MODULE 3

DEVELOPING VALID, RELIABLE AND FAIR STUDENT ASSESSMENT

A. BEST PRACTICES IN STUDENT ASSESSMENT

Quality assessment of student learning is one that is valid, reliable and fair. Validity is the extent to which a test, examination or any other assessment task measures the intended learning outcomes, while reliability is the degree to which the scores that students obtain from the test are free from errors. Fairness is determined by the extent to which the assessment method meets students' rights, responsibilities and expectations, which should be communicated to students and agreed upon by them prior to the assessment exercise.

Given the above definitions, the following criteria determine whether or not quality assessment has been practiced: **(i)** purpose of a particular test, **(ii)** procedures for developing a content-valid test, **(iii)** procedures to establish reliable test scores, **(iv)** procedures to communicate assessment requirements to students, and finally **(v)** evidence to support the test's quality.

i. Purpose of Test

This is a very crucial criterion as the purpose of assessment determines the learning outcomes, contents, format, and procedures for the validation and administration of the test. In general, the purpose of classroom assessment is to evaluate the achievement of learning outcomes as stated in the course outline. Any assessment task given to students should be able to address ***at least one*** of the targeted learning outcomes specified in the outline.

ANNIVERSARY

Assessment can be used to provide either summative or formative evidence of students' achievement in a given course. ***Formative assessment*** (such as lab work) contributes to the assessment ***for*** learning which provides opportunities to motivate and inspire students to learn and achieve more. On the other hand, ***summative assessment*** (such as a final examination) offers evidence of student achievement crucial for institutional accountability and public consumption. Some of the specific purposes of a final examination, which is the focus of this section, are to:

a. certify students' mastery of major concepts covered in a course,

b. certify students' acquisition of important skills imparted in a course,

c. check whether the goals of a course have been reached, and how far they have been reached.

d. evaluate the effectiveness of instruction.

Since comprehensive tests given as final examinations are summative in nature, the results obtained from them cannot be used by students to improve their mastery of the course contents, hence their achievement of specified learning outcomes. This type of assessment does not provide room for feedback that is crucial for students to analyze and correct errors in order to improve their competency. In addition, the common paper-and-pencil examinations hardly allow assessment to reach beyond the cognitive domain of learning. The use of performance-based assessment tools (or alternative assessment) is necessary if educators really wish to measure the achievement of complex learning behaviors. In a nutshell, final examination is essential but not sufficient to fully assess student achievement.

ii. Procedures for Developing a Content-Valid Test

It is the responsibility of the instructor (test writer) to establish the content-related validity of his/her examination paper. Content validity refers to the judgment of how well the contents of a final examination paper match the course objectives, specified learning outcomes, and the concepts and skills covered in the course. A content-valid examination paper should sufficiently sample relevant items measuring the course contents and learning outcomes. A test that deviates from the course learning outcomes or merely samples a non-representative set of items is not useful to measure student achievement because it has a low degree of content-related validity. Therefore it is imperative for a test writer to apply standard procedures to establish a test's content validity. These procedures involve:

1. Listing the content and skills to be covered (also called curriculum analysis).
2. Identifying the learning outcomes to be tested.

3. Deciding the test's format and length.

4. Constructing the test blueprint (or Table of Test Specifications).

5. Writing the examination items, be they selected-response (eg. multiple-choice or true-false) or constructed-response (eg. essay) items.

6. Developing a scoring plan for the essay items as the items are being developed.

7. Vetting the examination paper (use of expert judgment).

The first three steps require the test writer to make concrete decisions about the examination paper, taking the course requirements into consideration. First, most courses require students to achieve higher-order cognitive competencies involving application, analysis, evaluation and creating, as well as the ability to solve novel problems. It is important for the test writer to determine at the onset the proportion of items and marks to be given to each category of learning outcomes. Second, most 3-credit hour courses are assigned a 3-hour sit-in examination, which usually accounts for 40% to 60% of the final grade. Given these constraints (and the knowledge about item format), the test writer should decide the learning outcomes and contents to be tested, the test format and the number of items to be assembled.

Next, the test writer should ***construct a test blueprint***, which is a plan for the examination paper. The decisions derived from the first three steps should enable the test writer to tabulate the distribution of test items and scores according to the relative importance of the course objectives and contents. This plan is called ***Table of Test Specifications*** (TTS). The following section shows how the table is constructed. Once the TTS has been constructed, it is a good practice to seek peer-judgment on the adequacy and thoroughness of the blueprint. In this sense, peer-judgment provides preliminary evidence for the content-related validity of the examination paper.

The sit-in paper-and-pencil examinations at institutions of higher education widely use multiple-choice items (MCQ), brief-response essay items (BRE) and extended-response essay items (ERE). Each

has its own advantages and limitations. The use of essay items, however, adds another requirement to the process of creating a content-valid test; the test writer is required to construct a scoring key (marking scheme) most desirably while the test items are being developed. The details involved in this process will be covered in **Module 6: Developing Essay Items.**

The final step serves to establish the most important support for the content-related validity of the examination. The step involves selecting a content expert, a person who is proficient with the subject matter, and seeking his/her opinion about the good-of-fit between the examination questions and the contents tabulated in the Table of Test Specifications (in addition to developing a scoring key in the case of essay-type items). This process provides the evidence that expert judgment has been sought to verify that the test adequately measures what it purports to measure. It is a good practice to document and record the details of the vetting process.

iii. Procedures to Establish Reliable Test Scores

Reliability is a precondition to validity. An examination is valid so long as the scores obtained by students from the test are reliable. As noted earlier, reliability is the extent to which the scores are free from errors. An examination paper that has not been systematically developed and validated may likely contain errors that undermine its reliability. Inconsistent scoring of test items is another source of error; in fact, inconsistent scoring of essay-type items is the single most significant source of measurement error. The malpractices listed in **Module 2: Issues in Assessment of Student Learning (Malpractices 10-18)**, which include the use of grammatically incorrect items and poorly formatted examination paper, reduce the reliability of test scores.

To establish greater reliability of test scores, the following steps should serve as the standards of practice:

a. Using peer review to evaluate and edit the content as well as the language and grammar of the examination questions;

b. Having a "***sit-in marking***" session to ensure consistent scoring among examiners when there are two or more examiners

involved in grading exam scripts using one scoring key;

c. Documenting the reliability index of test scores.

To be comfortable with the quality of examination results, it is a good practice to estimate and document the reliability index of the test scores. The index, with values ranging between 0 and 1, indicates the degree of reliability of the test scores. To put it differently, the index estimates the extent to which the scores are free from measurement errors. In most situations, the reliability index of a locally-developed examination paper is somewhere between being totally present (r = 1.0) and totally non-existent (r = 0). The literature suggests using a reliability index of 0.6 as the cut-score for reliable results. A set of test scores with a reliability index of less than 0.6 should not be used to make decisions about student learning.

iv. Procedures to Communicate Assessment Requirements to Students

In the context of a final examination, the perception of students – the test takers – endorses the fairness of the test. In general, a test is fair when the distribution of grades matches students' grade expectation. The use of transparent, consistent and impartial grading procedure also favourably affects the test fairness as perceived by students. Instructors should practice free-flow communication regarding what is to be tested, how it should be tested, and how a response is graded. On the contrary, an instructor who considers grading students' achievement as his or her exclusive right where the scoring process is conducted mysteriously, in fact, has chosen to practice unfair testing!

It should be stressed that every instructor must practice fairness in testing students' achievement by adopting the following practices:

1. Establish clear grading criteria;
2. Inform students in advance about the criteria;
3. Apply the same standards and criteria to everyone;
4. Be firm when you are right;

5. Take students' input seriously.

v. Evidence to Support the Relevance of Test Content

It is the responsibility of the test developer (i.e. course coordinator, instructor, and/or examiner) to establish and demonstrate the evidence of validity and reliability of the examination paper and results. To uphold this principle, each kulliyyah, center or department must therefore make it official to institute examination vetting mechanisms that are cost-effective, practical and flexible for the assessment process.

Test writers and course coordinators in particular are responsible for ensuring that the following documents and information for each course are appropriately compiled:

1. The ***course outlines*** approved by the Senate,
2. The peer-reviewed ***Table of Test Specifications,***
3. The vetted ***examination paper,***
4. The ***records*** of the conduct of examination
5. The vetted ***scoring key***
6. The ***"sit-in marking" procedure***
7. The ***reliability index*** of the results for each course

B. TABLE OF TEST SPECIFICATIONS

Content-related validity should be built into a test even before the test is constructed. One legitimate method to achieve this is to draw up a table of test specifications. It is a blueprint for deciding the contents of the test, the details of which are defined by the importance of the test. In its simplest form, the blueprint should list all the topics and learning outcomes to be tested, and the sampling distribution of items representing the relative importance of individual topics and content. One example is given in Table 3.1 - a test blueprint developed for a 40-item MCQ test for a course on educational testing and assessment.

Generally there are seven steps to be followed when constructing a test blueprint:

Step 1 : List the topics, concepts and skills to be tested

Step 2 : Identify the learning outcomes that correspond to the topics, concepts and skills

Step 3 : Determine the relative importance of each concept

Step 4 : Assign the relative importance (weight) of each topic or concept

Step 5 : Distribute the items according to weights

Step 6 : Distribute the items (for each topic/concept) according to learning outcomes

Step 7 : Seek peer judgment or opinion

The blueprint readily permits the assessment and validation of the examination paper by a panel of expert judges, comprising peers and administrators who are well-acquainted with the content of the course. Thus, the table of test specifications serves as the basis for making an objective comparison between the content of the test and the curriculum. In summary, the table of test specifications is a built-in checking mechanism that gauges how well the test actually samples the learning outcomes and content knowledge covered in the course.

ANNIVERSARY

Table 3.1

A Sample Table of Specifications for a 40-Item Multiple-Choice Midterm Test for an Educational Testing and Assessment Course

Topics	Learning Outcomes						Weight		No. of Items	
	R	U	Ap	An	E	C	Br*	Total	Br*	Total
1. Introduction to Statistics										
• Define measurement & evaluation	√						1		2	
• Differentiate between testing & assessment				√			1		2	
• State the purpose of testing & assessment	√						1		2	
• Identify types of classroom assessment	√						1		2	
• Interpret issues in classroom assessment		√					1	5	2	10
2. Validity										
• Define content validity	√						1		2	
• Identify steps in developing a valid test	√						1		2	
• Explain the relationship between Table of Test Specifications with test validity		√					1.5		3	
• Decide whether a given Table of Test Specifications is adequate					√		1		2	
• Identify practices that reduce test validity		√					2.5	7	5	14
3. Reliability										
• Define reliability	√						1		2	
• Identify sources of measurement error		√					3		6	
• Calculate the item discrimination index			√				1		2	
• Establish the relationship between reliability and validity				√			1		2	
• Relate given assessment practices with test reliability				√			2	8	4	16
Total	12	16	2	8	2		**Total**	20	**Total**	40

Note:

R = Remember, U = Understand, Ap = Apply, An = Analyze, E = Evaluate, C = Create (based on Bloom's Revised Taxonomy) Br = Breakdown