ISSN 2705-2559
E-ISSN 2705-2567

COMPARISON OF METHODS TO DETECT ITEM PARAMETER DRIFT IN NABTEB SSCE FOR 2012-2015 CHEMISTRY MULTIPLE CHOICE TESTS

BY
Dr. (Mrs.) Augustina Enogie Ighodaro: Department of Educational Evaluation and Counselling Psychology, Faculty of Education, University of Benin, Benin City, Edo State;
E-mail: usigbemwona@yahoo.com
&
Dr. (Mrs.) Chinelo Blessing Oribhabor: Department of Guidance and Counselling, Faculty of Arts and Education, University of Africa, Toru-Orua, Bayelsa State; E-mail: chinelo.oribhabor@uat.edu.ng; chiblessing42004@yahoo.co.uk

Abstract

This study investigated the comparison of methods to detect Item Parameter Drift in National Business and Technical Examination Board (NABTEB) Senior School Certificate Examinations (SSCE) for 2012 – 2015 Chemistry multiple choice tests. It specifically aimed to ascertain the levels of drifted items and compare the differences in the number of drifted items across the stated examinations years using four methods (Robust–z, 3-sigma IRT, Delta plot and Mantel–Haenszel methods) of detecting IPD. This study was guided by four research questions and one hypothesis. The study adopted the survey research design. The population of the study was 3,239 scores of candidates who sat for the National Business and Technical Examination Board, NABTEB, SSCE Chemistry multiple choice tests in the first administration and 8,605 Senior Secondary School SS3 Chemistry students considered for second administration, making a total population of 11,844 in Edo State, Nigeria. The sample size used for the study was 5,040 which comprised 3,239 and 1,801 candidates' scores in first and second administrations respectively. Multistage sampling technique was employed. The instruments used to generate data were 50-item National Business and Technical Examination Board, NABTEB, SSCE Chemistry multiple choice test items each for the four years (2012, 2013, 2014 and 2015) making a total of 200 items. Using Cronbach Alpha, the reliability coefficients of the instruments obtained were 0.83, 0.85, 0.89 and 0.87 for the respective years. However, the item parameters were estimated from candidates' responses using eIRT software. The four methods (Robust- z method, 3 -sigma IRT method, Delta plot method and Mantel -Haenszel methods) were used to establish the IPD, while descriptive statistics of frequency count and percentage were used to answer the research questions, and the hypothesis was tested using Chi-square statistics at 0.05 alpha level. The results that were obtained from the analysis showed that 20 items, 80 items, 65 items and 151 items drifted using Robust-z method, 3-sigma IRT method, Delta plot method and Mantel-Haenszel method respectively in 2012 to 2015 NABTEB SSCE Chemistry multiple choice tests, and it was also found that there is a significant difference in the number of drifted items in 2012 to 2015 NABTEB SSCE Chemistry multiple choice test items using Chi-square statistics. Furthermore, it was concluded that Robust-z method was the most stable method because it flagged the least number of drifted items of all the four methods. Among others, it was recommended that NABTEB and other examination bodies should use Robust -z method to detect drift items to avoid false identification of drifted items.

Keywords: Item Parameter Drift, Robust-z, 3-sigma IRT, Delta plot, and Mantel-Haenszel

Introduction

One of the ways in which educational objectives is achieved and the extent to which educational institutions serve the needs of the environment is through assessment. Assessment is a way of supporting learning; it helps teachers, learners, parents and others to understand the depth and breadth of learning. Assessment is a process of measuring learning outcomes and other proficiencies in order to make authentic and valid decisions about the individual. It is done with the intent to provide information that will enhance quality professional testing and measurement. Assessment is a more encompassing term than testing, it uses test as a means through which information is gathered to make decisions. Test as a tool for assessment is administered to measure the examinees' knowledge or proficiency. It is an instrument presented to examinees in order determine the extent to

ISSN 2705-2559
E-ISSN 2705-2567

which trait is present or absent in the respondent(s). There are many formats of test of which Multiple Choice Test (MCT) is one of them. MCT is considered one of the most enduring and successful forms of educational technology that remains in practice today. They are widely used in standardized assessment settings. Large scales exams conducted by national exam bodies because it is easy and quick to score also reduces some of the burden of large classes. Moreover, test items are characterized by certain parameters – item difficulty, discrimination, pseudo guessing and carelessness that are defined within two measurement theories; which are the Classical Test Theory (CTT) and the Item Response Theory (IRT).

The Classical Test Theory (CTT) assumes that observed score is made up of the true score plus error. Under this umbrella, item difficulty parameter is denoted by p-value and the discrimination parameter is represented by D, while the Item Response Theory (IRT) is based on the idea, that the probability of a correct response to an item is a function of the person and item parameters. In this case, the item parameters are item discrimination that is denoted by a- parameter, item difficulty represented by b or b- parameter, the pseudo guessing parameter designated by c- parameter and carelessness depicted by d- parameter, while the person's parameter is usually interpreted as a single latent trait which is not directly measurable and unobservable. However, these item parameters are stable over testing occasions and deviation from this premise results to item parameter drift. Item Parameter Drift (IPD) is a differential change of item parameters over time (Goldstein, 1983). Item Parameter Drift is a phenomenon that examines comparability of violated items across time or testing occasions (Orheruata, 2015). Item Parameter Drift represents an unanticipated change in item parameter values between testing forms or between testing occasions. Drift is likely to occur when an item pool is not maintained over time. Such effects may be expected because of frequent item exposure or over usage of items.

One of the most important properties of the Item Response Theory (IRT) is the invariance of model parameters. Invariance indicates that the item parameters estimate from different samples taken from the same population or from separate testing occasions of a single test, stay invariant up to a set of linear transformations given that there is appropriate model-data fit (Kolen & Brennan, 2006). This desirable property makes the IRT models a top choice over other measurement models for both measurement analysts and testing institutions. The invariance property implies that the values of model parameters between separate calibrations should be identical after being transformed to the same scale. Any violation of this property would jeopardize model parameter estimation, person ability, scoring and interpretation. If invariance does not hold, one possible consequence may be IPD. Sukin and Keller (2011) stated that IPD can be thought of as a special case of Differential Item Functioning (DIF) between test administrations. In the studies of DIF, a focal group is defined and often that group is thought of as being potentially disadvantaged by the assessment (example female) while the reference group is the set of examinees that the focal group is being compared to (for example male). The reference and focal groups can also be defined as 'Administration 1' and 'Administration 2' respectively where the first administration may be delivered first year and the second in another year. Parameter estimates for the same items that vary across test administrations are therefore considered to possess IPD and thus those items function differentially between testing occasions. Rupp and Zumbo (2016) reported that the study of IPD is related to differential item functioning in that both detect item bias and are rooted in measurement invariance. They further said the difference between IPD and DIF rests in the notion that the DIF examines difference between manifest groups while IPD is between testing occasions. In light of the conceptual similarities between IPD and DIF, many of the methods employed to assess DIF within a test can be applied to the assessment of IPD across test forms.

For large-scale achievement assessment such as Senior School Certificate Examination (SSCE) in Nigeria, (which is conducted by West African Examination Council (WAEC), National Examination Council (NECO) and National Business and Technical Examination Board (NABTEB) a set of items are often maintained and secured for repeated use. The typical item bank includes test items with varying item characteristics such as the difficulty and discrimination that are considered suitable for specific subject areas and instructional objectives. To ensure that the highest quality item statistics are used when developing tests, test constructors are encouraged to use the most recently developed items. These repeatedly administered items typically function as items for investigating changes in performance over time. The premise and justification for the repeated use of such items are that the items perform identically for the target population across repeated administrations. That is, the precision or the discrimination power and the difficulty level of the frequently used items remain stable over repeated administrations. Maintaining the item pool therefore, is not only important to ensure that items are relevant and secure, but it is also necessary in order to evaluate any change in the item parameters. Thus, IPD has become a major concern in large-scale standardized achievement tests for Senior School Certificate Examinations because when an item shows evidence of drift, it may violate a fundamental IRT assumption: examinees of the same ability level have the same probability of answering an item correctly (Babcock & Albano, 2012). This is concerning because IPD has the potential to impact the measurement precision of examinee ability estimates.

Examination bodies administer test items in all subjects including Chemistry. Chemistry as a branch of science deals with the study of structure and composition of matter. Ogunleye and Babajide (2011) stated that Chemistry is the foundation upon which the scientific and technological advancement of any nation rests. They went further to state that the subject is the foundation of scientific knowledge as it has contributed immensely to the existence and activities of man towards improved standard of living and growth in wealth. Modern Chemistry has become more important for daily life activities. There is hardly any field left where Chemistry has no role to play. Everything around us composes of atoms and molecules including our bodies. Chemistry is seen in our day to day activities like from food production to consumption, transportation means, communication means, clothes, photosynthesis many more. It helps us understand the world we see and experience. Chemistry is very important and significant to the achievement of our daily activities. There are varieties of methods used for detecting Item Parameter Drift, IPD; these methods are based on two measurement theories aforementioned. According to Sukin (2010), IPD detection methods can be categorized as either empirical-based or model-based. Empirical methods include delta plots and Mantel-Haenszel procedures. The model-based methods include plots of IRT parameter estimates, Stocking and Lord‘s Test Characteristic Curve (TCC) inverse, Raju‘s area measures, the Likelihood Ratio (LR) test, Differential Functioning of Items and Tests (DFIT) among others, which includes the 3-Sigma p-value method, the “0.3 logits” approach that involves the use of IRT- based parameter estimates and the 3-Sigma IRT approach, the 3-Sigma scaled IRT method, Robust Z , the “Area Between ICCs” method, Logistic Regression (LR) many more. Moreover, the versions of the IRT parameter models that can be used to detect IPD in dichotomously scored items are three IRT models, which are three-parameter logistic (3PL) models, two-parameter logistic (2PL) and one parameter logistic model (1PLM) or the Rasch model which describe the relationship between examinees ability level and the probability of answering an item correctly. They are used to determine if differences exist in parameter estimate across testing forms.

Item Parameter Drift can be detected using IRT item analysis procedures. (Orheruata, Omorogiuwa & Osunde, 2017). In a similar study, Abad, Oleo, Aguado, Ponsoda and Barrada (2010); Meng, Steinkamp and Matthew Lopez (2011) as well as Park, Lee and Xing (2016) used 3PL model to detect IPD. Additionally, Giordano, Subhiyah and Hess (2005); Melican (2009); Huynh and Meyer (2010); Rudner, Getson and Knight (2013) examined IPD using single methods. Nevertheless, four methods of detecting IPD were considered in this study. They are Delta plot, Mantel Haenszel (MH), Robust- z and 3-Sigma IRT approach.

Statement of the Problem

In the IRT framework, an examinee’s ability estimation is a function of item parameters. Thus ability estimates for examinees are expected to change if the item parameters change. Consequently, failing to monitor this change may lead to inaccurate score calculations, misclassifications of examinees and false decision in certification. Though drift is not completely unexpected in practice but the magnitude calls for concern. Some studies observed drift in the magnitude capable of causing scores misrepresentation positively or negatively. Orheruata, Omorogiuwa and Osunde (2017) found drifted items in 2012 to 2014 WAEC and NECO SSCE Agricultural Science multiple choice items enormous enough to cause passing advantage and jeopardize interpretation of tests. While some studies like Melican (2009); Hagge et al. (2011); Syke et al. (2012) as well as Stahl et al. (2012) used single method. However, the presence of IPD can be determined by many methods. The setback with several methods is the contradictory results that these methods generate and often times a method may flag similar items for IPD that might not be flagged by another method. (Karkee & Choi, 2005; McNamara & Roever, 2006). Moreover, researchers are faced with a confusing variety of criteria upon which specific items might be evaluated. This is of serious concern, thus, it becomes important to examine IPD methods that are stable and dependable that can be applied in determining IPD and also ensure drift free across different administrations of test or examination over time. The researcher therefore, deemed it necessary to empirically compare four methods; Robust- z, 3-Sigma IRT, Delta Plot and Mantel-Haenszel of detecting IPD using 2012 – 2015 NABTEB Chemistry multiple choice test items.

Research Questions

The following research questions were raised to guide the study:

1. What is the percentage of item parameter drift in 2012 – 2015 NABTEB Chemistry multiple choice test items using Robust z method?
2. What is the percentage of item parameter drift in 2012 – 2015 NABTEB Chemistry multiple choice test items using 3-Sigma IRT method?
3. What is the percentage of item parameter drift in 2012 – 2015 NABTEB Chemistry multiple choice test items using Delta Plot method?
4. What is the percentage of item parameter drift in 2012 – 2015 NABTEB Chemistry multiple choice test items using Mantel Haenszel method?

Hypothesis

1. There is no significant difference in the number of drifted items in 2012 – 2015 NABTEB Chemistry multiple choice test items using Robust z, 3-Sigma IRT, Delta Plot and Mantel-Haenszel methods.

Methodology

The design of the study adopted was the survey. A total population of 11,844 candidates’ responses to the National Business and Technical Examinations Board (NABTEB) SSCE for 2012, 2013, 2014 and 2015 May/June Chemistry multiple choice examinations in Edo State was used in the study. However, the statistical population was 50 items for each year making a total of 200 items for the four years of study. The total sample of candidates scores used in this study was 5,040. Multistage sampling technique was employed for effective selection of the sample in the study. At the first stage, census approach was used to obtain all the candidate responses in the four years NABTEB examinations in Edo State. At the second stage, simple random sampling technique was applied to select two senatorial districts (Edo South and Edo Central) from the three senatorial districts/zones in Edo State. However, the third stage was also by simple random sampling technique to select schools from the two senatorial zones earlier selected. Out of 1,408 senior secondary schools 140 schools were randomly selected.

The instrument used to gather data were 50-item each for 2012- 2015 May/June NABTEB Chemistry multiple choice tests making a total of 200 items. Proforma was also used to obtain candidates’ responses. The 50-multiple choice Chemistry test items for 2012, 2013, 2014 and 2015 were presumed to be validated and standardized by the Examinations Development Department, National Business and Technical Examination Board. Thus, the items were considered valid. Using Cronbach Alpha, the reliability coefficients of the instruments obtained were 0.83, 0.85, 0.89 and 0.87 for the respective years. The item parameters were estimated using eIRT software. The data collected were analyzed using the formulae of the IPD methods: Robust z, 3-Sigma IRT, Delta Plot and Mantel-Haenszel to establish the presence of drifted items. The descriptive statistics, frequency count and percentage were used to answer research questions 1to 4 while the hypothesis was tested with Chi square statistic at 0.05 alpha level.

ISSN 2705-2559
E-ISSN 2705-2567

Results

Research Question One: What is the percentage of Item Parameter Drift in 2012, 2013, 2014 and 2015 NABTEB Chemistry multiple choice test items using Robust z method?

Table 1: Percentage Distribution of IPD in the 2012 to 2015 NABTEB Chemistry Multiple Choice Test Items using Robust-z Statistics Method

Variables	Number of items	Percentage	Items
2012	4	8%	27, 31, 34, 47.
2013	5	10%	3, 12, 24, 27, 32.
2014	7	14%	4, 10, 12, 19, 27, 34, 50
2015	4	8%	1, 20, 27, 31

Table 1 shows that using Robust z method to detect drift in 2012, 2013, 2014 and 2015 NABTEB Chemistry multiple choice test items, only 4 items indicated the presence of IPD in 2012 while 5 items did show the presence of IPD in 2013, however, 7 items indicated IPD in 2014 and lastly 4 items exhibited drift in 2015 implying therefore that 8%, 10%, 14% and 8% of the items drifted in the respective examination years.

Research Question Two: What is the percentage of item parameter drift in 2012, 2013, 2014 and 2015 NABTEB Chemistry multiple choice test items using 3-Sigma IRT method?

Table 2: Distribution of Drifted Items in the 2012 to 2015 NABTEB Chemistry Multiple Choice Test Items using 3- Sigma IRT Method

Variables	Number of items	Percentage	Items
2012	19	38%	2, 9, 10, 11, 12, 13, 15, 19, 20, 21, 23, 26, 28, 29, 30, 39, 40, 44, 49
2013	17	34%	1, 2, 4, 6, 8, 10, 17, 19, 20, 23, 31, 32, 33, 35, 42, 46, 50
2014	21	42%	1, 4, 5, 6, 8, 9, 10, 13, 15, 17, 20, 21, 23, 24, 29, 31, 33, 35, 38, 46, 50
2015	23	46%	3, 4, 5, 6, 7, 9, 10, 11, 13, 14, 15, 16, 17, 18, 21, 23, 37, 40, 41, 42, 45, 46, 49

Table 2 shows that using 3- Sigma IRT method to detect drift in 2012, 2013, 2014 and 2015 NABTEB Chemistry multiple choice test items, only 19, 17, 21 and 23 items indicated the presence of IPD in 2012, 2013, 2014 and 2015 examination years with respective percentages of 38, 34, 42 and 46.

Research Question Three: What is the percentage of item parameter drift in 2012, 2013, 2014 and 2015 NABTEB Chemistry multiple choice test items using Delta Plot method?

Table 3: Distribution of Drifted Items in the 2012 to 2015 NABTEB Chemistry Multiple Choice Test Items using Delta Plot Method

Variables	Number of items	Percentage	Items
2012	15	30%	1, 3, 5, 6, 17, 24, 27, 33, 34, 35, 37, 43, 46, 47, 48
2013	19	38%	1, 3, 5, 6, 17, 24, 25, 27, 31, 32, 33, 34, 35, 37, 42, 43, 46, 47, 48
2014	16	32%	3, 18, 19, 22, 26, 27, 28, 30, 34, 37, 39, 40, 42, 44, 48, 49
2015	15	30%	1, 2, 15, 19, 20, 22, 26, 27, 29, 35, 36, 40, 43, 48, 50

Table 3 shows that in 2012, 15 items reflected the presence of IPD, 19 items in 203 exhibited IPD, in 2014, 16 items revealed IPD and in 2015, 15 items displayed IPD using Delta plot method. This implies that 30%, 38%, 32% and 30% of drift revealed IPD in the respective years of examination.

ISSN 2705-2559
E-ISSN 2705-2567

Research Question Four: What is the percentage of item parameter drift in 2012, 2013, 2014 and 2015 NABTEB Chemistry multiple choice test items using Mantel Haenszel method?

Table 4: Distribution of Drifted Items in the 2012, 2013, 2014 and 2015 NABTEB Chemistry Multiple Choice Test Items using Mantel-Haenszel Method

Variables	Number of items	Percentage	Items
2012	37	74%	1, 2, 3, 4, 5, 7, 8, 10, 11, 12, 13, 14, 16, 17, 18, 20, 21, 22, 23, 27, 28, 29,30, 31, 32, 33, 35, 36, 38, 40, 41, 42, 43, 44, 45, 46, 49
2013	31	62%	1, 3, 4, 5, 7, 8, 10, 12, 14, 16, 17, 18, 20, 23, 27, 28, 29, 30, 31, 32, 33, 35, 36, 38, 40, 41, 42, 43, 44, 45, 46
2014	46	92%	1, 2, 3, 6, 8, 9, 10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 21, 22, 23, 24, 25, 26, 27, 28, 29, 30, 31, 32, 33, 34, 35, 36, 37, 38, 39, 40, 41, 42, 43, 44, 45, 46, 47, 48, 49, 50
2015	37	74%	1, 2, 3, 4, 7, 8, 9, 10, 11, 12, 14, 16, 18, 19, 20, 22, 23, 26, 27, 28, 29, 30, 31, 32, 35, 36, 38, 39, 40, 42, 43, 45, 46, 47, 48, 49, 50

Table 4 shows that using Mantel Haenszel method to detect drift in 2012, 2013, 2014 and 2015 NABTEB Chemistry multiple choice test items, only 37 items revealed the presence of IPD in 2012, while 31 items showed IPD in 2013, 46 items indicated IPD in 2014 and 37 items exhibited IPD in 2015 representing 74%, 62%, 92% and 74% respectively.

Hypothesis One: There is no significant difference in the number of drifted items in 2012, 2013, 2014, and 2015 NABTEB Chemistry multiple choice test items using Robust z statistics, 3-Sigma IRT, Delta Plot and Mantel-Haenszel methods.

Table 5: Chi Square Analysis of Difference in the Number of Drifted Items in the 2012 to 2015 NABTEB Chemistry Multiple Choice Test Items using Robust z statistics, 3-Sigma IRT, Delta Plot and Mantel-Haenszel Methods

Year	Methods of Detecting Item Parameter Drift				Total	df	Chi-square (Calculated)
	Robust z	3 Sigma IRT	Delta Plot	Mantel Haenszel			
2012	4	19	15	37	75		
2013	5	17	19	31	72		
2014	7	21	16	46	90	9	17. 16
2015	4	23	15	37	79		
Total	20	80	65	151	316		

Chi-square critical value (table value) = 16.92

Table 5 shows the difference in the number of drifted items in the 2012, 2013, 2014 and 2015 NABTEB Chemistry multiple choice test items using Robust z statistics, 3-Sigma IRT, Delta Plot and Mantel-Haenszel methods. The table reveals that 20 items drifted in the four years (2012= 4 items; 2013= 5 items; 2014= 7 items and in 2015= 4 items) using Robust z statistics method. Using 3 Sigma IRT method, 80 items drifted in the four years (2012= 19 items; 2013= 17 items; 2014= 21 items and in 2015= 23 items). Using Delta plot method, 65 items drifted in the four years (2012= 15 items; 2013= 19 items; 2014= 16 items and in 2015= 15 items); while using Mantel Haenszel method, 151 items drifted in the fours (2012= 37 items; 2013= 31 items; 2014= 46 items and in 2015= 37 items), making Mantel Haenszel the method that detected the highest number of drifted items. Table 5 also shows that the chi-square calculated is 17. 16, while the chi-square critical value (table value) is 16.92. Testing the hypothesis at 0.05 significant level, the calculated value (17. 16) is greater than the t-critical (16.92) therefore, the null hypothesis that says there is no significant difference in the number of drifted items using the four methods is rejected. This means that there is a significant difference in the number of drifted items in 2012, 2013, 2014 and 2015 NABTEB Chemistry multiple choice test items using Robust z statistics, 3-Sigma IRT, Delta Plot and Mantel-Haenszel methods.

Discussion of Findings

In a nutshell, the four methods flagged the drifted items in an undulating pattern but in different magnitudes. It was also observed from the analysis that Robust- z method consistently flagged the fewest number of drifted items over the examination years compared to the other methods, hence it was considered the most stable method. Research question one revealed the percentages of item parameter drift in 2012, 2013, 2014 and 2015 NABTEB Chemistry multiple choice test items using Robust -z method as 8%, 10%, 14% and 8% of items in the respective years. The study also revealed that in the difficulty dimension, only four (4) items in 2012; five (5) items in 2013; seven (7) items in 2014 and four (4) items in 2015 exhibited IPD. The findings of this study is in agreement with the findings of Huynh and Meyer (2010) who used Robust- z statistics to detect item parameter drift in a set of archival data from a large-scale assessment program. Using the cut score of 1.96 for the Robust- z statistics, eight items on the whole were found drifted, two (2) items (ID = 26 with ZR = 4.261; and 38 with ZR = 2.88) were found to be 'unstable' (possess item parameter drift) along the slope dimension. The second set of Robust ZR statistics revealed that six (6) items were found to be 'unstable' along the location dimension. They are listed as follows: ID = 17 (ZR = 2.624); ID = 21 (ZR = 2.37); ID = 28 (ZR = 2.58); ID = 33 (ZR = 2.399); ID = 35 (ZR = 3.924); ID = 44 (ZR = 1.987).

Research question two revealed the percentages of item parameter drift in 2012 - 2015 NABTEB Chemistry multiple choice test items using 3-Sigma IRT method to be 38%, 34%, 42% and 46% for the respective years. The outcome of this study is not in line with the findings of Huang and Shyu (2003), who in their study used 3-sigma IRT to detect item parameter drift in their simulated study. The study found that the drifted items constituted more than half of the common item pool and this led to profound consequences such as affecting the mean and passing scores especially with a small sample size of 500.

Research question three revealed the percentages of item parameter drift in 2012 - 2015 NABTEB Chemistry multiple choice test items using Delta plot method to be 30%, 38%, 32% and 30% for the years. The finding of this study is similar to the findings of Hu, Rogers and Vukmirovic (2008) who used delta plot method and found instability in the difficulty parameters in which the common items resulted in larger systematic error in the equated scores.

Research question four revealed that the percentages of item parameter drift in 2012 - 2015 NABTEB Chemistry multiple choice test items using Mantel Haenszel method are 74%, 62%, 92% and 74% respectively. The result of this study is in accordance with the finding of Mehriban (2019), who in his study investigated the item parameter drift (IPD) using the Mantel-Haenszel (MH) method. Results of IPD analysis revealed that 76 out 81 of the mathematics items exhibited IPD values across the assessment modes.

Hypothesis one revealed that there is a significant difference in the number of drifted items in 2012 - 2015 NABTEB Chemistry multiple choice test items using Robust z, 3-Sigma IRT, Delta Plot and Mantel-Haenszel methods. The finding of this study is in accordance with the findings of Michaelides (2008) who compared Mantel-Haenszel and Delta plot methods of detecting IPD. The result revealed that in the dichotomous items, the Mantel-Haenszel procedure worked well because more items were flagged as exhibiting IPD than outliers identified by the Delta-plot method, while one of the items was flagged by both methods. This shows that there is a significant difference in the number of items that exhibited IPD identified by the two methods.

The finding of the hypothesis is also in agreement with the findings of Donoghue and Isham (1998) who used Monte Carlo methods to compare 3 types of measures of item parameter drift. The study found that there is a significant difference in the number of items that exhibited IPD among the three types of measures of item parameter drift (item response theory-based, Mantel-Haenszel based or NAEP BILOG/PARSCALE Item-Level

ISSN 2705-2559
E-ISSN 2705-2567

χ^2 statistics). This study is in agreement with the study of Liaw (2012), who used two methods, Robust- z statistics and the signed area between item characteristics curves to detect items that demonstrated item parameter drift. The study found that the item parameters were unstable and a significant difference in items identified by both techniques. Contrary to the current study, Li (2008) did study on an investigation of item parameter drift in the Examination for the Certificate of Proficiency in English Language (ECPE). Using IRT techniques, no significant difference in the item drift across the years was found.

Conclusion

The study concluded that the Robust- z method flagged the fewest number of drifted items compare to the other methods across the years and that there is a significant difference in the number of drifted items in 2012 - 2015 NABTEB SSCE Chemistry multiple choice test items using Robust- z, 3 -sigma IRT, Delta plot and Mantel Haenszel methods to detect IPD.

Recommendations

On the basis of the findings and conclusion drawn, the following recommendations were made:

1. Robust -z method should be used by examination bodies for IPD analysis in order to avoid false identification of items.
2. Examination bodies such as National Business and Technical Examinations Board (NABTEB), West African Examination Council (WAEC), National Examination Council (NECO) and Joint Admission and Matriculation Board (JAMB) should include Item Parameter Drift analysis as part of their item analysis to avoid measurement error and produce quality items.
3. Examination bodies and other stakeholders should further probe items that exhibit drift for revision, modification or total removal of such item to ensure near drift free items.

References

Abad, F. J., Oleo, J., Aguado, D., Ponsoda, V., & Barrada, J. R. (2010). Item parameter drift in computerized adaptive testing: Study with eCAT. *Psicothema, 2,* 340-347.

Babcock, R. D. & Albana, E. (2012). Full-information item factor analysis. *Applied Psychological Measurement, 12*(3), 261–280.

Babcock, B. & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied Psychological Measurement, 36,* 565-580.

Donoghue, J. R. & Isham, S. P. (2008). A comparison of procedures to detect item parameter drift. *Applied Psychological Measurement, 22*(1):33–51.

Goldstein, H. (1983). Measuring changes in educational attainment over time: Problems and possibilities. *Journal of Educational Measurement, 20,* 369-377

Giordano, C., Subhiyah, R. & Hess, B. (2005). An analysis of item exposure and item parameter drift on a take-home recertification exam. *Paper presented at the Annual Meeting of the American Educational Research Association. http://www.eric.ed.gov/PDFS /ED497708.pdf.*

Hagge, S., Woo, A. & Dickison, P. (2011). Impact of item drift on candidate ability estimation. *Paper presented at the annual conference of the International Association for Computerized Adaptive Testing*.

Hu, H., Rogers, W. T. & Vukmirovic, Z. (2008). Investigation of IRT-based equating methods in the presence of outlier common items. *Applied Psychological Measurement,* 32, 311-333.

Huang, C. Y. & Shyu, C. Y. (2003). The impact of item parameter drift on equating. *Paper presented at the Annual meeting of the National Council on Measurement in Education*.

Huynh, H. & Meyer, P. (2010). Use of Robust z in Detecting Unstable Items in Item Response Theory Models. *Practical Assessment, Research & Evaluation.*15 (2):1-8.

Karkee, T. & Choi, S. W. (2005). Items flagged from statistical criteria on test score classifications in common item equating. *Paper presented at the Annual meeting of the American Educational Research Association*.

Kolen, M. J. & Brennan, R. L. (2006). *Test equating, scaling and linking: Methods and practices.* (2nd ed.). Springer.

Li, X. (2008*)*. An investigation of item parameter drift in the Examination for the Certificate of Proficiency in English (ECPE). *Foreign Language Assessment*, *6*, 1-28.

Liaw, Y.L. (2012). *Stability of item parameters in equating items*. A thesis submitted in partial fulfillment of the requirements for the degree of Master of Education, University of Washington.

McNamara, T. & Roever, C. (2006). *Language testing: The social dimension.* Malden, M A & Oxford: Blackwell.

Mehriban, C. (2019). *Investigating item parameter drift across computer and paper-based assessment modes in PISA 2015 mathematics*. https://education.wsu.edu/ documents/2019/11/thesis-defense-mehriban-ceylan.pdf/

Melican, W. P. (2009). The effects of item parameter drift on equating test scores. *Paper presented at the annual meeting of the National Council on Measurement*.

Meng, H., Steinkamp, S. & Matthews-Lopez, J. (2011). *Practitioner's approach to identify item drift in CAT.* Paper presented at the annual conference of the International Association for Computerized Adaptive Testing.

Michaelides, M. A. (2008). An illustration of a Mantel-Haenszel procedure to flag misbehaving common items in test equating. *Educational Research and Reviews, 9*(17), 642-649.

Ogunleye, B. O. & Babajide, A. O. (2011). Secondary school students' assessment of innovative teaching strategies in enhancing achievement in Chemistry and Mathematics. *IOSR Journal of Research & Method in Education (IOSR-JRME)*, *3*(5), 6-11.

Orheruata, M. U. (2015). Item parameter drift in certificate examination and its implication on decision making. *African Journal of Theory and Practice of Educational Assessment 2*, 98-105.

Orheurata, M., Omorogiuwa, O. K. & Osunde, A.U. (2017). Assessing scale drift of WAEC and NECO SSCE Agricultural Science multiple choice items with Item Response Theory. *(ASSEREN) Journal of Education 2*(1), 15-23.

Park, Y.S.; Lee,Y. & Xing, K. (2016). Investigating the impact of item parameter drift for item response theory models with mixture distributions. *Frontiers in Psychology*, 7, 255.

Rudner, L. M., Getson, P. R. & Knight, D. L. (2013). Individual assessment accuracy. *Journal of Educational Measurement, 20*, 207-219.

Rupp, A. A. & Zumbo, B. D. (2016). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measurement, 66*, 63-84.

Stahl, J., Bergstrom, B. & Shneyderman, O. (2012). Impact of item drift on test-taker measurement. *Paper presented at the annual meeting of the American Educational Research Association*.

Sukin, T. M. (2010). *Item parameter drift as an indication of differential opportunity to learn: An exploration of item flagging methods & accurate classification of examinees*. ProQuest LLC, Ed. D. Dissertation, University of Massachusetts. www.eric.ed.gov.

Sukin, T. M. & Keller, L. H. (2011). Item parameter drift as an indication of differential opportunity to learn: An exploration of item flagging methods & accurate classification of examinees*. Paper presented at the annual meeting of the National Council on Measurement in Education*.

Sykes, R. C. & Fitzpatrick, A. R. (2012). The stability of IRT b values. *Journal of Educational Measurement, 29*(3), 201-211.