review
- Working party
- Objective – syllabus- assessment – criteria
- The process ensures the validity of the assessment
- Beneficial backwash

Language A2 Objectives

Standard level

At the end of the A2 course standard level candidates are expected to:

- communicate clearly and effectively in a wide range of situations
- understand and use accurately the oral and written forms of the language in a range of styles and situations
- understand and use a broad range of vocabulary and idiom
- select a register and style that are generally appropriate to the situation
- express ideas with clarity and fluency
- structure arguments in a focused and coherent ways, and support them with relevant examples

- understand and make use of moderately complex written and spoken texts
- engage in critical examination of a wide range of texts in different forms, styles and registers
- appreciate some subtleties of technique and style employed by writers and speakers of the language
- show sensitivity to the culture(s) related to the language studied

What do we want to assess in each component?

Language A2 – Paper 1

- Criterion A : Understanding and comparison of texts (10 marks)
- Criterion B: Presentation (10 marks)
- Criterion C: Language (10 marks)

Common criteria for different components

Cross-referencing

Reliability of assessment

- Standardisation meetings
- Moderation process

Using assessment criteria

- Levels of achievement
- Each criterion worth 5 or 10 marks
- Start with level 0
- High and low levels of achievement
- Scores across criteria

Training

- Examnet
- Examiner instructions
- Examiner orientation meetings
- Examiner feedback forms
- Subject reports

Marking and Moderation

- Assistant examiners
- Team leaders
- Principal examiners

Panellist 3: Pn. Khatija Mohd. Tahir

The MUET is intended to be criterion-reference. Prior to the MUET, the measure of student ability for English that was used was the Malaysian O' level or School

Certificate Examination (SCE) taken by Malaysian students at Form 5. This was norm-referenced and more geared towards achievement testing. The purposes and objectives of the SCE was also different from those of the MUET, which reports the students' performance according to behavioural objectives so that the end user could immediately identify students according to their specific language ability by the MUET Band Description. The aim of the test was to gauge the ability of students in English at pre-university level and to find out whether they have acquired the language skills required of them to pursue their courses at tertiary level. In line with the objective, the test was geared towards English for Academic Purposes.

What do we mean by criterion-reference?

In the context of language learning, the word criterion-reference carries a number of different meanings. According to Skehan (1989) it referred to tests or assessments, which are based on a behavioural domain. For example, in an oral interview, a student may be given a score on a rating scale, which is arrived from performance composites that form the criteria upon which the students should be assessed. These criteria may then be described more fully in a band or rating scale. Skehan notes that such descriptions may represent a general behaviour that relates the performance to some external criteria. A criterion-reference test then enables the end user (in this case the universities) to interpret the test score with reference to an ability that the student has or has not acquired. The task types are designed to be representative of specific levels of ability or domains. These test items are then designed according to how well they represent the ability and sample levels.

As in most criterion-based tests, the MUET has an important gate-keeping function. The inferences to be made from the test scores would have to be justified clearly, especially when it is also a high stakes test. All universities in Malaysia require students entering university for their first degree to have a



MUET score upon entry. Adding to its importance, some faculties, for example the faculty of Engineering, the Law faculty and the Medical faculty in the University of Malaya have targeted a minimum Band 4 or 5 upon entry. On the other hand, there are a few other universities, for example the University Technology MARA, which requires students to have the MUET as an exit requirement. Without a minimum Band 3 on the MUET, students would not be allowed to graduate.

Some constraints and issues

The purpose that the MUET serves requires that it provides a valid and reliable measure of a student's proficiency, both for general and for academic purposes. Because of the large number of students involved (about 100,000 every year), the test has to be readily and rapidly administered twice a year. As in most large-scale tests it is costly to administer the test efficiently, as well as to train item builders and examiners.

Another issue related to the MUET concerns the developing of the items or the questions used in the test. Experienced lecturers from various universities in Malaysia set the items. The lecturers normally do this at their own leisure within a stipulated time. They then convene to discuss the items set after which the selected items are typed and then set in for banking. Since MUET is a test of Academic English it is expected that the items set relate at least generally, to the type of texts students would expect to read at the university. However, because the proficiency level of students, even at pre-university and undergraduate level is relatively low, many of them find it difficult to understand the texts given to them. Getting appropriate texts that meet the criteria is not an easy task and selection of appropriate texts is an important issue that needs careful consideration. Because of this, sometimes compromises and pragmatic

decisions have to be made and these decisions may not be based on hardcore data but rather on intuition and experience of test developers.

Getting good and experienced raters who are familiar with rating scales is another challenge that the MEC faces. This is a problem because in many cases the raters have had little or no experience in marking criterion-based tests and will sometimes fall back on their own classroom experience. The marking of the writing component is done by school teachers and university lecturers. A two-day coordination session is conducted by the team leaders to standardise the marking of the scripts. However in spite of this, the raters having had greater experience with marking according to their own students' level may experience difficulties examining mastery or performance based marking. Although we have not carried out an empirical study on inter-rater reliability, based on the Chief Examiner's comments and report, we can come to a fair conclusion that there is inter-rater reliability. However, there have been cases when examiners have been known to be too strict or too lenient in the awarding of marks. Attempts are usually made to reduce the variability in the marking by coordinating and moderating the marking done by assistant examiners. This is however a monumental task and takes time, training and effort.

Another constraint relates to the earlier one. This time, however, it is the professional of the examiners that comes into the question. In the present environment, English is taught more as a foreign language than a second language in Malaysia. The language of instruction in schools is the national language, which is Bahasa Malaysia. Not all the English language teachers in Malaysia are highly proficiency in the target language. Some, probably because of their own attitude, are sadly lacking in proficiency and even though the Council tries to get only the best examiners as its raters, there have been cases where the examiner's proficiency is questionable. These examiners are however few and far between, and raters are carefully screened. However it is still a problem we have to consider when selecting examiners.

Conclusion

From this review of the MUET it can be said that what Skehan commented is true: *'criterion-referencing is an attractive ideal, but extremely difficult to achieve in practice'*. The MUET is a very 'young' examination when compared to other examinations in the same category for example the TOEFL or the IELTS. We recognise the problems in preparing a criterion-referenced test and in Malaysia especially, the issues of urban – rural ability among students is great. This dichotomy expands beyond language – going into cognitive ability and experience. Often this can be resolved by making expedient compromises and using our own testing knowledge, which sometimes can only be said to be based not on what has been researched but based on what we think is best.

Issues/concerns raised by participants and responses from presenter/panellists:

No.	Issues/Concerns raised by participants	Presenter's/Panellists' responses
1.	Ms Roxanne Vigneault (IBO) Is the CRT carried out in Delaware successful?	Dr. Thom: Not successful yet. NRT may not be easy but we can base decision on psychometric whereas in CRT we are not supposed to do that.
2.	Pn.Khatija Mohd. Tahir(MEC) Can language testing be CRT?	Dr. Thom: More difficult because it deals with vocabulary sets and discrete points. Have to determine relative difficulty.
3.	Marcia Fisk Ong (ELTC) If you set pass score at 40%, what's the pass score in IBO? Pn. Khatija How many students would you test?	Ms Roxanne: In language A2, for grades 1-7, pass/fail is at grade 3 to 4. If the exam is more difficult, the pass mark would be 12-13. 100,000



4.	Marcia Fisk Ong (ELTC) Is it possible for CRT?	Dr. Thom: It is possible to develop CRT for large audiences. Difficult to persevere to develop and refine the criteria until you get the sample when you have stakeholders.
----	---	---