ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)
The African Journal of Behavioural and Scale Development Research
AJB-SDR Vol. 5, No 1, 2023

Comparison of parameter logistic models of Raju differential functioning of dichotomous response items

Daniel Olutola Oyeniran
The University of Alabama

Abstract
The statistical strategies for detecting Differential Item Functioning (DIF) in the Raju model and other methods was developed due to bias in educational and psychology assessment. Most of DIF models are intended for comparing pre-defined focal and reference groups, such as female and male. A total of 480 students attempted 30 mathematics items from the five areas that have continuously been identified by West African Examinations Council (WAEC) chief examiners for the past ten years as challenging for the student to acquire appreciable results from the items related to them. The Raju method of DIF was utilized, and because it is based on Item Response Theory (IRT), the 1PL, 2PL, and 3PL models were compared to determine which items bias male and female samples. The results demonstrate that when 1PL is employed, two items have DIF, however when 2PL and 3PL are utilized, 26 items are biased. As a result, it is clear that those topics aren't as difficult as they appear, but students aren't scoring well because of the bias in those items. The bias detected in the items means that conclusions made from scores from the items are not reliable. Hence, it is essential that examination bodies ensure at least 2PL should be used to investigate bias and items that indicates bias towards a group is removed from the final test administered.
Keywords: *Item Response Theory, Raju, Differential Item Functioning, Parameter Logistic Model*

Introduction
The importance of examinations cannot be overstated in any academic setting where teaching and learning take place. As a result, examination is a broad phrase that encompasses written exercises, oral questions, and practical tasks that are designed to evaluate or assess a candidate's knowledge and skill after they have completed a specific task. The assessment entails both quantitative and qualitative descriptions of pupils' conduct, as well as a value judgment about the action's desirability. National examination evaluations are crucial in our educational system since they help calibrate grades for certification and provide indicators of educational quality, as well as for entrance to higher education institutions. English, mathematics, science subjects, commercial subjects, and technical subjects are all included in the national examinations. One of the goals of these national assessments is to make the assessment criteria as uniform as feasible across the country (Madu, 2012). Furthermore, the National Policy on Education (2014) said that national examinations should be as valid and equitable to all pupils as feasible. As a result, a good test should not contain biased items, as test bias occurs when a test or item generates systematic measurement errors (Schumacker, 2005).

It is necessary to conduct external examinations in the Senior Secondary School terminal class to achieve the goals of mathematics education. This goal is accomplished using various assessment forms, such as essays and objective tests. The (Mathematics) objective test, which is the subject of this study, is one of the evaluation tools used by the National Examinations Council to test or assess students' academic achievement (NECO). Students are asked and forced to select the best possible answer (or replies) from the options provided in objective assessments, such as multiple-choice questions (Okoro, 2006).

The usefulness of using multiple-choice items to assess examinees cannot be overstated because they can cover representative samples of the universe of the topic of interest without having to extend the testing duration. Because of its objectivity in assessing the examinees' responses, it is used to supplement constructed-response exams (Ayanwale, 2019). The evaluation of students' mathematical abilities through multiple-choice assessments is a common practice, as it allows for

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)

comprehensive coverage of course material by posing numerous questions (Okoro, 2006). To maintain competitiveness in the field of assessments, particularly in Nigeria, test developers must adopt innovative strategies for constructing these multiple-choice test items in Mathematics. It is imperative that these test items possess the necessary psychometric qualities to ensure that any achievement examination is both valid and reliable in measuring the intended learning outcomes (Ayanwale, 2019). There are many factors which could influence the reliability, validity, and fairness of test items among which include gender, race, social economic status among others.

Gender serves as a significant moderating factor within this context, as it has been recognized as an influential element affecting student achievement (Furner & Duffy, 2002; Workman & Heyder, 2020). However, the research findings regarding the impact of gender on student achievement have been somewhat inconsistent (Awofela, 2017; Daher et al. 2021; Hazari & Potvin, 2005; Laura, 2006). For example, in a study conducted and Iroegbu (1998), a noticeable gender effect was observed, with male students outperforming their female counterparts. This trend was also reflected in the findings of Quiaiser-Poul and Lehman (2002). However, Oladipo (2012) study showed a difference in academic achievement in basic science between male and female students in the junior secondary classes. In contrast, studies such as those and Arigbabu and Mji (2004) did not identify significant differences in cognitive, emotional, or psychomotor abilities outcomes based on gender.

Differential Item Functioning (DIF) has been described as a variation in the performance of test takers or survey respondents on a specific item, which is dependent on their group membership while being matched on a common latent trait (Abbas and Enayatollah, 2010; Finch and French, 2008). What is the difference in how items work for groups that are matched on similar ability. In other words, DIF arises when answers of individuals with the same ability of interest show systematic variances simply because of their membership in a particular group, such as geography, gender, or other characteristics (Ibrahim, 2016). Differential item functioning (DIF) is a concern to comparability that appears when one set of test-takers finds an item simpler than another after controlling for overall ability (Kahraman et al, 2009). The concept of DIF has emerged as a crucial component of test validation and fairness research. The definition of DIF for surveys, which are used to measure attitudes, is slightly different because respondents are matched based on their level of agreement rather than their ability (Dodeen, 2004). Inferential DIF detection methods are commonly used to identify DIF. To assess if an item has DIF, inferential DIF detection methods apply a significance test. The correct answer in the cognitive environment, according to Dodeen and Johansson (2003), is like the positive impacts of attitude toward the object. DIF can appear on items on ability and attitude tests for a variety of reasons. To discover the cause of DIF, item developers must examine the item with DIF.

Differential item functioning (DIF) can manifest in two distinct ways: uniform and non-uniform. In the case of uniform DIF, one group consistently outperforms the other across all levels of ability. Karami (2012) explains that in cases of uniform DIF, almost every member of one group performs better than members of the other group. On the other hand, non-uniform DIF occurs when the likelihood of answering a particular item correctly varies across skill levels for members of a group (Camilli & Shepard, 1994; Zumbo, 1999). To put it another way, there is a relationship between grouping and ability levels (Karami, 2012). Based on the test theory under consideration, there are two major groups of DIF. As a result, DIF can be based on either Classical Test Theory (CTT) or Item Response Theory (IRT). Some DIF approaches, such as Mantel-Haenszel, Lord, and others, are based on CTT, whereas others, such as Likelihood-Ratio Test, Raju, and others, are based on IRT.

One popular approach is the Raju method, which uses a regression-based method to identify DIF and estimate the effect size of DIF (Raju et al., 1995). In a study by Omorogiuwa and Iro-

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)

Agbedo (2016) using Raju approach show there were variations in performance between male and female examinees in seventeen items, which accounts for 34% of the total items. In contrast, there were no discernible differences in performance in 33 items, making up the remaining 66%. Among the seventeen items with differential performance, six items favored male students, while the remaining eleven items favored female students. The evaluation of the area between the item characteristic curves of the reference and focus groups is a commonly used method for detecting DIF in items (Magis et al., 2010).

Considering the Raju (1988) estimation, the formula below can be used to derive the DIF using 1pl, 2pl and 3pl.

$$1\ PL\ Area = 2D\ln(1+\exp(D(b_2 - b_1))) - (b_2 - b_1) \quad (1)$$

Holding the discrimination and guessing parameter constant in equation 1, b1: difficulty parameter for males (reference group), b2: difficulty parameter for females (focal group), D = 1.7 (constant: scaling factor)

$$2\ \text{PL Area} = 2\frac{a_2 - a_1}{Da_1a_2} In(1+\exp(\frac{Da_1a_2(b_2 - b_1)}{a_2 - a_1})) - (b_2 - b_1) \quad (2)$$

Holding the guessing parameter constant in equation 2, a1: discrimination parameter for males (reference group), a2: discrimination parameter for females (focal group).

$$3\text{PL Area} = (1 - c_1 - c_2)\ \left|2\frac{a_2 - a_1}{Da_1a_2} ln(1+\exp(\frac{Da_1a_2(b_2 - b_1)}{a_2 - a_1})) - (b_2 - b_1)\right| \quad (3)$$

While in equation 3, all the parameters are allowed to vary, c1: guessing parameter for males (reference group), c2: guessing parameter for females (focal group).

A zero value for the area between the curves indicates the absence of DIF, while increasing values indicate greater bias in the item (Lord, 1980; Raju, 1988). Various methods can be employed to measure DIF, including weighted and unweighted marked and unmarked area indices, marked, and unmarked area indices, and weighted and unweighted marked and unmarked area indices (Crocker & Algina 1986; Raju & Arenson 2002).

Research Questions

The following questions will be investigated in this study;

1. What is the dimensionality structure of the test?
2. Is there significant difference between the mathematics achievement score of males and females?
3. Are there any items that exhibit DIF between male and female test-takers? Using
   i. One-parameter Logistic Model
   ii. Two-Parameter logistic Model
   iii. Three-Parameter Logistic Model

Method

Design and Participants

The study used a survey research design, 480 students comprising 286 (59.6) females and 194 (40.4) males were selected from senior secondary school three class, the student selection is multistage from six local government areas of Oyo state. The state was stratified into three senatorial districts from which one senatorial district was randomly selected, then from the 11 local governments in the senatorial district five was selected and two schools were selected from each local government. Male serves as reference group in this study giving the previous evidence of male having more ability that female in numerical related field.

Instrument

The instrument used for the study comprises 30 items after trial testing 100 items which comprise Integration of simple algebraic functions (4 items), Proof of some basic theorem

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)

(8 items), Arithmetic of finance (6 items), Logical reasoning (7 items), and construction (5 items). A stem and a group of options make up a multiple-choice item. The stem of a multiple-choice test is a list of potential responses to a set of problems (questions) (the accurate response is considered the key, while the incorrect responses are referred to as distractors).

Analysis tools

The data was analyzed using RStudio open-source software version 4.0.2, which was released in June 2020, and the difR package version 5.1, which was released in June 2020. The validateR package was used to examine the Kuder Richardson 20 (kr20) reliability estimate, which yielded a good estimate of 0.84.

Results

Research question 1: What is the dimensionality structure of the test?

When it was determined that the students have covered that component of the curriculum breakdown as it should be studied on a term-by-term basis and in the class they are currently in, the accomplishment exam was given. It was looked at using the Raju DIF technique, which is based on IRT and meets the essential assumptions of unidimensionality and local item independence.

Furthermore, the Raju model, like many IRT models, is based on two fundamental assumptions: unidimensionality and local independence. The assumption of unidimensionality states that the set of items in the test measures only one underlying construct. That is, the test just looks at one aspect. The assumption of local independence states that an examinee's response to one question has no bearing on his or her response to any other item. As a result, the items must not provide any information on the proper response to another item.

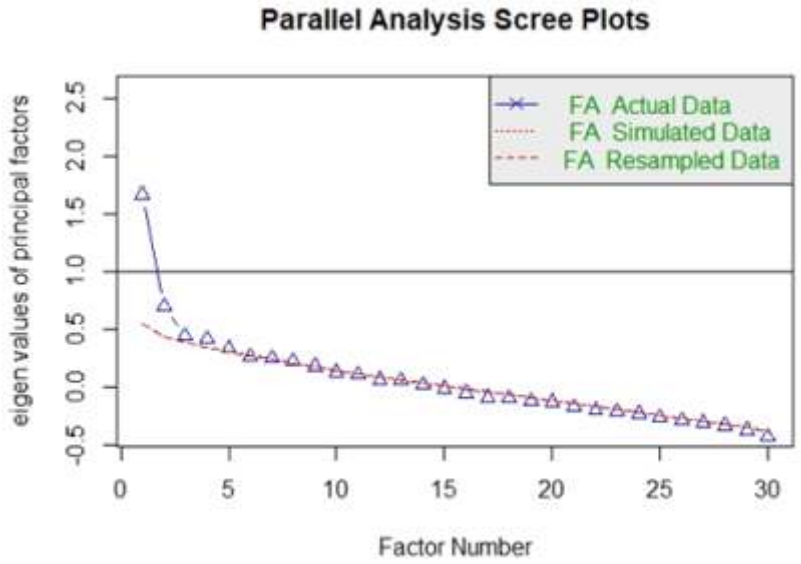

Figure 1: Scree Plot displaying instrument factor structure using Kaiser > 1 approach

The factor analysis minimal residual approach in RStudio using the Psych package was used to check for unidimensionality. What's needed is for there to be a single dominant component that explains the items' shared line of covariance (Hambleton et al., 1991). As a result, if the first extracted factor explains a substantially larger proportion of total variation than the secondary dimensions, unidimensionality will hold.

Research question two: Is there significant difference between the mathematics achievement score of male and female?

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)

Table 1: Independent Samples Test between male and female on Mathematics achievement

Independent Samples Test							
		Levene's Test for Equality of Variances		t-test for Equality of Means			
		F	Sig.	t	df	Significance	
						One-Sided p	Two-Sided p
Score	Equal variances assumed	0.012	0.913	0.825	478	0.205	0.41
	Equal variances not assumed			0.839	437.817	0.201	0.402

Levene's test is a statistical test used to assess the equality of variances across different groups or samples. In this case, the Levene's test produced an F value of 0.012 and a p-value of 0.913. The null hypothesis of Levene's test is that the variances male and female are equal. Therefore, a high p-value (greater than 0.05) indicates that there is not enough evidence to reject the null hypothesis, and we can conclude that the variances of male and female are not significantly different.

In this case, the p-value of 0.913 is much greater than the significance level of 0.05, which means that we fail to reject the null hypothesis. Thus, we can interpret the result as evidence that there is no significant difference in variances between male and female. Overall, the result suggests that the assumption of equal variances is met, which is important for t-test that require equal variances across groups.

The t-statistic value is 0.825 and the associated p-value is 0.205. The one-sided p-value is 0.41, and the two-sided p-value is also 0.205. The null hypothesis for an independent samples t-test is that the means of male and female are equal. The alternative hypothesis is that they are not equal. In this case, since the p-value (0.205) is greater than the significance level (0.05), we fail to reject the null hypothesis. This means that we do not have sufficient evidence to conclude that there is a significant difference in the means of male and female being compared.

Ensuring similarity of group is essential for estimating differential item functioning adequately which is the next research question.

Research question three: Are there any items that exhibit DIF between male and female test-takers? Using 1-PL, 2-PL and 3-PL model

With 30 items and 480 respondents, the Raju Differential Item Functioning (DIF) method of 1PL, 2PL, and 3PL analysis was performed on the same instrument. The 1PL consists solely of difficulty, the 2PL of difficulty and discrimination, and the 3PL of difficulty, discrimination, and guessing. The DIF detection threshold is between -1.96 and 1.96, regardless of the parameter logistic model. As a result, any item with a statistics value between -1.96 and 1.96, as well as a significant p-value (0.05), can be said to be biased or function differently between the target group and the reference group.

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)
The African Journal of Behavioural and Scale Development Research
AJB-SDR Vol. 5, No 1, 2023

Table 1: DIF analysis of items for 1, 2, and 3 parameters logistic model of IRT

	1PL		2PL		3PL	
Items	**Stat**	**P-Value**	**Stat**	**P-Value**	**Stat**	**P-Value**
1	-0.0932	0.9258	**25.8906**	**0.0000**	**2.7864**	**0.0053**
2	1.3956	0.1628	**-15.543**	**0.0000**	**2.2155**	**0.0267**
3	0.9659	0.3341	0.142	0.8871	-0.2446	0.8067
4	-0.4869	0.6263	**70.4**	**0.0000**	**6.4898**	**0.0000**
5	1.1475	0.2512	-0.6138	0.5393	0.4308	0.6666
6	-1.4057	0.1598	**-12.369**	**0.0000**	-0.6587	0.5101
7	1.3649	0.1723	**28.8685**	**0.0000**	0.9163	0.3595
8	-0.9527	0.3407	**41.0394**	**0.0000**	**5.0181**	**0.0000**
9	**2.2261**	**0.026**	**9.9073**	**0.0000**	-0.9537	0.3402
10	-0.8718	0.3833	**17.7656**	**0.0000**	1.9083	0.0564
11	0.3241	0.7459	**-14.313**	**0.0000**	1.6083	0.1078
12	0.747	0.455	0.0983	0.9217	-0.4540	0.6498
13	-1.0966	0.2728	**13.9401**	**0.0000**	1.5290	0.1263
14	0.545	0.5858	**22.3306**	**0.0000**	-0.8889	0.3741
15	0.3158	0.7521	**7.4995**	**0.0000**	-0.9212	0.3569
16	-0.4823	0.6296	**-11.102**	**0.0000**	0.8176	0.4136
17	1.4552	0.1456	**109.057**	**0.0000**	**-10.3290**	**0.0000**
18	-0.1257	0.9	**49.1285**	**0.0000**	**3.2365**	**0.0012**
19	0.0302	0.9759	**24.9028**	**0.0000**	**2.7347**	**0.0062**
20	-1.0556	0.2912	**71.2966**	**0.0000**	**-7.2725**	**0.0000**
21	0.5564	0.5779	**12.4244**	**0.0000**	-1.5351	0.1248
22	1.8109	0.0702	**111.631**	**0.0000**	**-10.3049**	**0.0000**
23	-0.3717	0.7101	**11.7431**	**0.0000**	-1.0790	0.2806
24	-0.0115	0.9909	**59.1744**	**0.0000**	**5.7674**	**0.0000**
25	-1.0628	0.2879	**83.2943**	**0.0000**	**4.8774**	**0.0000**
26	-1.3823	0.1669	**47.4883**	**0.0000**	**2.3750**	**0.0175**
27	-1.653	0.0983	**8.7315**	**0.0000**	**-3.0773**	**0.0021**
28	**-2.0565**	**0.0397**	**30.631**	**0.0000**	**2.4354**	**0.0070**
29	-0.644	0.5196	-0.8473	0.3968	0.4354	0.6633
30	1.0815	0.2795	**10.2162**	**0.0000**	0.9377	0.3484

Bold mean presence of DIF

Detection thresholds:- -1.96 to 1.96 (significance level: 0.05)

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)

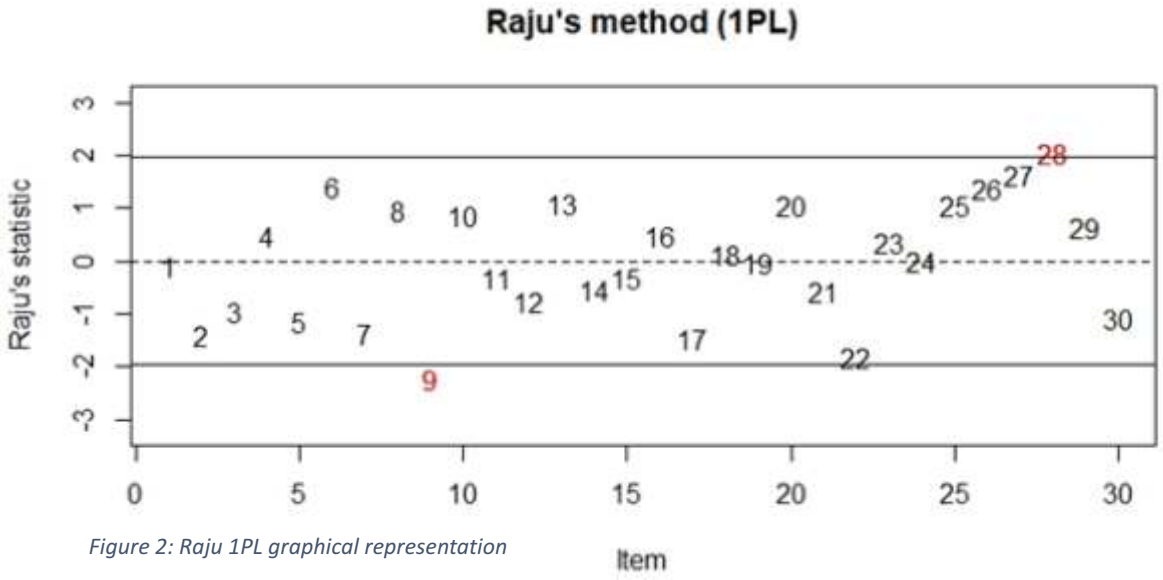

Figure 2: Raju 1PL graphical representation

illustrates the items that exhibit DIF when the Raju method's 1PL, 2PL, and 3PL models are employed. When 1PL was used, only two items (9 and 28) show DIF, whereas when 2PL was used, twenty-six items (1, 2, 4, 6, 7, 8, 9, 10, 11, 13, 14, 15, 16, 17, 18, 19, 20, 21, 22, 23, 24, 25, 26, 27, 28, 30) was flagged as having while for 3PL 14 items were flagged as having
DIF which are item 1, 2, 4, 8, 17, 18, 19, 20, 22, 24, 25, 26, 27, 28.
Figure 2 illustrates the objects that are within the detection threshold, indicating that they do not contain DIF, but those outside the threshold with the red hue have DIF when utilizing the 1PL Raju approach. Using this methodology, it is possible to find a significant number of good items that are not biased based on the gender of the test takers. With the 1PL model, only 2 items were flagged as having DIF, indicating that there may be some differences in item difficulty or discrimination between the groups being compared, but the differences are not significant for most items.

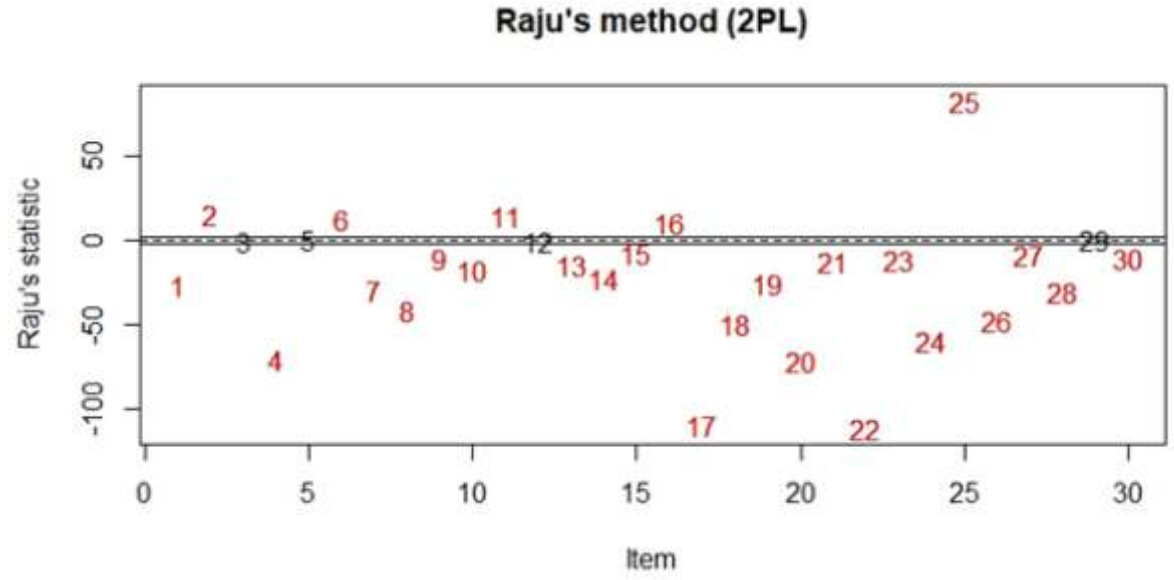

Figure 3: Raju 2PL graphical illustration

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)
The African Journal of Behavioural and Scale Development Research
AJB-SDR Vol. 5, No 1, 2023

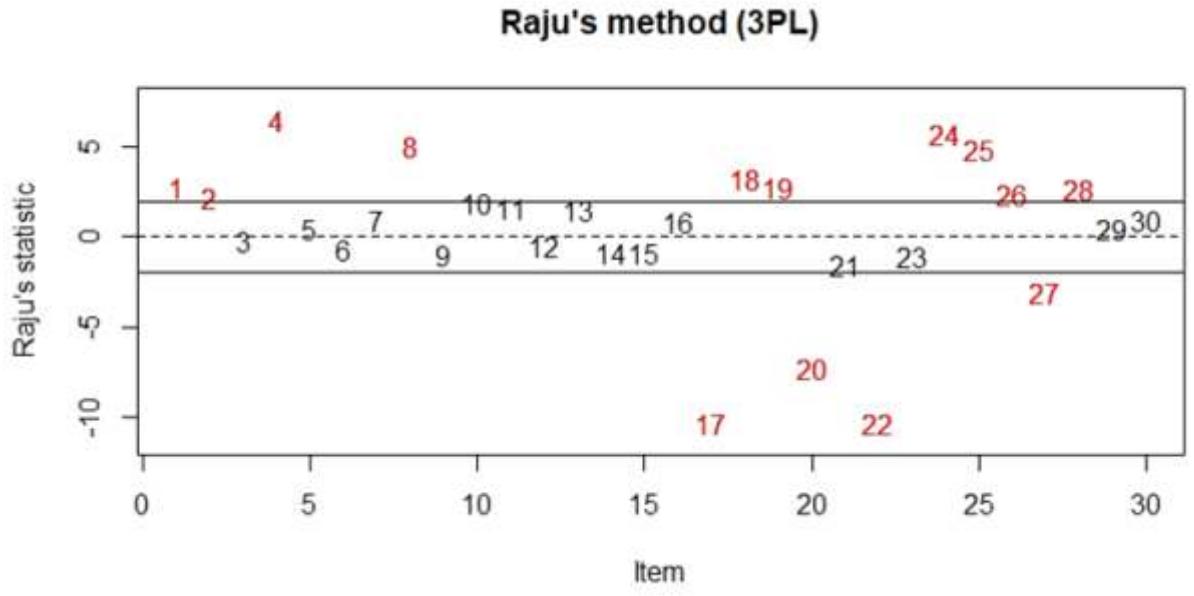

Figure 4: Raju 3PL graphical illustration

Figure 4 also demonstrates that 16 items do not have DIF while 14 items have. With the 3PL model, 14 items were flagged as having DIF, indicating that some of the differences observed with the 2PL model may be due to group differences in guessing or in the probability of a correct response at the lower end of the ability scale.

The results suggest that the 3PL model may provide a more nuanced and accurate picture of the sources of DIF than the 1PL or 2PL models, as it allows for group differences in guessing or in the probability of a correct response that may affect item performance differently across the ability range.

Discussion and Conclusion

The items that make up the instrument used to test the students are drawn from five primary themes that examination authorities in Nigeria regard to as challenging topics for students taking the exam. Those topics, on the other hand, are covered in class prior to the certificate class, where the final external examination is given.

The choice of item response theory (IRT) model can have an impact on how Differential Item Functioning (DIF) is detected and addressed (Choi, 2010). The 1-parameter logistic (1PL) model assumes that all items have the same discrimination parameter, which may not be realistic for many tests. The 2-parameter logistic (2PL) model allows for differences in item discrimination, and the 3-parameter logistic (3PL) model adds an additional parameter to account for guessing or careless responding. Meanwhile, the 2PL and 3PL models are generally considered to be better than the 1PL model for analyzing DIF due to their ability to account for differences in item difficulty and discrimination across groups (Embretson & Reise, 2000).

The 2PL and 3PL models may be better than the 1PL model, especially for DIF. The 2PL and 3PL models are more precise and flexible than the 1PL model and can account for more sources of variation in item performance. This can result in more precise measurement of individual differences in ability and better detection of DIF. Also, the 2PL and 3PL models can help to better detect DIF by allowing for differential item discrimination and guessing parameters. This can lead to more accurate identification of items that function differently for different groups of people. And the 2PL and 3PL models can better calibrate items by accounting for item-specific parameters such as discrimination and guessing, resulting in more accurate measurement of individual differences in ability, especially for difficult or less frequently answered items.

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)

Differential item functioning analysis, according to Camilli (2006), focuses on the performance of two or more diverse groups. As a result, such an investigation is unable to reveal the presence of bias against specific individuals. There has been conflicting information about how gender effects a student's test performance. Males generally do better than female in any component that is mathematically related, whereas females generally perform better in any part that is language usage based, according to most of the study. When 1PL was used, this was not well supported, but it was highly clear when 2PL and 3PL were used (Jackman & Morrain-Webb, 2019; Workman & Heyder, 2020).
Furthermore, the Raju method, which is commonly used for DIF analysis, is based on the 2PL or 3PL model and can provide more accurate estimates of DIF effects and impact than the 1PL-based methods (Raju, van der Linden, & Fleer, 1995). The study used the Raju Item Response Theory model of 1PL, 2PL, and 3PL models to investigate the differential item functioning of 30 Mathematics achievement exams administered to 480 students based on gender (male and female). The findings were compared to determine which items from each model had DIF.

The results show that two items (9 and 28) function differently between males and females when 1PL is used, but 26 items function differently between males and females when 2PL was used, while 14 items have DIF when 3PL was used as they do not fall within the expected range of -1.96 to 1.96.

Although the study does not explore reasons for differences in the number of items reported to possess DIF, however given that the only changes introduce in the analysis are the parameters it might suggest that 3PL provides more balanced result between 1PL and 2PL as it takes into consideration the possibility of guessing by the test takers. This shows that only a sophisticated differential item detection approach can identify the existence or absence of DIF in an instrument by checking the items more attentively. However, using 3PL will give further information regarding DIF in an item beyond what 1PL or 2PL reveals as 1PL underestimate by not considering discrimination and possibility of guessing while 2PL overestimate by excluding guessing parameter from the DIF test.

As a result, although 1PL only showed two items that distinguish between male and female, 2PL highlighted biases in 26 and 3PL show bias in 14 items of the 30 questions, which will account for disparities in the test's level of accomplishment between males and female. This is a significant topic to address when creating achievement items since it poses a threat and limits the scores of a group of pupils in comparison to one another. The gender groups had varied probabilities of approving the test items, according to DIF data. According to the DIF results, 26 of the 30 items had DIF-flagged items when utilizing the 2PL and 14 when using 3PL models. This suggests that test scores are not free of construct-irrelevant variance. Hence, it does not support the argument for construct validity.
The finding that the 3-parameter logistic (3PL) model outperforms the 2-parameter logistic (2PL) model in detecting Differential Item Functioning (DIF) is consistent with the results reported by Osterlind and Everson (2009). This supports the validity and reliability of the Raju method as a useful tool for detecting DIF in tests and improving the fairness and equity of test scores across different groups.

The superiority of the 3PL model over the 2PL model can be attributed to several factors. First, the 3PL model allows for the estimation of a separate parameter for the guessing probability, which may affect the item performance differently across the ability range and between the groups. This can help to identify and adjust for DIF items that may be related to guessing or careless errors rather than to real differences in knowledge or skill (de Ayala, 2009).

Second, the 3PL model allows for the estimation of a parameter for the upper asymptote, which may capture ceiling effects or saturation of item responses at the high end of the ability scale. This can help to improve the precision and accuracy of the estimates of item difficulty and discrimination, and to reduce the bias and error in the DIF detection (Embretson, & Reise, 2013).

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)

Third, the 3PL model is more flexible and adaptive to the data than the 2PL model, as it allows for the estimation of the item parameters and the latent trait simultaneously, and it can accommodate different item response formats and scoring rules. This can help to increase the sensitivity and specificity of the DIF detection, and to reduce the false positive and false negative rates (Chen, & Thissen, 1997).

Overall, the finding that the 3PL model provides more balanced result than the 2PL model in detecting DIF is a significant contribution to the field of educational and psychological measurement, and it has important implications for test developers, educators, policymakers, and researchers. The Raju method and the 3PL model can be used to improve the validity and fairness of tests, to reduce the potential biases and errors in the test scores, and to enhance the equity and access to educational and employment opportunities for diverse populations.

Recommendations

For ensuring fairness in test items between different groups taking such test it is essential to ensure such items are not biased and favour a group over the other to ensure comparability of the scores. Also, reviewing the item content and language may help identify any potential sources of DIF. The wording and response options provided for test items can affect the performance of different groups of people. Hence, reviewing the wording and response options for the items flagged as DIF may help identify any issues that could be causing the problem.

It may be helpful to conduct additional statistical analyses to identify the specific sources of DIF. This can help identify any specific item characteristics that are contributing to the problem and provide guidance for resolving the issue. Adding more test items that are specifically designed to assess the knowledge and skills of different groups of test takers may help improve the accuracy and fairness of the test. Overall, addressing DIF in a mathematics test requires careful consideration of the specific test items and the groups of people who are taking the test. By carefully reviewing the test items and conducting additional analyses as needed, it may be possible to identify and address the sources of DIF to improve the accuracy and fairness of the test.

Limitations and Suggestion for further studies

This study does not consider other factors which could contribute to DIF reported, hence it is essential to consider various factors which may lead to or contribute to reported Differential Item Functioning (DIF). Therefore, it is imperative to consider these factors in future research endeavors to gain a more comprehensive understanding of the efficacy of the Raju approach in detecting DIF. Additionally, does not provide a comparative analysis of the Raju approach with other Item Response Theory (IRT) methods that have been previously employed in the literature. Inclusion of such comparisons would be invaluable in assessing how the Raju approach performs relative to these other established methodologies.

References

Arigbabu, A.A. & Mji, A. (2004) Is Gender a Factor in Mathematics Performance Among Nigerian Pre-service Teachers? Sex Role, Nigeria. 51, 11 & 12.

Awofala, A.O.A. (2017). Assessing senior secondary school students' mathematical proficiency as related to gender and performance in mathematics in Nigeria. *International Journal of Research in Education and Science* (IJRES), 3(2), 488-502. DOI: 10.21890/ijres.327908

Ayanwale, M. A. (2019). Efficacy of Item Response Theory in score ranking and concurrent validity of dichotomous and polytomous response mathematics achievement test in Osun state, Nigeria. Unpublished Ph.D. Thesis. Institute of Education. The University of Ibadan.

Ayva-Yörü, F. G. and Atar, H. Y. (2019). Determination of differential item functioning (DIF) according to SIBTEST, Lord's 2, Raju's area measurement and, Breslow-Day

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)

methods. *Journal of Pedagogical Research, 3*(3), 139-150. http://dx.doi.org/10.33902/jpr.v3i3.137

Camilli, G., and Shepard, L. A. (1994). *Methods for identifying biased test items.* Newbury Park, CA: Sage.

Chen, W. H., and Thissen, D. (1997). Local dependence indexes for item pairs using item response theory. *Journal of Educational and Behavioral Statistics*, 22(3), 265-289. https://doi.org/10.2307/1165285

Choi, S. W. (2010). Differential item functioning analysis. In S. E. Embretson & S. L. Hershberger (Eds.), *The Wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test development* (pp. 709-735). Wiley.

Crocker, L., and Algina J. (1986). *Introduction to classical and modern test theory.* Orlando: Harcourt Brace Jovanovich Inc.

Daher, W., Essa Alfahel, E., & Anabousy, A. (2021). Moderating the relationship between students's gender and science motivation. *Journal of Mathematics, Science and Technology Education*, 17(5), 1-16. https://doi.org/10.29333/ejmste/10829

de Ayala, R. J. (2009). *The theory and practice of item response theory.* Guilford Press.

Desjardins, C. (2023). validateR: Psychometric validity and reliability statistics in R. R package version 0.1.0.

Embretson, S. E., and Reise, S. P. (2000). *Item response theory for psychologists*. Lawrence Erlbaum Associates.

Embretson, S. E., and Reise, S. P. (2013). I*tem response theory for psychologists. Psychology Press.*

Eren, B., Gündüz, T., & Tan, Ş. (2023). Comparison of methods used in detection of DIF in cognitive diagnostic models with traditional methods: Applications in TIMSS 2011. *Journal of Measurement and Evaluation in Education and Psychology*, 14(1), 76-94. https://doi.org/10.21031/epod.1218144

Furner, J. M & Duffy, M. L. (2002). Equity for all students in the new millennium: Disabling math anxiety. Intervention in School and Clinic 38(2) 67-74

Hambleton, R. K., Swaminathan, H., and Rogers, H. J. (1991). *Fundamentals of item response theory.* Newbury Park, CA: Sage

Hazari, Z. & Potvin, G. (2005). Views on female under-representation in physics: Retraining women or reinventing physics?, *Electronic Journal of Science Education*, 10(1). Retrieved June 20, 2008, fromhttp://wolfweb.unr.edu/homepage/crowther/ejse/potvin.pdf

Iroegbu, T. O. (1998). Problem based learning, numerical ability and gender as determinants of achievements problems solving line graphing skills in senior secondary physics in Ibadan. PhD. Thesis. University of Ibadan, Ibadan.

Jackman, W. M., and Morrain-Webb, J. (2019). Exploring gender differences in Achievement through student's voice: Critical insight and analyses, Cogent Education, 6:1, https://doi.org/10/1080/2331186X.2019.1567895

Karami, H. (2012). An introduction to differential item functioning. *The International Journal of Educational and Psychological Assessment*, *11*(2), 59-76.

Laura, A. R. (2006). Why are there so few female physicists? The Physics Teacher, 44, 177-180

Magis, D., Beland, S., Teurlinckx, F., and Boeck, P. (2010). A general framework and an R package for the detection of dichotomous differential item functioning. *Behavior Research Methods, 42*(3), 847-862.

Okoro, O. (2006). Principles and Methods in Vocational and Technical Education. Nsukka: University Trust Publishers

Oladipo, D. I. (2012). Gender difference in Nigerian junior secondary students' academic achievement in basic science. *Journal of Educational and Social Research*, 2(1), 93-99. Doi:

ISSN: 2714-2965 (Print)
ISSN: 2714-3449 (Online)

10.5901/jesr.2012.02.0

Omorogiuwa, K. O. and Iro-Agbedo, E. P. (2016). Determination of differential item functioning by gender in the national business and technical examinations board (NABTEB) 2015 mathematics multiple choice examination. *International Journal of Education, Learning and Development* 4(10), 25-35

Osterlind, S. J., and Everson, H. T. (2009). Differential item functioning analysis with 2- and 3-parameter logistic models: DIFdetect and difwithpar. *Applied Measurement in Education*, 22(4), 331-347. https://doi.org/10.1080/08957340903245435

Quaiser-Pohl, C. and Lehman, W (2002). Girls' spatial abilities: charting the contributions of experiences and attitudes in different academic groups. *British Journal of Educational Psychology* 72(2), 245-260.

Raju, N. S. (1988). The area between two item characteristic curves. *Psychometrika, 53*(4), 495-502.

Raju, N. S., and Arenson, E. (2002). Developing a common metric in item response theory: An area-minimization approach. Paper presented at the annual meeting of the *National Council on Measurement in Education*, New Orleans, LA.

Raju, N. S., van der Linden, W. J., and Fleer, P. F. (1995). IRT-based internal measures of differential functioning of items and tests. In K. F. Geisinger (Ed.), Psychological testing of Hispanics. *American Psychological Association*. 271-297

Revelle W (2021). psych: Procedures for Psychological, Psychometric, and Personality Research. Northwestern University, Evanston, Illinois. R package version 2.1.6, https://CRAN.R-project.org/package=psych

RStudio Team (2020). RStudio: Integrated Development for R. RStudio, PBC, Boston, MA URL http://www.rstudio.com/.

Schumacker, R. (2005). Test bias and differential item functioning retrieved on September 2021 from http://www.appliedmeasurementassociates.com/WhitePapers/TEST-Bias-and-Differential-Item-Functioning.pdf

Workman, J., and Heyder, A. (2020). Gender achievement gaps: the role of social costs to trying hard in high school. *Soc Psychol Educ*, 23, 1407-1427. DOI:10.1007/s11218-020-09588-6