The Federation for Self-financing Tertiary Education (FSTE)

Project on Teacher Competency Framework

Brief Notes and Suggested Reading for

Module 3A: What is educational assessment and the most common assessment tool: Written Test

and

Module 3B: Constructing assessment task in practice

This brief note intends to give participants the blueprints of the contents covered in the modules. Suggested readings are listed to facilitate participants in accessing the sources and to obtain more detail information.

These two modules intends to provide answers to the following four key questions in the context of post secondary education:

1. What is assessment (in the context of education)?
2. What are the functions of assessment in education?
3. What are the general available assessment tools in education?
4. How can the quality of assessment be assured?

These two modules provide the big picture of assessment in the context of education. Theory and skills in the construction of written test will be discussed as a taster to the planning, design, implementation, interpretation and application of assessment. Extended modules and workshops will be provided as appropriate.

I. What is assessment? Assessment for learning or assessment of learning?

1. It is essential to clarify the differences between assessment, test and measurement:

- ➢ Assessment is the collection of data for educational decision making. It includes a full range of procedures (or tools) to gain information about student learning (observations, rating of performances or projects, skill tests, paper and pencil tests etc) and the formation of value judgments concerning learning process.

- Test is a particular type of assessment that typically consists of a set of questions administered during a fixed period of time under reasonable comparable conditions for all students.

- Measurement is the assigning of numbers, or grades, to the results of a test or other type of assessment according to a specific rule (e.g. counting correct answers or awarding points or grades for particular aspects of an essay).

2. The following guidelines were developed by a group of assessment practitioners under the auspices of the American Association for Higher Education (AAHE) Assessment Forum as "Principles Of Good Practice For Assessing Student Learning":

   a. The assessment of student learning begins with educational values.

   b. Assessment is most effective when it reflects an understanding of learning as multidimensional, integrated, and performance over time.

   c. Assessment works best when the programs it seeks to improve have clear, explicitly stated purposes.

   d. Assessment requires attention to outcomes but also and equally to the experiences that lead to those outcomes.

   e. Assessment works best when it is ongoing, not episodic.

   f. Assessment fosters wider improvement when representatives from across the educational community are involved.

   g. Assessment makes a difference when it begins with issues of use and illuminates questions that people really care about.

   h. Assessment most likely leads to improvement when it is part of a larger set of conditions that promote change.

   i. Through assessment, educators meet responsibilities to students and to the public.

http://www.mcli.dist.maricopa.edu/LF/Spr99/assidere1.html

3. The Holmes Group[i] also has the following expectation: "Professional teachers" are described as "skilled diagnosticians of children's learning needs." They are expected to interpret the understandings students bring to and develop during lessons … and to identify students' misconceptions and question their surface responses that mask true learning.

http://www.eric.ed.gov/PDFS/ED270454.pdf
http://www.personal.psu.edu/rsw136/blogs/rebecca_west_burns/2010/06/the-holmes-group-1986-tomorrows-teachers-a-report-of-the-holmes-group-michigan-the-holmes-group.html

4. National Board of Professional Teacher Standards (NBPTS)[ii] in US indicates that: Teachers are responsible for managing and monitoring student learning. The teachers we intend to recognize know how to create, enrich, maintain, and later instructional settings to capture and sustain student interests. They use many methods to measure student growth and understanding.

http://www.nbpts.org/about_us

II. What are the functions of assessment in education? Establish the needs and clarify the purpose.

5. "Assessment is essential not only to guide the development of individual students but also to monitor and continuously improve the quality of programs, inform prospective students and their parents, and provide evidence of accountability to those who pay our way."

-- Redesigning Higher Education: Producing Dramatic Gains in Student Learning by Lion F. Gardiner; ASHE-ERIC Higher Education Report Volume 23, No. 7, p. 109

6. Universities, for example Loyola University Chicago (LUC), view assessment as a natural concern of the scholar as teacher. They want to know what their students have learned, the means by which they learned, and the effectiveness of the learning process. As teacher-scholars, we must ask: "What evidence might we

gather that our students, taken as a group, are in fact acceptably achieving the learning outcomes that we, the faculty of a given program, intend?"   The pursuit of this question is how we learn what our students know and what they are able to do as the result of their course of study.

Aside from the overarching necessity for assessment as a means of meeting both institutional and programmatic accreditation requirements, there are two main reasons that assessment is important in higher education:

a. Accountability

- To students and their families
- To government
- To society

b. Instruction and Program Improvement

Evidence-based program improvement soundly answers questions such as:

- How well are student learning outcomes being met?
- Which outcomes need to be revised?
- Which programs/services/courses need to be revised to better fit the outcomes?
- Which programs/services/courses are no longer congruent with the mission and goals of the department?

c. Certification

To qualify individuals with knowledge, skills and attitudes to undertake :

- Further learning
- Work
- Particular profession

http://assessment.uconn.edu/why/index.html
http://sites.google.com/site/luctwtguide/why-assessment-is-important

III. How to prepare a classroom assessment? Whether the results of assessment reflect reliably the intended learning outcomes?

7. The process of preparing, administering and using assessment to improved learning and instruction can be broadly summarized in the following 8 steps:

8. Interpreting and using the results
7. Appraising the assessment
6. Administering the assessment
5. Assembling the assessment
4. Preparing relevant assessment tasks
3. Selecting appropriate assessment tasks
2. Developing specifications
1. Determining the purpose of assessment

Item types of classroom assessment

8. Item types of assessment can be broadly classified into two categories:

A. Objective test items

a. Objective test items have a common feature: they present students with a highly structured task that limits the type of response they can make. To obtain correct answer, students must demonstrate the specific knowledge, understanding, or skill called for in the item; they are not free to redefine the problem or to organize and present the answer in their own words.

b. The positive side of this item type is that it contributes to objective scoring that is quick, easy, and accurate.

c. The negative side is that it is inappropriate for measuring the ability to select, organize, and integrate ideas.

d. Examples of objective test items

Supply types:	Short answers
	Fill in blank

Selection types: True-false or alternative responses
Matching
Multiple-choice

B. Performance assessment tasks

a. Performance assessment tasks allow students to decide which facts they think are most pertinent, to select their own method of organization, and to write as much as seems necessary for a comprehensive answer.

b. The positive side is that such tasks tend to reveal the ability to evaluate ideas, to relate them coherently, and to express them succinctly. They also reflect individual differences in attitudes, values and creativity.

c. The negative side is that (1) they are inefficient in measuring knowledge of factual material; and (2) scoring is difficult and apt to be less reliable.

d. Examples of performance assessment tasks

- Extended-response essay questions
- Restricted-response essay questions
- Oral presentations
- Project assessment
- Use of equipment or scientific instruments
- Playing a musical instrument

Alignment of Assessment with Learning Objective

9. Different learning objectives (or learning outcomes) will need different assessment tools to identify the evidences of achieving. As such, it is essential to align assessment with learning objective (or learning outcome)

- **Learning objectives:** What do I want students to know how to do when they leave this course?

✧ **Assessments:** What kinds of tasks will reveal whether students have achieved the learning objectives I have identified?

Type of learning objective	**Examples of appropriate assessments**
Recall **Recognize** **Identify**	Objective test items such as fill-in-the-blank, matching, labeling, or multiple-choice questions that require students to: • recall or recognize terms, facts, and concepts
Interpret **Exemplify** **Classify** **Summarize** **Infer** **Compare** **Explain**	Activities such as papers, exams, problem sets, class discussions, or concept maps that require students to: • summarize readings, films, or speeches • compare and contrast two or more theories, events, or processes • classify or categorize cases, elements, or events using established criteria • paraphrase documents or speeches • find or identify examples or illustrations of a concept or principle
Apply **Execute** **Implement**	Activities such as problem sets, performances, labs, prototyping, or simulations that require students to: • use procedures to solve or complete familiar or unfamiliar tasks • determine which procedure(s) are most appropriate for a given task
Analyze **Differentiate** **Organize** **Attribute**	Activities such as case studies, critiques, labs, papers, projects, debates, or concept maps that require students to: • discriminate or select relevant and irrelevant parts • determine how elements function together • determine bias, values, or underlying intent in presented material
Evaluate **Check** **Critique** **Assess**	Activities such as journals, diaries, critiques, problem sets, product reviews, or studies that require students to: • test, monitor, judge, or critique readings, performances, or

	products against established criteria or standards
Create **Generate** **Plan** **Produce** **Design**	Activities such as research projects, musical compositions, performances, essays, business plans, website designs, or set designs that require students to: • make, build, design or generate something new

Table of Specifications

10, The purpose of a Table of Specifications is to identify the achievement domains being measured and to ensure that a fair and representative sample of questions appear on the test. Teachers cannot measure every topic or objective and cannot ask every question they might wish to ask. A Table of Specifications allows the teacher to construct a test which focuses on the key areas and weights those different areas based on their importance. A Table of Specifications provides the teacher with evidence that a test has content validity, that it covers what should be covered.

A Table of Specifications helps to ensure that there is a match between what is taught and what is tested.

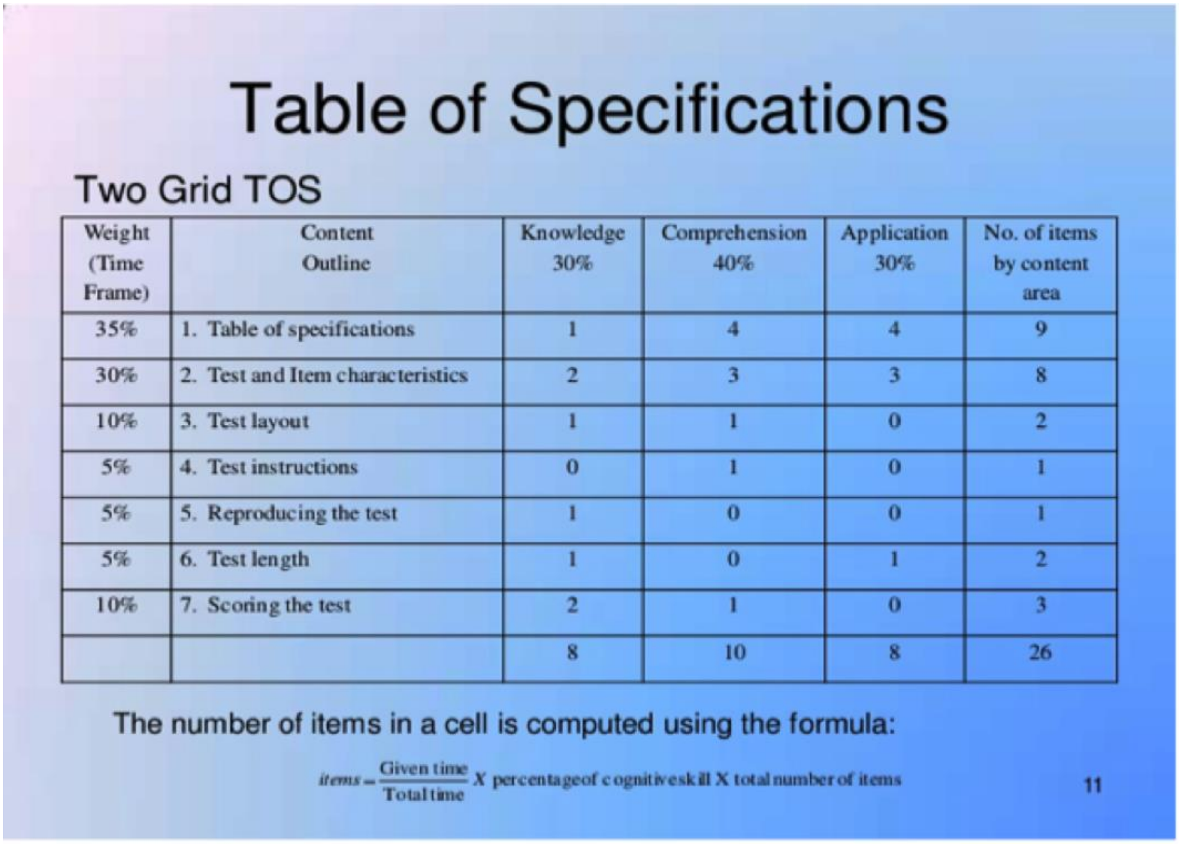

Table of Specifications

Two Grid TOS

Weight (Time Frame)	Content Outline	Knowledge 30%	Comprehension 40%	Application 30%	No. of items by content area
35%	1. Table of specifications	1	4	4	9
30%	2. Test and Item characteristics	2	3	3	8
10%	3. Test layout	1	1	0	2
5%	4. Test instructions	0	1	0	1
5%	5. Reproducing the test	1	0	0	1
5%	6. Test length	1	0	1	2
10%	7. Scoring the test	2	1	0	3
		8	10	8	26

The number of items in a cell is computed using the formula:

$$items = \frac{\text{Given time}}{\text{Total time}} X \text{ percentageof cognitiveskill X totalnumberof items}$$

11

IV. How can the Quality of Assessment be assured?

Quality indicator of a test: Reliability and Validity

11,. The two factors governing the quality of assessment are: (a) Reliability and (b) Validity. Both reliability and validity refer to the results obtained with an assessment instrument and NOT to the instrument itself.

a. Reliability refers to the consistency of measurement; that is, how consistent test scores or other assessment results are from one measurement to another.

   - ➢ Reliability is primarily statistical

   - ➢ Examples of estimating reliability:

     i. Test-retest method (measure of stability)
     ii. Split-half method (measure of internal consistency)
     iii. Kuder-Richardson method and coefficient Alpha (measure of internal consistency)
     iv. Inter-rater method (measure of consistency of rating)

b. Validity refers to whether the assessment measures what intended to measure, i.e. the adequacy and appropriateness of the interpretations made from assessments, with regard to a particular use.

   - ➢ Major considerations:

     i. Content validity – How well the sample of assessment tasks represents the domain of tasks to be measured
     ii. Construct validity – How well performance on the assessment can be interpreted as a meaningful measure of some characteristic or quality
     iii. Test-Criterion Relationship – How well performance on the assessment predicts future performances or estimates current performance on some valued measures other than the test itself.

Robert L. Linn, Norman E. Gronlund “Measurement and Assessment in Teaching”7th ed Prentice-Hall Inc. 1995.

http://www.eric.ed.gov/PDFS/ED270454.pdf
http://en.wikipedia.org/wiki/Educational_assessment
http://en.wikipedia.org/wiki/Classical_test_theory

Quality indicators of an item: Item difficulty index and item discrimination index

12. The purpose of item analysis is to improve the quality of an exam by identifying items that are candidates for retention, revision, or removal.

In addition to qualitative procedures, item analysis also includes a number of quantitative procedures. Two common numerical indicators are often derived during an item analysis: item difficulty, item discrimination.

a. Item Difficulty Index (p)

The item difficulty statistic is an appropriate choice for achievement or aptitude tests when the items are scored dichotomously (i.e., correct vs. incorrect). Thus, it can be derived for true-false, multiple-choice, and matching items, and even for essay items, where the instructor can convert the range of possible point values into the categories "passing" and "failing."

The item difficulty index, symbolized p, can be computed simply by dividing the number of test takers who answered the item correctly by the total number of students who answered the item. As a proportion, p can range between 0.00, obtained when no examinees answered the item correctly, and 1.00, obtained when all examinees answered the item correctly. Notice that no test item need have only one p value. Not only may the p value vary with each class group that takes the test, an instructor may gain insight by computing the item difficulty level for a number of different subgroups within a class, such as those who did well on the exam overall and those who performed more poorly.

b. Item Discrimination Index (D)

Item discrimination analysis deals with the fact that often different test takers will answer a test item in different ways. As such, it addresses questions of considerable interest to most faculty, such as, "does the test item differentiate those who did well on the exam overall from those who did not?" or "does the test item differentiate those who know the material from those who do not?" In a more technical sense then, item discrimination analysis addresses the validity of the items on a test, that is, the extent to which the items tap the attributes they were intended to assess. As with item difficulty, item discrimination analysis involves a family of techniques. Which one to use depends on the type of testing situation and the nature of the items. I'm going to look at only one of those, the item discrimination index, symbolized D. The index parallels the difficulty index in that it can be used whenever items can be scored dichotomously, as correct or incorrect, and hence it is most appropriate for true-false, multiple-choice, and matching items, and for those essay items which the instructor can score as "pass" or "fail."

The item discrimination index is calculated in the following way:

- Divide the group of test takers into two groups, high scoring and low scoring. Ordinarily, this is done by dividing the examinees into those scoring above and those scoring below the median. (Alternatively, one could create groups made up of the top and bottom quintiles or quartiles or even deciles.)
- Compute the item difficulty levels separately for the upper (p upper) and lower (p lower) scoring groups.
- Subtract the two difficulty levels such that D = p (upper)- p (lower).
- How is the item discrimination index interpreted? Unlike the item difficulty level p , the item discrimination index can take on negative values and can range between -1.00 and 1.00. Consider the following situation: suppose that overall, half of the examinees answered a particular item correctly, and that all of the examinees who scored above the median on the exam answered the item correctly and all of the examinees who scored below the median answered incorrectly. In such a situation p(upper) = 1.00 and p (lower) = 0.00. As such, the value of the item discrimination index D is 1.00 and the item is said to be a perfect positive discriminator. Many would regard this outcome as ideal. It suggests that those who knew the material and were well-prepared passed the item while all others failed it.

http://faculty.mansfield.edu/lfeil/201/item-analysis-explained.htm
http://www.omet.pitt.edu/docs/OMET%20Test%20and%20Item%20Analysis.pdf

Annex: Classical test theory

Classical test theory assumes that each person has a *true score*, *T*, that would be obtained if there were no errors in measurement. A person's true score is defined as the expected number-correct score over an infinite number of independent administrations of the test. Unfortunately, test users never observe a person's true score, only an *observed score*, *X*. It is assumed that *observed score* = *true score* plus some *error*:

$$\underset{\text{observed score}}{X} \quad = \quad \underset{\text{true score}}{T} \quad + \quad \underset{\text{error}}{E}$$

FSTE August 2015

[i] The Holmes Group(1986), a consortium of deans and a number of chief academic officers from research institutions in each of the 50 states in US, is organized around the twin goals of the reform of teacher education and the reform of the teaching profession.

[ii] NBPTS is an independent, nonprofit, nonpartisan and nongovernmental organization. It was formed in 1987 to advance the quality of teaching and learning by developing professional standards for accomplished teaching, creating a voluntary system to certify teachers who meet those standards and integrating certified teachers into educational reform efforts.