AJSTME, Volume. 10, Issue 3. 433-439, March, 2024. https://www.ajstme.com.ng
ISSN: 2251-0141 (Print), 2971-6233 (Online)

Effect of Item Drift on Item Parameter Estimation of Multiple Choice Items of Basic Education Certificate Examination in Basic Science

Oguguo, B. E. C & Nwani, Samuel Uchenna
Department of Science Education, University of Nigeria, Nsukka
Corresponding author: samuel.nwani.pg93202@unn.edu.ng

Abstract

The study examined the effect of item parameter drift on test item parameters. Instrumentation research design was adopted. A sample of 508 JSS 3 students from Nsukka Education Zone was drawn out of a population of 2,876 JSS 3 students for the study. Three parallel forms of Basic science examinations were used for the study. Three anchor items from Past Questions where identified and added to the instrument. Reliability co-efficient of 0.87, 0.86 and 0.88 were obtained respectively for the instruments. The instruments were administered three times at various intervals on the spot with the help of five research assistants who were also postgraduate students of University of Nigeria, Nsukka. The data collected were marked using Microsoft excel. The data collected was analyzed using Normal Ogive Harmonic Robust Method (NOHARM) and Multidimensional Three Parameter Logistic Model of Item Response Theory using R programming language. The findings of the study revealed that exposing students to the same items on various occasions of test administrations causes variation in the parameters of items, item parameters are also affected by various test occasions, test items tend to become easier with respect to any administration, thereby leading to faulty assessment and decisions based on the outcome of the examination. The study recommended among others that test item security must be taken serious because one of the major causes of item parameter drift is question leakage, hence serious care must be put in place to prevent students from seeing the questions before the actual testing.

Keywords: Item parameter drift, item parameter estimation, Anchor Items

Introduction

The ability to establish stable correlations between assessments over time is crucial for tracking trends in student achievement. In large-scale assessments, this is usually achieved by using the item response theory (IRT) model. The main benefit of IRT modeling is its ability to hold the property of parameter invariance, which states that item parameters are invariant to the specific examinee sample used to estimate them and vice versa (Babcock & Albano, 2011). In certain situations, however, this property may not hold true because the parameter values for the same test items may change systematically over multiple testing occasions, a phenomenon known as item parameter drift (IPD). In certificate examinations, the validity of inferences drawn from the parameter invariance

The examination results must accurately reflect the examinee's knowledge in order to draw conclusions about their abilities. Exam results that do not accurately reflect an examinee's aptitude constitute a misrepresentation that may jeopardize their validity and fairness (Park, Lee & Xing, 2016). Item parameter drift is one phenomenon that, in the words of Kyung and Famnmin (2016), has the potential to result in this kind of score distortion. Good quality items may be selected and secured carefully by large-scale examination bodies but systematic influences may cause the item parameters estimates to change or become theoretically unstable over time to the extent that the item becomes easier and less discriminating, causing errors in proficiency estimation using the original item parameters (Xin, 2018). This change in item parameters over time is referred to as item parameter drift (IPD).

Item Parameter Drift (IPD) refers to a change in parameter value(s) of the anchor item across several sequential testing occasions or test levels. Item parameter drift is simply the shift of item parameter estimates from the acceptable theoretical scale. In other words, the concept of item parameter drift in measurement holds when there is a violation in the

stability of the parameter scale; it also examines the comparability of violated items across time or testing occasions (Orheruata, 2015). Xin (2018) maintained that the presence of Item Parameter Drift (IPD) poses a threat to the fairness and validity of test scores, thereby introducing trait-irrelevant differences that impact performance on the item over time. This threat according to De Ayala (2022) could amount to some aftereffect that will jeopardize the fair interpretation of test scores from year to year. When sizeable magnitude of IPD exists in an achievement instrument according to Orheruata, (2015), the amount of measurement error in scores produced by that instrument increases, thereby leading to a reduction in test reliability. This in turn increases the potential for misclassifying candidates whose true scores fall at or near the passing score. Item parameter drift can have impact on examinees' classification accuracy (De Ayala, 2022) which may lead to false decision in certificate examination. Modifications to item parameters could be dangerous for estimating the latent construct. A fair score discussion utilizing linking items across several administrations may be seriously jeopardized by drift of anchor item specifications in particular, which could result in incorrect conclusions in certification and licensing exams.

Anchor items are specific items or tasks within a test that are strategically placed to serve as reference points for the difficulty or complexity of the entire test. Anchor items also serve as quality control measures during the test development process. They help in identifying items that might be too easy or too difficult relative to the intended difficulty level of the test (Albano & Rodriguez, 2012). Anchor items help test developers calibrate the difficulty of the entire test. By including items that are known to be easy, moderate, and difficult, they ensure that the test effectively measures a range of abilities. These items according to Park, Lee and Xing (2016) serve as a quality assurance measure during test development; they help in identifying items that may be too easy or too difficult relative to the intended difficulty level of the test. Anchor items aid in standardizing the difficulty across different versions or forms of the test. This is particularly important in standardized tests where multiple versions are administered, ensuring fairness and comparability of scores. In standardized testing, scores from different test versions need to be comparable. Anchor items facilitate score equating, where statistical methods adjust scores to account for differences in difficulty between test forms. After testing, anchor items are used to analyze the performance of test-takers. This analysis helps evaluate the effectiveness of individual test items and inform decisions about item retention or revision. Anchor items also play significant role in identifying drifts in item parameters,

Many factors could cause the parameters for the same item(s) to drift upwards or downwards; some of these factors, like item exposure or cheating, curricular misalignment with state standards, are undesirable, while others, like assessment-driven curriculum changes or more focused instruction, are desirable (Albano, & Rodriguez, 2012). IPD may have a variety of causes, but multiple studies have demonstrated that it can result in skewed ability assessments and, ultimately, incorrect examinee classification (Aolfazi & Erfan, 2018). This could have significant implications when test results are used for key decisions. The potential source of IPD was content and context effect such as change in content and curriculum coverage, mass media, lack of item pool maintenance, increase in teaching and exercise, Immense teaching-to-test, item exposure, item position or location, security breaches, test preparation and historic events could change how an item originally performs. Excessive item exposure or inadequate security control can also cause test items to exhibit IPD (Gaertner & Briggs, 2009). Parameter drift might also result from modifications made to the item in between administrations. This could entail altering the item's scoring, presentation style, placement on the form, formatting, and other aspects of its appearance. For instance, between 32 and 49% of items were marked for drift while switching between pre-test and operational use over a five-year period, according to an investigation of drift for a nationwide computerized adaptive exam (Bergstrom, Stahl & Netzky, 2001). A few potential causes of IPD during the useful life of the scale include curriculum modifications based on assessment results, more focused instruction, item exposure, cheating, or curriculum misalignment with certain standards (Kyung & Fanmin, 2011). Additional research by Kyung and Fanmin (2011) indicated that a change in the physics curriculum may be the cause of the relationship between item content and relative direction of drift. Research in applied psychology by Albano & Rodriguez (2012) also indicated that time had an impact on the efficacy of the Armed Services Vocational Aptitude Battery and that certain cognitive-ability measures were more vulnerable to time impact than other item types.

AJSTME, Volume. 10, Issue 3. 433-439, March, 2024. https://www.ajstme.com.ng
ISSN: 2251-0141 (Print), 2971-6233 (Online)

Several studies have been carried out on item parameter drift some of which include; Albano and Rodriguez (2012) who reported that students may not likely remember their responses in previous test administration, hence majority of the scales in measurement applications were maintained beyond an initial test administrations. Park, Lee & Xing (2016) found out that there are considerable modifications in the distribution of examinee ability between latent classes over the three administrations and the requirement to caution and assess IPD using a mixed IRT framework to comprehend its effects on item parameters and examinee ability. Xin (2018) claimed that the results confirm that the potential influence of multidimensionality was identified connected with the item parameter drift (IPD) for the same set of common items utilizing four distinct data dimensional compositions. Kyung, and Fanmin, (2011) also reported that the short-term effect of IPD on item calibration turned out to be very limited and fairly inconsequential under the studied conditions. The present study tends to examine the effect of item drift on item parameter estimation. Specifically, the study tends to determine the:

1. difference in the item parameters of the three anchor items of Basic Education Certificate Examination in Basic Science
2. trend in the item parameters of the three anchor items of Basic Education Certificate Examination in Basic Science
3. difference in the item characteristic curve of item parameters of the three anchor items of Basic Education Certificate Examination in Basic Science

Research Questions

The following research questions guided the study

1. What is the difference in the item parameters of the three anchor items of Basic Education Certificate Examination in Basic Science?
2. What is the trend in the item parameters of the three anchor items of Basic Education Certificate Examination in Basic Science?
3. What is the difference in the item characteristic curve of item parameters of the three anchor items of Basic Education Certificate Examination in Basic Science?

Methodology

The study adopted instrumentation research design. A sample of 508 JSS 3 students from Nsukka Education Zone was drawn out of a population of 2.876 JSS 3 students for the study. Three parallel forms of Basic Education Certificate Examination in Basic science (Form A, B and C) were used for the study. Three anchor items from Past Questions where identified and added to the instrument. Two criteria informed the decision for identifying anchor items. First was that the questions were repeated verbatim from previous questions. Secondly, the questions were pulled from the same content area; hence the way the items were written made them the same items. The instruments were subjected to estimate of temporal stability and reliability co-efficient of 0.87, 0.86 and 0.88 was obtained respectively. The instruments were administered three times at various intervals. The Instruments were administered to the students on the spot with the help of five research assistants who were also postgraduate students of University of Nigeria, Nsukka. The data collected were coded and marked using Microsoft excel. The data collected was analyzed using Normal Ogive Harmonic Robust Method (NOHARM) and Multidimensional Three Parameter Logistic Model of Item Response Theory using R programming language.

Results

Table 1: Difference in the item parameters of the four anchor items of Basic Education Certificate Examination in Basic Science and Technology

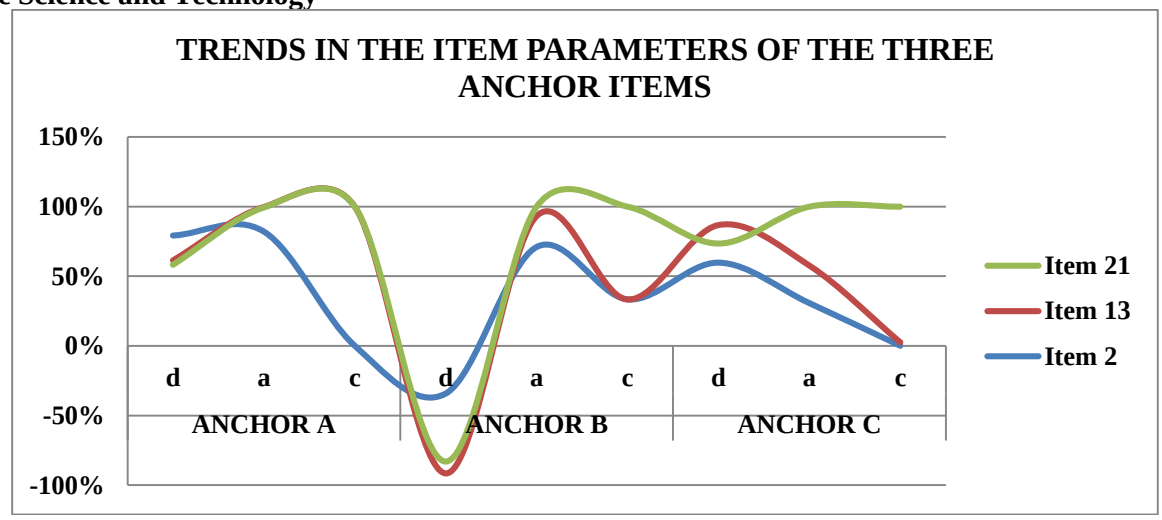
S/N	Anchor A			Anchor B			Anchor C		
	d	***a***	***c***	***d***	***a***	***c***	***d***	***a***	***c***
Item 2	48.25	53.41	0.00	-1.09	2.83	0.03	3.88	2.57	0.00
Item 13	-10.842	11.58	0.136	-1.82	0.89	0.00	1.76	2.27	0.01
Item 21	-1.869	-0.179	0.00	0.27	0.27	0.06	-0.86	3.55	0.38

d = difficulty index, a = discrimination index, c = guessing index

AJSTME, Volume. 10, Issue 3. 433-439, March, 2024. https://www.ajstme.com.ng
ISSN: 2251-0141 (Print), 2971-6233 (Online)

Table 1 shows the differences in the item parameters of the four anchor items of Basic Education Certificate Examination in Basic Science and Technology. The table shows variations that have occurred in all the item parameters of the three anchor items, from the difficulty indices, discrimination indices and the guessing indices of the three anchor items. This is an indication that exposing the same items to the students in various occasions causes a variation in the parameters of items. Hence, also influence the quality of the items varies alongside.

Figure 1: Trend in the item parameters of the three anchor items of Basic Education Certificate Examination in Basic Science and Technology



d = difficulty index, a = discrimination index, c = guessing index

Figure 1 show the trend in item parameters of the three anchor items of Basic Education Certificate Examination in Basic Science and Technology. The trend shows significant variations that have taken place as a result of exposing the same items to the same group of students over a number of test administrations. Hence, the item parameters are also affected by the various test occasions.

AJSTME, Volume. 10, Issue 3. 433-439, March, 2024. https://www.ajstme.com.ng
ISSN: 2251-0141 (Print), 2971-6233 (Online)

Figure 2: Differences in the item characteristics curve of item parameters of the three anchor items of Basic Education Certificate Examination in Basic Science and Technology

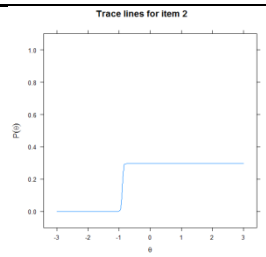
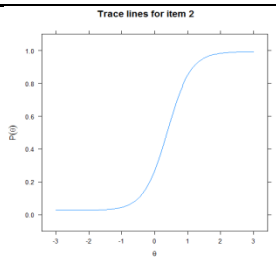
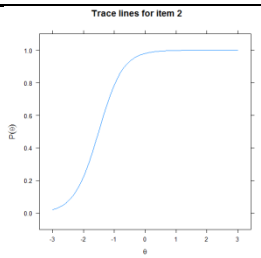
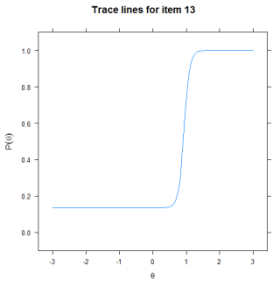
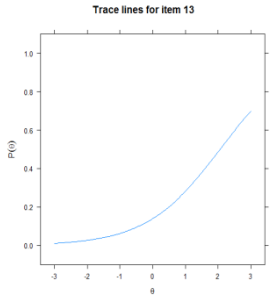
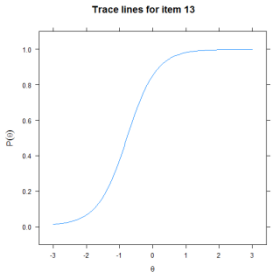
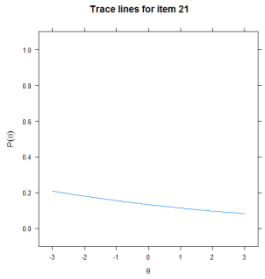
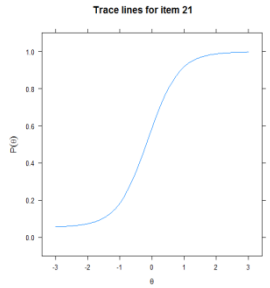
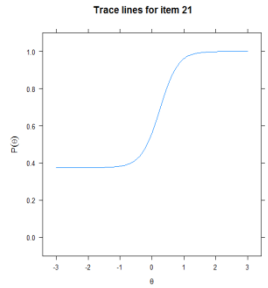
S/N	Anchor A	Anchor B	Anchor C
Item 2			
Item 13			
Item 21			

Figure 2 shows the differences in the item characteristics curve of the three anchor items of the Basic Education Certificate Examination in Basic Science and Technology. All items, 2, 13 and 21 show variations in the three different occasions of test administrations. Item 2 shows a clear discrimination between the bright and dull students in anchor A, however, the bright students did not score high in that item, but in anchor B, the bright students were able to score higher in the same item compared to anchor A, whereas many other students score higher in the same item in anchor C when compared with anchor B. similar result plays in items 13 and 21 respectively. This is indicative that the more and more students are been exposed to the same item(s), the item(s) tend to become easier with respect to any administration, thereby leading to faulty assessment and decisions based on the outcome of the examination.

Discussion of Findings

The findings of the result show that exposing the same items to the students in various occasions causes a variation in the parameters of items. Hence, also influence the quality of the items varies alongside, item parameters are also affected by the various test occasions and that the more and more students are been exposed to the same item(s), the item(s) tend to become easier with respect to any administration, thereby leading to faulty assessment and decisions based on the outcome of the examination. This means that the variation in parameter value(s) of repeated item across several sequential testing occasions or test levels will bring about violation in the stability of the parameter scale; it will pose threats to the fairness and validity of test scores thereby introducing trait-irrelevant differences that impact performance on the item over time. This threat will definitely amount to some after-effect that will jeopardize the fair interpretation of test scores.

Measurement error will also manifest in the testing process and when measurement error in scores produced by test instrument increases it will lead to reduction in test reliability. This in turn increases the potential for misclassifying candidates whose true scores fall at or near the passing score. This study's findings are consistent with those of Park, Lee, and Xing (2016), who found that examinee ability distributions between latent classes varied significantly over the course of the three administrations. They also suggested that item parameter drift should be carefully considered and assessed using a mixed IRT framework in order to comprehend the effects of drift on examinee ability and item parameters. The results were consistent with those of Xin (2018), who showed that, for the same set of common items, four distinct data dimensional compositions were used to confirm the possible effect of multidimensionality on item parameter drift. The result is also in tandem with the report of Kyung, and Fanmin, (2011) also reported that the short-term effect of item parameter drift on item calibration turned out to be very limited and fairly inconsequential under the studied conditions. the findings of the study however contradict with the finding of Albano and Rodriguez (2012) who reported that students may not likely remember their responses in previous test administrations, hence majority of the scales in measurement applications were maintained beyond test administrations.

Conclusion

Exposing students to the same items on various occasions of test administrations causes variation in the parameters of items, which negatively influences the quality of the items as well as lead to faulty assessment and decisions based on the outcome of the examination. It also bring about violation in the stability of the parameter scale and poses serious threats to the fairness and validity of test scores thereby introducing trait-irrelevant differences that impact performance on the item over time which have the capacity of jeopardizing the fair interpretation of test scores.

Recommendations

Based on the findings of the study, the following recommendations were made:

1. Policies that enable periodic determination of the parameters of items in item pool (bank) should be implemented so as to understand items that may potentially led to drift in parameters.
2. Test item security must be taken serious because one of the major causes of item parameter drift is question leakage. Serious care must be put in place to prevent students from seeing the questions before the actual testing.

References

Abolfaz G. & Erfan, H. (2018). ITEM Parameter Drift: Concepts, Methodology and identifying. Rooyesh-e- Ravanshenasi Journal (RRJ), 7(6), 163 – 182

Albano, A. D., & Rodriguez, M. C. (2012). Multilevel modeling of item parameter drift. Paper presented at the annual meeting of the National Council on Measurement in Education, Vancouver, BC.

Amery, D. W., Zhen, L., Siok, L. N., & Bruno, D. Z. (2008). Investigating and comparing the item Parameter Drift in Mathematics anchor items in TIMSS between Singapore and U.S.A. *TIMSS Technical Report*, *1*(6), 59 – 64.

Babcock, B., & Albano, A. D. (2012). Rasch scale stability in the presence of item parameter and trait drift. *Applied. Psychological. Measurement. 36*,565–580.doi: 10.1177/0146621612455090

Clark, A (2013). Review of parameter drift methodology and implications for operational testing. Paper presented at the Annual Meeting of the America Education Research Association, San Diego. June 13th 2013.

Gaertner, M. N. & Briggs, D. C. (2009). Detecting and addressing item parameter drift in IRT Test Equating Contexts. Retrieved from: https://www.researchgate.net/publication/32361 6531

Gaertner, M. N. & Briggs, D. C. (2009). Detecting and Addressing Item Parameter Drift in IRT Test Equating Contexts. Retrieved from: https://www.researchgate.net/publication/32361 6531

Guo, F., & Han, K.T. (2011). Potential impact of IPD due to practice and curriculum change in item calibration in CAT. *GMAC Research Report* RR 11-02.

Guo, F.,& Han, K.T. (2011). Potential impact of IPD due to practice and curriculum change in item calibration in CAT. GMAC Research Report RR 11-02.

Han, K. T. (2010). Simul CAT: Simulation software for computerized adaptive testing [computer program]. Retrieved March 20, 2010, from http://www.hantest.net/

Kyung, T. H & Fanmin, G (2011) Potential Impact of Item Parameter Drift Due to Practice and Curriculum Change on Item Calibration in Computerized Adaptive Testing. GMAC ® Research Reports, 11-02

Makransky,G.,Schnohr,C.W.,Torsheim,T.,andCurrie,C .(2014).Equating the HBSC family affluence scale across survey years: a method to account for item parameter drift using the Rasch model. *Q. Life Res.* 23, 2899 – 2907. doi: 10.1007/s11136-014-0728-2

Orheruata, M. U (2015). Item parameter drift of 2012 to 2015 WAEC and NECO SSCE Agricultural Science multiple choice items using item response theory. Ph.D Dissertation. University of Benin, Benin City, Nigeria.

Park, Y. S. Lee, Y. & Xing, K. (2016). Investigating the Impact of Item Parameter Drift for Item Response Theory Models with Mixture Distributions. Front. Psychol. 7:255. doi: 10.3389/fpsyg.2016.00255

Wei, X. & Meyers, J. P. (2013, April). Evaluation Of four robust z procedures for detecting Item parameter Drift in The 3PLM. Paper presented At the Annual meeting Of the National Council of Measurement in Education, San Francisco, CA.

Xin, L (2018). An Investigation of the Item Parameter Drift in the Examination for the Certificate of Proficiency in English (ECPE). Spaan Fellow Working Papers in Second or Foreign Language Assessment, 6, 1–28.