GENDER BIAS IN 2020 USE OF ENGLISH MULTIPLE CHOICE EXAMINATION ITEMS OF SULE LAMIDO UNIVERSITY, KAFIN-HAUSA, JIGAWA STATE, NIGERIA

DR. MUHAMMAD GADDAFI SHUAIBU & DR. SANI AHMAD KATSAYAL
[1]Department of Education, Sule Lamido University Kafin Hausa, Jigawa State
[2]Department of General Studies, School of General Studies. Federal University of Transportation Daura, Katsina State.
muhammadgaddafishuaibu12@gmail.com
saniakatsayal@gmail.com

Abstract

This study investigated the gender bias in 2020 GSP 121 Use of English multiple choice examination items of Sule Lamido University Kafin Hausa Jigawa State Using Item Response Theory. Two objectives were raised, with two research hypotheses. Ex-post facto research design was employed and 518 student's responses were the total population of the study. Enumeration sampling technique was used as sampling technique. All the population of the study was used as sample of the study. The instrument used for this study was the responses of students in GSP 121 Use of English multiple choice examination of SLUK. The analyses were done using Item Response Theory PRO software and t-test independent sample. The findings revealed that no statistically significant gender difference exists in the parameter b and a regarding items on the 2020 GSP121 Examination. The recommendations given include that: School of General and Entrepreneurship Studies to publish standard textbooks, which will guide lecturers and at the same time provide the students with standard reading materials. The School of General and Entrepreneurship Studies (SGES) should constitute an examination evaluation team to encourage pilot testing of multiple choice items examination during the test construction stage prior to administering it.

Keywords: Gender Bias, Item Response Theory, Item Difficulty, Item Discrimination

Introduction

Information on test scores and the abilities of test takers at different levels has been a persistent problem in educational and psychological measurement. The measurement of individuals' traits or mental properties, such as abilities and attitudes, has been a long-lasting quest that dates back to 1882 with Galton's pioneering work developing rating scales and questionnaires and Thorndike's contributions to psychometric theory and its application to educational measurement. Currently, the measurement of students' academic achievement has a prominent position and influencing not only classroom practices but also testing at state and national levels. Measurement in educational settings serves several purposes, namely planning, monitoring and evaluation. Moreover, achievement data influence educational decision making. That in addition to improving teaching and learning, the information that test scores provides has great impacts in the classification, selection, placement and promotion of test takers (Adistutrisno, 2022).

One of the principal roles of a teacher in the education enterprise is to effectively instruct learners in order to bring about desirable changes in their behaviors. How does the teacher discover whether or not the desired changes in the learners' behaviors have actually taken place and in what direction? A way of ascertaining whether desirable changes in the learners' behaviors have been achieved is through processes that involved testing, measuring, assessing and evaluating.

Items in the test should be analyzed both qualitatively and quantitatively. Qualitative analysis is done

in terms of their content and form, including the consideration of content validity. It is done in terms of their statistical properties (Boopathiraj & Challamani, 2023). Basically, it includes the item difficulty and item discrimination. High reliability and validity can be ensured in a test in advance through item analysis. Item analysis makes it possible to shorten a test and at the same time increase its validity and reliability (Sidhu, 2012). Item difficulty refers to the level of the difficulty of the item of a test. This is determined by the number or percentage of the students who get the item right or wrong, and this is refers to item indices (Demars, Baschov & Socha. 2023). It should be noted that very low indices implies very difficult item, while very high indices implies very easy item. How difficult was each item to the students? If the difficulty indices to a student is <0.50 is refers to an easy items and the difficulty indices of > 0.50 is refers to difficult items (Baker, 2021). Did the test items adequately discriminate between high and low achievers? The discrimination indices of an item indicate how well an item discriminates between the strong and weak students. In this study an Item Difficulty parameter (b): Ranges from -3 to +3 was used in finding out difficulty in dices. And Item discrimination parameter (a) was based on 0.01 as item with no discrimination power, 0.01 – 0.34 as very low, 0.35 – 0.64 low, 0.65-1.34 moderate, 01.35- 1.69 high, > 1.70 very high, and +∞ as perfect discrimination indices.

Item analysis is about how difficult an item is and how well it can discriminate between the good and the poor students. In other words, item analysis provides a numerical assessment of item difficulty and item discrimination. Item analysis is the assessment of the essential qualities of the test items which helps in building reliability and validity into the test from the start (Zumbo 2019). Item analysis can be both qualitative and quantitative. Qualitative item analysis focuses on issues related to the content of the test e.g. Test analysis. Quantitative analysis involves measurement of item difficulty and item discrimination (Ebel & Fresbie, 2011). The outcome of item analysis helps the teacher to improve on item selection for the test by eliminating unreliable items, substituting for poor items, or by recasting poorly stated questions for better effect. The essential qualities normally considered in item analysis are item difficulty/easiness, item discrimination and distracter analysis (Anikweze, 2012). Borsboom, (2016) opined that item analysis provides three kinds of important information about the quality of test items. Item difficulty is a measure of whether an item was too easy or too difficult. Item discrimination is a measure of whether an item discriminated between candidates who knew the test well and candidates who did not. Clarity in the items involved no language error; whether basic rules of item writing are followed or not (Urbina, 2014). Quantitative- The psychometric properties of the items are assessed with the use of statistical procedures (Urbina, 2014).

According to DeMars (2010) difficulty is defined in both CTT and IRT in terms of the likelihood of correct response, not in terms of the perceived difficulty or amount of effort required. In CCT, the difficulty index, P, is the proportion of examinees who answer the item correctly (sometimes P-value). In IRT, the difficulty index, b, is on the same metric as the proficiency distribution in a designated group has a mean of 0 and standard deviation of 1. The item difficulty identifies to which about 50% of the examinees (or a little more, depending on the model) are expected to answer the item correctly. It is denoted by ‘p’ and the subscript is used to denote the item number, for example p1. The difficulty level of an item ranges from 0 (no one got the correct response) to 1 (everyone got the correct response) and is different for each item. The higher difficulty level items with high trait levels and lower difficulty items target lower trait levels thus helping in differentiating between the responses (Penfield, 2013). Therefore, optimal levels for difficulty generally are midway between 100 percent respondents scoring correctly and those expected by chance alone (Kaplan &Saccuzzo, 2009). The second item parameter is discrimination; the discrimination indices of an item indicate whether or not the item is measuring the same ability as the test measures. It shows how well an item discriminates between the

strong and weak students. It is a measure of correlation between the item and total test score. Like the coefficient of correlation, the DI ranges between -1.00 and +1.00 (Sidhu, 2012).

According Gao and Strokes, (2018) Item Response Theory (IRT) is one of the Latent Trait theories. It has three/four parameter models, namely:

1) 1-prameter model (also known as the Rasch Model) ascribes only the difficulty level of an item as the trait level required to correctly answer a question.

2) 2-parameter model deals with the discrimination parameter of an item in addition to the items difficulty parameter.

3) 3-parameter model gives the probability of an individual with ability, responding correctly to an item with a difficulty index, discrimination index and a guessing index. The model assumes that the three parameters (difficulty, discrimination and guessing) are necessary for an estimate a valid relationship between the probability of a correct response of an item and the trait level (ability) of an individual.

4) 4-parameter: The four-parameter logistic (4PL) item response model is a generalization of the 3PL model, which includes an additional parameter di that accommodates slipping effects.

As is true when comparing other statistical models, the choice of Rasch, 1PL, 2PL, or 3PL should be based on considerations of theory, model assumptions, and sample size. Because of its simplicity and lower sample size requirements, the Rasch model is commonly used in small-scale achievement and aptitude testing, for example, with assessments developed and used at the district level or instruments designed for use in research or lower-stakes decision-making (Vander & Linden 2016). The MAP tests published by Northwest Evaluation Association are also based on the Rasch model. Some consider the Rasch model most appropriate for theoretical reasons. In this case, it is argued that we should seek to develop tests that have items that discriminate equally well; items that differ in discrimination should be replaced with ones that do not. Others utilize the Rasch model as a simplified IRT model, where the sample size needed to accurately estimate different item discriminations and lower asymptotes cannot be obtained. Either way, when using the Rasch model, we should be confident in our assumption that differences between items in discrimination and lower asymptote are negligible (Rutkowski, et al., 2013).

The 2PL and 3PL models are often used in larger-scale testing situations, for example, on high-stakes tests, such as the GRE and ACT. The large samples available with these tests support the additional estimation required by these models. And proponents of the two-parameter and three-parameter models often argue that it is unreasonable to assume zero lower asymptote or equal discriminations across items (Foy & Yin, 2016).

There are four models in IRT based on a number of parameters, namely one parameter logistic model (1pl), two parameter logistic model (2pl), three parameter logistic model (3pl) and four parameter logistic model (4pl). These models differ from each other in at least two important ways. One important difference among the measurement models is in the terms of the item characteristics or the parameters that are included in the models. A second important difference among measurement models is in terms the response option format.

The origin of the General Studies programme in Nigerian universities is through the approval of a minimal standard for academic activities that the National Universities Commission (NUC) launched the programme into University's curricula. This was done in order to satisfy the longing for students in Nigerian Universities to be well grounded and perhaps well rounded as well in interdisciplinary studies so as to compare conveniently with their mates in universities in other parts of the globe.

Hence, Use of English was brought in as a finishing course in Nigerian Universities with the following

main objectives as outlined by National University Commission:

a) To offer students in all departments a sound foundation and functional mastery of the English Language in its various uses.

b) To breed able and inspiring users of English who can assert themselves as expected.

c) To enable students understand adequate approaches of organizing time, taking, organizing and developing notes.

Students' much success in the 2020 GSP 121 Use of English also motivated the researcher to investigate the causes of its much success. Use of English is among the compulsory courses that are offered under the General Studies Programme in Nigerian universities and other tertiary institutions due to incessant complaints by lecturers and employers of labor that many undergraduates, and even graduates themselves, lacked the ability to express themselves competently in the English language. It was rather shameful and actually disheartening that Nigerian graduates could not even write common application letters for employment in a language considered a national language (Nweye & Nwoye, 2016).

The study investigated gender differences in the two parameters i.e. b-difficulty and a-discrimination indices using IRT.

Statement of the Problem

The School of General and Entrepreneurship Studies experienced much success of students in the 2020/2021 GSP 121 (Use of English) Examination in Sule Lamido University Kafin Hausa, Jigawa State. Ideally, good examination items are expected to discriminate between high ability examinees and low ability examinees, failure of discrimination between the examinees by any examination items really affect the validity and reliability of the examination items. The result of the 2020 GSP 121 (Use of English) examination of SLUK showed that about 95% of the students that sat for the examination passed the course, and 20.2% got A, 27.8% got B grade, 32% got C, 14.6% got D grade and only 5.4% of the students failed. Consequently, it's against this background that the study analyzed the difficulty level and discrimination indices of the multiple choice test items of the 2020 (Use of English Examination) among students in different Faculties of SLUK Jigawa State.

Objectives of the Study

The study was guided by the following objectives;

1. To find out the influence of gender on b parameter on items of 2020 GSP 121 in SLUK, Jigawa State.
2. To find out the influence of gender on a parameter on items of 2020 GSP 121 in SLUK, Jigawa State.

Research Hypotheses

The research formulated two hypotheses for the study:

H01: There is no significant influence of gender difference in the b parameter on items of 2020 GSP 121 in SLUK, Jigawa State.

H02: There is no significant influence of gender difference in the a parameter on items of 2020 GSP 121 in SLUK, Jigawa State.

Methodology

The researcher employed ex post facto research design since the focus of this study was to analyzed a-parameter and b-parameter of 2020 GSP 121 (Use of English) of SLUK. The population of the study are 518 students that sat for 2020 GSP 121 examination, which are in two categories: The first category is the observation unit, which contains all the undergraduate students of SLUK who sat for 2020 GSP

121 Use of English. Second category of the population is the Unit of analysis which was the items in the GSP 121 Use of English of 2020. The researcher used the entire population of the study as sample of the study. This means the researcher used five hundred and eighteen (518) students as sample of the study. The researcher used census/enumeration sampling technique.

The data collection instruments of this study are the examination items of GSP 121 Use of English of 2020. It is Multiple Choice Examination set and administered by the School of General and entrepreneurship studies, SLUK, Jigawa State. This examination consisted of 60 multiple choice items, it also has five (5) options (A, B, C, D and E) and the administration time for this examination was two hours (2hrs). The analysis was computed through the means of IRT PRO 2.1 for windows and t-test independent sample.

Results

Hypotheses Tested

HO1: There is no significant influence of gender difference in b-parameter on the items of the 2020 GSP121 Examination in SLUK, Jigawa State.

Table 1: Result of t-test of Gender Difference in Parameter b of the 2020 GSP 121Use of English Examination of SLUK, Jigawa State

Item by Gender	N	Mean	S.D	Std. Error Mean	t-value	Df	P value	Remarks
Male	60	.18	4.015	.518				
					0.119	118	0.906	Not Sig.
Female	60	.28	5.142	.664				

An independent sample t-test was performed in examining difference in the gender difference of parameter b of the 2020 GSP121 Examination in SLUK, Jigawa State. From the Table above, the mean b parameter in respect to male was .18 and that of the female .28, respectively. The computed t value was 0.119 with sig value of 0.906 which is greater than the alpha value of 0.05, meaning that the obtained mean values do not significantly differ by gender. Based on the obtained result, the stated Null Hypothesis was therefore retained. Thus, it is concluded that no statistically significant gender difference exists in the influence of parameter b regarding the items on the 2020 GSP121 Examination in SLUK, Jigawa State.

HO2: There is no significant gender difference in a-parameter on the items of the 2020 GSP121 Examination in SLUK, Jigawa State.

Table 2: Result of t-test of Gender Difference in Parameter a of the 2020 GSP 121 Examination of SLUK, Jigawa State

Item by Gender	N	Mean	S.D	Std. Error Mean	t-value	Df	P value	Remarks
Male	60	.45	.746	.096				
					0.503	118	0.616	Not Sig.
Female	60	.55	1.346	.174				

An independent sample t-test was performed in examining the gender difference of parameter a of the 2020 GSP121 Examination between male and female candidates in SLUK, Jigawa State. From the Table above, the mean a parameter in respect to males was 0.45 and of the females was 0.55, respectively. The computed t value was 0.503 with sig value of 0.616, which is greater than the alpha value of 0.05 meaning that the obtained mean values do not significantly differ by gender. Based on the obtained result, the stated Null Hypothesis was therefore accepted. Thus, it is concluded that no statistically significant gender difference exists in the parameter a regarding items on the 2020 GSP121

Examination in SLUK, Jigawa State.

Discussion

Findings of Hypothesis one found no statistically significant influence of gender difference exists in the parameter b on the items of 2020 GSP 121 Examination in SLUK, Jigawa State. This finding contradicts the findings of Mustapha (2018) who carried out a study comparing the difficulty and discriminatory indices of boys and girls in the matching and completion test format of chemistry achievement test. The result of the analysis showed a is significant difference between the performance of male and female students in the completion test format of chemistry test; there is a significant difference between the performance in the matching test format of the chemistry achievement test.

The finding of Hypothesis two indicated no statistically significant influence of gender difference exists in the parameter a in the items of the 2020 the GSP 121 Examination in SLUK, Jigawa State. This is supports the findings of Rabi,u (2021) which revealed that items on the GSP 1201 Use of English multiple choice examination of the 2016/2017 Academic session of Bayero university, Kano in 2017 showed no significant differences in the mean discrimination indices of the items by GSP 1201 (Form A and B).

Conclusion

Based on the obtained result, it is concluded that no statistically significant gender difference exists in the parameter b regarding the items on the 2020 GSP121 Examination in SLUK, JigawaState; Based on the obtained result it is concluded that no statistically significant gender difference exists in the parameter a regarding items on the 2020 GSP121 Examination in SLUK, Jigawa State.

Recommendations

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

School of General and Entrepreneurship Studies to publish standard textbooks, which will guide lecturers and at the same time provide the students with standard reading materials.

The School of General and Entrepreneurship Studies (SGES) should constitute an examination evaluation team to encourage pilot testing of multiple choice items examination during the test construction stage prior to administering it.

References

Adisutrisno, W. D. (2022). Multiple Choice English Grammar Test Items That Aid English Learning for Students of English as a Foreign Language. https://doi.org/10.1037/edu0000396.

Ahmed, M. & Shaheed, E. (2018). Analysis of item difficulty and item discrimination as quality indicators of Physiology MCQ Examination at the Faculty of Medicine, Khartoum University.(Doctoral Thesis Khartoum University.)

Badamasi, N. T. (2018). An IRT Analysis of Psychometric Properties of the 2015 Kano State Mathematics senior secondary certificate Qualifying Examination in Dala Education Zone. (M.Ed. Dissertation, Bayero University Kano.)

Baker, F. B. (2021) Basics of Item Response Theory (2nd ed.). ERIC Clearinghouse on Assessment and Evaluation.

Bichi, M.Y. (2015). Introduction to research methods and statistics. Kano: Debis Co. Press and Publishing Co. Ltd.

Borsboom, D. (2016). Attack of the psychometrians. Psychometrika 71(3) 425-440.

DeMars, C., Bashcov H., & Socha R. (2023). Item Response Theory: Series in Understanding Statistics. Oxford University Press.

Ebel, R., L. & Freshbie, D. A (2011) Essentials of Educational Measurement (5th ed.). Prentice Hall of India Private Limited.

Gao, J. & Stokes, S.L. (2018). Bayessian IRT guessing models for partial guessing behaviours Psychometrika , 73(2).

Mustapha, S. H. (2018). Analysis of Items On 2014 Jigawa State Mathematics Qualifying Examination in Hadejah Education Zone. (M.Ed. Dissertation, Bayero University Kano State.)

Rabi,u A. (2021). Assessment of quality of general studies programme of GSP 1201 (Use of English) Examination of bayero University Kano using Item Response Theory. (M.Ed. Dissertation, Bayero University Kano State.)

Zumbo B. D. (2019). A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert-like (ordinal) item scores. Directorate of Human Resources Research and Evaluation. Ottawa, Canada.