EQUATING OF THE SENIOR SECONDARY SCHOOL WASSCE ECONOMICS MULTIPLE CHOICE TEST PAPERS (2011 AND 2015)

Tomilola R. AWOMOKUN
Department of Social Sciences Education,
University of Ilorin, Ilorin.
awomokuntomilola@gmail.com

&

Henry O. OWOLABI
Department of Adult and Primary Education
University of Ilorin, Ilorin

Abstract

Proportion of candidates passing at each grade of Senior School Certificate Examination (SSCE) in Economics varies from year to year. It is important that test scores obtained by candidates with equal abilities be comparable from one test form used in one year to another. This has necessitated a study of the comparability of test scores from different test forms over time. This study therefore focused on equating the Senior Secondary School WASSCE Economics Multiple Choice Test papers for 2011 and 2015. Specifically, this study investigated obtained empirical evidence of the (i) ability level of students as reflected by their scores in the 2011 and the 2015 multiple choice tests and (ii) horizontal equating of 2011 and 2015 WASSCE Economics multiple-choice test items. The Non-Equivalent Anchor Test Design of test score equating was used for this study. The sample for this study consisted of 650 students in SS2 selected from 14 senior secondary schools in Kwara Central Senatorial District through the use of multi-stage sampling technique. The 2011 and 2015 Multiple Choice Economics Test papers were adopted and used as instruments for data collection in this study. The data collected were analysed with the use of mean-mean and mean-sigma statistical tools. The findings of this study revealed that (i) the mean ability of sampled students in the 2011 and 2015 WASSCE Economics MCTs were of 0.12 and 0.13 respectively and (ii) a score of 4 in 2011 was found to be equivalent to 2.33 in 2015, a score of 10 in 2011 was equivalent to 7.54 in 2015, while 23 in 2011 was equivalent to a score of 22.63 for the MS in 2015. Based on the findings from this study, it was recommended that the exam body should continue to ensure that their scores are comparable in terms of item parameters while students and teachers of Economics should ensure that students are moved from medium to high ability levels.

Keywords: Ability estimates, Test, Equating, Horizontal Equating, SSCE

Introduction

Assessment is very useful for monitoring and evaluating the educational system and it serves as motivation for the teaching and learning processes. Test, as an assessment technique, is a means of assessing and getting feedback from students thereby making some inference about the attributes of the examinees. It is used for obtaining information from students as well as to describe the behaviour in terms of scores or categories.

Owolabi (2004) defined a test as a sample of behaviour drawn to ascertain the presence or absence of certain traits, characteristics or skills and the extent to which these are present or absent in an individual or a group of persons. Tests are administered to students to determine the extent to which they have attained previously identified objectives in a learning situation. Therefore, testing is a fundamental part of the teaching-learning process used not only as a basis for ranking student at the end of the teaching/learning process but to guide teaching, aid curriculum development, determine needs, learning difficulties and as well promote mastery and eliminate differences among learners.

Abiri (2007), based on purpose and use classified tests as teacher-made and standardized. A teacher-made test is one which is internally prepared, administered and scored by classroom teachers for a student or a group of students in a particular class. A standardized test on the other hand is any form of test that requires all test takers to answer the same questions, or a selection of questions from common banked items, in the same way, and that is scored in a "standard" or consistent manner, which makes it possible to compare the relative performance of examinees or groups of students. Examples of standardized tests are Unified Tertiary Matriculation Examination (UTME), West African Senior School Certificate Examinations (WASSCE) and Senior School Certificate Examinations (SSCE) conducted by JAMB, WAEC and NECO respectively. Examples of different types of test that are used in education include achievement test, aptitude test, intelligence test, psychological test and personality test.

Standardized tests are widely used in educational institutions for determining achievement, evaluating and comparing tests takers' abilities in a specific content area. They are administered and scored in a predetermined, set manner that is consistent for all test takers (Xuan & Rochellan 2011). According to Xuan and Rochellan (2011), in order for standardized tests to have consistency in score interpretation when there are different forms of a test, test scores are often transformed into a set of values different from the raw score points obtained directly from the test. However, these transformed test scores (called scaled scores) are usually reported to have consistent meaning to all test takers.

Most standardized tests require scores that can be compared across different forms. In order for different stakeholders to make consistent and fair decisions for assessment results, the score reported from standardized tests must be comparable. This implies that they must carry the same meaning regardless of which form was administered. It can be simply said that different forms of the test should indicate the same level of performance no matter which form the test

taker attempted (Xuan & Rochellan, 2011). This applies to tests conducted by public examination bodies conducted yearly for candidates which must assess at the same level.

The primary purpose for conducting a test under standardized conditions is to provide a means of measuring or evaluating a group of examinees' skills that is as fair and objective as possible (Linda, Cook & Eignor 1991). The scores obtained from tests provides primarily two types of information. One is the degree to which a student has attained criterion performance that is, determining whether the student can solve a certain number of problems. The other is the relative ordering of individuals with respect to their test performance within group. However, raw scores need to be accurate and fair because they provide the basic information for all other types of transformed scores (Grolund, 1998).

The West African Examinations Council (WAEC), National Examinations Council (NECO) and National Business and Technical Examination Board (NABTEB) are bodies that conduct standardized tests for Nigerian secondary school students. These tests, according to Abiri (2007), are to possess acceptable psychometric properties before they could be judged worthwhile. It is assumed that they have been carefully prepared and standardized using large samples which help to ascertain their validity and reliability for use. Psychometrics is primarily concerned with the construction and validation of measuring instruments such as test, questionnaires and personality inventories. The psychometric properties of multiple-choice items include distractor effectiveness, difficulty and discrimination indices. According to Schumacher (2005), classical test theory (CTT) utilizes traditional item difficulty and discrimination estimates, distractor analysis, test item inter correlations and a variety of related statistics.

Olatunji (2007) described the difficulty of an item as the extent to which it has been answered correctly by the testees. That is, the percentage of the testee that chose the right option. The closer to 1 the value of the difficulty index the easier the item and the closer the value to 0, the more difficult the item. Discrimination power of multiple-choice items, on the other hand, is the ability of the test items to discriminate between the brilliant students and the poor students. Discrimination power of a test item ranges from zero to one (0-1). The closer the value is to one (1), the better the item. The effectiveness of a Multiple-choice test is judged by how plausible the distractor is if five percent of the testees selected it. Baker (2001) also defined the difficulty of an item as a location index. The difficulty of an item describes where the item functions along the ability scale. For example, an easy item functions among the low-ability examinees and a hard item functions among the high-ability examinees.

Since test scores serve as the major yardsticks for determining students' achievement, it is critical for them to be comparable from test to test and year to year. Scores on tests are often used as one piece of information in making important decisions. Any type of decision that is to be made should be based on the most accurate information. Making decisions in many of these contexts require that tests be administered more than once. A test form is a set of test questions that is built according to content and statistical test specifications (Millman & Greene, 1989). The use of different test forms leads to the concern that the forms might differ somewhat in terms of their difficulty levels. This challenge is usually addressed through test equating.

Equating is a statistical process that is applied to test forms so that scores on them can be used interchangeably. Equating also adjusts for differences in difficulty among forms that are built to be similar in difficulty and content. When test forms are created to be similar, equating is a process of making scores across different forms of the same test interchangeable. As a result, test forms are considered to be essentially the same, so scores on the two forms can be used interchangeably after equating has adjusted for differences in difficulty.

The need for test equating arises when there are two or more tests of the same construct or subject that can yield different scores for the same examinee. The goal of test equating is to allow the scores on different forms of the same tests to be used and interpreted interchangeably. Test equating, as it is currently practiced, requires some type of control for differential examinee ability, or proficiency in the assessment of and adjustment for differential test difficulty. Equating is thus the procedure for establishing comparable scores from different test forms. It is used to provide solutions to the challenge of maintaining comparability of scores from different forms of a test and/or from different test administrations (Doran, Moses & Eignor, 2010).

Peterson, Cook and Stocking (1989) defined equating as the process of ensuring that scores resulting from the administration of the multiple forms can be used interchangeably. Angoff (1971) has defined the equating of tests as a process of converting the system of units of one form to the system of units of the other so that the scores obtained from one form could be compared directly with the scores obtained from the other form. In many test administration situations, more than one form of a test is used for security reasons **(**Jaeger, 1980; Jaeger, 1981; Kolen & Whitney, 1982; Cook & Eignor, 1991; Kolen, 1994).

When reporting evidence of change in performance between two forms of a test, it is important that a distinction be made between differences in difficulty of the test forms used to assess the students and real gains or losses in achievement between the two years. The purpose of equating is to adjust for test difficulty differences so that only real differences in performance are reported (Xuan & Rochellen 2011). There are two ways by which equating could be done: Vertical (also referred to as across-grade), and Horizontal (also referred to as within-grade) equating (Xuan & Rochellen 2011).

The use of different tests aimed at measuring the same constructs from year to year by such different examination bodies in Nigeria as the Joint Admissions and Matriculations Board (JAMB), West African Examinations Council (WAEC), National Examinations Council (NECO) and National Business and Technical Examination Board (NABTEB) which conduct different examinations for candidates raises the issue of comparability of scores across years or across the bodies.

A common feature of the results being released by these examination bodies is non-stability of the percentages of failure and passes each year. Reports have shown that there has been a continuous decline in the performance of candidates sitting for SSCE examinations (Adedoyin, 2010). Eguridu (2015) noted that in 2012, 2013 and 2014, 38.31%, 36.57% and 31.28% of the respondents had credits in WASSEC. Adeyemi (1990) gave difficulty level of test items as one of the several reasons for the decline in students' performance This is one of the motivating factors for carrying out this present study. Olatunji (2007) studied the difficulty of

economics items used by WAEC and NECO and found no significant difference existing among them. Also, Olatunji (2015) carried out analysis of linear and equipercentile equating of Senior School Certificate Economics Multiple-choice Papers. It was found that the linear equating method was preferable to the equipercentile equating method. Series of studies have been carried out on comparability of test scores from different test forms, but WASSCE Economics Multiple Choice test papers for 2011 and 2015 were yet to be equated. Hence, the purpose of this study is to equate the Senior Secondary School WASSCE Economics Multiple-choice Examination Papers (2011 & 2015).

Research Questions

1. What are the ability levels of students in 2011 and 2015 WAEC multiple choice Economics items?
2. What is the result of horizontal equating of 2011 and 2015 WAEC Economics multiple-choice Examination?

Methodology

The Non-Equivalent Anchor Test design was used for this study. This design is also called the common-items nonequivalent groups design or simply the common items or the anchor test design. The design is used when equating of two test forms requires linking items (Holland & Dorans, 2006). The Non-Equivalent Anchor Test (NEAT) Design according to Kolen and Brennan (2004) is such that two equivalent samples are taken from a common population P, one is tested with Test Form **A** and the other with Test Form **B** and the two groups take an Anchor test. The anchor test measures and quantifies the ability differences between two distinct, but not necessarily equivalent samples of examinees. In the Non-Equivalent Anchor Test (NEAT) design, the procedure for gathering data is carried out and achieved through achievement test (Kolen & Brennan 2004).

The population under study consisted of all Secondary School Students in Kwara Central Senatorial District and the target population comprised all SS3 Students in the District. A sample of 650 SS3 students was selected from the total of 7689 SS3 students through the multi-stage sampling technique.

At the first stage, proportionate sampling technique was used to 14 select schools from the four local government areas in Kwara Central Senatorial District. Ten percent of the schools in each Local Government Area participated in the study (Kwara State Ministry of Education, 2015). Kwara Central Senatorial District has 95 public Senior Secondary Schools. Asa Local Government Area has 21 public senior secondary Schools, Ilorin East Local Government Area has 26 public secondary schools, Ilorin South Local Government has 22 public senior secondary schools and Ilorin West Local Government has 26 public senior secondary schools.

At the second stage, purposive sampling technique was used to select SS3 students from the selected schools. Senior Secondary three students were chosen because of their level of preparedness for their final examination and also because it is assumed that they have covered a wide range of the Economics syllabus. At the third stage, proportionate sampling technique was

used to select students who offer Economics from the Art, Commercial and Science classes in each school.

The 2011 and 2015 May/June WASSCE Multiple-choice Economics items were adopted in their original forms for this study. These test forms were administered by the researcher and with the help of Research Assistants to the students from the selected schools on the dates approved by the Principals.

The test forms were scored dichotomously and analysed. The ability index of each respondent to the test items and the item parameter (item difficulty) were generated with the use of Wingen IRT software package and analysed to provide answers to research question one. Also, the mean-mean and mean-sigma statistical tools of IRTEQ IRT software package was used to carry out horizontal equating of the two tests to answer research question two.

Results

Research Question One: *What are the ability levels of students in 2011 and 2015 WASSCE Economics Multiple-Choice items?*

Under Item Response Theory (IRT), ability levels range from -3 to +3. The estimated examinees' abilities which were generated are summarised in Table 1.

Table 1: Examinees' Ability (Theta) in 2011 and 2015 Economics Multiple-Choice Items

WASSCE	N	Mean	S.D.	Minimum	Maximum
2011	325	0.12	1.08	-3.04	2.86
2015	325	0.13	0.97	-2.89	2.93

Table 1 shows the estimated abilities of the respondents to the test items. The mean ability (theta) of the examinees in WASSCE 2011 is 0.121 with a range of -3.04 to 2.86 and standard deviation (SD) of 1.083 while the mean ability (theta) of the examinees in WASSCE 2015 is 0.128 with a range of -2.89 to 2.93 and SD of 0.970. Baker (2001) explained that ability ranges from -3 to +3 and that the closer it is to 0, the more moderate the ability which is also an indication of a normal distribution. This shows that there is no difference in the ability of the SS3 examinees who responded to the test items of 2011 and 2015 WASSCE and that they are also of medium ability.

Research Question Two: *What is the result of horizontal equating of 2011 and 2015 WAEC Economics Multiple-Choice Examination?*

The results of horizontal equating of 2011 and 2015 WAEC Economics Multiple-Choice Examinations were generated through mean-sigma in IRTEQ Equating Software and presented in Table 4. The mean-sigma equating parameter uses the mean and standard deviation of each score to carry out the equating process. The Mean-sigma equating parameter was generated for each corresponding score.

Table 2: Summary of Horizontal Equating of 2011 and 2015 Economics Examination

Scores Test 1 (2011)	Mean Sigma (Test 2-2015 rescale
1	0.48
4	2.33
10	7.54
23	22.63
32	33.59
49	49.36
50	49.99

From Table 4, it is shown that a score of 4 in 2011 is equivalent to a score of 2.33 for the Mean-Sigma in 2015, This also means that an examinee who earns a score of 4 in 2011 is considered to be at the same achievement level as an examinee who earns a score of 2.33 (MS) in 2015. An examinee who earns a score of 10 in 2011 is at the same achievement level as an examinee who earns a score of 7.54 (MS) in 2015. Also, a score of 23 in 2011 is equivalent to a score of 22.63 for the MS in 2015. The implication of this is that any student who did well in the first test form will also do well in the second test form. This is supported by the findings of Keller et al. (2004) who evaluated the ability of four equating methods (Concurrent Calibration (CC), Fixed Common Item Parameter (FCIP), Stocking and Lord Test Characteristic Curve (TCC) and Mean/Sigma (M/S)) to recover changes in the examinee ability distribution using simulated data based on a standard normal-ability distribution and found that M/S performed the best while FCIP performed the least.

Discussion of Findings

The findings of this study revealed that the SS3 examinees who took the 2011 and 2015 WASSCE Economics tests were of equivalent ability (theta). The examinees who took WASSCE 2011 had mean ability (theta) of 0.121 and SD of 1.083 and the examinees who took WASSCE 2015 had mean ability (theta) of 0.128 and SD of 0.970. Baker (2001) explained that ability ranges from -3 to +3 and that the closer it is to 0, the more moderate the ability which is also an indication of a normal distribution. This shows that there is no difference in the ability of the SS3 examinees who responded to the test items of 2011 and 2015 WASSCE and that they are also of medium ability.

This result supports Pido (2012) who found students to be of medium ability level on the average. This implies that SS3 students were of average ability level because majority of the students are usually average students as stated by Pido (2001) and Amidu (2015)

The result of horizontal equating of 2011 and 2015 WASSCE Economics multiple-choice Examinations generated through IRTEQ Equating Software revealed that a score of 4 in 2011 was found to be equivalent to a score of 2.33 for the MS in 2015, a score of 10 in 2011 is equivalent to a score of 7.54 for the MS in 2015, also, a score of 23 in 2011 is equivalent to a score of 22.63 for the MS in 2015. The implication of this is that any student who did well in the first test form will also do well in the second test form. This means that the scores on the two test forms can be used interchangeably given the Mean-Sigma equating procedure.

This is because the Mean-Sigma (MS) equating procedure uses the mean and standard deviations (SD) of the b-parameter (difficulty) to transform the scale of form one to the scale of form two. It was also observed that the Mean Sigma (MS) equating procedure. This is supported by the findings of Keller et al. (2004) who evaluated the ability of four equating methods (CC, M/S, TCC and FCIP) to recover changes in the examinee ability distribution using simulated data based on a standard normal-ability distribution and found that M/S performed the best while FCIP performed the worst. Hu et al. (2008) also added in a simulation study to investigate ten variations of four equating methods (CC, M/S, TCC and FCIP) in the absence and presence of outliers in the set of common items. They concluded that "TCC and M/S transformations performed the best. This means that, using the Mean-Sigma equating procedure, the possibility of comparing and using students' scores on two test forms interchangeably can be achieved.

Conclusion

Based on the findings from this study, it can be concluded that the examinees who took the WAEC 2011 and WAEC 2015 Economics test items were of the same ability and that they are of medium ability (the mean ability (theta) of the examinees in WASSCE 2011 is 0.121 with a range of -3.04 to 2.86 and standard deviation (SD) of 1.083 while the mean ability (theta) of the examinees in WASSCE 2015 is 0.128 with a range of -2.89 to 2.93 and SD of 0.970.).

The result of horizontal equating of 2011 and 2015 WAEC Economics multiple-choice Examinations generated through IRTEQ Equating Software revealed that the scores on the two test forms can be used interchangeably given the Mean-Sigma equating procedure.

Recommendations

Based on the findings and conclusions in this study, the following recommendations are made:

i. WAEC Economics papers are comparable across the years. This is not only an indication of comparable validity indices but also that of their reliabilities. Therefore, the exam body should continue to ensure that their scores are comparable in terms of item parameters.
ii. There is need to move students from medium to high ability. Students and teachers of Economics should ensure this.

References

Abiri, J. O. O. (2007). *Elements of evaluation measurement and statistical techniques in education.* Unilorin Press: University of Ilorin, Nigeria.

Adedoyin, O. O. (2010). Using IRT approach to Detect Gender Biased Items in Public Examinations. *Educational Research and Reviews Academic Journal,* 5(7), 385-399

Angoff, W. H. (1971). *Scales, norms and equivalent scores.* In R. L. Thorndike (Ed.), Educational measurement (2nd ed.). Washington, DC: American Council on Education. Belmont, CA: Wadsworth Group, 508–600.

Cook, L. L. & Eignor, D. R. (1991). An NCME instructional module on IRT equating methods. *Educational Measurement: Issues and Practice, 10,* 37-45.

Doran, N. J., Moses, T. P. & Eignor, D. R. (2010). *Principles and practices of test score equating.* New Jersey: ETC Princeton.

Grolund, N. E. (1998). *Assessment of student achievement* (7th ed.). Boston: Allyn and Bacon.

Jaegar, R. M. (1981). *Some exploratory indices for selection of a test equating method. Journal of Educational Measurement, 18,* 23-38.

Jaeger, R. M. (1980). Some exploratory indices for selection of a test equating method. *Journal of Educational Measurement, 18,* 23-38.

Kolen, M. J. & Brennan, R. L. (1995). *Test equating: Methods and practices*. NewYork: Springer Verlag.

Kolen, M. J. & Whitney, D. R. (1982). *Comparison of four procedures for equating the tests general educational development*. *Journal of Educational Measurement, 19*(4), 279–293.

Kolen, U. (1994). Conditional standard errors of measurement for scale scores using IRT. *Paper presented at the annual meeting of the National Council on Measurement in Education, New Orleans.*

Millman, J. & Greene, J. (1989). The specification and development of tests of achievement and ability. In R.L. Linn (Ed.), *Educational measurement* (3rd ed., pp.335-366). New York: American Council on Education & Macmillan.

Olatunji, D. (2015). Analysis of Linear and Equipercentile Equating of Nigerian Senior School Certificate Examination Economics Multiple Choice Papers in Kwara State, Nigeria. *Unpublished Ph.D. Thesis, University of Ilorin, Ilorin.*

Olatunji, D. S. (2007). *Effects of Number of Options on Psychometric Properties of Multiple Choice Tests in Economics. Unpublished M.Ed. Thesis, University of Ilorin, Ilorin.*

Owolabi, H .O. (2004). Assessment in the classroom. In E.O. Ogunsakin (Ed). *Teaching in Tertiary Institutions.* (pp.93-99). Faculty of Education, University of Ilorin, Ilorin.

Petersen, N.S. Cook, L.L. & Stocking, M.L. (1989). *IRT versus Conventional EquatingMethods: A Comparative Study of Scale Stability. Journal of Educational Statistics, 8.* 137-156.

Schumacker, R. E. (2005). *Test equating*. Retrieved March 21st 2009 from http://www.appliedmeasurementassociates.com/ white%20papers/TEST%EQUATING.Pdf.

Xuan, T. & Rochellen, M. (2011). *Why do standard testing programme report scaled score:* Retrieved from http://www.ets.org/understandingtesting/glossary/.