International Association for Educational Assessment (IAEA) 34th Annual Conference, Cambridge, 7th – 12th September, 2008.

Theme: Re-Interpreting Assessment: Society, Measurement and Meaning

Sub Theme: Evaluating the Quality of Assessment

Title: Meaningful Learner Assessment: Ensuring Quality of Test Items

By

Uche Mercy Okonkwo (PhD)
School of Education,
National Open University of Nigeria,
Victoria Island, Lagos.

and

Mrs Charity Akuadi Okonkwo (PhD)
School of Education,
National Open University of Nigeria,
Victoria Island, Lagos.

Abstract

Learner assessment is very critical in any instructional programme, therefore assessment processes should be based on best practices. This paper presents the outcome of an evaluation of samples of examination papers of an Open and Distance Learning (ODL) university. The analysis covered issues as instructions, item types and formats, quality of items and domains of learning tested. Result showed that but for a few papers, most failed to satisfy the requirements for constructing quality test items. For instance, most papers did not give clear instructions while some did not have appropriate stems for multiple choice items. This could have confused the examinees as well as cost them time figuring out what was expected of them. Some papers failed to group similar items, a situation that could have affected the examinees' line of thought. A major weakness of many papers is that items focused on recall of facts, a pattern of questioning that could produce learners who regurgitate facts instead of learners who are challenged to apply facts to solving problems.The implications of these findings for improvement of learner assessment in the university and similar institutions of learning are discussed.

Introduction

Learner assessment is an integral part of every instructional programme. Learners are assessed to ascertain how much they know of a subject matter and how well they can use the knowledge. Assessment could be formal or informal and usually consists of questions to which learners respond. Questions could be in objective or essay format or both. The purpose of every question should be to elicit a given response from the learners. The design and content of each question therefore, must be clear and unambiguous as to the task the examinee is expected to perform in responding to the question. On the critical importance of question design, Mandernach (2003, 2) says:

> While learning goals and the taxonomy of learning provide an excellent structure for designing assessment items, the educational impact of any assessment rests in the content of individual questions.

Ebel (1972, 147), agrees with the above urging that

> the examinee's task be defined as completely and specifically as possible without interfering with measurement of the achievement intended.

Ensuring a meaningful measure of learner performance therefore, requires that the design and content of the test questions are of acceptable quality.

Statement of the Problem

The report presented in this paper is the outcome of a study that analysed a sample of test instruments used in examining students in an ODL university. The study was prompted by an observation made by the presenters concerning instructions in some of the question papers during the end of semester examination in 2007. The not so good quality of these instructions made the presenters to select a sample of question papers for evaluation.

The Purpose of the Study

The purpose of the study was to analyse the papers sampled to identify areas of weaknesses and use the findings to improve test development procedures in the university.

Test Development Procedure at the University

A uniform format was used by all the four schools in the University for developing examination questions. The format required that each paper had two sections. One section was made up of 20 objective items. These items could be in multiple choice, completion or true/false or a combination of these three formats. Each item was one mark, making the total for the section 20%. The second section contained essay questions made of 5 or 4 out of which examinees were to answer 3 or 2 questions respectively. The total score for the essay section was 50%. The tutor marked assignment was a take home assignment marked over 30%. The overall total score for assessment was 100%.

The Study Sample

The question papers were administered at the end of the first semester in 2007. The students that sat for the examination were mostly in 100 level with a few in 200 level of the degree programmes in the university. Question papers in three out of the four schools in the institution were used as well as those used in testing the general courses. The three schools were School of

Education, School of Arts and Social Sciences and School of Science and Technology. The fourth school, School of Business and Human Resources Management was not included because it runs programmes that the presenters are not very familiar with. Below is a table presenting the number and percentage of courses examined in each school and the sample of question papers for analysis.

Table 1: Schools, Courses Examined and Percentage of Question Papers Analysed

School	Number of Courses Examined	Number of Question Papers Sampled	Percentage of Question Papers Sampled
School of Education(SED)	15	6	40%
School of Arts and Social Sciences(SASS)	27	11	41%
School of Science and Technology(SST)	30	7	23%
General Courses(GST)	4	3	75%

Table 1 above shows that out of 15 courses examined in the School of Education, 6 (40%) were analysed. The School of Arts and Social Sciences examined 27 courses out of which 11 (41%) were analysed. Out of the 30 courses examined by the School of Science and Technology, 7 (23%) were analysed. Finally, of the four general courses examined, 3 (75%) were analysed.

Framework for Analysis of Question Papers

The analysis covered the following areas:

- Quality of instructions (general and specific);
- Quality of item types and formats;
- Organization of item types;
- Quality of items e.g. clarity of task(s) to perform;
- Quality of distractors in the case of multiple choice items;
- Number of essay questions;
- Domains of learning examined.

Analysis and Interpretation of Data

The above framework was used to do an in depth analysis of each of the question papers sampled. The framework was developed using test construction standards as obtained in the literature. (Gronlund, 1981, Ebel, 1972)

Quality of Instructions

The general and specific instructions were analysed regarding their suitability in properly guiding the students as to what they were to do. Most of the question papers analysed failed to give sufficient instructions for guidance. This was the case for the objective section of most of the papers sampled. These papers had general instructions about the examination but not specific instructions about how examinees would respond to the items in the objective section. Here are some examples for illustrations:

Poorly written instructions	Comments
Example 1. " Instruction: Answer ALL questions in Section A, and any two in Section B" Note: There was no further instruction.	Section A has a set of multiple choice items but there is no further instruction to the examinees on how to record answers for instance whether by circling the chosen option in the examination paper or by writing the letter of the option chosen in the answer booklet.
Example 2. This is another paper with a similar instruction as in the above example.	This is a case of giving completion items without the prerequisite instruction to students to complete the blanks provided.
Example 3. Unnecessary repetition of instruction – e.g. "State whether the statement is true or false"	This is a case where for every one of fifteen items in the objective section, the examinees had to read the same opening statement. This is repetitive, monotonous and could have cost examinees time.

Quality of Item Types and Formats

The study assessed the quality of item types and formats. Some papers were grossly inadequate in this regard. Here are some examples:

Poor Item Types and Formats	Comments
Example 1: Lack of item stem for multiple choice items- e.g. "Choose the correct option only one is correct amongst the given options a. Napoleon was the Emperor of France b. Napoleon was the Emperor of Italy c. Napoleon was the Emperor of Africa d. Napoleon was the Emperor of Europe"	There are three things wrong with this item. The first is the poor quality of instruction. The second is the absence of an item stem. The third problem with this item is that the distractors are not of the same form. While a. and b. are continents c. and d. are countries. This item could be restructured thus: Napoleon was the emperor of … a) Austria b) England c) France d)Italy
Example 2: This is a case of using short answer questions where objective questions were required. Here is an example: "What do you understand by ionosphere?" "Mention four human activities that encourage desertification"	There were 20 of this question type and the instruction asked students to answer all the questions and each was 1mark. The demand of each question calls for a short answer. The task required by these questions would definitely take much of the examinees' time. Each of the twenty questions in this section could have conveniently been framed in objective form.

Organization of Item Types

The study looked at how item types were organized within sections. A major finding was that some papers mixed item types instead of having similar items together. This style of grouping items could have affected the examinees' line of thought in their attempt to switch from the demand of one type of item to another. Another consequence of this arrangement could be loss of time.

Instances of Poor Item Organization	Comments
Having a set of multiple choice items followed by a set of completion items and then followed again by multiple choice items etc	This kind of item arrangement would likely affect examinees line of thought as well as cost examinees time.

Quality of Items

Each item in both the objective and essay sections was analysed for clarity as regards the task(s) examinees were expected to perform.

Instances of Poorly Written Items	Comments
Example 1: " Complete the following statement: To improve the society you must improve …. (a) the men (b) the women (c) the schools (d) the hospitals Note: the correct answer option for the above item is schools.	This is a case of poor stem construction giving room to more than one plausible answer. Any one of the distractors could complete the statement. The stem of this item could be reframed as follows to make only one option the appropriate answer: " Complete the following statement with the best option from the list provided: The … provide the skills and knowledge needed for improving the society" With the reframing of the item, the only option out of the four that can complete the above statement is "schools".
Example 2: "Attempt a detailed explanation of the concept 'education' "	This item is too general because it fails to direct the examinee as to the specific task s/ he was required to perform. Education is a loaded term that could be understood from many perspectives by different people. One wonders therefore the perspective(s) the examiner had in mind when constructing this item. An item like this could result in variability of the examinees' performance as well as in scoring by different scorers.

Example 3: "Who has the control and financing of education in France?" Note: This is an essay question marked over 20	This item is more like a one word answer question than an essay question that it was supposed to be. The tester failed to specifically convey the task that the examinees would accomplish to qualify their answer as an essay.

Quality of Distractors in the Case of Multiple Choice Items

The quality of distractors was considered from two points namely, similarity in form as well as order of arrangement. Some papers failed to observe the rules governing the above. Below are some examples**.**

Instances of Poor Quality Distractors	**Comments**
Example 1: " In drama, miming is the same as a) basic acting skill b) acting without words c) acting without costume d) acting without make-up"	In the first example the first answer option is different from the other three in form. The examinee would very likely ignore this option by process of elimination which reduces the range of options from four to three.
Example 2: "The connotation of a word refers to a) the dictionary meaning of the word b) the correct application of the word c) the additional meaning of the word d) humour"	Example 2 is also a case of mismatch of distractors. Whereas the first three options are of the same form the fourth one is not. Here the examinees could easily have ignored the last option "humour" by the process of elimination since it was shorter in form than the other options and does not contain an important clue to the correct answer namely, "word" which is in the stem as well as in the other three options.
Example 3: "The disease that is highly fatal, very contagious and most deadly among cattle is: a) Rinderpest b) Foot and mouth disease c) Bruccicosis d) Maille fever	The weakness of this item is that option "b" which is the correct answer is longer in form and is the only one that is not in scientific term. Once again the examinees would use the process of elimination to get the correct answer.

Number of Essay Questions

The number of questions in the essay section of each paper was considered as to its adequacy. The analysis revealed that every one of the question papers had in the essay section an array of questions for examinees to choose from ranging from five to do three to four to do two. This was an end of semester examination whose goal was to assess the learners' attainment of course objectives. It was an achievement test that called for comparison of examinees' performance.

With this wide range of options, an objective comparison of examinees was compromised because the basis of comparison was not the same.

Domains of Learning Examined

The observation here was that in every question paper analysed, all the items under the objective section were of lower order level of recall of facts. The questions in the essay section of most of the papers analysed were also of the level of recall. Here are some examples:

Instances of Poor Essay Questions	**Comments**
Example 1: "a. What is juvenile delinquent behaviour? b. State five remediation procedures for juvenile delinquent behaviour."	In example 1, the demand of both "a" and "b" is mere recall of facts. This item could be modified to get a question of a higher order level as follows: "a. What is juvenile delinquent behaviour? b. Name and describe an example of a juvenile delinquent behaviour c. Describe with illustrations five procedures you would use to remedy the delinquent behaviour".
Example 2: "a) Define the term cooperatives b) State the 1995 cooperative Principles c) What are the types of cooperatives d) State the characteristics of informal cooperatives e) Give five benefits of farmers as cooperators"	All of the five sub items in this question are simply on recall of facts. This question could be reframed to challenge the examinees to apply what they know about cooperatives, thus: " Suppose your organization asked you to set up a cooperative society, write a proposal describing types of cooperatives and based on the 1995 cooperative principles recommend with reasons a particular type of cooperative to your organisation"

Discussion of Findings

The analysis carried out in this study focused on areas of weaknesses of all the question papers analysed. An in depth analysis of each question was done using test construction standards as recorded in the literature. The findings revealed a sizable number of flaws in most of the papers analysed. This is not to say that all the questions papers were of poor quality. There were indeed some objective and essay questions that were of acceptable standard. These were however very few. A brief discussion of the findings is presented here.

Quality of instructions
Most papers failed to give adequate instructions. This was especially the case for the objective section of the papers where questions were not introduced with instructions on how students should record their answers. Instructions are very important because they serve as guide for examinees. Insufficient, ambiguous and superfluous instructions could negatively affect examinees' performance.

Quality of item types and formats
The item types used in the objective sections of some of the papers, namely multiple choice, fill in /completion and true/false were alright for the learning objectives being measured. There were some papers, however that used wrong item types and formats to measure learning. A flaw with most of the items was the absence of item stem. This could confuse the examinees and cost them time trying to figure out what the task in an item is. The stem of the item is very important. It should be meaningful by itself and should present a definite problem.(Gronlund, 1981).

Organization of item types
The items in most of the papers were poorly arranged. For instance, true/false and completion items were mixed with multiple choice questions. An important guideline for constructing tests is to ensure that similar items are grouped together. This arrangement enables examinees to retain the same mental set throughout a given section. The arrangement also facilitates scoring.

Quality of items e.g. clarity of task(s) to perform
The analysis revealed inadequacies in both objective and essay questions. There were instances of objective items with more than one correct answer. The quality of the objective items was also affected by insufficient instructions as earlier mentioned. The analysis also revealed problem with most essay questions. Examples are essay questions that were too broad and too general. Questions like these are open to a wide array of interpretations by both examinees and scorers. There is bound to be variability in examinees' performance and scorers' marking. Gronlund (1981, 230), signals the danger in administering poorly written essay items thus:

> Since it is impossible to determine which of the incorrect answers are due to misinterpretation and which to lack of achievement, the results are worse than worthless.

Quality of Distractors.
A major finding here is that many questions were matched with distractors that were not of the same form. Cases like this would lead students to get the correct answer choice by the process of elimination. Functional distractors must not only distract the uninformed but must be homogeneous.

Number of Essay Questions
Every one of the paper analsyed had a list of items in the essay section to choose either two or three to write on. Giving examinees options would not allow for an objective

comparison of their performance because "the basis for the comparability of their scores is weakened." (Ebel, 1972; 147)

Domains of learning examined.

A major weakness of most of the papers analysed was that both objective and essay questions were of the lower order of learning. They were all on recall of facts. Such questions are not intellectually challenging. If the trend continues we would be producing learners who regurgitate facts instead of learners who use facts to solve problems.

Summary and Conclusion

Assessment is an important component of any instructional programme. The import of it is even more so when the purpose is to ascertain learners' attainment of course objectives and when learners are compared by their performance. Every care must therefore be taken to ensure that the instruments used in assessing learners are developed based on acceptable principles and techniques. The findings revealed by this study show that most of the papers analysed were largely not developed based on acceptable standards. This calls for a critical review of item development procedures of this institution to ensure that the flaws revealed by this study are attended to in order to avoid occurrence in the future. The review is even more critical considering that examinations in an ODL institution are conducted not in one location as found in conventional institutions but in several study centres located in different geographical regions of the country. There is a need therefore, for assessment instruments used in this institution as well as in all similar ODL institutions to be as much as possible devoid of weaknesses that could lead to misinterpretations by different groups of examinees and test administrators. The revelations from this study also have implications for test development and administration at all levels of education, from primary to tertiary. Classroom tests/ examinations at all these levels are usually developed and administered by the teachers who teach the subjects. These tests are usually not subjected to any form of evaluation. What this study therefore has pointed out is the need to train classroom teachers on the importance of evaluating every test paper before administrating it.

References

Ebel, R. L. (1972). Essentials of Educational Measurement. Prentice-Hall, Inc. New Jersey

Gronlund, N. E. (1981). Measurement and Evaluation in Teaching. Macmillan Publishers, New York.

Mandernach, B. J. (2003). Using the Taxonomy of Educational Objectives to Create Effective Assessments. Retrieved November 12, 2007 from Park University Faculty Development Quick Tips.