West African Journal of Education
Vol. XL 2020 ISSN: 0043-2997
Website: http//www.instituteofeducui.com

Assessing Teacher Competence in Items Development through Evidence of Convergent Validity of Test Scores from Alternate Examinations

ADEYEMO Emily Oluseyi
Department of Educational Foundations and Counselling,
Faculty of Education, Obafemi Awolowo University, Ile-Ife.
seyiadeyemo2007@yahoo.com ; +2348033907463

Submitted: January 19, 2021; Revised: February 25, 2021; Accepted: April 8, 2021

Abstract

This examined the extent to which two different measures of examination items prepared for two consecutive years measured a common construct as evidence of teachers' competence in item development. The study compared item parameter estimates of the 2016 and 2017 BECE items and established the convergent validity with a few of providing information on teachers' competencies in item development. The research design was ex-post factor while the population consisted of 34,899 and 35,859 students who sat for the Osun State Basic Education Certificate Examination (BECE) in 2016 and 2017 respectively. Comparability of these items was made in terms of their item parameters, reliability estimates and evidence of convergent validity. Data were analyzed using test score statistics, transformed score and correlation coefficient .The results showed that reliability estimate of the BECE Mathematics items revealed Alpha Coefficient $\alpha = 0.09$ for 2016 and 0.53 for 2017. The coefficient of Correlation of the test score was low, $r = 0.191, p < 0.05$. The implication was that the items had different reliabilities estimates and the convergent validity of the tests was low. Students who sat for the 2016 examinations stood at advantageous than 2017. The items of the two examinations did not capture the same construct, indication of the facts that more training was required for teachers in item development.

Key words :- Teachers competence, convergent validity, test development, test items.

Introduction

Comparison among different test items is usually conducted by test developers, educational researchers and psychometricians. Equating, calibrations are different terms used by educational researchers to describe the comparisons. Equating is a process of deriving a function of a test to the scale of the anchor from such that after equating, any given scale would have the same meaning regardless of which test form was administered. Equating is conducted to establish score of different versions of a

test, allowing them to be used interchangeably in the process of assessment. Teachers' competence is specified by standard for educational assessment of students as adopted by UNESCO. This is a development model about educational factor that aims at identifying the broad competence of the teachers in the teaching and learning process across grade levels. It also includes content areas showing the aspect of each ability as typically developed from beginning to advance performance in teaching (UNESCO 1990). Standard are specific expectations for assessing knowledge or skills which is required of teachers to possess in order to perform well in their evaluation efforts. The standard required of teachers is to demonstrate skills at selecting, applying and evaluating students' assessment information and assessment practices.

Teachers' quality matters a lot in developing items. Assessment, specifically the area of test construction forms a critical part of the teaching and learning. This area of teachers' responsibility has opined by Frank and Amoako (2018) has been questioned by several authorities in contemporary times. Quality teachers are those who exhibit desirable traits that bring about students learning. Greenstein (2010) mentioned that the most important factor that influence students' achievement is teachers' quality. Teachers' quality includes teacher teaching experience, the extent of their preparation in subject matter, qualification in area of expertise and their ongoing professional development. If teachers possess low level of knowledge in assessment, they may not be able to help improve students' learning, however, teachers' qualification matter a lot in development of test since professional teachers should have a clear understanding of ways of developing credible instruments for assessment and also practice it appropriately.

The importance of assessment cannot be overemphasized. It is used for the improvement of teaching and learning and improvement of educational system in general since education is useful for everybody including policy makers and educators. Also, it cannot be denied that school-based assessment is an approach to re-design assessment in schools as it gives all educational stakeholders the power to improve teaching and learning practices. Assessment also gives immediate feedback to educational stakeholders and curriculum planner on academic progress of the students. Despites the relevance, studies like Onuka and Ogbebor (2013), Frank (2018), have shown that most teachers do not have necessary skill for test development and they used substandard and un-validated instruments to assess students, alongside, some teachers do not plan scoring procedure before teaching neither do they record scores properly during assessment all these could affect the credibility of the assessment procedures.

Developing assessment competency refers to ways or procedures that teachers use in improving themselves in assessment knowledge and skill so as to make education better. Good educational assessment can be done by teachers who have enough knowledge and skills in assessment. Among the skills as mentioned by Anhere (2009), Frank (2018), Onuka and Ogbebor (2013) were ability to choose appropriate assessment method, developing appropriate assessment method, administering tests,

interpreting results and using assessment results to make decisions. Teachers should be skilled in using assessment results when making decisions about individual students, planning teaching, developing curriculum and social improvement. Also they should be able to communicate assessment results to students, parents and other educators and recognize unethical, illegal and otherwise inappropriate assessment methods and uses of assessment information. There are various ways to address these issues and make teachers competent in developing quality items for assessing students. One of them is to allow teachers to attend professional development workshops and other training in area of assessment which allows entire academic program to engage in best practices of assessment. Through this approach, teachers gain experience from expertise in area of assessment. Alongside, educational system could provide a blueprint for all teachers to follow in teaching, so that there will be no issue that a teacher is bias in assessment from one school to another and it will definitely bring uniformity among teachers in terms of assessment practices.

In Nigeria, generally, teachers are trained in assessment of which test construction is an important component especially in the curriculum of Colleges of Education. For instance, students are taken through a full course of educational assessment. The course contents allow these students to have practical knowledge of test construction and assessment. Similarly, Nigerian universities that trained teachers also offer courses in education for assessment of potential teachers needed to be trained in assessment, this course also enlightens students in the construction of test items. Ololube (2008) has been able to find out that professional qualified teachers construct effective evaluative instruments more than the non-professional teachers and that they also keep assessment records accurately by employing various evaluative techniques, the attributes that are not common among the non-professionals. Alongside, though on the contrary, Anhwere, (2009), Amedahe, (1989) and Ebinye, (2001) confirmed that even among teachers that were trained in school assessment which include test construction, most of them do not adhere to the rules governing these practices, hence these has contributed to poorly drafted items.

The question that is always asked about developing a set of items for school-based assessment is whether the items is valid or not. The issue of validity spread across all facets of test in terms of content, construct and concurrent phases. Studies have revealed that relationship exist between teachers' knowledge of test items construction and the internal validity of the test. Teachers who have knowledge of test construction are supposed to prepare qualitative items than those who do not have. A teacher level of competence according to Darlin-Hammed (2000) is one of the factors that directly affect the quality of his/her test items.

Convergent and discriminant validity are components of a larger scientific measurement known as construct validity. Convergent validity is established when the score obtained from two different instruments measuring the same concept are highly

correlated. A convergent coefficient measures the correlation between an assessment and other measure of the same construct. According to Crocker and Algina (2006) convergent validity reflects the extent to which two measures capture a common construct. Alternate measure that provides less than perfect convergent validity introduces ambiguities that interfere with the development of meaningful interpretations of findings within and across studies. Convergent validity does not address the construct validity of any measure directly, instead it reflects the extent to which two measures capture the same information. The more similar the information they capture, the more likely they are to produce equivalent research results. Although construct is not equivalent to instructional validity but it is a form of evidence used to judge the construct validity of a measure.

The evidence of convergent validity is commonly assessed using the magnitude of the zero –order correlation between the proxy and another closely related measure. If two measures are hypothesized to represent the same construct, a strong correlation between these measures suggest that the measure capture the construct and correlation close to one (r = 1) which indicates strong convergent validity. However, strong convergent validity is a necessary condition for construct validity but it is not sufficient. When convergent validity is weak, it implied that one or more variable(s) do not capture the intended construct well. Consequently, evidence of weak convergent validity will introduce ambiguity into the meaning of research results and hence diverge. Divergent validity on the other hand referred to as discriminant validity is established when two variables are predicted to be uncorrelated and the scores obtained by measuring them are empirically found to be different. Also, a discriminate validity coefficient measures the correlation between an assessment and a measure of a different construct. This is expected to be low according to Crocker & Alginia (2006).

Validity and reliability are important qualities that any test must have. A test must be reliable if it will be interpretable and must be valid if it must be used at all. Invalid test or instrument would expectedly yield invalid results. Validity emphasized contents and coverage and it is seen to be very fundamental. Reliability on the other hand is extremely important. High reliability is an indication of a good instrument and it is practically meaningful if it is based on the use of valid instrument. If a test is valid but not reliable, it is suitable for exercise of test construction only. If on the other hand a test is reliable but not valid, it can be likened to doing a wrong thing consistently and accurately.

Several researchers (Ashichia 2010), Adebule (2004) have identified various factors that affect students' performance. Prominent among these factors are the nature of the test items and learners' characteristics. Experiences and observations have revealed that performance of students is never the same across subjects. Variation in students' performance may be attributed to various factors such as the contents, quality of the testing procedure and learners' characteristics such as ability of students. Anyone or combination of any of these could affect students' performance.

The purpose of this study therefore was to explore the convergent validity of the Basic Education Certificate Examination (BECE) across two consecutive years given the speculation that identical items with alternative use have equivalent psychometric properties which in some way should render the two set of items test administered to two different set of examinees equivalent: an attribute expected from professional competence teachers. Thus the study considered the comparative analysis of 2016 and 2017 BECE Mathematics items which necessitates the psychometric analysis in order to establish the convergent validity of the items and to examine the extent to which the two measures of BECE items prepare for the two consecutive years capture a common construct as evidence of teacher's competency in test development

Methodology

The ex-post facto research design was adopted for the study. The population of the study comprised 34,889 and 35,857 public school candidates who sat for the 2016 and 2017 BECE Mathematics items for 2016 and 2017. The sample for the study comprised of 60 multiple –choice items for 2016 and 2017 each, making a total of 120 items and 1000 response of candidate as contained in the OMR sheets for the 2016 and 2017 respectively. These were selected using systematic random sampling technique, making a total of 2000 responses. The research instruments were the candidate OMR sheets for 2016 and 2017 Mathematics items. Data were analyzed using both correlation coefficient and t-test statistics. descriptive and inferential statistics.

Research Question 1: What are the differences in the item parameters of the 2016 and 2017 BECE Mathematics test items?

Table I: Descriptive of 2016 and 2017 BECE Mathematics Test Items Parameters

	BECE 2016		BECE 2017	
	Mdisc	**Mdiff**	**Mdisc**	**Mdiff**
$\bar{X}$	5.15	1.97	11.40	0.30
SD	7.89	1.44	22.26	1.13

Table I showed that 2016 BECE Mathematics test item parameters. (mean = 1.97, SD =1.44) were more difficult than the 2017 (mean = 0.30, SD =1.13) and that 2016 items discriminated better (mean = 5.15, SD =7.89) than 2017 (mean = 11.40, SD = 22.26). The items parameters of 2016 were better than the 2017, this implied that the test were of different version.

Ho_1: There is no significant difference in the psychometric properties of 2016 and 2017 Osun State BECE Mathematics test items.

To test the hypothesis, the reliability estimates of 2016 and 2017 OSBECE Mathematics tests were obtained and the estimated reliability coefficient were thereafter compared. The reliability estimates were calculated using Kuder Richardson K-R20 and the estimates were compared using dependent alpha. This statistic was implemented in cocron package of the R language and environment for statistical computing. The result was as presented on Table 2.

Table 2: Comparability of Reliability Estimate of 2016/2017 Osun State BECE Mathematics Items

Year	Reliability coefficien t K-R20	df	t	p-value
2016	0.90	998	8.6393	0.000
2017	0.83			

p < 0.05 significant

Table 2 showed that 2016 OSBECE Mathematics multiple test was more reliable (α = 0.90) than 2017 (α = 0.83). The difference observed in the reliability estimate was significant with t = 8.6393, p < 0.05. Therefore, the hypothesis was rejected. The result showed that reliability of 2016 BECE Mathematics items was different from that of 2017. The implication of this is that 2016 and 2017 BECE mathematics items had different reliabilities.

Ho_2: The convergent validity of the 2016 and 2017 BECE Mathematics items are not significantly different.

To test this hypothesis, three levels of analyses were conducted (i) the ability of the examinees in the test were estimated, (ii) the estimated ability in the two tests were respectively converted to the true score T, (iii) and the converted score on the two tests were correlated. The correlation coefficient obtained represent the convergent validity of the test. The result is as presented in Table 3

Table 3: Correlation Coefficient of Test Score of Examinees in 2016 and 2017 Osun Sate BECE Mathematics Items

Year	N	Mean	Sd	r	p
2016	1000	19.7562	10.276	0.191	0.000
2017	1000	32.5469	11.191		

p < 0.05

The coefficient of correlation showed that the relationship which existed between the test scores of 2016 and 2017 Mathematics items was low and significant r = 0.191, p< 0.05., Therefore the hypothesis was rejected. This implied that quantum of traits measured by the two tests are similar but low. The implication is that the convergent validity of the test was low.

Discussion

The high parallel form of reliability coefficient together with the significant difference obtained in the reliability coefficients of the test scores of the Mathematics items indicate that the different forms of the test are not similar but virtually different, the implication of this is that it made large difference which version of the test a person takes, a student for a particular year (2016) stands advantageous than the other year (2017) This shortcoming arose as a result of inability to prepare different set of items with the same psychometric properties .This could be lack of skills from test development according to Onuka and Ogbebor (2013) or lack of professionalism as opined by Anhere (2009). On the other hand, a low parallel form reliability estimate obtained suggested that the different forms were probably not comparable, they might be measuring different things and hence cannot be used interchangeably.

This confirmed the opinion of Ololube (2008) that some teachers have limited skills in the construction of end of term examinations as this was evident when issues were found with the content representativeness and relevance of the test, reliability and fairness of assessment task which were evaluated. He also found out that teachers who had knowledge of test constriction prepared qualitative items than those who do not have. Also supporting this view is the opinion of Darlin- Hammed (2000) that teachers' level of competence was one of the factors that directly affected the quality of items. It should also be borne in mind that good educational assessment can only be done by teachers with adequate knowledge in skills of assessment and ready to adhere to the rules that govern assessment practices.

Without prejudice, these could only be recorded from dedicated and committed teachers, ready to plan the scoring procedures before teaching and record scores properly and on time. Convergent validity does not address the construct of any measure directly, instead it reflects the extent to which two measures capture the same information. The weak convergent validity as presented by the weak correlation between the items resulted in increased divergence of the magnitude of the correlation between the two measured. This presumed that as the divergent validity continue to fall, the magnitude of the frequency of divergent increases and by extension, the risk of inappropriate interpretations of results. When the convergent validity is weak the implication of this is that the variables under consideration do not capture the intended construct very well.

Conclusion/recommendations

Teachers' quality matters a lot in development of school-based assessment. Quality of learning is determined by quality of assessment practices in the classroom. Since assessment is viewed as integral part of students learning, and helps to improve learning by the students, teachers should therefore adhere to develop a range of assessment that is suitable and of quality standard for instructional planning. Alongside, government and school administrators should take up the challenges of inviting resource persons from recognize academic institutions to organize workshops for teachers on a regular basis to sharpen their skills on effective test construction practices.

References

Adebule, S. O. (2004). Gender difference on locally standardized anxiety rating scale in Mathematics for Nigerian Secondary Schools. *Journal of Counselling and Applied Psychology 1, 22-29.*

Amedahe, F. K. (1989). Testing Practices in Secondary School in the Central Region of Ghana. *Unpublished thesis*

Anhwere, Y. M. (2009). Assessment Practices of Teacher Training College Tutors in Ghana, *Unpublished Masters' Thesis*.

Ashikia, O. A. (2010). Students and Teachers' perception of the cause of poor academic performance in Ogun State Schools 9Nigeria). *European Journal of Social Science, 13(2) 229-242.*

Carlson, K. D. & Herdman, A. O. (2012). Understanding the impact of Convergent Validity on Research Results. *Organizational Research Methods 15(1) 17-21*. Sage Publication.

Crocker, L. & Algina, J. (2006). Introduction to Classical and Modern Test theory. *Mason Ohio. Thompson Wentworth.*

Doran, W. J. Moses, T. P. & Eignor, D. E. (2010). Principles and Practice of Score Equating. *Educational Testing Service* (ETS) Princeton. New Jersey

Duong, M. (2004). Introduction to Item Response Theory and its applications, Research Development.ttps;www.msu.edu/dwong/student work. *Archive/CEp900F04.RDP/mihn- Item Response Theory.htm.*

Ebinye P. O. (2001). Problems of testing under the Continuous Assessment Programme *Quarterly Journal of Education*. 4(1):12-19.

Frank, Q, I. Amoako, F A. (2018) Teacher's Test Construction Skills in Senior High Schools in Ghana. *International Journal of Assessment tools in Education. Doi.21449/ijate 481164*

Frank Q. I. (2018). Attitudes of Senior High School Teachers Towards Test Construction. ISSN 22245766 (paper) ISSN 2225-0484 *(online) vol. 8 No 1.*

Greenstein, L. (2010) What Teachers Need to Know about Formative Assessment. Alexandria , VA:ASCD.

Ogbebor. U. C. (2015) Mock Economics test construction using classical test and item response theories in Delta state Nigeria. Unpublished PhD proposal at the International Centre for Educational Evaluation, Ibadan, Nigeria.

Ololube N. P. (2008) Evaluation Competencies of Professional and Non professional Teachers in Nigeria. *Studies in Educational Evaluation, vol. 34 (1)*

Onuka, A. & Ogbebor, U. (2013) Differential Item Functioning Method as an Item bias indicator. *Research International Journal*, 4(4),367-373.`

UNESCO (1990) World Conference on Education for All:- Meeting Basic Learning Needs. *United Nation Development Programme, Jomtien International Agency Commission*.