INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN PROGRESSIVE EDUCATION & DEVELOPMENT

OPEN ACCESS JOURNAL

Vol 12, Issue 2, (2023) E-ISSN: 2226-6348

Detecting the Differential Item Functioning of Numerical Ability Test in the Gulf Multiple Mental Abilities Scale by Mental-Haenszel and Likelihood Ratio Test

Mohammed Al Ajmi[1], Siti Salina Mustakim[1], Samsilah Roslan[1], Rashid Almehrizi[2]

[1]Faculty of Educational Studies, Universiti Putra Malaysia, 43400 Serdang, Selangor, Malaysia, [2] College of Education, Sultan Qaboos University, Alhouz, Muscat, Sultanate of Oman

Corresponding Author Email: Email: mohd7010@gmail.com

To Link this Article: http://dx.doi.org/10.6007/IJARPED/v12-i2/17661 DOI:10.6007/IJARPED/v12-i2/17661

Published Online: 18 June 2023

Abstract

The current study aimed to examine differential item functioning (DIF) of a numerical ability test for Gulf state school students. This study examined the DIF items by gender and country using two DIF analysis methods; the Likelihood Ratio Test (LRT) and the Mantel-Haenszel (MH). The sample size was 2689 individuals throughout grades 5 and 6, and the researchers used MH with the SPSS and LRT with the BILOG-MG. The study used the classification stability coefficient kappa (κ) to compare how well the two methods agreed to examine DIF. Regarding gender, LRT yielded precise results for 30% of the items, and in terms of country, for 43.3%. Also, analyzing DIF with MH found that 26.7% of items exhibited DIF based on gender and country. For gender, there was strong concordance (0.925) between the MH method and the LRT. For the country, the MH and LRT agreement was also high (0.683). The study suggested investigating the causes of test items' differential performance and comparing the DIF in two test types, paper-and-pencil and computer-based.

Keywords: Differential Item Function, Numerical Ability, Coefficient kappa, Mantel-Haenszel Method, Likelihood Ratio Test.

Introduction

The study of cognitive abilities is one of the psychological developments in the twentieth century. Psychometric methods for measuring these abilities appeared at the beginning of 1904 when the Simon and Bennett scale of intelligence and other measures appeared. The measures used for that were built based on various definitions of cognitive abilities provided by psychologists (Al Nafouri, 2015). As a result, organisations concerned with educational and psychological assessments, such as the American Psychological Association and the European Union of Psychological Societies, have continued to require assessing cognitive abilities (Warnimont, 2010).

One of these tests used to assess cognitive ability is the Gulf Multiple Mental Abilities Scale (GMMAS). This scale is based on the premise that general mental ability is a multi-dimensional skill in three domains: verbal, numerical, and spatial. Higher cognitive mental processes manage mental activity by perceiving external world stimuli, remembering events, reasoning and analysing diverse situations, and inference (Alzayat et al., 2011).

Measurement scholars distinguish between two main approaches to analysing and grading items. The first approach: is the traditional theory of measurement, which is called the theory of true and false degree, as this theory is used to determine the factors that affect the degree that an individual obtains in the test, and the second approach is the item response theory that emerged as a result of criticism that guided the traditional measurement theory.

The traditional measurement theory provided solutions to some of the problems facing researchers in constructing and developing tests. However, it failed to solve other problems, as it assumes that the standard measurement error is equal for all subjects, and this assumption needs to be more accurate (Jabrayilov et al., 2016; Jumadi et al., 2023). The expression of an individual's ability is through the degree of the truth evident through his performance on the test as a whole and not at the level of the items. Therefore, the status of the individual's ability will change according to the change in the test level. Moreover, the test and the items change their characteristics with the change of the characteristics of the individuals, just as the characteristics of the individuals change with the change of the characteristics of the test in terms of difficulty and ease. Hence, the item response theory (IRT) came to overcome that IRT evaluates the teste's performance by employing the item as a measurement unit (Bichi et al., 2019).

Researchers' efforts have centred on establishing and developing tests to extract the effectiveness of items in terms of difficulty, discrimination, and guessing, whether in the CCT or the IRT. Despite their importance, these parameters are insufficient to judge the validity of the test items because the response on these items may be affected by other factors such as the bias of one group against another, not based on the ability of the examinee, but based on gender or socio-economic level, which negatively affects the accuracy of the results (Diaz et al., 2021). Thus, the item of the scale is described as biased. If a scale item differs between groups of persons of equivalent ability due to variables other than the measured attribute, the item has differential functioning (Aryadoust, 2018; Geramipour, 2020). Such requirements are significant in developing the scale and evaluating its fairness (Geramipour & Shahmirzadi, 2019), as well as a prerequisite for the construction of tests employed in decision-making since they affect the parameters of test items (Nawafleh, 2017).

Resolving item bias and test fairness issues is critical work in psychometrics (Diaz et al., 2021), and the presence of unequal performance in the tests is one of the threats to the internal validity of the test (Gómez-Benito et al., 2018). As a result, international organisations concerned with educational and psychological test preparation, such as the American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council for Measurement in Education (NCME), have deemed differential item functioning (DIF) a necessary standard when developing and publishing tests (Geramipour, 2020).

The differential item functioning gives a statistical indicator used to express the differences in the probability of correct response on the items between two subgroups of the respondents of a statistical population who have the same level of ability, one of which is called the reference group, while the other is called the focal group (Sayed et al., 2022). That is to say, the likelihood that two students with identical exam scores but different subgroups

(such as male and female or scientific and literary) will arrive at the exact correct response varies. Test validity is jeopardised, and comparing groups may be more difficult when items demonstrate DIF across groups. That is because their test results can indicate something different than what the scale is designed to measure (Krabbe, 2017). Hence, the present study aims to verify that the GMMAS numerical ability scale does not operate differently for different types of students (males against females and Oman versus the rest of the Gulf countries).

Literature Review

The issue of building scales and their fairness receives the attention of specialists in the field of measurement and evaluation due to the importance of the decisions that are taken in the light of the results reached from the application of scales, as there must be an acceptable degree of validity and reliability for those scales (Zakri, 2020). So, many scholars lately went on to investigate a crucial psychometric property for achieving fairness and equality in examinations, known as the differential item Function (DIF) (Abu Shindi & Kazem, 2018).

The detection of the differential item Function on the items of the psychological scales is essential because of the psychometric problems that result from it, represented in the fact that it may be an indicator of the bias of these items and thus leads to the failure to achieve the fairness of the test. It may also affect the validity and reliability of the tests. Zakri (2020) confirmed that the test, in order to be fair, should not be biased against or in favour of any group of test subjects due to their differences in gender, race, country or otherwise. That is, no group has a preference over other groups in answering the scale's items except concerning their possession of the measured trait or ability.

To determine whether DIF should be deleted or revised, researchers must collect information about the causes of DIF occurrence. If the information obtained about DIF items is limited, DIF items cannot be appropriately processed. There were cases where researchers decided to remove DIF-marked items from the item bank, while in other cases, researchers decided to dig deeper into the data (Huang et al., 2012). However, Cho et al (2016) revealed that 30% of studies eliminated the DIF items, and 26% ignored them after examining 27 studies on DIF item treatment.

The degree of the differential functioning is categorised as small, medium, or large based on different indices of effect size employed to detect it. In the event of small differential functioning, no action is usually taken, while in the instance of significant differential functioning, the item should be deleted or revised (Al Sawalmeh & Al Ajlouni, 2019).

Given the possibility that the presence of differential performance of the items threatens the validity of the test, its detection and removal are necessary for a good measurement that is free from bias in the items (Salman & Thatha, 2022). Hence, many statistical methods emerged to detect differential item functioning, including those that depend on the classical measurement theory, such as the Transformed Item Difficulty Method (TID), Analysis of Variance ANOVA, Item Discrimination Method (IDM), and the method of camouflage analysis. In addition, methods that are based on the item response theory (IRT), like the Item Characteristic Curve (ICC), b - Parameter Difference Method, and Likelihood Ratio Method (Almaskari & Almehrizi, 2021; Zakri, 2020; Oalla & Matarneh, 2018). However, there is a scarcity of studies that attempted to identify the presence of differential item functioning in the scale that was built in the Arab environment, which calls for more research efforts to cover this gap, and this is what the current research aims to achieve in revealing the

differential item functioning of numerical ability items in the GMMAS scale of mental abilities concerning gender (male and female) and country (Oman, the rest of the Gulf countries).

The current study relied on two methods to detect the existence of the DIF. The first method is the Likelihood Ratio Test (Cohen et al., 1996; Thissen et al., 1988) which is based on the item response theory (IRT) of measurement. A likelihood ratio test can be conducted to examine the potential for bias between two groups (reference and focus) by calibrating the data as one group (Al Ajmi et al., 2022). According to the chosen IRT model, the likelihood ratio test method can deal with dichotomous and polytomous data (Yildirim, 2006).

The Second method is the Mantel- Haenszel Chi-Square (Mantel & Haenszel, 1959; Holland & Thayer, 1988). It is one of the most widespread methods of traditional theory in detecting differential item functioning due to the ease of its calculations and procedures. It is based on comparing the averages of the groups in performance on the test as a whole or its equivalent items by examining the bias between two groups, one of which is called: the reference group and the other is called the focal group, which is the group affected by items bias (Fidalgo & Madeira, 2008). The estimate of the Mantel-Haenszel statistic is determined using the following equation, which requires a square binary matrix containing the number of persons who answered correctly and wrongly to the item from the two groups.

$$MHx^2 = \frac{(|\sum A_t - \sum E(A_t)| - 0.5)^{\ 2}}{\sum var(A_t)}$$

(A_t): the number of members of the reference group who answered the item correctly at the ability level t. Where E(At) is the expected value of the At. It is calculated from the following equation:

$$E(A_t) = \frac{(N_{Rt}\ N_{Ft})}{N_t}$$

Where N_{rt} is the number of individuals who answered the item with the same ability level t in the reference group, N_{ft} is the number of individuals who answered the item with the same ability level t in the focal group, and N_t is the number of individuals who answered the item, with ability level t. Similarly, var (A_t) is the variance of A_t and calculated from the following equation,

$$var\ (A_t) = \frac{N_{rt}\ N_{ft}\ N_{1t}\ N_{0t})}{N_t^2(N_t - 1)}$$

N_{1t} is the number of individuals who answered the item correctly from both groups at the ability level t, and N_{0t} is the number of individuals who could not answer the item correctly from both groups at ability level t (Al Ajmi et al., 2023).

Statement of Problem

It is noted that when applying tests of all kinds, different results appear for each test on the groups targeted by that test, which raises several questions about whether this discrepancy in scores is due to the nature and levels of the trait to be measured or is it due to the nature of the test and the characteristics of its items, as some of the items may include a differential item functional that makes him confused, which would threaten the test's validity and reliability (Salman & Thatha, 2022).

In addition, the items' properties must remain consistent across various examinees. For example, some items may favour males over females, and such items must be identified (and possibly eliminated) to ensure fair measurement (Magis et al., 2017).

The Gulf Multiple Mental Abilities Scale (GMMAS) is one of the newest and most essential measures in the Gulf Cooperation Council countries and its developments by (Alzayat et al., 2011). The scale's premise is that intelligence manifests differently in the linguistic,

quantitative, and spatial spheres. Higher cognitive thinking processes control all mental activities; they include taking in information from the outside world, remembering past experiences, and engaging in abstract thought processes like analysis and inference. The scale is based on the premise that general mental ability is a multi-dimensional ability that manifests itself in three areas: verbal, numerical, and spatial. Higher cognitive mental processes govern mental activity, represented by perceiving external world stimuli, remembering events, reasoning and analysis of various situations, and inference.

Alzayat et al (2011) found that there were disparities in the numerical ability of fifth and sixth-class students in the Arab Gulf countries according to the grade and country in the three levels of the GMMAS scale, and the existence of these differences may indicate the existence of differential functioning of the scale items. Given the Gulf scale's widespread use in assessing quantitative skills and guiding diagnostic choices, it is crucial to look for evidence of differential functioning between the sexes and between Oman and the rest of the Gulf countries. Given the Gulf scale's widespread use in assessing quantitative skills and guiding diagnostic choices, it is crucial to look for evidence of differential functioning between the sexes and between Oman and the rest of the Gulf countries. Thus, the following questions define the research problem

1. Which items in the GMMAS numerical ability test demonstrate differential item functioning based on gender and country using the likelihood ratio test method?
2. Which items in the GMMAS numerical ability test demonstrate differential item functioning based on gender and country using the Mantel-Haenszel method?
3. How well do the likelihood ratio test method and the Mantel-Haenszel method agree in finding differential item functioning in the GMMAS numerical ability test according to gender and country variables?

Significance of the Study

The importance of the study emerges theoretically through shedding light on the differential item functioning (DIF) of the numerical ability test items in the GMMAS. It is expected that revealing the DIF of the items of the numerical ability test will contribute to improving the validity and fairness of the test. It will help the authors of the test to improve the psychometric properties of the test through the presence of several indicators and evidence that improve the process of formulating good items and review the differential item functioning of them, and thus the validity of the test and the fairness of interpretation of its results and increase confidence in them. This study also contributes to the rare national efforts to explore differential functioning and bias in psychological scales used in Arab countries, which may encourage researchers to carry out such studies. Practically, it is hoped that this study will provide statistical data through the use of the likelihood ratio and the Mantel-Haenszel methods in detecting the DIF of the numerical ability test according to some variables (gender and country), which would help workers in the field of constructing items to detect about the differential performance of the items to exclude or modify the items that show DIF from the test items. Hence, the present study aims to verify that the GMMAS numerical ability scale does not operate differently for different types of students (males against females and Oman versus the rest of the Gulf countries).

Method

Participants

This work employs quantitative research utilising a descriptive approach to characterise the statistical characteristics of the numerical ability exam in GMMAS employing differential item functioning. The researcher relies on secondary data acquired from the Arab Office for the Gulf States' GMMAS standardisation in 2011. This sample was collected in the fifth and sixth grades, with pupils ranging in age from nine years and three months to twelve years and three months. The sample size was 2689 people, with 1273 females and 1416 males. The student's ages range from nine and three months to twelve and three months.

Instrumentation

The study uses the numerical ability test in GMMAS prepared by (Alzayat et al., 2011). It consists of three tests measuring verbal, numerical, and spatial abilities. This study focused on the numerical ability test consisting of 30 items of a multiple-choice type. Numerical ability is measured by counting, addition, subtraction, multiplication, division, numerical relations, numerical reasoning and arithmetic problems. The correct answer is given one score, while the wrong answer is given zero, so the total score ranges between 0 and 30.

The predictive validity of the numerical ability test was confirmed by calculating the correlation coefficients between numerical ability and achievement in mathematics at all levels in the State of Kuwait only. The correlation coefficient between numerical ability and academic achievement in mathematics in the fifth grade came with a value of 0.63, which is statistically significant at the level of significance of 0.05. The correlation coefficient between numerical ability and academic achievement in mathematics in the sixth grade came with a value of 0.38, which is statistically significant at the level of significance is 0.05. We note that these values were statistically significant, despite the small size of the samples within each academic. The Raven successive matrices also confirmed the construct validity of the numerical ability test. The results showed that the correlation coefficients are positive and statistically significant, indicating the validity of the test construction (Alzayat et al., 2011).

The test re-test reliability coefficient for numerical ability was 0.89. Internal consistency was high for numerical ability across all grade levels, with Cronbach alpha coefficients ranging from 0.82 to 0.87 for Gulf countries (Alzayat et al., 2011).

Study Procedures

Verification of the assumptions of item response theory

The assumptions of the item response theory in the third-level numerical ability test of the GMMAS were verified as follows

Unidimensionality Assumption

Unidimensional means that the test items measure one trait that explains the individual's performance on the test items. To verify the unidimensional assumption, the following was done:

Exploratory Factor Analysis (EFA)

The Kaiser-Mayer-Olkin (KMO) and Bartlett's tests confirmed the sample size's adequacy for exploratory factor analysis. The estimated chi-square value was (10426.066), a function at the level (0.001) and degree of freedom (435), confirming the test's unidimensionality. After that, the 30 items on the numeric ability scale were subjected to exploratory factor analysis using

the principal components of the correlation matrix. The analysis revealed four latent root factors with eigenvalues greater than one, which account for 34.99% of the total variance. Dividing the first eigenvalue (5.48) by the second eigenvalue (1.57), which equals 3.49 and is greater than 2, is a sign of unidimensionality (Reckase, 1997, cited in Oalla, 2015). The first component accounts for 52.17 percent of the total explained variation. Given that it achieves the 20% threshold established by Reckase (1979), it means calling this a unidimensional test (cited in Lee, 2004). In addition, using Cattell's scree plot test (1966) for the 30-items factor analysis. Figure 3 shows distinguishing the first factor from the others and ensuring the test is unidimensional.

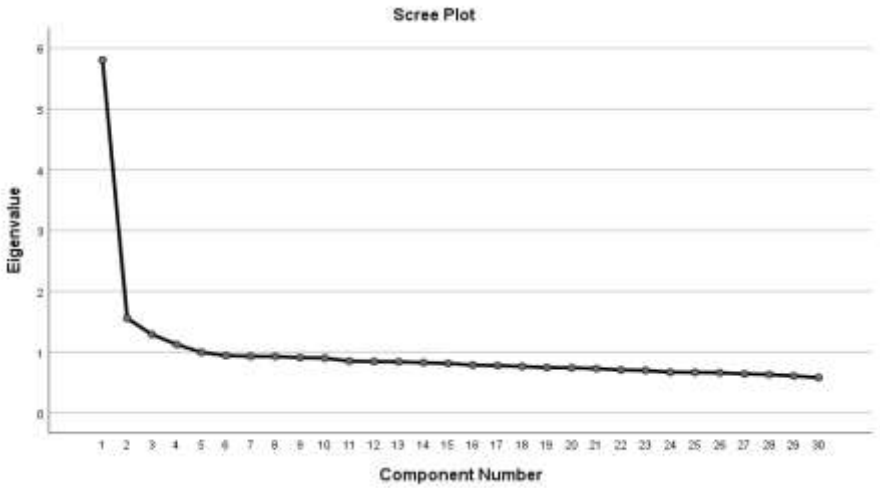

Figure 3. Factor scree plots from principal component analysis of 30 items

Confirmatory factor analysis (CFA)

The AMOS programme was used to calculate the Root Mean Square of Residuals (RMSEA) and Tanaka Index (GFI) as an additional indicator of the data's conformity with the unidimensionality assumption. The results reveal that the Root Mean Square of Residuals (RMSEA) is equal to.036, which satisfies the criterion established by Browne and Cudeck (1993) that an RMSEA of.05 or less indicates a good fit. In addition, the value of the GFI is (0.95), which fits the criteria established by (Tanaka and Huba, 1985).

Local Independence

The second assumption is local independence (LI), which is described by Hambleton and Swaminathan (1985) as the idea that a person's performance to test items of the same ability is statistically independent and should not affect their replies to other items. Local item independence implies unidimensionality, as proved by (Hambleton & Swaminathan, 1985; Allam, 2005). This means that if a scale ensures unidimensionality, it also assures local item independence. However, the researcher has used Yen (1993)'s statistical indicator, the correlation coefficient between the residuals for a pair of items after modifying the individual's ability, to verify LI. The computer program for Local Dependence Indices for Dichotomous Items (LDID) was used to verify the assumption of the numerical test's local independence. It is common to use a uniform critical value of 0.2 for the absolute value of Q3 (Chen & Thissen, 1997; Kim et al., 2005).

In this case, the results showed that Q3 values were generally smaller than (0.144), which indicates the assumption of local independence being met among the test items. Moreover,

the results show that all possible pairs of items on the numerical ability test were truly independent, proving that test takers' answers achieve local independence.

Freedom from speed

The researcher allowed sufficient time to complete the items, so that lack of ability rather than a time constraint was the determining factor in poor performance. No one who took the exam expressed concern about the time limit. The researcher further validated the freedom from speed by calculating the percentage of students who completed all test items, which was 100%, confirming the assumption of speed performance.

Psychometric properties of the numerical test according to the IRT

Choosing the Model

To figure out which unidimensional logarithmic models fit the test data best, the -2 log Likelihood (−2LL), the Akaike Information Criterion (AIC; Akaike, 1974), the Bayesian Information Criterion (BIC; Schwarz, 1978), the Index of the values of the information function, and the Root Mean Square Standard Errors of Estimates (RMSE) were used as criteria to determine that. Table 10 shows the values of the model fit indices for choosing the appropriate model for the numerical ability test data.

Table 10

The values of the indicators for choosing the appropriate model for the numerical ability test data.

S	Indicators	Model		
		1PL	2PL	3PL
1	-2 log Likelihood	97253.2598	96800.9758	96539.1533
	Model Differences		452.284*	261.822*
2	AIC	97280.4	96886.8	96590.5
3	BIC	97463.2	97240.6	97121.2
4	Average Test Information	5.192	5.61	6.275
5	RMSE	0.4071	0.3955	0.4355

Note: IPL, one parameter logarithmic model; 2PL, two parameter logarithmic model; 3PL, three parameter logarithmic model; -2LL, -2 log-likelihood; AIC, Akaike's information criterion; BIC, Bayesian information criterion; RMSE, root mean Square Errors

It is clear from table 10 that the best appropriate model for the numerical test data is the three-parameter logarithmic model (3PL), which takes difficulty, and discrimination and guessing parameters into account.

Reliability of Numerical Ability Test

Three test reliability coefficients were extracted according to the item response theory:

- *Test information function:* It shows how reliably the test's items are used to judge the test-taker's abilities. The more information the test provides, the better it can evaluate the assessed trait. With an ability level of -0.125, the maximum value of the information function on the numerical ability test is 6.869, with a standard error of 1.25. This indicates that the average ability level is where the test information function is most informative.
- *Test Reliability coefficient:* The reliability coefficient measures how consistent an individual's estimations of their abilities concern the trait being assessed; it is derived from the variation

of those estimates relative to the mean function of the data obtained from the assessment. There was a high level of reliability in the test's ability estimates, with an index of 0.851.

-*The empirical reliability of the test:* Empirical reliability is defined as the ratio of the error variance of individuals' ability estimates to the variance of individuals' ability estimates. It is thus a measure of how closely the estimated ability using the response theory models corresponds to the actual ability of individuals. The exam was shown to have a reasonable degree of reliability for estimating individuals' abilities, with an empirical reliability coefficient of 0.847.

Results

1. Which items in the numerical ability test show differential item functioning based on gender and country using the likelihood ratio test method?

Table 2 presents the difference in likelihood ratio between the reference and focal groups and their standard errors of the estimate for the numerical ability test to the gender and country variable.

Table 2 shows that nine out of thirty items (30%) have DIF based on gender, with three items (3, 5, and 27) favouring females and four items (11, 14, 17,18, 22 and 28) favouring males. Table 2 showed that 13 items (43.3%) had DIF based on country, with DIF against Oman in items 3, 6, 17, 19, 26, and 27, and DIF in Oman's favour in items 1, 9, 11, 12, 14, 15, and 28.

Table 2
Likelihood ratio test and their standard errors for the numerical ability test according to gender and country variables

	Gender		Country	
ITEM	Estimate	SE	Estimate	SE
1	0.22	0.161	0.637*	0.271
2	0.126	0.103	-0.031	0.14
3	0.313*	0.09	-0.281*	0.118
4	0.034	0.095	-0.042	0.191
5	0.376*	0.084	0.064	0.125
6	0.166	0.093	-0.206*	0.14
7	-0.039	0.087	0.24	0.152
8	-0.103	0.069	0.061	0.123
9	0.069	0.073	0.244*	0.125
10	0.524	0.466	0.258	0.168
11	-0.327*	0.084	0.174*	0.147
12	-0.067	0.07	0.231*	0.093
13	0.011	0.071	-0.003	0.095
14	-0.246*	0.085	0.433*	0.129
15	-0.156	0.106	0.273*	0.177
16	-0.015	0.081	0	0.137
17	-0.396*	0.076	-0.338*	0.127
18	-0.21*	0.079	-0.177	0.172
19	-0.146	0.089	-0.26*	0.097
20	-0.012	0.083	0.012	0.084
21	0.231	0.132	-0.053	0.138
22	-0.221*	0.089	-0.073	0.178
23	0.177	0.148	0.385	0.141
24	-0.08	0.11	-0.164	0.147
25	0.026	0.098	0.164	0.09
26	-0.185	0.216	-1.165*	0.321
27	0.529*	0.212	-0.501*	0.205
28	-0.428*	0.148	0.283*	0.098
29	-0.1	0.135	0.058	0.11
30	-0.071	0.105	-0.224	0.15

* Indicates significant DIF

Which items in the numerical ability test show differential item functioning based on gender and country using the Mantel-Haenszel method?

Table 3 shows the Chi-square test results of a Mantel-Haenszel, a D-value, a probability value, an odds ratio, and a Mantel-Haenszel test by gender in the numerical ability test. Mantel and Haenszel's chi-squared values varied from 0.001 to 25.203. Eight items (26.7% of the total) on the numerical ability test showed significant differences between sexes. Based on the D index, there was a medium degree of DIF for items (17) and a weak degree of DIF for items (11), (14), (18), and (28) in favour of males. Items 3 and 27 showed a weak degree of DIF in favour of females, whereas item 5 showed a medium degree of DIF.

For DIF by country, chi-squared test results for Mantel and Haenszel ranged from 0.002 to 28.893. The results suggested that eight items (26.7%) of the numerical ability test showed DIF according to the student's country. Differential item functioning (DIF) was seen for three items against Oman; item 19 had low DIF, and items 17 and 26 had medium DIF. In contrast, five items demonstrated DIF in favour of Oman, with a low DIF for items 9 and 12 and medium DIF for items 1, 14, and 28 (as measured by the D-index).

Table 3
Chi-squared test values of Mantel and Haenszel, the probability value, the odds ratio, and the D value for the numerical ability test according to the gender variable

Item	Gender				Country			
	$MH\chi 2$	αMH	D	Strength & direction	$MH\chi 2$	αMH	D	Strength & direction
1	5.489	0.792	0.547	-	19.686*	0.555	1.383	MO
2	1.929	0.884	0.29	-	0.002	0.988	0.028	-
3	12.892*	0.726	0.752	WF	6.445	1.383	-0.76	-
4	0.432	0.935	0.159	-	0.553	1.124	-0.27	-
5	22.579*	0.651	1.01	MF	3.072	0.789	0.558	-
6	3.714	0.840	0.41	-	5.649	1.346	-0.7	-
7	0.076	1.030	-0.069	-	2.884	0.806	0.507	-
8	1.801	1.141	-0.311	-	0.228	0.926	0.179	-
9	1.218	0.898	0.253	-	7.260*	0.696	0.85	WO
10	6.009	0.754	0.664	-	4.336	0.710	0.805	-
11	13.893*	1.405	-0.8	WM	2.566	0.814	0.484	-
12	0.580	1.078	-0.176	-	9.322*	0.667	0.953	WO
13	0.327*	0.944	0.135	-	0.008	1.020	-0.05	-
14	7.013	1.274	-0.569	WM	12.841*	0.645	1.029	MO
15	0.782	1.084	-0.189	-	3.709	0.789	0.557	-
16	0.014	0.985	0.035	-	0.131	1.053	-0.12	-
17	25.203*	1.606	-1.113	MM	11.906*	1.610	-1.12	MG
18	7.055*	1.306	-0.627	WM	2.047	1.239	-0.5	-
19	2.763	1.173	-0.374	-	8.967*	1.475	-0.91	WG
20	0.001	0.998	0.004	-	0.005	0.983	0.041	-
21	5.024	0.813	0.485	-	0.450	1.089	-0.2	-
22	4.040	1.205	-0.438	-	0.032	1.032	-0.07	-
23	2.316	0.869	0.331	-	3.037	0.808	0.5	-
24	0.102	1.033	-0.077	-	2.043	1.188	-0.41	-
25	0.181	0.958	0.102	-	0.030	1.028	-0.07	-
26	0.585	1.072	-0.163	-	28.893	1.838*	-1.43	MG
27	12.351*	0.731	0.736	WF	6.483	1.353	-0.71	-
28	14.049*	1.463	-0.895	WM	10.144	0.639*	1.051	MO
29	0.619	1.082	-0.185	-	0.237	1.070	-0.16	-
30	0.100	1.032	-0.075	-	2.666	1.218	-0.46	-

WM: weak for male; MM: medium for male; WF: weak for female; MF: medium for female; WG: weak for gulf; MG: medium for gulf: medium for Oman; WO: weak for Oman.

What is the degree of agreement between the likelihood ratio test method and the Mantel-Haenszel method in detecting the differential functioning of the verbal ability test items according to gender and country variables?

The researcher used the percentage and the classification stability coefficient kappa to determine whether the likelihood ratio test method and the Mantel-Haenszel method agreed in identifying the differential functioning of the numerical ability test items according to gender and country variables.

Table 4

Differential item functioning of numerical ability test in GMMAS according to the likelihood ratio test and Mantel- Haenszel methods.

		MH of Gender				MH of Country	
		No DIF	DIF			No DIF	DIF
LR of Gender	No DIF	21	0	LR of Country	No DIF	17	0
	DIF	1	8		DIF	5	8
kappa	0.925			kappa	0.683		
Agreement	96.7%			Agreement		83.3%	

Table 4 demonstrates that the two methods agree in identifying DIF by gender for 29 items. Twenty-one items had concordant findings, indicating the absence of differential item functioning (DIF). In contrast, three items indicated DIF in favour of the focal group (females), and five indicated DIF in favour of the reference group (males). Regarding the other items, the Likelihood Ratio Test method finds a DIF favouring males for item 22, whereas the Mantel-Haenszel method finds no DIF for either gender.

The classification stability coefficient kappa, which was used to find out the degree of agreement between the two methods of DIF for gender, was 0.925, which was statistically significant at the $\alpha = 0.05$ level, and the percentage of agreement between the two methods was 96.7%. These numbers suggested considerable concordance between the two DIF methods for gender (Landis & Koch, 1977).

Also, table 4 clearly shows that the two methods agree on the presence of 25 items with the DIF for the country variable. There was consensus between the two methods that DIF does not exist on 17 items, that DIF does exist on 3 items for the reference group (the other Gulf countries), and that DIF does exist on 5 items concerning the focal group (Oman). In contrast, five items remained where the two methods disagreed; the Likelihood Ratio Test method found a DIF favouring Oman on items 11 and 15, while the Mantel-Haenszel method found no DIF for any country. Mantel-Haenszel finds no DIF for either sex, but the Likelihood Ratio Test finds a DIF towards the other Gulf states in items 3, 6, and 27.

The classification stability coefficient kappa, which was used to find out the degree of agreement between the two methods of DIF for gender, was 0.683, which was statistically significant at the $\alpha = 0.05$ level, and the percentage of agreement between the two methods was 83.3%. These numbers suggested a good level of concordance between the two DIF methods for gender (Landis & Koch, 1977).

Discussion

According to the likelihood ratio test and the Mantel-Hansel method, the study found no significant gender-based differences in the performance of the GMMAS scale's numerical

ability items. This finding lends support to the validity of the GMMAS numerical ability test, in line with the assertion made by Benito et al (2018) that the absence of DIF provides evidence of internal structure validity depending on psychological testing standards.

The lack of differential functioning in the numerical ability test may be attributed that the content of the test items was built according to precise standards so that they were appropriate to the school curriculum and the levels of the mental and chronological age of the students. Precision was considered in formulating camouflage alternatives to questions so that they do not have a clear role in showing the differential functioning of one social type at the expense of another. All of this confirms that the cultural differences of both sexes have been considered.

The results also found no significant country-based differences in the performance of the GMMAS scale numerical ability items according to the likelihood ratio test and the Mantel-Hansel method. This result may be because the content of the items of numerical ability relied on numbers more than words. Therefore, no items contained unfamiliar words among the Gulf countries' students, leading to differential item functioning among them.

In detecting differential item functioning in numerical ability tests according to gender and country, the likelihood ratio test method was more stringent than the Mantel-Haenszel method, with 30% and 43.3% of items with DIF by the likelihood ratio test method, respectively, and 26.7% by the Mantel-Haenszel method for both variables. Contrary to what was shown by (2006), the Mantel-Haenszel approach revealed 65% of items with DIF according to the country variable, and the likelihood ratio test method revealed 50% of items. In addition, the percentage of agreement between the two methods ranged from 96.7% for the gender variable to 83.3% for the country variable. The study's findings are consistent with those of Giray &Yildirim (2007), who said that the percentage of agreement between these methodologies was 82% in the PISA test but 48% in the TIMSS test. This conclusion confirms the importance of using item response theory when preparing psychological and educational measures to obtain valid and reliable measures in measuring the trait to be measured.

It is also necessary to mention that the present study involves some limitations. First, we applied the 2-PL model, which fit the data, although it did not consider a guessing parameter so that the 3-PL model could be tested. Second, the study used the Mantel-Haenszel Procedure in classical test theory and the likelihood ratio test in item response theory to find differential item functioning. Alternative approaches such as Lord's Chi-squared (LC), Logistic Regression (LR), and Item Characteristic Curve can be utilised.

Differential item functioning was observed across gender and country on the GMMAS scale's numerical ability test items, suggesting that further research is required to investigate bias in these items. It also needs research that studies the reasons that led to the emergence of differential performance in the test items to avoid and treat them. It is also in the future to compare the differential item functioning between the two versions of the test, the paper-and-pencil and the computer-based version, to find out the Effect of the type of test on the degree of existence of the differential performance.

Conclusions

The current study employed the likelihood ratio test and the Mantel-Haenszel method to investigate item functioning on the GMMAS numerical ability test in relation to students' gender and country. The findings demonstrated the presence of differential item functioning (DIF) among a subset of items, indicating variations in performance across different genders and countries, ranging from weak to moderate effects. Nevertheless, despite the identified

DIF, the study affirms the validity and effectiveness of the GMMAS numerical test as a reliable measure of cognitive abilities in the Gulf states. Moreover, the study revealed that the likelihood ratio test was a more stringent approach for detecting item bias compared to the Mantel-Haenszel method. These findings highlight the importance of acknowledging potential biases and employing appropriate statistical methods when evaluating item functioning in cognitive assessments.

References

Abu Shindi, Y. A., & Kazem, A. M. (2018). Sex differential item functioning for Mathematics test in cognitive development program in Sultanate of Oman by Mental-Haenszel and item characteristic curve methods. *International Journal of Learning and Management Systems*, 6(2), 61-73.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, 19, 716–723. https://doi: 10.1109/TAC.1974.1100705

Al Ajmi, M., Mustakim, S. S., Roslan, S., & Almehrizi, R. (2023). Differential Item Functioning of Verbal Ability Test in the Gulf Multiple Mental Abilities Scale by Mental-Haenszel and Likelihood Ratio Test. *International Journal of Academic Research in Business and Social Sciences*, 13(1), 1038–1056.

Almaskari, H. A., Almehrizi, R. S., & Hassan, A. S. (2021). Differential item functioning of verbal ability test in Gulf multiple mental ability scale for GCC students according to gender and country. *Journal of Educational and Psychological Studies*, 15(1), 120-137.

Al Nafouri, L. (2015). W-J. Test of Cognitive Abilities III Standard Battery. Unpublished thesis dissertation.

Alsawalmeh, Y., & Al Ajlouni, J. (2019). The relationship between differential distractors functioning and differential item functioning in a multiple-choice mathematics test. *Jordanian Journal of Educational Sciences*, 15(1), 49-63.

Alzayat, F., Almehrizi, R., Arshad, A., Fathi, K., Albaili, M., Dogan, A., Asiri, A., Hadi, F., & Jassim, A. (2011). Technical report of the Gulf scale for multiple mental abilities (GMMAS). Arab Gulf University, Bahrain.

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing.* Washington, DC: Author.

Aryadoust, V. (2018). Using recursive partitioning Rasch trees to investigate differential item functioning in second language reading tests. *Studies in Educational Evaluation*, 56, 197–204. https://doi: 10.1016/j.stueduc.2018.01.003

Bichi, E., Embong, R., Talib, R., Salleh, S., & Ibrahim, A. (2019). Comparative analysis of classical test theory and item response theory using Chemistry test data. *International Journal of Engineering and Advanced Technology, 8*(5), 1260-1266. https://doi:10.35940/ijeat.e1179.0585c19.

Diaz, E., Brooks, G., & Johanson, G. (2021). Detecting Differential Item Functioning: Item Response Theory Methods Versus the Mantel-Haenszel Procedure. *International Journal of Assessment Tools in Education, 8*(2), 376–393. https://doi:10.21449/ijate.730141.

Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel-Haenszel Methods for Differential Item Functioning Detection. *Educational and Psychological Measurement*. https://doi.org/10.1177/0013164408315265

Geramipour, M. (2020). Item-focused trees approach in differential item functioning (DIF) analysis: a case study of an EFL reading comprehension test. *Journal of Modern Research in English Language Studies*, 7(2), 123-147. https:/doi: 10.30479/jmrels.2019.11061.1379

Geramipour, M., & Shahmirzadi, N. (2019). A gender–related differential item functioning study of an English test. *Journal of Asia TEFL, 16*(2), 674.

Giray, B., Yildirim, H. (2007). *The DIF analyses of PISA2003 mathematics items via likelihood ratio, Mantel-Haenszel and restricted factor analysis procedures*. The report, Retrieved on Jan 4, 2010, from http:// www. Etd.lib.metu.edu.tr.

Gomez-Benito, J., Sireci, S., Padilla, J.-L., Hidalgo, M. D., & Benitez, I. (2018). Differential item functioning: Beyond validity evidence based on internal structure. *Psicothema, 30*(1), 104–109.

Huang, T. W., Wu, P. C., & Mok, M. M. C. (2022). Examination of Gender-Related Differential Item Functioning Through Poly-BW Indices. *Frontiers in psychology, 13*, 821459. https://doi.org/10.3389/fpsyg.2022.821459.

Jabrayilov, R., Emons, W., & Sijtsma, K. (2016). Comparison of Classical Test Theory and Item Response Theory in Individual Change Assessment. *Applied Psychological Measurement 40*. 1-14. https://doi.org/ 10.1177/0146621616664046.

Kim, S., Cohen, A., & Lin, Y. (2005). LDID: *A Computer program for local dependence indices for dichotomous Items*. Version 1.0.

Krabbe, P.F. (2017). *The measurement of health and health status: Concepts, methods and applications from a multidisciplinary perspective.* Elsevier.

Landis, J. R., Koch, G. (1977). *The measurement of observer agreement for categorical data. Biometrics, 33*(1), 159–174. doi:10.2307/2529310

Magis, D., Yan, D., & Von Davier, A. A. (2017). *Computerised adaptive and multistage testing with R: Using packages catR and mstR*. Springer.

Nawafleh, A. (2017). The Effect of paragraphs with differential functioning of uniform on estimating paragraphs parameters and persons using a stimulated data according to the Three parameters model. *Educational science studies*, 44(4), 187–207.

Oalla, B., Matarneh, A. (2018). Differential performance of the items of the University level Test for the English language among the students of Mutah University. *Journal of Educational and Psychological Sciences, 19*(2), 449- 475.

Salman, M., & Thatha, H. (2022). The Differential Performance of the Items of the National Test to Control the Quality of Education in the Subjects of Science, Mathematics, Arabic and English for the Eighth Grade in Jordan. *Jerash for Research and Studies Journal 23*(2).

Sayed, M., Bakhoum, R., Moussa, M., & Mohamed, M. (2022). Detecting the differential item function of gender on the emotional balance scale using mantel Hansel method According to the assumptions of the item response theory. *Journal of Research in Education and Psychology*, 37(1), 361–396.

Schwarz, G. (1978). Estimating the Dimension of a Model. Annals Stat. 6, 461–464. doi: 10.1214/aos/1176345415.

Warnimont, C. S. (2010). *The Relationship between Students' Performance on the Cognitive Abilities Test (CogAT) and the Fourth and Fifth Grade Reading and Math Achievement Tests in Ohio*. Unpublished doctoral dissertation. Bowling Green State University.

Yildirim, H. H. (2006). *The Differential Item Functioning (DIF) Analysis of Mathematics Items in The International Assessment Programs*. Unpublished thesis dissertation.

Zakri, A. (2020). Identifying differential item functioning of the "EMBU" test of parental rearing styles among a sample of secondary school students. *Journal of Education College, 3*(186). 676–720.