EFFECT OF TEST FACETS ON THE CONSTRUCT VALIDITY OF ECONOMICS ACHIEVEMENT TESTS IN OSUN STATE SECONDARY SCHOOLS

Yusuf O. Shogbesan & Bamidele A. Faleye
Department of Educational Foundations and Counseling,
Faculty of Education,
Obafemi Awolowo University, Ile-Ife

ABSTRACT

The study determines the construct validity of the tests used by Economics teachers in the secondary schools in the study area and further examined the influence of test item format on the construct validity of Economics achievement tests in Osun state Secondary schools. To achieve this, the study adopted descriptive survey research design. The sampling technique used was multi-stage sampling used to select randomly a sample size of 300 Senior Secondary School II Economics students and 36 Economics teachers. The study made use of two research instruments: Economics Achievement Test (EAT), and a score record sheet. The EAT was used to collect information used to determine the construct validity of Economics achievement test and was validated using the test blueprint to ensure content validity and the cronbach's alpha reliability which yielded coefficient value of 0.68. The data collected were analyzed using Principal component analysis, scree plot and one-way analysis of variance. The result showed that Economics tests constructed by the Economics teachers as used in their classroom assessment have construct validity ($r=0.414$, $df=298$, $p<0.05$) and that there is a significant influence of test item format on the construct validity of the Economics Achievement Test in Osun state secondary schools. Finally, there is a significant difference in the performances of test-takers across various test formats used in the Economics Achievement Test ($F=290.25$, $p<0.05$).The study concluded that test item formats as a facet have an effect on the construct validity of the Economics Achievement Test in Osun state secondary schools and recommend that test constructors and classroom teachers must understand the characteristics of each format and select the best format which most appropriately serves the purpose of a test in each context.

Keywords: Test Facets, Test Formats, Construct Validity, Achievement Test.

INTRODUCTION

Test refers to a method that is use to determine a student's ability to complete certain tasks or demonstrate mastery of a skill or knowledge of content (Kizlik, 2012). Tests are meant to elicit information about a latent ability of an individual with respect to a particular variable of interest. This variable of interest is to assess student learning and to provide evidence so that educational decisions can be made. These decisions when made provide information about students on whether they have reached a particular level of skill and knowledge. It may help us evaluate a teaching programme or to make decisions about the next aspect of teaching for particular students. Also, classroom testing can provide teachers with valuable feedback about what students do and do not know, and teachers in turn can encourage students to change their study behaviour. However, making valid educational decisions with test depend upon the validity and therefore reliability of the measures. It is also important to realize that the most vital quality of a test is that it exhibit

construct validity. Validity is an overall evaluative judgment of the degree to which empirical evidence and theoretical rationales support the adequacy and appropriateness of inferences and actions based on test scores (Messick, 1993). Emerging paradigms replace prior distinctions of face, content, and criterion validity with the unitary concept "construct validity," as the degree to which a score can be interpreted as representing the intended underlying construct. But when the measurement of a particular trait in a test is affected by the method of testing and its related facets, the interpretation of scores obtained from such tests as well as validity evidence is questionable. Students might not understand the test instructions, question formats and structure, eventually limiting their scores and will underestimate their actual learning. When a candidate answers a question purely through familiarity with the question style, rather than through an understanding of the subject content then it could be argued that they have not developed the cognitive skills that the test is designed to measure. It is therefore essential to understand that the validity of tests are affected by the procedures and format used to measure the intended construct when test- takers are not familiar with the style and format of testing.

Generally, given that the goal of testing is to assure the extent to which learners have achieved the instructional goals during a course then the development of valid tests would be a troublesome task to be accomplished if all the major factors affecting learner's performance are not considered. However, tests may cover relevant construct but may not provide valid scores because of the method used in the test or procedure of test construction. This is a problem in testing known as "construct-irrelevant variance" (Haladyna and Downing, 2004). This is because construct-irrelevant variance produces variation in examinees' marks that does not come from their knowledge of the ideas being tested. Specifically, factors which include; examinee's interest in the test and effort on the test, test anxiety and related conditions, prior exposure to test items or other relevant preparation and training, test scoring, cheating etc have the potential to yield construct-irrelevant variance in achievement test scores, threatening the validity of these tests (David, Benjamin and Alexander, 2011). Also, the construct irrelevance variance can be affected by several factors which may include but not limited to their knowledge, gender, and culture, as well as the test facet, therefore weakening the test's validity for them (Roya and Nazli, 2012). Mousavi (2009) added that when a test measures only the abilities it is supposed to measure, we can say the test has construct validity.

In a study that was conducted to explore the effect of test method facet on the validity of the grammar sub-tests of high stakes tests with a view toward test fairness. The results of the factorial analysis revealed that the different formats of grammar tests produced different levels of construct validity, with the TOEFL format enjoying the highest and the SAT format the lowest level of construct validity. This indicated that the no-error option of the SAT error identification test might have played a role in its low level of validity, thus testifying to the presence of bias in this kind of testing (Roya and Nazli, 2012). Also, it was discovered that although all of the tests measured the same construct and had been made on the basis of the same table of specifications, there existed only moderate correlations between them, with the highest between the TOEFL and MC tests (0.69), and the lowest between the SAT and the MC tests of grammar (0.417). This also indicated that the test format should have played a significant role in producing moderate correlations among the pairs of scores on the different tests. Moreover, the test facets are, in fact, of the greatest importance in determining the effects of the test on the learner's performance (Zahra and Abdolreza, 2012). The test facets refer to all aspect of a test format that affects student scores or

performances. Test method facets characterize the operational framework of tests and different types of test method facets tap into different aspects of the construct (Bachman, 1990; Chalhoub-Devile, 1995; 1997). Backman (1990) as cited by Eom and Minhee (2006), proposed the description of test method facets that includes setting, rubric, input, expected response and relationship between input and response. The framework of test facets may provide test developers a detailed basis for the description of different tests for purposes of selection, and for the design of specific tests. It is also an appropriate means of codifying and describing the tests they are developing, using, or researching. Also, different test format have different test rubrics and are administered under different testing conditions. The test facet allows for the formulation of hypotheses for testing research to investigate the factors that affect performance on tests. It examines the effects of specific facets, either by themselves, or as they interact with other facets. As these facets have a role in performance, test- developer should make use of them effectively while preparing tests (Tanyer, 2015). According to Zahra and Abdolreza (2012) in their study which investigated whether test participants' performances were different with respect to the different test facets and if these performances had some effects on the construct validity of the tests carried out among selected 50 Iranian EFL students revealed that significant differences existed in the test facets among the performances of Iranian EFL students. The test facets included the integrative forms such as cloze-test, c-test, and discrete test items such as multiple choice and true/false, the study concluded that of the entire facet used to measure several abilities and mental strategies, the cloze-test was the most difficult form of testing (Zahra and Abdolreza, 2012). This finding was in agreement with Weir (1990) who also believes that the integrative tests, such as cloze test and c-test only demonstrate a view of the learners' knowledge, and they fail to illicit the learners' language performance.

In testing, by applying different test facets, we can examine much knowledge of the students and through different test forms, students learn to study and understand the material comprehensively in different ways and it allows them to tap their strategies to various test facets in different administrations. These features of test methods, alternately termed facets affect test performance (Salehi and Bagbesi, 2013). It is partly owing to the fact that the individual characteristics i:e cognitive and affective styles interact with the aspects of the test methods (Bachman,1990). Bachman further asserted that performance of test takers in language tests is the outcome of interaction between a testee's language ability and other variables not targeted by the research such as cognitive and affective characteristics and features of the test facets. Similarly, Signi, Ajideh and Esfandiari (2009) conducted a study in order to investigate and compare two tests formats, the multiple-choice test and cloze test. First, they administered a test to homogenize the participants and the contents of the two tests had been kept constant. They concluded that in testing the proficiency of a group of learners, the achieved scores on the multiple-choice lexical tests were much similar to the cloze test scores. Although two tests were seemingly different, there was a high correlation between the two types of test formats on vocabulary-discrete-point item, and integrative cloze test. An interesting point from the finding was that those who acted better on cloze tests could also perform better on discrete-point tests. This is an indication that the test format ensures convergent validity evidence of tests measuring similar construct. Specifically, the measurement of the different construct of the subject matter of Economics requires that the students have much knowledge because of it broader cognitive requirement. Economics is a subject taught is both theoretical and mathematical level in senior classes with the use of syllabuses. The teaching of economics required specific skills just as the teaching of other subjects.

It is one of the subjects that require thinking and reasoning, in that it takes long time for one to understand it especially in the calculation (Mackson and Oruta, 2000). As a result, the Economics teacher-made test constructed will make use of different test facets to measure different intended behavioural objectives and content domain. It is in furtherance of this background that the present study concerns itself with the investigation of the effect of tests facets on the construct validity of tests. In classroom testing, the various formats are used to measure various levels of cognitive domains as it interest the classroom teacher. All these formats have various facets that can be use to explain them. But, the concern of the classroom teacher will centre around which format(s) can be used to effectively measure a particular domain of learning with a great consideration on how it affect every facets of the test. Since a given trait can be measured through different formats, these methods can have different effects on that trait and the test-takers' scores. The item format may limit or prevent certain construct elements from being included in the test, or otherwise interfere with it, causing distortions in the scores with the possible result that they no longer reflect the construct very well (Gergely, 2007). The resulting effect of the distortion may result into the test been seen as a bias or unfair measure of certain attribute. The respondents that are familiar with the test format are likely to perform better on the test than respondent that are not familiar with the testing method. Similarly, a test taker may perform better in a test item format probably because of the educational domain the format is capable of measuring or that it is measuring. All these are the most important factors affecting fairness of a test. It can therefore be seen that format of a test not only affect the performance of the test takers but also considered as a possible source of bias that is capable of raising fairness issue as organized around four stages of the assessment process: design, development, administration and use.

In a study conducted by Shohamy (1997), she reviews two major sources of bias or systematic errors as those associated with the test itself in language testing, such as the format effects and those associated with the consequences and uses of tests. As she illustrates, it may examine whether success on a reading comprehension test implies that the test score is more dependent on how it is being measured, e.g. by multiple choice, open ended questions, than on the trait of reading comprehension. Therefore, if the test takers had been tested by a different method, they would have obtained different scores. Studies on the format effect indicate that the test itself affects the scores that test takers obtain on tests. Thus, aspects related to item format, testing tasks are found to affect test taker scores resulting in bias against certain test takers. According to Shohamy (1997), the bias related to test use implies that a language test is being used for unfair purposes. Hence, testers and test users need to consider ways of minimizing these unethical uses of language tests.
Also, students sometimes have the same understanding of a given test, but the way in which the test is administered leads to different consequences (Roya and Nazli, 2012). Since the examinees are usually familiar with the test item format frequently used in the classroom during assessment. They may perform well in such test as the answer to a question can be purely through familiarity with the question style and not depends on understanding of the subject content. Then, such test could be argued not to have developed the cognitive skills that the test is really designed to measure. Conversely, a student, who is not experienced enough in various formats of testing, should not be expected to do well in unknown formats as opposed to more known ones (Roya and Nazli, 2012). In Nigeria, student performance in the last decade in Economics has been considered low. This can be deduced from the report of the WAEC chief examiners report (2003) which indicates that although performance in Economics had been just fair, yet it could have been better if the candidates could do better in the mathematical/statistical aspect of the subject. The report

further indicates that the candidates are unable to draw and label diagrams well and that they often misinterpret questions. An undated WAEC statistics on student performance in Economics (as cited in Onuka, 2009) show that credit passes recorded from 1994 to 2004 were: 27.9%, 15%, 19.6%, 14%, 22%, 21.7%, 35.4%, 28.2%, 22.3%, 42.98% and 38.2% respectively. This result implies that much is still left to be desired in the performance of students in SSCE Economics, unless the performance level was improved (Onuka, 2009). It is important to note that these low performance in Economics can also be as a result of characteristics of the persons and that are independent of the items such as illness, boredom, fatigue, lack of interest etc. and the properties of the items which are independent of the characteristics of the person. It consist of the amount of the trait possessed by the person i.e. the person's ability and the amount of the trait necessary to provide response to a given stimulus. i.e. item difficulty (Rasch, Schumaker, Mount and Marcoulides, 2005, Adebowale, 2007).

The influence of numerous factors on learners' performances in tests has been investigated, but it is not clear whether a particular objective item type or a particular essay test item type will be better able to validly measure economics achievement among Osun State secondary school students. The main objective of this study was to determine the construct validity of the achievement tests used by Economics teachers in the secondary schools in the study area; investigate whether test facets affect construct validity of the test or not and determine whether participants' performances were different with respect to the different test facets. To achieve the above objectives, the following research question and hypotheses were formulated;

Research Question

a. What is the construct validity of tests used by economics teachers in their classroom assessment?
b.

Research Hypotheses

a. There is no significant influence of test item format on the construct validity of Economics Achievement Test in Osun state secondary schools.
b. There is no significant difference in the performances of test-takers across various test formats used in the Economics Achievement Test.
c.

METHOD

The study adopted the survey research design. The study population comprise public senior secondary school Economics teachers and students in Osun State. The multi-stage sampling technique was used to select the study sample which comprise of 36 Economics teachers and 300 Senior Secondary School Two (SSSII) Economics students. From each of the three senatorial districts in Osun State, two Local Government Areas (LGAs) was selected using simple random sampling technique while two schools was further selected using simple random sampling technique in each of the selected LGAs. The study makes use of two research instruments: Economics Achievement Test (EAT) and a score record sheet. The EAT, which was used to collect

information on the construct validity of the achievement test and comprised of five items each of multiple-choice, binary choice, completion, short-answer and essay item formats while the score record sheet was used to record previous examination scores (2015/2016 academic session) of the selected students. The data gathered was analysed using Pearson product moment correlation coefficient to establish a convergent/ discriminant validity as a measure of construct validity of Economics test used by Economics teachers in their classroom assessment, while the principal component analysis (PCA) and the Scree plot diagram were further used to explore and confirm the research hypothesis one of the study to determine whether the test item format significantly influence the construct validity of the EAT. Finally, Analysis of Variance (ANOVA) was used to determine whether there is a significant difference in the test-takers performance across the various format used in the EAT.

RESULTS

Research Question one: What is the construct validity of tests used by economics teachers in their classroom assessment?

To answer this research question, the scores obtained in the 2015/2016 first term examination by the selected Economics students used for the study was obtained from the score record sheet provided by the Economics teachers. Also, the total scores obtained in the Economics Achievement Test (EAT) were computed and used as a criterion measure. The scores of each student on the both Economics test were then correlated using Pearson product moment correlation to establish whether the Economics test constructed by the Economics teachers as used in their classroom testing have either convergence/divergent validity as a measure of construct validity. The summary of this is presented in table 1 below.

Table 1: The correlation coefficient of scores obtained from test used by Economics teachers in their classroom assessment and the Economics Achievement Test (EAT)

Variables	N	$\overline{X}$	S.D	r	df	P
School exam scores	300	50.52	13.54			
EAT total scores	300	17.76	8.03	.414	298	<0.05

From table 1 above, there exists convergence validity between the scores obtained by Economics students in the scores obtained from test used by Economics teachers in their classroom assessment and the Economics Achievement Test (EAT) with a moderate correlation coefficient of 0.414. Hence, it can be concluded that tests used by Economics teachers in their classroom assessment have exhibit construct validity.

Hypotheses testing

Research Hypothesis One:

There is no significant influence of test item format on the construct validity of the Economics Achievement Test in Osun state secondary schools.

To test this hypothesis, respondents' scores on each of the test item format on the EAT were subjected to Cronbach Alpha as a measure of internal consistency. The results are shown in Table 2 below.

Table 2: Cronbach Alpha of test item formats in the EAT

Test item format	N	Cronbach Alpha
True/ false	300	0.026
Short answer	300	0.801
Completion	300	0.699
Multiple choice	300	0.425
Essay items	300	0.515

As shown in table 6 above, the Short answer test item format (0.801), Completion test item format (0.699), multiple choice test item format (0.425) and Essay test item format (0.515) enjoyed acceptable to good levels of reliability. While the True/False test item format has the lowest level of reliability (0.026) which is not acceptable. If a test enjoys strong internal consistency, most measurement experts agree that it should show only moderate correlation among items. For exploratory purposes 0.60 is accepted; for confirmatory purposes 0.70 is accepted; and 0.80 is considered good (Garson, 2010). Here, the short answer test item format (0.801) had the highest level of reliability and the True/False test item format has the lowest level of reliability.

Furthermore, to investigate the construct validity of the Economics Achievement Test in Osun state secondary schools, the scores of the respondents on the EAT were subjected to a factor analysis. This analysis was done to determine whether all these measures shared some common variance and, thus, could be said to tap the same underlying construct using the principal components analysis (PCA) to extract the initial factors. The extraction of factors was base on the suggestion by Sharma (1996) and Zwick and Velicer (1986) that the eigenvalue-greater-than-one should be selected as the extraction rule. This rule suggests that those factors whose eigenvalues (sum of squared loadings) are less than unity be excluded from the analysis. The results of the factor analysis are presented in table 3 and 4 below and figure 1 below.

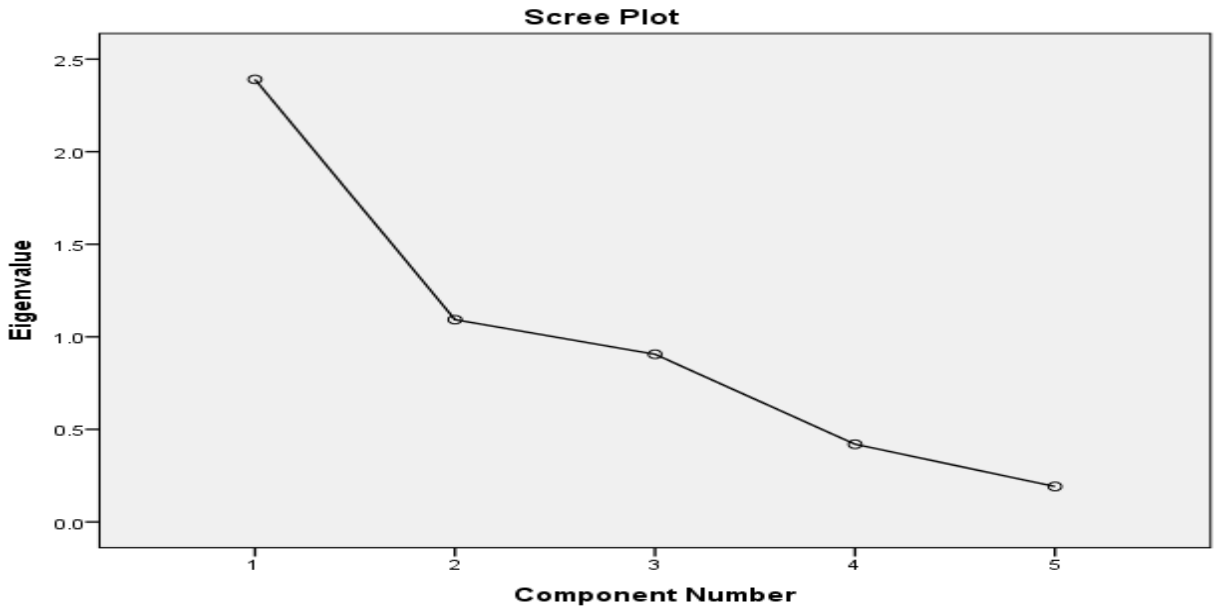
Table 3: Total Variance Explained by result of factor analysis

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings
	Total	% of Variance	Cumulative %	Total	% of Variance	Cumulative %	Total
1	2.391	47.813	47.813	2.391	47.813	47.813	2.365
2	1.093	21.852	69.665	1.093	21.852	69.665	1.124
3	.906	18.115	87.780				
4	.420	8.394	96.174				
5	.191	3.826	100.000				

Extraction method: principal component anlysis

From table 3 above, the result of the pricipal component analysis as a measure of construct validity indicated that only two factors loaded with factor 1 explaining 47.81% of the variance and factor 2 explaining 21.85% of the total variance in the EAT across the test scores. The two factors extracted have eigenvalues greater than one. This is shown using a Scree plot in figure 1 below.

Figure1: scree plot of factor analysis result using principal component analysis



From figure 1 above, it can be seen that the first factor have an eigenvalue of 2.391(>1.000) while the second factor have an eigenvalue of 1.093(>1.000) with both factors accounting for a total variance of 69.67% of the total variance explained by the scores obtained on each of the test format on the EAT.
Also, the factor loadings of each format on both factor 1 and 2 was obtained. This is shown in table 4 below.

Table 4: Component matrix of factor loadings of each test format

Test Item Format	Factor 1	Factor 2
True/False	0.369	-0.561
Short answer	0.835	-0.324
Completion	0.835	0.005
Multiple-choice	0.422	0.805
Essay items	0.826	0.161

Extraction Method: Principal Component Analysis.

Table 4 above indicates that the multiple- choice test item format (0.826), short answer item format (0.802), essay items (0.709) and completion test item (0.697) enjoyed high loadings on Factor 1. The highest loading on factor 1 belonged to the multiple- choice test item format (0.826) and the lowest to the true/false test item format (0.450). While only the multiple- choice test item format (0.805) enjoyed a high factor loadings on Factor 2 extracted.

Although all the tests item formats measured the same construct to a large extent especially as observed from their respective factor loadings on factor 1, it seems that the true/false test item format (0.450) significantly reduced its construct validity. Hence, there exists a significant influence of test item format on the construct validity of the Economics Achievement Test in Osun state secondary schools.

Research Hypothesis Two: There is no significant difference in the performances of test-takers across various test formats used in the Economics Achievement Test.
To test this hypothesis, respondents' scores on each of the test item format on the EAT were computed as a single score. A one-way analysis of variance (ANOVA) and a post-hoc test (Scheffe test) were carried out to see if there was any significant difference among the respondents' performances on each of the test formats used in the EAT. The result is presented in table 5 below;

Table 5: ANOVA of respondents' scores on each of the test formats used in the EAT

	Sum of Squares	df	Mean Square	F	p
Between Groups	7699.131	4	1924.783	290.254	.000
Within Groups	9913.917	1495	6.631		
Total	17613.047	1499			

Table 5 shows results acquired in the one-way ANOVA to find if there was a significant difference in the means of performances of the subjects across the test formats used in the EAT. Also, the mean differences across all the formats were significant (F=290.254, P< 0.05). This made the researcher claim that the meaningful differences could be attributed to the test formats used in the study. So, the null hypothesis was rejected and therefore concludes that there is a significant difference in the performances of test-takers across various test formats used in the Economics Achievement Test.
However, another analytic method 'a Scheffe test' was used to pinpoint the exact location of the difference among the means. The result is presented in table 6 below;

Table 6 : Multiple Comparisons of mean scores across each test format used on the EAT

(I)	(J)	Mean Difference (I-J)	Std. Error	Sig.	95% Confidence Interval	
					Lower Bound	Upper Bound
A	B	1.25000*	.21026	.000	.6015	1.8985
	C	1.72333*	.21026	.000	1.0749	2.3718
	D	.93000*	.21026	.001	.2815	1.5785
	E	-4.51000*	.21026	.000	-5.1585	-3.8615
B	A	-1.25000*	.21026	.000	-1.8985	-.6015
	C	.47333	.21026	.281	-.1751	1.1218
	D	-.32000	.21026	.678	-.9685	.3285
	E	-5.76000*	.21026	.000	-6.4085	-5.1115
C	A	-1.72333*	.21026	.000	-2.3718	-1.0749
	B	-.47333	.21026	.281	-1.1218	.1751
	D	-.79333*	.21026	.007	-1.4418	-.1449
	E	-6.23333*	.21026	.000	-6.8818	-5.5849

D	A	-.93000*	.21026	.001	-1.5785	-.2815
	B	.32000	.21026	.678	-.3285	.9685
	C	.79333*	.21026	.007	.1449	1.4418
	E	-5.44000*	.21026	.000	-6.0885	-4.7915
E	A	4.51000*	.21026	.000	3.8615	5.1585
	B	5.76000*	.21026	.000	5.1115	6.4085
	C	6.23333*	.21026	.000	5.5849	6.8818
	D	5.44000*	.21026	.000	4.7915	6.0885

*. The mean difference is significant at the 0.05 level.

NB:

- A: True/False test item format
- B: Short answer test item format
- C: Completion test item format
- D: Multiple choice test item format
- E: Essay test item form

The highlighted values in Table 6 above delineate the differences between the students' means on different tests. By comparing the mean differences of scores obtained by respondent among the test formats, the researcher found that the mean differences were significant at the level of 0.05. The difference was found to significantly exist between the true/ false test formats and short answer format, completion item format, and multiple-choice format respectively. Also, the difference was found to significantly exist between the essay item format and true/false test format, short answer format, completion item format, and multiple-choice format respectively. So, the null hypotheses which claimed that the results of the each test do not differ significantly with the results of other test formats was rejected, because the mean differences are significant at $p<0.05$.

DISCUSSION OF FINDINGS

The first finding of the present study establish that the Economics test constructed by the Economics teachers as used in their classroom testing have convergence validity as a measure of construct validity with a moderate correlation coefficient of 0.414 with the Economics Achievement Test (EAT) constructed by the researcher. The result of the analysis of the hypothesis one of the present study indicated that there exists a significant influence of test item format on the construct validity of the Economics Achievement Test in Osun state secondary schools with the Short answer test item format, Completion test item format, multiple choice test item format and Essay test item format enjoyed acceptable to good levels of reliability. While the True/False test item format has the lowest level of reliability in the EAT. The result further shows that only two factors were extracted which explained the total variance of 69.67%. The variance produced by the measures found to be due to both method and trait factors, which can best be interpreted as accounting for students' knowledge of the subject matter (Economics) as a construct. Similarly, the result of the study conducted by Kyoii and Paydarnia (2011) where only one factor was extracted, suggested that from among all the three test formats used in the study, the multiple-choice test had the highest loading on this factor suggesting that it was capable of fulfilling many of the requirements of a suitable test in terms of construct validity. The statistical characteristics, particularly the high construct validity of the Short answer test item format, Completion test item format, multiple choice test item format and Essay test in the EAT used in this study revealed that

this format could be the most appropriate for testing the knowledge of the subject matter of Economics. While in multiple-choice questions students select a response rather than construct their own, which may lower test anxiety for test-takers, allowing them to make the best use of their knowledge. However, the True/False test item format was found to reduce the construct validity of Economics Achievement Test to a considerable degree despite the fact that in True/False test items, the chance of answering is 50% to 50%, allowing the students to answer the items by simplicity or by chance. The low construct validity suggests that the format does not measure the subject matter of Economics as a construct very truly well because of its nature that allows for negative suggestibility. This is similar to results of various researchers (Roediger & Marsh, 2005; Fazio, Agarwal, Marsh, & Roediger, 2010; Marsh, Roediger, Bjork, & Bjork, 2007; the SAT; Marsh, Agarwal, & Roediger, 2009; Butler & Roediger, 2008) on test items which shows that negative suggestibility is real, at least on true/false and multiple-choice tests. Roediger & Marsh, (2005) discovered that when they answered erroneously, the negative suggestibility effect occurred, thereby affecting their performance.

The results of the analysis of the hypothesis two of the present study showed that there is a significant difference in the performances of test-takers across various test formats used in the Economics Achievement Test. This difference was accounted for by the true/false format and essay test item formats when compared with other formats. The finding supported the claim of Kyoii and Paydarnia, (2011) in their findings which claim that test-takers perform differently across different response types which indicated that different test formats produce different results. Also, numerous research studies specifically in language testing have demonstrated that the methods which are used to measure the language ability influence performance on the related tests (Bachman, 1990).

CONCLUSION

This study concluded that the Economics tests constructed by the Economics teachers as used in their classroom assessment have construct validity and that test item format have an effect on the construct validity of the Economics Achievement Test in Osun state secondary schools. Finally, there is a significant difference in the performances of test-takers across various test formats used in the Economics Achievement Test.

REFERENCES

Adebowale, O. F. (2007). A study of differential item functioning (dif) in physics examinations in selected secondary schools in lagos state. An Unpublished Master Thesis, Faculty of Education, Obafemi Awolowo University, Ile-Ife, Nigeria.

Bachman, L.F. (1990). Fundamental considerations in language testing. Oxford: Oxford University Press.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. Memory & Cognition, 36, 604–616.

Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. Memory & Cognition, 38, 407–418.

Gergely. D. (2007). Investigating the Performance of Alternative Types of Grammar Items [Electronic version], Language Testing, 24: 1, 65-97

Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. Psychonomic Bulletin & Review, 14, 194–199.

Marsh, E. J., Agarwal, P. K., & Roediger, H. L. (2009). Memorial consequences of answering SAT II questions. Journal of Experimental Psychology: Applied, 15, 1–11.

Messick, S. (1993), Validity, in R. L. Linn, (1993). Educational Measurement, American Council on Education, Oryx Press, New York.

Mousavi, A. (2009). An Encyclopedic Dictionary of Language Testing. (4th Ed). Tehran: Rahnama, Press, I. R. Iran.

Roediger, H. L., & Marsh, E. J. (2005). The positive and negative consequences of multiple choice testing. Journal of Experimental Psychology: Learning, Memory and Cognitive, 31, 1155–1159.

Weir, C. J. (1990). Communicative Language Testing. Englewood Cliffs, NJ. : Prentice Hall.