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

By

AUGUSTINA ENOGIE IGHODARO PhD
Dept. of Educational Evaluation and Counselling Psychology, University of Benin
Email: usigbemwona@yahoo.com. Contact: 08062413730

and

M.U. ORHERUATA PhD
Dept. of Educational Evaluation and Counselling Psychology, University of Benin

Abstract

This study investigated the comparison of methods to detect Item Parameter Drift (IPD) in National Business and Technical Examination Board (NABTEB) Senior School Certificate Examinations (SSCE) for 2012 – 2015 Chemistry multiple choice tests. The main purpose was to examine the percentages of item drift and compare the differences in the number of drifted items across the stated examinations years using two IRT techniques (Robust –z and 3 -sigma IRT) of detecting IPD. This study was guided by three research questions and one hypothesis. The study adopted the Survey research design. The population of the study was 11,844 scores of candidates who sat for the National Business and Technical Examination Board, NABTEB, SSCE Chemistry multiple choice tests in Edo State, Nigeria. The sample size used for the study was 5,040 candidates' scores. 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. The instruments being standardized by their source were considered valid and reliable. However, the item parameters were estimated from candidates' responses using EIRT (Item Response Theory Assistant for Excel) software. The two methods (Robust- z method and 3 -sigma IRT method) were

used to establish the IPD, 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 on the whole 20 items and 80 items drifted using Robust- z method and 3 -sigma IRT method respectively in 2012 – 2015 NABTEB SSCE Chemistry multiple choice tests and it was also found out that there is no significant difference in the number of drifted items in 2012 – 2015 NABTEB SSCE Chemistry multiple choice test items using Chi- square statistics. Furthermore, it was concluded that Robust -z method was the more stable method because it flagged the least number of drifted items between the two methods. Among others, it was recommended that NABTEB and other examination bodies should use Robust -z and 3 -sigma IRT methods to detect drift to avoid false identification of drifted items.

Keywords: Item Parameter Drift, Chemistry, Multiple Choice, Test Items, IRT Techniques

Background to the Study

Assessment is a process of measuring learning outcomes and other proficiencies in order to make authentic and valid decisions about the individual. It is also the process of describing, collecting, recording, scoring and interpreting information about learning. The goal of assessment is to make improvement as opposed to simply being judged. It is done with the intent to provide information that will enhance quality professional testing and measurement. It helps in the improvement of educational activities across the educational system in a cycle of continuous upgrade.

Items in the test are characterized by certain parameters – item difficulty, discrimination, pseudo guessing and carelessness that are defined within two measurement theories, among them is the Item Response Theory (IRT). 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. It is founded on the premise that the probability of a correct response to a test question is a mathematical function of parameters such as a person's latent traits or abilities and item characteristics (such as difficulty, "guess ability," and discrimination).

In the Item Response Theory (IRT) measurement model, although the item parameter estimates for common items are treated as fixed or

unchanging after their first exposure, the parameters may fluctuate over subsequent administrations; this phenomenon is referred to as Item Parameter Drift. Thus IPD refers to change in parameter values of item across several testing occasions. According to Wells, Hambleton and Meng (2014), deviations in item parameters from true value in its successive testing administrations are known as item parameter drift. They added that IPD occurs when invariance no longer holds.

According to the invariance property of IRT, item parameters estimated from different samples of the same population are supposed to be invariant, even over different measurement occasions (Wells, Subkoviak & Serlin, 2012). Parameter invariance is the equivalence of item and person parameters belonging to different populations and measurement applications (Rupp & Zumbo, 2016).

Drift is likely to occur when an item pool is not maintained over time even though good quality items are selected and secured carefully. Such effects may be expected because of frequent item exposure or over usage of items. Changes in curriculum, instructional variation and increasing practice effect can cause item to drift. Also inappropriate test-wise training, increase in teaching and exercise, immense teaching–to-test, changes in item position or location, security breaches, test preparation and historic event may cause drift to occur. Items may also perform differently across years due to changes in the construct or content. Such changes can threaten the validity of test scores by introducing trait-irrelevant differences on ability estimates. Item Parameter Drift poses a threat to measurement applications that require a stable scale (Wells, Subkoviak & Serlin, 2012).

For standardized assessment such as conducted for 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. These repeatedly administered items typically function as items for investigating changes in performance over time.

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.

However, there are varieties of methods used for detecting Item Parameter Drift, IPD under IRT framework; these methods according to Gaertner and Briggs (2009) are stated as follows, the "0.3 logits" approach that involves the use of IRT- based parameter estimates, the 3-Sigma IRT approach, the 3-Sigma scaled IRT method. DIF-based methods: Robust Z is an IRT - based DIF method, "Area Between ICCs" method among others. 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.

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) did a study on item parameter drift (IPD) using of 2012 to 2014 WAEC and NECO SSCE Agricultural Science multiple choice items and found drifted items in both examinations 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. 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 two methods; Robust- z and 3-Sigma IRT of detecting IPD using 2012 – 2015 NABTEB Chemistry multiple choice test items.

The following research questions and hypothesis 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. Is there a difference in the number of drifted items in 2012 – 2015 NABTEB Chemistry multiple choice test items using Robust z and 3-Sigma IRT?

Hypothesis

There is no significant difference in the number of drifted items in 2012 – 2015 NABTEB Chemistry multiple choice test items using Robust z and 3-Sigma IRT. **Methods.**

This study adopted the survey research design. A total population of 11,844 candidates' responses to National Business and Technical Examinations Board (NABTEB) 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 candidates' scores of 2012, 2013, 2014 and 2015 NABTEB May/June examinations. 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 candidates responses in the four NABTEB examinations in Edo State in the first administration. 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 (Edo South, Edo Central and Edo North) 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 NABTEB Chemistry multiple choice test for the four years making a total of 200 items. The instruments were presumed to be validated and standardized by the Examinations Development Department, National Business and Technical Examination Board. The instruments being standardized by

NABTEB, a national examination board were considered reliable. The research data were obtained in a soft copy of the candidates' responses to the NABTEB 2012 – 2015 multiple choice Chemistry test items in the excel form of students' matrix scores from Information and Communication Technology (ICT) Department,

The item parameters were estimated using EIRT software. The data collected were analyzed using the formulae of the IPD methods: Robust z and 3-Sigma IRT to establish the presence of drifted items. The descriptive statistics, frequency count and percentage were used to answer research questions 1 to 3 while the hypothesis was tested with Chi square statistic at 0.05 alpha level.

Table1: Distribution of IPD in the 2012, 2013, 2014 and 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. It however showed unit shift in the drifted items in the four years, that is the drift increased by one item from 2012 to 2013 and then increased by two items in the next year (2014), thereafter showed a decrease in the last year (2015) by three items.

Table 2: Distribution of Drifted Items in the 2012, 2013, 2014 and 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 examiniation years with respective percentages of 38, 34, 42 and 46.

Table 3: Chi Square Analysis of Difference in the Number of Drifted Items in the 2012, 2013, 2014 and 2015 NABTEB Chemistry Multiple Choice Test Items using Robust z statistics and 3-Sigma IRT Methods

Year	Methods of Detecting Item Parameter Drift		Total	df	Chi-square (Calculated)
	Robust z	3 Sigma IRT			
2012	4	19	23		
2013	5	17	22		
2014	7	21	28	3	1. 101
2015	4	23	27		
Total	**20**	**80**	**100**		

Chi-square critical value (table value) = 7.82

Table 3 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 and 3-Sigma IRT 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 while 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) making 3-Sigma IRT the method that detected the highest number of drifted items.

Table 3 also shows that the chi-square calculated is 1. 101, while the chi-square critical value (table value) is 7.82. Testing the hypothesis at 0.05 significant level, the calculated value (1. 101) is less than the t-critical (7.82) therefore, the null hypothesis that says there is no significant difference in the number of drifted items using the two methods is retained. In other words, 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 and 3-Sigma IRT methods.

Discussion of Findings

In a nutshell, the two methods identified the drifted items in an undulating pattern but in different magnitudes. It was also observed from the analysis that Robust z method identified drifted items within the range of four to seven items across the years and it showed element of consistency because Robust- z method consistently flagged the fewest number of drifted items over the examination years compared to the other method, 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. The data came from the administration of Mathematics test to 5th grade students. PARSCALE was used to estimate the item parameters for each test form. 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). The result of this study is also in agreement with the findings of Yuan-Ling (2012), who used two methods, Robust- z statistics and the signed area between Item Characteristics Curves (ICC) to detect items that demonstrated item parameter drift. The result showed that twelve items were seen as flagged items by signed area between two ICCs and few (4) items were flagged by Robust- z statistics. The result of this finding is also similar to that of Rahmawati and Djemari (2015) who used Robust -z method to detect item parameter drift in a two simulated data which were in the form of responses of 40,000 students on 40 dichotomous items generated by stimulating six variables. The result revealed that Robust z is accurate to detect the b and ab-drifting.

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 in line with Li (2008) 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. In the same vein, Yoon, Young-Sun and Kuan (2016) study on investigating the impact of item parameter drift for IRT models with mixture distribution. They examined instability in item parameter estimation using mixture of IRT techniques and found that there is no significant difference in the IPD items over the testing administrations. Contrary to the current study, Huang and Shyu (2003) used 3-sigma IRT to detect item parameter drift in 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. The finding of this study is not in agreement with the study of Melican (2009) who used 3-sigma IRT to reveal that IPD was found in a very small number of items, even over the four-year period in a

CAT program, whereas, the present study revealed that the numbers of drifted items in the 2012, 2013, 2014 and 2015 NABTEB Chemistry examinations to be aggregate of 80 items.

Hypothesis one revealed that there is no significant difference in the number of drifted items in 2012 - 2015 NABTEB Chemistry multiple choice test items using Robust z and 3-Sigma IRT methods. The result of the study also revealed that 3 sigma IRT method detected the highest number of drifted items in the order of 19, 17, 21 and 23, which summed up to 80 items across the years of examination. Robust z statistics method is the method that detected the least drifted items (20).

This study is not in agreement with the study of Liaw (2012), who investigates the item factors that may cause item parameter instability. The data for the study was obtained from the state-level Washington Assessment of Student Learning (WASL) tenth grade mathematics exams administered from 1999 to 2004. Two methods, Robust- z statistics and the signed area between item characteristics curves were used 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.

The finding of the hypothesis is also not in line with the findings of Donoghue and Isham (1998) used Monte Carlo methods to compare 3 types of measures of item parameter drift. These measures were item response theory-based, Mantel-Haenszel based or NAEP BILOG/PARSCALE Item-Level χ^2 statistics. Number of examinees, number of items and number of drift items in the test were manipulated. 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 χ^2 statistics).

This study is contrary to Masters, Muckle and Bontempo (2009) who compared methods to recalibrate drifting items in Computer Adaptive Testing (CAT), using empirical data with 450 examinees and 152 operational items. They examined whether applying the displacement statistic to drifted items could account for the drift. The authors assessed whether calculating a new difficulty value (adding the displacement to the original calibration) or recalibrating the item in another pretest better accounted for drift. They then compared the adjusted calibrations to the

new calibrations. Their results showed a high correlation between the adjusted and new calibrations for drifted items. The difficulty measures of 40 of the 152 items were statistically significantly different.

Conclusion

The study concluded that NABTEB May/June SSCE Chemistry multiple choice tests for 2012 - 2015 showed that drift was present. It was also concluded that the Robust- z method flagged the fewest number of drifted items across the years. Lastly, that there is no significant difference in the number of drifted items in 2012 - 2015 NABTEB SSCE Chemistry multiple choice test items using Robust- z and 3 -sigma IRT methods to detect IPD.

Recommendations

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

- Robust -z and 3 -sigma IRT methods should be used by examination bodies for IPD analysis in order to avoid false identification of items.
- 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 make Item Parameter Drift analysis as part of their item analysis to avert measurement error and produce quality items.
- Stakeholders should carry out further examination on items that exhibit drift for revision, modification or total removal of such item to ensure near drift free items.

References

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

Gaertner, M. N. & Briggs, D. C. (2009). *Detecting and Addressing Item Parameter Drift in IRT Test Equating Contexts*. www.researchgate.net/publication/3236616531.

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*.

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.

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.

Masters, J. S, Muckle, T. J. & Bontempo, B. (2009). Comparing Methods to Recalibrate Drifting Items in Computerized Adaptive Testing. *Paper presented at the annual meeting of the American Educational Research Association*.

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*.

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.

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.

Rahmawati, R. & Djemari, M. (2015). Modified Robust z method for equating and detecting item parameter drift. *Research and Evaluation in Education, 1*(1): 100-113.

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*.

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

Wells, C. S., Hambleton, R. K. & Meng, Y. (2014). An examination of two procedures for identifying consequential item parameter drift. *Applied Measurement Education, 27,* 214-231

Well, C. S., Subkoviak, M. J. & Serlin, R. C. (2012). The effect of item parameter drift on examinee ability estimates. *Applied Psychological Measurement, 26*: 77-87.

Yuan-Ling, 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.