African Journal of Interdisciplinary Studies
Vol. 11, No. 1, pp 1-7, June 2018
Print ISSN: 0855-9724

Investigating flaw items of WASSCE Agricultural Science multiple choice items

Matilda U. Orheruata*[1] & Amen V. Uyigue[2]

[1 & 2] Dept. of Educational Evaluation and Counselling Psychology, Faculty of Education, University of Benin, Benin City, Edo State, Nigeria. *Email: mati.orheruata@uniben.edu*

Corresponding Author

Abstract

The study investigated flawed items of the West African Senior School Certificate Examination (WASSCE) Agricultural Science multiple choice items across 2012 to 2014 to determine the level of flaw in the item parameters across the stated examination years using Item Response Theory (IRT).Survey research design that adopted multistage sampling technique was used in selecting a sample of 3,744 of senior secondary three (SS3) Agricultural Science students from Edo South Senatorial District The instruments used were 2012 to 2014 WASSCE Agricultural Science multiple choice test items. The instruments (test items) were assumed to be valid and reliable by nature of standardized instrument administered by WAEC. The items were calibrated using EIRT computer software programmes to determine item difficulty (b), item discrimination (a) and guessing (c) parameter estimates. From these estimates, items with parameter value not within the IRT theoretical scale were flagged as flawed items for analysis. The results established that the condition of the items with flawed difficulty parameter estimates showed a percentage of 88.2, 50 and 71.4 pointed easier across 2012, 2013 and 2014 respectively. There was no significant difference between the flawed items and the standard items with regards to the item difficulty, item discrimination and guessing parameters across the years. On the basis of the, it becomes necessary that WAEC should ensure that as items are re-used or repeated, response parameters must be updated and made more accurate to acceptable criteria before use.

Key Words: Flaw items, Standard items, Item response theory, Item parameters, WASSCE

Introduction

Examination is a fundamental part of the teaching – learning process used not only as a basis for ranking students at the end of the teaching –learning process but to guide teaching, and aid in the development of curriculum, as well as in the assessment of needs, learning difficulty, level of mastery and differences among students. Large scale examinations that are used for certification decision are designed to measure knowledge acquired with reference to specified norms. Within a liable educational system that seeks to accurately capture students' performance and growth, the concern of item parameters to ensure theoretical stability becomes very necessary or else may introduce additional sources of error. Item parameters are statistical indicators that define the quality of an item in the instrument employed, otherwise known as psychometric properties (item difficulty; item discrimination and guessing parameter) of an item useful in item selection. The manual "Standards" for educational and psychological testing has specified that to ensure proper accountability, there is the need to conduct periodic checks of the stability of test items on which scores are reported (American Educational Research Association-AERA; American Psychological Association-APA & National Council on Measurement in Education-NCME, 1999). This has prompted researchers to evaluate scale stability in many large-scale assessments.

In large scale examinations, perhaps, of greatest importance is the validity of inferences that can be made from their test scores. In order to make valid inferences regarding examinee's ability, test scores must accurately reflect examinees' knowledge. Year after year, examining bodies award certificates that are assumed to be equal in academic abilities but when the item parameters of the items in their examination instruments suffer flaws, this assumption is therefore in error. A flawed item is operationally define as the violation indices of the item parameters of a test from the acceptable theoretical scale while items that within the bounds of the acceptable theoretical scale for item selection are regarded as standard items. The quality of the assessment instrument conducted is dependent on the psychometric properties of the assessment and the psychometric properties of test items are classified as either standard or flaw. Item flaws introduces systematic error of construct irrelevant variance to assessments, thereby reducing validity evidence of the test scores (Downing, 2004).Flaw items can have impact on examinees classification accuracy and could complicate the comparison being made of examinees performance over time if not checked. Therefore, as reported by Clark (2013);Wise and Kingsbury (2006) it is important to address test item flaws so as to avoid systematic errors that distort the inferences regarding the interpretation and use of test scores and to ensure that flaws are within bounds that are expected due to sampling error.

In Nigeria, one of the agencies external to the school system is the West African Examinations Council (WAEC). This body conducts the West African Senior School Certificate Examination (WASSCE) amongst other functions. The item pool of this examination body consists of a set of items in which the item parameters have been calibrated. However, as pointed out by Demars (2004), the item parameter in large- scale examinations could become less theoretically stable especially with testing programmes that rely on a large bank of items to select from when building assessments. Items with flawed item Parameter estimates threatens the fairness and validity of test scores, thereby jeopardizing the fair interpretation of test scores from year to year. When sizeable magnitude of such items exists in an achievement instrument, the amount of measurement error in scores produced by that instrument increases, thereby leading to reduction in test reliability. This in turn increases the potential for misclassifying candidates whose true scores fall at or near the passing score (Orheruata, 2015). If item parameters on WASSCE be found to exhibit such evidence of flaw, Inferences made from their test scores may be questionable.

One big change in the field of educational measurement under the influence of innovation is the new measurement theory for item selection known as item response theory (IRT). Item response theory is conceptualized as a paradigm for design, analysis and scoring of test, questionnaires and similar instruments measuring abilities, attitudes or other variables (Hambleton, 2000). IRT is an item-level focus theory that attempts to model the relationship between an observed variable, usually, conceptualized as an examinee's ability and the probability of the examinee correctly responding to any particular test item.IRT models the response of each examinee of a given ability to each item in the test. IRT is based on the idea that the probability of a correct response to an item is a mathematical function of person and item parameters. Where the person parameter is construed as a single latent trait or dimension while the parameters on which the items are characterized include their difficulty, discrimination and a pseudo guessing parameter (Alphen, Halfens & Imbos, 1994 and Ostini & Nering, 2006).The difficulty parameter defines how easy or how difficult an item is, the discrimination parameter shows how efficiently an item can distinguish between examinees with high and low test scores while the guessing parameter shows how likely the examinees are to obtain the correct answers by guessing. A variety of models have been developed from the IRT perspective and these models are known as one, two and three parameter models, The one - parameter logistic model (IPLM) usually called the Rasch model is the simplest IRT model for a dichotomous item and has only one item parameter (item difficulty). In this model, the probability of an examinee responding correctly or positively to item modeled as a function of an item parameter bi representing item difficulty, and a person parameter, θn representing the person's magnitude of the latent trait. The two–parameter logistic model predicts the probability of a correct response to any test for ability and two item parameters – item difficulty and discrimination parameter. The three–parameter logistic

model further adds a "pseudo-guessing" parameter with the intent of accounting for observed performance of these persons with very low levels of the latent trait.

Enu and Okwilagwe (2015) calibrated mathematics and geography items for JPSCPE in Nigeria. The study revealed only one item falling within the guessing parameter value in the geography test calibrated while a good number of the items were guessed. They concluded that the geography items were found to be simple at the level of the students that the mathematics items as could be inferred from the very high scores of the students in the geography ability. Bock, Muraki, and Pfeiffenberger (1998) investigated the stability of item parameter estimates in the 3- parameter logistic IRT model for College Board Physics achievement test over a period of ten years using Anova design. The study revealed that 21 of 29 items were flagged for evidence of parameter instability. Of these items, 10 became differentially harder, while 11 became differentially easier. The change in difficulty was attributed to a change in the focus of the physics curriculum across that span of time. The authors found that there was a statistically significant instability in item difficulty across time. The authors performed a similar analysis of the College Board English achievement test, and found no evidence of instability. Keller, Egar and Schneider (2010) on detecting item parameter drift in a science achievement test, the test reported no significant difference in drifted items across the three years studied neither across the three geographical locations studied. Kingsbury and Wise (2002) investigated the stability of item parameter estimates with 1-parameter logistic IRT model for 50 mathematics items and 40 reading items administered to students in 10 schools over 2 years. Results in their study showed no substantial evidence of instability and they concluded from their study that the measurement scales examined were stable across time though some items fluctuated noticeably from the original calibration which has no potential impact on classification accuracy.

To this end, this study investigated the flaws in the item parameter estimates of WASSCE Agricultural Science multiple choice test items from 2012 to 2014 examination years using the Three -Parameter Logistic Model (3PLM) of the Item Response Theory (IRT) with focus on students from Edo South senatorial District of Nigeria. Three- Parameter Logistic Model (3PLM) because it provides much more encompassing information about an item than the other two models. Also, the instruments for this study being multiple choice test item format, they are susceptible to guessing error and as such, the 3-PLM has the applicability to account for the effect of the guessing.

Research questions

In the course of the study, the following research question was raised:

1. What are the conditions of the flawed IRT item difficulty parameter estimates of 2012 to 2014 WASSCE Agricultural Science multiple choice test items?

Research Hypothesis

The following hypotheses were formulated to guide the study:

1. There is no statistically significant difference in the flawed IRT item difficulty parameter estimates of WASSCE Agricultural Science multiple choice test items across the years.
2. There is no statistically significant difference in the flawed IRT item discrimination parameter estimates of WASSCE Agricultural Science multiple choice test items across the years.
3. There is no statistically significant difference in the flawed IRT guessing parameter estimates of WASSCE Agricultural Science multiple choice test items across the years.

Methodology

The study adopted the survey research design because it relies on enquiring into data, with the intent of providing information about the conditions of the independent variables. That is, the item parameters under study were critically examined for instability at a given time. The population of the study comprised of all the Senior Secondary School three (SS3) Agricultural science students in the118 public senior secondary schools in Edo South Senatorial District. A sample size of three thousand, seven hundred and forty-four (3,744) senior secondary school three (SS3) Agricultural Science students was selected using the simple random sampling

technique. The instruments for data collection in this study were WAEC SSCE Agricultural Science objective paper (paper 2) of 2012, 2013 and 2014.The instruments for this study are standardized tests of West African Examination Council (WAEC) which have undergone the procedure of validation. As such the items were appropriate in terms of subject contents, instructional objectives and are reliable. Having retrieved all the response sheets from the sampled students, the response sheets were sorted out according to their examination years. Correct responses were scored dichotomously as "1" and incorrect responses as "0". The data were analyzed to determine the IRT parameter estimates (a_i, b_i and c_i) using the IRT statistical software: EIRT-Item Response Theory Assistant for Excel by Germain, Valois, & Abdous (2007). Items that did satisfy the IRT statistical conditions were flagged for flawed items. Furthermore, the items flagged in each item parameter and for each year were analyzed with Chi-square (X^2) test for independence.

Results

The results of the analysis as they relate to the research question and hypotheses are presented as follows:

Table 1: Percentage condition of the flawed item difficulty parameter estimates

Years	Flawed items	Easier (%)	Harder %
2012	17	88.2(15)	11.8(02)
2013	08	50(04)	50(04)
2014	14	71.4(10)	28.6(04)

Table 2 shows that, from the condition of the flawed item difficulty estimates, 88.2 percent of the flawed items tend to be easier and 11.8 percent tend to be harder for 2012 test. Equal percentages (50 percent each tend easier and harder in 2013 test while for 2014 test 71.4 percent tend easier and 28.6 percent of the flawed items were harder.

Table 2: Chi-square test of independence of the flawed items across years in the difficulty parameter estimates.

Parameter	Year	Flawed items	Standard items	Total	χ^2	Sig
B	2012	17(13.8)	43(46.2)	60		
	2013	08(13.8)	52(46.2)	60	4.844	.089
	2014	14(11.5)	36(38.5)	50		
	Total	39	131	170		

Expected frequencies in parenthesis

In Table 2, the Chi-square value of 4.844 was found to be non-significant ($p>0.05$). It therefore means that the null hypothesis is retained. This indicates no significant difference in the flawed item difficulty parameter estimates across the 2012 to 2014 examination years in the WASSCE.

Table 3:Chi-square test of independence of the flawed items across years in the discrimination parameter estimates.

Parameter	Year	Flawed items	Standard items	Total	χ^2	Sig
a	2012	2(1.8)	58(58.2)	60		
	2013	2(1.8)	58(58.2)	60	0.220	.896
	2014	1(1.5)	49(48.5)	50		
	Total	05	165	170		

Expected frequencies in parenthesis

The result in Table 3 showed the Chi-Square value of 0.220 and was found to be non-significant ($p>0.05$). It therefore means that the null hypothesis is retained. This implies there is no significant difference in the flawed item discrimination parameter estimates across 2012 -2014 WASSCE.

Table 4: Chi-square test of independence of the flawed items across years in the guessing parameter estimates.

Parameter	Year	Flawed items	Standard items	Total	χ^2	Sig
c	2012	8(6.4)	52(53.6)	60		
	2013	4(6.4)	56(53.6)	60	1.558	.459
	2014	6(5.3)	44(44.7)	50		
	Total	18	152	170		

Expected frequencies in parenthesis

Table 4 showed the Chi – square value of 1.558 and found to be non-significant ($p>0.05$). It therefore means that the null hypothesis is retained, hence there was no significant difference in the flawed guessing parameter estimates across 2012 to 2014 WASSCE.

Discussion

Findings from the condition of the flawed item difficulty parameter estimates given in Table 2 indicate that of the seventeen flawed items in 2012 test, fifteen of the items (88.2 percent) tend easier and fourteen of the flawed items in 2014 test, ten of the items(71.4 percent) tend easier while 50 percent of the flawed items tend easier in 2013 test. Thus, the resultant effect on examinees score for 2012 and 2014 test will be obvious as it will jeopardize fair score interpretation than the 2013 test that recorded equal flawed condition. This findings of the item difficulty flaws agrees with Bock et al., (1998) who investigated the stability of item parameter estimates in the 3- parameter logistic IRT model for College Board Physics achievement test and found that 21 of 29 items were flagged for evidence of parameter instability. Of these items, 10 became differentially harder, while 11 became differentially easier.

Generally, the condition of the flawed item difficulty parameter estimates of WASSCE showed that more of the flawed items were non –subtle and therefore reflect obvious violation of the well-established criteria of effective multiple choice item construction. This flaw in difficulty parameter is not surprising given that examinees in this studied subject and in this locality are presumably test wise. Guo and Wang (2005) and Michaelides (2010) queried that some examinees focus too much time and effort on test taking strategies rather than on skills and knowledge that the test will measure. As a result some items may become easier to examinees who practice specific type of test items simply due to familiarity with the items and not because they improve their proficiency in the tested skill. Test-wiseness in this regard can be blamed on lack of proper maintenance of the

item bank from which these items were drawn for the examinations. This finding is educationally important and need urgent attention in the measurement community where an examinee's score is a true reflection of his/her knowledge.

The test of the flawed item difficulty parameter estimate in WASSCE showed no significant difference across the years (Table 2). The non-significance implied that irrespective of the studied years, the flawed items of the difficulty parameter estimates did not vary from each other. Though in absolute values the flawed item difficulty parameter estimates were different across the years but statistically they were not. The finding agrees with the study of Keller, Egar and Schneider (2010) which reported no significant difference in unstable items across the three years studied neither across the three geographical locations studied. The finding is in disagreement with the studies of Kingsbury and Wise (2002) who investigated the stability of item parameter estimates with 1-parameter logistic IRT model for 50 mathematics items and 40 reading items administered to students in 10 schools over 2 years. Results in their study showed no substantial evidence of instability and they concluded from their study that the measurement scales examined were stable across time though some items fluctuated noticeably from the original calibration which has no potential impact on classification accuracy.

The test of the flawed item discrimination parameter estimate in WASSCE across the years showed no significant difference as presented in Table 3. This implies that irrespective of the years, the flawed item discrimination parameter estimates were statistically different across the years. In other words, the flawed item discrimination parameter estimates across the years is not time based. In absolute values, the number of item flaws in the discrimination parameter were very small across the years. Thus, indicating that the WASSCE Agricultural Science multiple choice test item discriminated well between the examinees with trait levels below and above the threshold across the years studied. The finding is an attestation that high discriminating power contribute more to measurement precision than items with low discriminating values (Nworgu & Ajah, 2012; Ojerinde, Popoola & Onyeneho, 2012).

The study further tested the flawed guessing parameter estimates of WASSCE to detect if differences exist in the flawed guessing parameter estimates across the years. The result showed no significant as presented Table 4. This suggests that the flawed guessing parameter estimates across the years studied were not time based. Generally, the absolute number of the flawed items were low, thus suggesting that the "weak" examinees were not easily influenced nor tricked to guessing the items. The findings is an indication that the "weak" examinees did not risk guessing, as a higher numbers of the flawed items across the studied years were of the easier class. The finding corroborates the studies of Enu and Okwilagwe (2015) which revealed only one item falling within the guessing parameter value in the geography test calibrated. This findings also confirm the report of Chernyshenko; Stark; Chan; Drasgow & Williams (2001)where they pointed out that difficult items or items with implausible distracters are more susceptible to guessing error. Based on the report of Chernyshenko et al., and also going by the findings of this research, it is therefore evident that the agricultural science multiple choice test items in the WASSCE of 2012 to 2014 used for this study were well-targeted, well- timed and distracters were effectively designed to reduce the flaws in the guessing parameter estimates.

Conclusion and Recommendations

The item pool of large scale examination bodies consist of a set of items in which the item parameter are have been calibrated. To ensure continuous quality, the calibrated items could be recalibrated with modern techniques such as item response theory to ensure that the items when reused are valid, reliable and interpretable. Based on the study, it was concluded that the condition of the flawed difficulty parameter estimates showed that88.2, 50 and 71.4 percent of the flawed items pointed easier in across the years. This could introduce trait-irrelevant differences on ability estimates and as such result in misclassifying candidates positively. With the no significant difference in the drift of the item difficulty, discrimination and guessing parameters tested in the examination types across the years, it is clear that the drifts were not peculiar to an examination year. It was therefore recommended that the workers of West Africa Examination council should ensure that as items are re-used or repeated, response parameter must be updated and made more accurate to stated criteria before use.

References

Alphen, V., Halphen, R. & Imbos, T. (1994). Likert or Rasch? Nothing is more applicable 8than good theory. *Journal of Advance Nursing.* 20,196-201.

American Educational Research Association; American psychological Association and National Council of Measurement in Education (1999). *Standard for Education and Psychological testing.* Washington, DC: AERA.

Bock, R.D, Muraki, E. & Pfeiffenberger, W. (1998). Item pool maintenance in the presence of item parameter Drift. *Journal of Educational Measurement* 25 (94), 275-278.

Chernyshenko, O. S., Stark, S., Chan, K., Drasgow, F & Williams, B. (2001). Fitting item response theory model to two personality inventories: Issues and sight. *Multivariate Behavior research.* 36(4), 523-526.

Demars, C. E. (2004). Detection of item parameter drift over multiple test administrations. *Applied Measurement in Education*, 17(3),265–300.

Downing, S.M. (2002). Construct-irrelevant variance and flawed test questions. *Academic Medicine* 7(7), 103-104.

Enu, V. O. & Okwilagwe, E. A. (2015). Calibration of Mathematics and geography items for joint command schools promotion examination of Nigerian army education corps in Nigeria. *African Journal of theory and Practice of educational Assessment*2, 40-60

Germain, S.,Valois, P. & Abdous, B. (2007). Eirt-Item response theory assistance for excel (freeware). Retrieved May 27, 2015 from http://libirt.sf.net

Guo, F. & Wang, L. (2005). Evaluating scale stability of a computer adaptive testing system. *GMAC Research Report*, RR-05-12.

Hambleton, R. K. (2000). Emergence of item response modeling in instrument development and data analysis. *Medical Care*, 38 (9 Supplement), 60-65.

Keller, L. A., Egan, K. L. & Schneider, M. C. (2010). *Item parameter drift in anchor items-detection and consequences: An analysis of simulated and operational test data.* CTB/McGraw-Hill: Monterey, CA.

Kingsbury, G. & Wise, S.L. (2002). Creating a k-12 adaptive test: Examining the stability of item parameter estimate and measurement scales. *JALT Testing Evaluation.* 12(2), 23-29.

Michaelides, P.M. (2010). A review of the effects on IRT item parameter estimates with focus on misbehaving common items in test equating. *Journal on Frontier in Psychology.* 1 (167), 167-171.

Nworgu, B. G. & Agah, J. J. (2012). Application of 3plm in the calibration of a mathematical achievement test. *Journal of Educational Assessment, 29* (2), 168-172.

Ojerinde, D., Popoola, K., Ojo, F. & Onyeneho, P. (2012). *Introduction to item response theory, parameter modeles, estimation and application.* Abuja: Marielouse Mike Press Ltd.

Orheruata, M. U. (2015).Item Parameter Drift in Certificate Examinations and it's implication on Decision Making. *African Journal of Theory and Practice of Educational assessment. Educational Assessment.*2, 98-105.

Ostini, R., & Nering, M. (2006). *Polytomous item response theory models:* Quantitative Applications in the Social Sciences, 144. Thousand Oaks, CA: Sage Publications.

Wise, S. L. & Kingsbury, G. (2006). Practical Issues in Developing and maintaining computerized adaptive testing program. *Psychological.* 2 (11),135-155.