Taxonomial Analysis of Selected Teacher-Made Multiple Choice Tests in Obafemi Awolowo University, Nigeria

Bamidele Abiodun Faleye

Department of Educational, Foundations and Counselling,
Faculty of Education, Obafemi Awolowo University, Ile-Ife
Email: bamidelefaleye@yahoo.com, bafaleye@oauife.edu.ng

Oluwole Rufus Ayeloja

Department of Educational, Foundations and Counselling,
Faculty of Education, Obafemi Awolowo University, Ile-Ife

Doi:10.5901/jesr.2014.v4n3p315

Abstract

The study determined the content and cognitive validity of the selected undergraduate objective tests used in Obafemi Awolowo University, Ile-Ife, Nigeria. The population of the study included all the undergraduate objective tests in Obafemi Awolowo University, lecturers from four selected universities and part one students. Sample consisted of tests of four courses purposefully selected. Ten lecturers as raters and 50 students participated in the study. Data were analyzed using percentage and Content Validity Ratio (CVR). The results showed that only one course has a very high content coverage and CVR of 94.60 and 0.91respectively. Majority of the items on the tests measured lower cognitive outcomes (knowledge, comprehension and application). Some items were either too simple or too difficult. The study concluded that objective tests in Obafemi Awolowo University have content validity but some items of tests constructed for use are superficial, sampling candidates' lower level of mental process.

Keywords: *Taxonomial analysis, content validity, content relevance, teacher-made tests*

1. Introduction

A current subject of public concern in Nigeria in the recent past is the issue of general poor performance of candidates in tests, especially public examinations (Alli, 2013; Adepoju and Oluchukwu, 2011 and Oweh, 2014). This situation becomes relevant point of reference because candidates sitting for public examinations are expected to have been taught and prepared adequately by their teachers through several pedagogical activities in the classroom. One of the myriads of these activities is classroom achievement tests, which are to be used in obtaining information (in form of responses or answers to questions) that will enable the teacher to place the candidate to particular achievement level (in terms of results) after results are released. This then serve as feedback for the candidates' homes and government, both of which are stakeholder in the day-to-day running of the school system.

Classroom achievement tests do not only provide basis for setting students' level of achievement in particular school subject, they also helps the teacher in monitoring the likelihood or otherwise of achieving the goals of teaching the subject at that particular level of schooling. When results of teacher-made classroom achievement tests are released, all those involved in the procedures that led to the release of that result (the students, teachers/school, government and students' homes) are able to evaluate the quality of their contributions and thereafter, remediation (when results are poor) or consolidation (when results are good) activities are planned (Faleye and Afolabi, 2007).

One of the characteristics of teacher-made tests is the flexibility it confers on the teacher in the choice of words and structure of options (in the case of multiple-choice objective items) to be used in item design. As a result, series of test item formats in terms of length of words (for the item stem), number of options, plausibility of the options and clarity of grammar and tense used abound. Not only this, a number of the items used in classroom evaluation across departments in Obafemi Awolowo University (Nigeria) (as could be the case elsewhere since most of the tests are not validated before use) have been found to be of varying degrees of psychometric qualities. The nature of the test formats used during the regular classroom evaluation is an important issue in classroom assessment. Thus, the quality of the

items used by the teacher (in terms of the extent to which it samples the domains of cognitive behavior) will have impact on the performance of the students during public examinations (Faleye, 2005).

Bloom (1956) proposed the six levels of students' cognitive domains of behavior. The six levels represent the depth of activities required for the proper assimilation of the content of any classroom instruction. The six levels according to Bloom include knowledge, comprehension, application, analysis, synthesis and evaluation. The first three levels are regarded as lower level of mental process and most assessment activities at the primary and junior secondary schools (both of which are now Basic school) are expected to terminate at this level. This is because students at the basic level of schooling have not yet mature for engagement in critical or abstract reasoning (and it will not be fair to make them perform assessment tasks requiring abstract reasoning (Crowl; Kaminsky and Podell, 1997). The last three levels of cognitive domain is regarded as the higher level of mental process.

One of the characteristics of a good Multiple Choice Question (MCQ) is the its ability to cover the six domains of students' cognitive behavior as submitted by Bloom in his Taxonomy of Educational Objectives. As posited by Popham (2002), "…if the assessment domain representing … is not satisfactorily represented by the teacher's final examination, the teacher's score-based interpretations … are apt to be in error. [p.53]" A good test is thus expected to cover the six levels of cognitive domain in reasonable proportion with the first and last levels having the fewest items and the levels at the middle of the range having the most of the items set in a test (Faleye, 2007). This feat could hardly be achieved by most teachers (not minding their levels of cognitive ability and teaching experience) without the use of Test Blueprint (also called Table of Specification). However, the use of test blueprint has not been popular among teachers (NECO, 2001). Irrespective of the busy schedule of teachers, the use of test blueprint is a *sine qua non* for preparing tests of high quality in terms of the sampling of the various levels of cognitive levels.

Thus, the foregoing provides a justification for the present study, the objectives of which are to

i. determine the content validity of the selected undergraduate objective tests used in Obafemi Awolowo University;
ii. investigate the spread of the test items across the levels of cognitive domain of Bloom's Taxonomy and;
iii. determine the plausibility of the options used in the objective tests.

2. Methodology

The study employed the survey design. The population of this study comprised all the objective tests used in Obafemi Awolowo University, Nigeria. There are about 23,000 students in the university. Seventy five percent are undergraduates across 13 faculties and over 90 departments (The Postgraduate College, 2010). There is no statistical data as to the number of objective tests used for assessment in available courses during each session. However, from the oral interview conducted by the researcher it was gathered that the factor that determined the choice of examination format is the size of the class. Study sample consisted of four objective tests. Purposive sampling technique was used to select the tests of courses that attract large students from different departments. The researcher employed this non—probability sampling technique because it is the most appropriate one for the study which knowledge of population is not known (Nworgu, 2006).

In this study, the sampling technique allows the researchers to select courses that are offered by students from different departments. Two courses, Introductory Chemistry I coded as CHE 101 and General Physics II coded as PHY 102 were selected from sciences. The courses are offered by the students in the following faculties: pharmacy, clinical sciences, basic medical, and nursing; engineering environmental design and management and sciences respectively. Two courses that were selected from faculty of social sciences are Elements of Economic Theory and Principles coded as SSC 102 and Mathematics for Social Scientists coded as SSC 105. SSC 102 is offered by all the departments in faculty of social sciences. Using this sample, the researcher believed that relevant information on assessment of larger proportion of the university students could be inferred.

Two research instruments were used for data collection in this study. They are:

i. Content validity questionnaire: a questionnaire was developed for each course making four different questionnaires on content validity. For SSC105, the questionnaire consisted of the 75 items- test used for 2010/2011 rain semester examination, course content for that semester and schedule having three columns for rating each item in terms of if the item is essential, useful but not essential or not essential with respect of the content. The development of questionnaires used for SSC 102, CHE 101 and PHY 102 followed the same process however their test items were 60, 40 and 45 respectively, as they were used in Obafemi Awolowo University for the same semester. The raters were asked to rate each item by ticking the appropriate columns

of relevance of each item to the course content.

ii. Cognitive domain questionnaire: This questionnaire was developed for each of the courses making four different questionnaires in this respect. Each questionnaire consisted of the test used for 2010/2011 rain semester examination and two columns showing whether the item of the test is measuring lower cognitive domain or higher cognitive domain of Bloom's Taxonomy. The four questionnaires were different from each other only in terms of the number of the test items. As stated above, SSC 105 contained 75 questions, 60 for SSC102, 40 for CHE 101 and 45 for PHY 102. The raters were also asked to indicate the particular cognitive level each item measured.

Furthermore, the convergent validity and internal consistency reliability (Kuder-Richadson KR-20) were determined by administering 2006 and 2012 versions of objective tests of the four courses on Part One students of the University of Nigeria, Nsukka. Coefficients of convergent validity of the two versions of each of the tests for the four courses and their coefficients of reliability were presented in Table 1 and Table 2.

Table 1: Courses and Coefficients of Convergent Validity.

No	Courses	Convergent validity coefficients
1	CHE 101	0.46
2	PHY 102	0.51
3	SSC 102	0.67
4	SSC105	0.71

According to James (1978) and William (2006), convergent validity should be as high as possible while discriminant validity should be as low as possible. The coefficients in Table 1 show that PHY 102, SSC102 and SSC 105 have a moderately high convergent validity for the two tests (i. e, 2006 and 2012 objective tests). Only CHE 101 has low convergent validity.

2.1 Reliability

Table 2: Courses and the Coefficients of Reliability.

No	Courses	Coefficient of reliability
1	CHE 101	0.60
2	PHY 102	0.52
3	SSC102	0.62
4	SSC105	0.68

Reliability indices range from 0.00 to $\pm$ 1. A value close to 0.00 means the test is measuring many unknown factors. The test is measuring a single factor when KR-20 is near I.00, (Aiken, 1985). School test should have a KR-20 of 0.60 or better to be accepted (Oladunni, 1995). It follows that CHE 101, SSC102 and SSC105 are more reliable.

The data on course contents and the tests were collected from the offices of the heads of the departments of the concerned courses. Copies of the scale on content validity were given to ten lecturers in Chemistry, Physics, Mathematics and Economics from Ekiti State University, Ado Ekiti; Adekunle Ajasin Universiy, Akungba Akoko; College of Education, Ikere Ekiti and Adeyemi College of Education, Ondo. The researcher ensured that only the lecturers who haves had at least five years teaching experience were made to participate. The participating raters were asked to respond by ticking the sub-topic each item measured. In the same way, the copies of scale on cognitive domain were given to ten experts in tests and measurement to determine the levels of the cognition measured by each item. The lecturers included possess Doctorate Degree in Tests and Measurement participated in the exercise. The lecturers were from Ekiti State University, Ado Ekiti; Adekunle Ajasin Universiy, Akungba Akoko; University of Benin, Benin, Ambrose Ali University, Ekpoma; College of Education, Ikere Ekiti and Adeyemi College of Education, Ondo. The completed questionnaires were collected a week after.

Responses from the experts were analyzed using frequency tables. Percentages were computed and the content validity ratio (CVR) developed by Lawshe (1979) was also determined. The mean percentages and the mean CVRs calculated were used to determine the content validity of the four tests. Concerning the cognitive levels of the test items, the same process of entry of responses of raters was ensured. Mean percentages were used to judged tests that

measured lower and higher cognition. Item analyses were carried out to determine items that were not plausible.

3. Results and Conclusion

Results are presented in accordance with the ordering of the research questions raised for the study.

3.1 Research Question 1: What is the content validity of the objective test items used in Obafemi Awolowo University?

In answering this question, the frequency distributions of the rating of the test items in terms of essential, useful but not essential and unessential were computed and the summary of mean percentages and mean CVRs for each course were presented in Table 3.

Table 3: Summary of Responses of the Raters to Test Items on the Four Courses.

Course	Essential (%)	Useful but not Essential	Unessential	No response	A CVR
PHY 102	73.1	18	6.88	2	0.49
CHE 101	80.5	9.25	9.5	0.75	0.62
SSC 105	96.8	2.4	0.8	0.13	0.91
SSC 102	90	7.67	0.87	1.17	0.83

Column one of Table 3 shows the mean percentages of essential items of the four tests. Column two and three contain the mean percentages of useful but not essential items and the unessential items. The mean percentages of the items that were not responded to were computed in column four while column five shows mean content validity ratio. PHY102 about seventy-three percent of the items sampled the content. Eighteen percent of the items were useful but are not essential with respect to the content they purport to measure. About six percent of items were declared totally unessential. The content validity ratio (CVR) of 0.49 indicates that the test items have some level of content validity though a good number of the items did not sample the content.

Concerning CHE101, 80.5% of the items sampled the content. About nine percent of the items were useful but not essential with respect of the instructional content also 9.5 % of the items sampled other things entirely while 0.75% of the items were not responded to by the item raters and the mean CVR of 0.62 indicates that the test items have content validity.

Concerning SSC105, 96.8% of the items sampled the content. Two percent of the items were useful but not essential with respect of the content and 0.8% of the items did not sample the content while 0.13% of the items were not responded to by the raters. The value of 0.91 of mean CVR indicates that almost all the items sampled the content.

For SSC102, 90% of the items were essential, 7.7% of the items were useful but not essential with respect of the content while 0.87% of the items did not sample the content. The percentage of the items not responded to was 1.17% and the test produced a hefty mean CVR of 0.83 meaning that the test items have strong content validity.

3.2 Research Question 2: To what extent were the items spread across the levels of cognitive domain of Bloom's Taxonomy?

In answering question 2, the frequency distributions of the rating of the test items in terms of lower level, higher level and undecided were computed and the summary of mean percentages for each course were presented in Table 4.

Table 4: The responses of item raters of the cognitive levels of the test items

Course	Lower cognition (%)	Higher cognition (%)	Undecided (%)
SSC105	43.3	56.6	0
CHE 101	51.13	48	0.08
SSC 102	64	35.33	0.67
PHY 102	46.8	51.56	0.02

In Table 4, *knowledge, comprehension and application* were referred to as the lower level of cognition while *analysis, synthesis and evaluation* were referred to as higher level of cognition.

Column one of Table 4 contains the courses. Columns two and three contain the percentages of items that measured lower and higher cognition while column 4 contains percentages of items not responded to by the raters. From Table 10, 43.3% of SSC 105 items measured lower cognition and 56.6% measured higher cognition. Though the test items engaged students on all the levels of cognition, ten more items measured analysis, synthesis and evaluation.

Concerning CHE 101, 51.13% of the test items sampled lower cognition, while 48 percent measured lower cognition and 0.08% were seen to sample neither lower cognition nor higher cognition. It means that the test engaged students in all the levels of cognitive domain..

Concerning SSC 102, the responses of the item raters showed that 64 percent of the items measured lower cognition, 35.33% measured higher cognition while 0.666 measured neither. This implies that two-third of the test items sampled lower cognition. It is interpreted to mean that the test engaged students more on lower levels of cognition

Results for PHY 102 showed that 46.8% of the items measured lower cognition, 51.56 measured higher cognition while 0.02 measured neither. In the case of CHE 101, PHY 102 almost had equal number of the items sampling lower cognition and higher cognition. It means that the test engaged students in all the levels of cognition.

3.3 Research Question 3: How plausible the options used for the tests?

All the items of the four tests were examined for possible option flaws. Item analyses were also carried out to further determine the plausibility of the options.

Results of the items identified with option flaws were presented in Table 5.

Table 5: Response Option Flaws and the Probability of Getting the Item Right

Course	Number of Items	Number of options	Number of option flaws	Probability of getting the item right with flaws	Probability of getting the item right without the flaws
SSC 105	13	5	13	1/5	¼
SSC102	10	4	14	¼	1/3
PHY 102	45	5	45	1/5	¼
CHE 101	2	4	3	¼	1/3

The first column on Table 5 presents the number of items with option flaws in each test. Column 2 refers to the number of response options each item had. Column 3 shows the number of option flaws in each test. Column 5 is on the probability of getting the item right if it had flaws and column 6 shows probability of getting the item right if the option with flaw was eliminated by the brilliant ones.

Table 5 shows that SSC105 has thirteen items written in five-multiple choice items containing one flaw each. That increased the chance of getting item right for brilliant students from 1/5 to 1/4. Fourteen options of 10 items were written with flaws in four-multiple choice test of SSC102. The probability of getting the item right increased from 1/4 to 1/3. The entire items of PHY 102 contained one item option flaw each. The probability of getting the items right increased from 1/5 to 1/4. Only three flaws were observed in options of two items in CHE 101. The probability of getting the items right increased from 1/4 to 1/3.

Item analyses were carried out on SSC 102and the results were presented in Table 6.

Table 6: Analysis of Response Options of Selected Items in SSC 102.

Item No	Response option	Key	Frequency of patronage (%)	Item Difficulty level	Item Discrimination Index
1	A		8 (16) *16.67*	0.22	0.25
	B	Key	8 (16) *16.67*		
	C		15 (31) *31.25*		
	D		17 (33) *35.42*		
2	A		8 (16)	0.22	0.16
	B	Key	8 (16)		
	C		16 (32)		
	D		18 (36)		
3	A	Key	38 (80)	0.75	0.5
	B		4 (8)*		
	C		4 (8)*		
	D		2 (4)*		
4	A		3 (6)*	0.38	0.42
	B		11 (22)		
	C	Key	19 (38)		
	D		17 (33)		
5	A		17 (35)	0.24	0.17
	B		6 (12)		
	C		15 (29)		
	D	Key	12 (24)		

Column 2 on Table 6 shows the percentages of the examinees that responded to each option. Columns 4 and 5 present the difficulty levels and the discrimination indices of the correct answer. Examinee responses are well distributed over the options in item 2. However, there is need to suggest flaw in option D chosen by 36 percent on an item answered correctly by 16 percent with a lower discrimination of 0.16. Options A to D of item 3 are plausible. The responses of low scorers equally distributed over the three distracters. The correct option produced a hefty discrimination of 0.50. Option A of item 4 is not popular at all. It was chosen by 5.5 percent of the students. However the discrimination of 0.42 makes the correct option a plausible one. There is a problem in option A of item 5. It was chosen by more people including the brilliant one. In SSC102 only four options flanked with asterisk are not plausible. The summary of item analysis carried out on CHE 101 were presented in Table 7

Table 7: Analysis of Options of Selected Items in CHE 101

Item No	Response options	Key	Frequency of patronage (%)	Item Difficulty Level	Item Discrimination Index
1	A	Key	9 (27)	0.27	0.47
	B		15 (45)		
	C		2 (5)		
	D		7 (23)		
2	A	Key	9 (27)	0.27	0.51
	B		10 (32)		
	C		10 (32)		
	D		3 (9)*		
3	A	Key	5 (17)	0.17	0.13
	B		13(42)		
	C		10 (33)		
	D		5 (17)		
4	A		15 (45)	0.41	0.23
	B	Key	13 (41)		
	C		3 (9)*		
	D		2 (5) *		
5	A		5(15)	0.42	0.07
	B	Key	14(45)		
	C		8 (25)		
	D		5 (15)		

Table 7, items and options were presented in the first column. Column 2 shows the correct option. Columns 3 to 5 show the responses to each item, the difficulty level and the discrimination index of the correct answer. For item 1, option B and C are not good. Option C is not popular even among the low scorers. The difficulty level indicates that the option attracts even the high scorers. The distribution of the responses over the options in item 3 is plausible. However option D is not popular given the level of response to it. The low discrimination of 0.13 of option A in item 3 is not good. Options B and C were very popular by the high level of patronage by testees.

There is a problem in option A of item 4. 45% of the examinees chose it however the level of discrimination shows that majority of the examinees that chose option A are low scorers. The responses on item 4 distributed properly over the distracters. 0.16 discrimination of the correct answer is not encouraging. In all only three options flanked with asterisk are not plausible.

Table 8: Analysis of Options of Elected Items in SSC 105

Item No	Response options	Key	Frequency of patronage (%)	Item Difficulty Level	Item Discrimination Index
1	A		0 (00)*		
	B	Key	27 (62)	0.62	0
	C		3 (8)*		
	D		14 (31)		
2	A		1 (2.2)*		
	B		12 (29)	0.35	0.72
	C	Key	31(72)		
	D		16(29)		
3	A		16 (36)		
	B		16(36)	0.29	0.42
	C	Key	12(29)		
	D		0 (0)*		
4	A	Key	1 (44)*		
	B		0(0)*	1	1
	C		0(0)*		
	D		0(0)*		
5	A		0(0)*		
	B	Key	19(43)	0.43	0.57
	C		24(57)		
	D		0(0)*		

In Table 8, options A and C of item1 were not popular. None of the examinees selected option A. Option C recorded a low response of 7.7 % of the examinees. Distracters B and D are plausible. Only option A suggests option flaw in item 2. Option D of item 3 is not popular at all because it was not responded to by any examinee. However, the distribution of the responses is plausible. Item 4 is to easy because all the examinees got the answer correctly. Options B, C and D are not plausible at all. Item 5 is a little better than item 4, however options A and D not responded to should be avoided. In SSC102, ten options flanked with asterisk are not plausible.

Table 9: Analysis of Options of Selected Items in PHY 102

Item No	Response options	Key	Frequency of patronage (%)	Item Difficulty Level	Item Discrimination Index
1	A	Key	20(72)		
	B		1(3)		
	C		2(7)	0.28	0.40
	D		3 (9)		
	E		3(9)		
2	A	Key	20 (72)		
	B		1 (3)		0.04
	C		2 (7)	0.28	
	D		3 (9)		
	E		4 (12)		
3	A		1 (3)*		
	B		0(0)*	0.02	0.00
			1(3)*		
	D	Key	28 (94)		
	E		0 (0)*		
4	A		1 (3)*		
	B		415		
	C		2(5)	0.30	0.09
	D	Key	21(70)		
	E		3 (8)		

In item 1 of Table 9, 72% responded to the correct option. The responses of other 27 percent were well distributed over the distracters B to D. The discrimination of the correct option is also good. The discrimination index of item 2 indicates that those who did poor in the test got the item correct. However the response distribution on other options is good. Options A, B C and E of item 3 are not plausible. Options B and E were not chosen by any examinee. Only 1% of the examinees chose options A and C. Given the level of discrimination of the correct option, equal number of high and low scorers selected the option. There is problem with options A, C and E. They attracted few examinees. In PHY 102, 5 options flanked with asterisks were not plausible.

4. Discussion

This present work focuses on content validity and levels of cognition of the objective test in Obafemi Awolowo University. Within the large framework of content validity, the researchers aimed to investigate if the objective tests used in this school actually measured the content and the levels of cognition.

First of all, the mean percentages of the relevance of the items to the contents range from 71 to 96 percent. As was expected, SSC 105 which is mathematics had the highest mean percentage, thereby confirming the previous studies that achievement of high content validity is easy with mathematics. Singamaneni (2011) in his study on exploring content validity, item level analysis and predictive validity for two algebra progress monitoring measured observed that high content validity can be achieved in mathematics than can be achieved in personality trait.

With regards to the sampling of the content by the test items, the present study revealed high content sampling indicated by the values of the mean percentages of the four courses. These results confirm previous findings of research on investigation of content validity for courses involving calculation (Coaley, 2010). With the use of CVR, item with positive value measures the content (University of Pretoria, 2001 and Coaley, 2010). The values of ACVRs of the four tests indicate that the test items of the four courses measured the content. The high content validity coefficients from this study do not support the view of many that teachers/lecturers outside the field of education just set objective questions without taking cognizance of the content Mehret (2008).

Also worthy of note is the level of cognitive behaviour measured by the items. Bloom (1956) classified cognitive domain into six levels. However, Nwana (2007), classified them into two levels-lower and higher levels- and maintained the fact that objective test format can be used to measure higher other mental processes.

This study has revealed that the tests measured lower and higher cognition but at different rates. This finding confirms the previous studies that objective tests can also measure high cognitive level if properly constructed (Bark, 2011). Further, the results from SSC105 and PHY 102 in which higher cognition was emphasized ran contrary to the

view of Kohn (2000) and Jacobs (2004) who believed that objective tests correlated with shallow knowledge and could only measure isolated knowledge of facts and assess learning in that respect. However, SSC 102 which had more of its items sampling lower cognition supports their view.

Findings emanating from this study also reveal that virtually all option Es in PHY are not plausible (NBME, 2010; Arisona State University, 2002 and Oladuuni, 1995). This pattern of answer increased the probability of getting the items right for the test wise examinees. The researcher discovered that six items in SSC102 were three-option multiple choice items and one of the options was written as either undecided or as uncertain and thus increased the chance of getting such item right from 1/3 to 1/2. In SSC 105 It was observed that option "none of the above" was used for items with five options. Where the options were four "none of the above" was not used. The researcher was of the view that the test developer just added such options there for the impression that the test was five-option multiple-choice test. It was observed that in all the items where the option was used in the test it was made a correct answer of only one item. The researcher further observed that "none of the above" used as the fifth option for all the items of PHY 102 was not made a correct answer of any of the item.

Findings of this study show that six options out of 20 options were not responded to by any examinee in SSC105. Ten options in SSSC102, Two options in CHE 101, four options in SSC 102 and five options in PHY 102 are not plausible. The options were not functioning with respect of the construct been measured (Kevin and Charles, 2005). The options did not contribute to the discrimination and difficulty of the items. Kevin and Charles (2005) advised that they should be removed or restructured for subsequent examination

5. Conclusion

The study concluded by declaring that the four courses considered had high content validity. The items sampled most of the content areas of each of the courses. The study also concluded that the items of CHE 101 and PHY 102 were appropriately spread across the six levels of cognitive domain. Questions measuring lower levels of cognitive domain were pronounced in SSC 102 while questions measuring higher levels of cognitive behaviour were more pronounced in SSC 105. This shows that though SSC 102 had high content validity, its items might not have measured all the dimensions of cognitive behaviour that should lead to reliable judgments about the students.

6. Recommendations

Item writers of tests in the universities (or similar institutions) should develop their tests in accordance to the topics listed in the course contents. They should also increase the number of the items per course to between 70 and 100. This will increase the test reliability and make it to actually mirror the objectives of the content. They should also increase the number of items designed to measure high cognitive behaviour so that students certified to have passed the courses could demonstrate high level of proficiency.

Irrelevant options such as "none of the above" and "all of the above" that benefit only the testwise examinees should be avoided. It is important that all test item writers carry out item analysis in order to determine the psychometric property of each item and thus have genuine assurance of the appropriateness of the items used in testing undergraduate students.

References

Adepoju, T. L. and Oluchukwu, E. E. (2011). A Study of Secondary School Students' Academic Performance at the Senior School Certificate Examinations and Implications for Educational Planning and Policy in Nigeria. African Research Review, 5(6). Available: http://www.ajol.info/index.php/afrrev/article/view/72369

Alli, U. (2013). Tackling the challenges in Nigeria's public examinations. [Online] *Sunday Trust,* 23rd June, 2013. Available: http://sundaytrust.com.ng/index.php/comment-debate/13415-tackling-the-challenges-in-nigeria-s-public-examinations (17/02/2014)

Arizona State University and the Arizona Board of Regents (2002). *Forced–Choice Testing Formats.* Arizona State University.

Bark A. R. (2011). *A Consumer's Guide to Multiple-Choice Item Formats that Measure Complex Cognitive Outcomes*. School of Nursing, Johns Hopkins University.

Bloom, B. S. (ed.) (1956). *Taxonomy of Educational Objectives*. Vols. 1 and 2. New York: David McKay.

Crowl, T. K.; Kaminsky, S. and Podell, D. M. (1997). *Educational Psychology: Windows on Teaching*. London: Brown & Benchmark.

Faleye, B. A. & E. R. I. Afolabi (2007). *Continuous Assessment Practices in Osun State (Nigeria) Secondary Schools: From Policy to Practice*. International Journal of Learning, (Australia), 12 (12), 11-16. Available at: http://www.Learning-Journal.com

Faleye, B. A. (2005). *Establishing and Maintaining Standards in Nigeria's Senior School Certificate Examination: The Challenge of Malpractice. Paper presented at the at the 2005 Conference of the International Council on Education for Teaching* (ICET) held at the University of Pretoria, South Africa. Yearbook of the July 2005 Conference of International Council on Education for Teaching. Illinois: ICET (CD Rom).

Faleye, B. A. (2007). *How to Construct Table of Specification*. Unpublished Paper Presented at a Workshop Organised for Teachers of Obafemi Awolowo University Staff School on 30th October, 2007 at the Science Laboratory, Obafemi Awolowo University Staff School, Obafemi Awolowo University, Ile-Ife, Nigeria.

Jacobs L . C . (2004). *How to Write Better Tests: Handbook for Improving Test Construction Skill*. Indiana University, Indiana.

James A. R. (1978). *Evidence of Convergent Validity of the Dimension of Affect*. Journal of personality and Social Psychology, 36 (10), 1165.

Kevin R. M. & Charles O. D. (2005). *Psychological Testing : Principles and Applications*. New Jersey: Pearson Education, Inc., Upper Saddle River.

Kohn A. (2000). *Standardized Testing and Its Victims*. [Online] Available: http://www.alfiekohn.org/teaching/edweek/staiv.htm (17/02/2014).

Mehret, A. (2008). *An Assessment of the Content Validity of English Language Tests*. Addis Ababa: Addis Ababa University Press

Nworgu B. G. (2006). *Educational Research: Basic Issues and Methodology*. Ibadan: Wisdom Publishers Limited.

Obafemi Awolowo University (2010). *The Postgraduate College Handbook*. Ile-Ife: Obafemi Awolowo Press.

Oladunni M. O. (1995). *Introduction to Research Methods and Statistics in Education*. Ikere--Ekiti: Tafak Publications Nig. Ent.

Oweh, I. (2014). *Unending Worries over Mass Failures in Public Examinations in Nigeria*. [Online] Daily Independent, February 17, 2014. Available: http://dailyindependentnig.com/2012/12/unending-worries-over-mass-failures-in-public-examinations-in-nigeria/ (17/02/2014).

Popham, W. J. (2002). *Classroom Assessment: What Teachers Need to Know*. Boston: Allyn and Bacon.

Singamaneni S. (2011). *Exploring Content Validity Item Analysis and Predictive Validity for Two Algebra Progress Monitoring Measure*. Unpublished M.Sc. Thesis, Iowa State University, Iowa. Available at: http://lib.dr.iastate.edu/etd/11929/

The Postgraduate College (2010). *The Postgraduate Handbook*. Obafemi Awolowo University, Ile-Ife Osun State, Nigeria.

University of Pretoria (2001). *A Validated Model of the South Africa Labour Relation Systems*. Pretoria (South Africa): University of Pretoria.

William M. K. (2006). *Measurement Validity Types*. [Online] Available: http://www.socialresearchmethods.net/kb/measval.php (17/02/2014).