INTERNATIONAL JOURNAL OF ACADEMIC RESEARCH IN BUSINESS & SOCIAL SCIENCES

Published Online: 25 January 2023 Vol 13, Issue 1, (2023) E-ISSN: 2222-6990

Differential Item Functioning of Verbal 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's Email: mohd7010@gmail.com

Abstract

The purpose of this study was to examine the differential item functioning (DIF) of verbal ability test items by gender (male vs female) and country (Oman vs the rest of the Gulf countries) using the Mantel-Haenszel (MH) and the Likelihood Ratio Test (LRT) methods which will be reflected on the accuracy of the test results. The sample was 2688 students in grades five and six and to achieve the study's objectives, MH was applied using the SPSS program and LRT using the BILOG-MG program. The classification stability coefficient kappa (κ) used to know the agreement ratio between the two methods was calculated to detect differential performance. The results using MH showed that 16.7% of items exhibited DIF in relation to gender, and 33.3% regarding country. Additionally, results showed that DIF utilizing LRT was evident for 10% of the items with respect to gender and 30% of to the country. The agreement between the MH approach and LRT for gender was quite high (κ = 0.725). The agreement between the MH approach and LRT for the country was also quite high (κ = 0.655). The study recommended further study to investigate of the causes of the differential functioning of some items of the verbal ability test.

Keywords: Differential Item Function, Verbal Ability, Item Response Theory, Mantel-Haenszel Method, Likelihood Ratio Test.

Introduction

Cognitive ability is related to many important variables in life, such as academic achievement, critical thinking and problem solving (Smith, 2011; Warnimont, 2010; Tinajero et al., 2012), and thus measuring cognitive ability has remained a necessary requirement for organizations that are concerned with educational and psychological tests such as the American Psychological Association and the European Union of Psychological Societies. As a result, measurements of cognitive ability assist educators in helping more students achieve by giving teachers dependable information on each student's cognitive abilities and how to use this information to focus more on structure for learning (Warnimont, 2010).

The Gulf Multiple Mental Abilities Scale (GMMAS) is another test used to measure cognitive abilities (Alzayat et al., 2011). The idea of this scale is that general mental ability is a multi-dimensional ability that expresses itself through three domains: Verbal, numerical and spatial. Mental activity is governed by higher cognitive mental processes, represented by the perception of the stimuli of the external world, the recollection of experiences that pass by the individual, thinking and analyzing different situations and inference.

Many tests have been built to measure verbal ability, including the current one, and were built under the assumptions of the classical theory test (CTT). Although the classical theory of tests dominated measurement methodologies throughout the last century, it contained some defects that were addressed in several studies, such as (Ojerind, 2013; Eleje et al., 2018; Bichi et al., 2019; Jabrayilov et al., 2016; Kiany & Jalali, 2009). The most critical pieces of information in CTT are based on the total scores, and the individual and item statistics (item difficulty and item discrimination) were related to the sample to which the test was applied.

Classical test theory has its drawbacks, so 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). In addition, it is more theory-based and its models are the probabilistic distribution of examinees' item performance. As the name implies, IRT focuses on item-level data; item parameters could involve difficulty (location), discrimination (slope), and guessing. The efforts of researchers, whether in the CCT or the IRT, have focused on building and developing tests on extracting the effectiveness of items in terms of difficulty, discrimination and guessing. Despite the importance of these characteristics, they are not sufficient to judge the validity of test items as the items may be affected by other factors such as gender, social and economic level, in addition to the ability of the examinees, which negatively affects the results and thus behaves biased towards one group against another, and accordingly, the item of the scale is described as biased. If the scale item shows a difference between groups of individuals of equal ability due to characteristics other than the measured trait, then the item has a differential functioning (Aryadoust, 2018; Geramipour, 2020). Such requirements are an important requirement in building the scale and verifying its fairness (Geramipour & Shahmirzadi, 2019) and a prerequisite for the development of tests used in making decisions as it affects the parameters of test items (Nawafleh, 2017).

International organizations concerned with test preparation in education and psychology such as the American Educational Research Association (AERA), the American Psychological Association (APA) and the National Council for Measurement in Education (NCME), have considered the differential item functioning (DIF) a necessary standard when preparing and publishing tests (Geramipour, 2020). The presence of differential performance in the tests is one of the threats to internal validity of the test (Gómez-Benito et al., 2018).

The differential item functioning is a statistical indicator for expressing the differences in the probability of a correct response to the item among the different groups of equal ability (Sayed et al., 2022). DIF means that the way an item works for two different groups of respondents is different. In other words, students who score the same on a test but belong to different subgroups (such as male vs female or scientific vs literary) have different chances of answering a question. When items show DIF across groups, they pose a serious threat to the validity of a test and may make it harder to compare groups. The reason is that their scores might show things other than what the scale is meant to show (Krabbe, 2017). So, the idea of the current research is to ensure that there is no differential functioning of the type of student (males vs females and Oman vs the rest of the Gulf countries) on testing verbal ability in the GMMAS scale.

Research Background

Many researchers recently went to examine an important psychometric property to achieve the principle of fairness and equity in the tests, which is the differential item Function (DIF). DIF ensures that the scores of the examinees reflect their ability and are not affected by other variables such as race, culture, nationality, and gender (Abu Shindi & Kazem, 2018). According to Alquraan and Alkuwaiti (2017), finding DIF entails determining the extent to which two test respondents with equivalent standing on the latent trait but from different groups (e.g., female and male) have the same likelihood of selecting the same item option. In other words, DIF occurs when items operate differently when students with equal ability for the construct under study produce diverse responses due to belonging to various sub-groups. DIF occurs when test items behave differently for the reference group than the focal group, even after controlling for student proficiency (Shanmugam, 2020).

Liu (2011) states that the scientist Mellenberg presented the concept of uniform DIF and nonuniform DIF in 1982. Rashwan (2021) explains that uniform DIF is present when there is no interaction between the individual's performance level on the item and the individual's belonging to the group. The probability of a correct answer on the item is always greater for one of the two groups (reference or focal) over all levels of ability. In other words, it can be said that it is shown by the non-intersection of the two item characteristics curves (ICC) along the ability intervals. In contrast to uniform DIF, nonuniform DIF is present as interaction occurs between the individual's performance level on the item and the individual's belonging to the group. The differential Function appears once in favour of the reference group for a specific level of ability and once in favour of the focal group for another level of the ability. Graphically, it can be said that it appears through the intersection of two item characteristics curves, as shown in Figure 1.

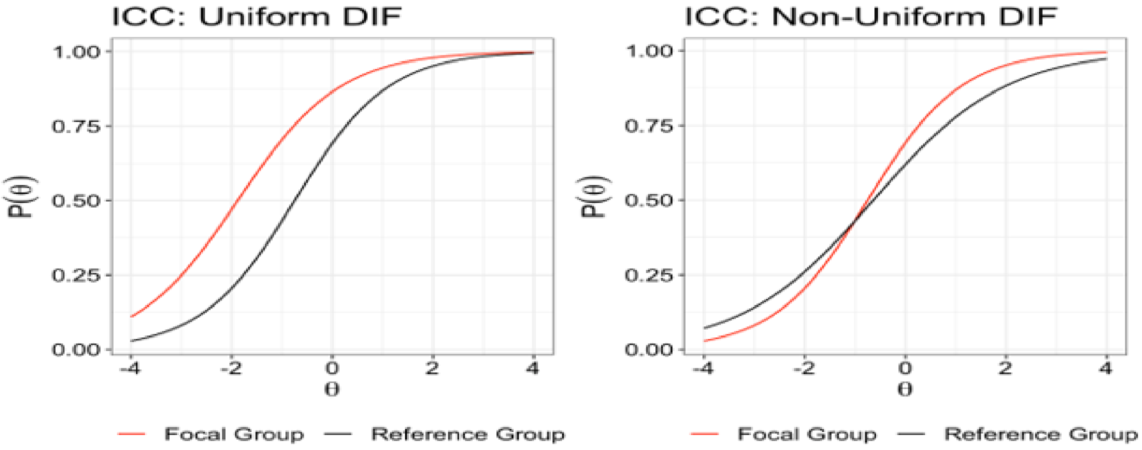

Figure 1. Example of Uniform and Nonuniform DIF item (Ayala, 2009)

The size of the differential functioning is classified: small, medium, and large, using different measures of effect size depending on the method used to detect it. Usually, no action is taken on the item in the case of small differential functioning, and in the case of large differential functioning, it is advised to delete or revise the item. The main point of the review process is to identify the possible reasons for the differential functioning, especially concerning the item itself (Al Sawalmeh & Al Ajlouni, 2019).

Psychometricians have developed several methods that help the researcher detect differential functioning in test items. The essential methods that are used to detect the differential functioning of an item are: Mantel Haenszel Method (MH), Transformed Item

Difficulty Method (TID), Analysis of Variance ANOVA, Logistic Regression Method, Item Discrimination Method (IDM), Chi-Square Method, Distracter Response Analysis, Item Characteristic Curve (ICC), b - Parameter Difference Method, and Likelihood Ratio Method (Almaskari & Almehrizi, 2021; Zakri, 2020; Oalla & Matarneh, 2018).

The current study will depend on the Likelihood Ratio Test method (LRT) and Mantel Haenszel (MH) method to detect differential item functioning.

Likelihood Ratio Test (Cohen et al., 1996; Thissen et al., 1988).

The likelihood ratio test method is based on the item response theory (IRT) of measurement and investigates the bias between two groups (reference vs focal) and calibrating the data as one group, i.e., combining the data of the different groups into one group and calculating the likelihood ratio values using appropriate software such as BILOG-MG.

There are three phases to the IRT-LR DIF study used to identify DIF. First, compute the likelihood deviation G_C^2 (= -2 log-likelihood) of the maximum likelihood estimates while estimating the compact IRT model, in which all items are bound to have the same parameters in both groups. Second, compute the likelihood deviance G_A^2 by estimating the augmented model, in which all items except the one being studied examination are constrained to have the same parameters in both groups. Third, calculate the difference in likelihood deviances between the compact and augmented models as $G^2 = G_C^2 - G_A^2$, and then run a chi-square test with degrees of freedom equal to the difference in the number of estimated parameters in the two models. If G^2 is statistically significant, then the item under study has DIF. To identify DIF across many test items, this procedure must be done for each item (Liu, 2011).

Mantel- Haenszel Chi-Square (Mantel & Haenszel, 1959; Holland & Thayer, 1988)

The Mantel-Haenszel method is based on the classical theory of measurement and investigates the bias between the reference group and the focal group, which is the group affected by the item’s bias (Allabadi, 2008). To estimate the differential functioning by the Mantel-Haenszel method. a square binary matrix containing the number of individuals who responded correctly and incorrectly to the item from the two groups is to be computed, and then the value of the Mantel-Hansel statistic is calculated according to the following equation:

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

(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.

The Mantel-Haenszel index follows a chi-square distribution with degrees of freedom of 1. The odds ratio's differential functioning trend of the item was judged through the value of the odds ratio. To interpret the results, the items were divided into three types: items that do not show functioning differentially (1 = α_{MH} and not statistically significant), and the items with differential functioning in favour of the group Reference (1 < α_{MH} and statistically significant), and items with differential functioning in favour of the focal group (1 > α_{MH} and statistically significant) (Al Bursan, 2013). The α_{MH} calculated from the following equation:

$$\alpha MH = \frac{\sum_{t=1}^{s} \frac{A_t \, D_t}{N_t}}{\sum_{t=1}^{s} \frac{B_t \, C_t}{N_t}}$$

Where s is Number of ability levels, A_t is the number of members of the reference group who answered the item correctly at ability level t, B_t is the number of members of the reference group who could not answer the item correctly at ability level t, C_t is the number of members of the focal group who answered the item correctly at ability level t, and D_t is the number of members of the focal group who could not answer the item correctly at ability level t.

To judge the strength of the differential functioning of the item in the case of a differential functioning of the item according to the value of the odds ratio, a criterion used by the centre for Educational Testing and Services (ETS) can be used by calculating the value of delta (D) according to the following equation

$$D = \beta MH = -2.35 * \ln(\alpha MH)$$

βMH is called the signed index, as it is inferred from its reference to the direction of the differential functioning of the item. The item is classified considering this index in terms of the strength of its differential functioning into three types in absolute magnitude, namely: an item with weak differential functioning (0≤ D < 1), Medium differential functioning (1≤ D < 1.5) and high differential functioning (D < 1.5) (Thissen, et. al., 1988).

Statement of Problem

Measurement efforts in many areas of education such as national and international tests, as well as in measuring various psychological traits such as intelligence, anxiety, stress, and others are directed equally to the improvement of education quality and its outcomes. If efforts in these areas are to be placed on a sound scientific basis, then we must rely on methods to find indicators of validity, reliability, and effectiveness of items in terms of their level of difficulty and distinction to develop tools whose results can be trusted (Krabbe, 2017). Also, it is very critical to maintain consistency in the item attributes across several groups of examinees. For instance, it is possible that some items favour (e.g., are more accessible for) males over females (allowing for equivalent skill levels), and such items must be identified (and maybe removed) to ensure fair measurement (Magis et al., 2017).

Tests, and especially cognitive abilities tests are extensively used in different context such as schools, university, and health (hospital). Test results for various purposes, such as the screening and selection of individuals, assessment of student learning development or evaluation of the effectiveness of education systems, can be used in educational and psychological measurement (Alodat & Zumberg, 2018). The Gulf Multiple Mental Abilities Scale (GMMAS) is one of the most recent and important measures in the Gulf Cooperation Council countries and was prepared and codified by Alzayat et.al (2011) as a research grant funded by the Education Office for the Arab Gulf States in 2011. The idea of the scale is that general mental ability is a multi-dimensional ability that expresses itself through three domains: verbal, numerical and spatial. Mental activity is governed by higher cognitive mental processes, represented by the perception of the stimuli of the external world, the recollection of experiences that pass by the individual, thinking and analysis of different situations and inference.

Alzayat et al (2011) indicated that there were differences in verbal ability for fifth and sixth grade students in the Arab Gulf countries according to the country in the three levels of the GMMAS scale, and the presence of these differences may indicate the existence of a differential functioning of the scale items. Given the importance of the Gulf scale in measuring verbal ability and making decisions related to diagnostic procedures, checking the presence of differential functioning | for its components is essential with respect to gender (male/ female) and the country (Oman/ the rest of the Gulf countries) is important. Therefore, the problem of the study is determined in the following questions:
1. What are the items in verbal ability test of GMMAS showing differential item functioning according to gender and the country using likelihood ratio test method?
2. What are the items in verbal ability test of GMMAS showing differential item functioning according to gender and the country using Mantel-Haenszel method?
3. What is the degree of agreement between the likelihood ratio test method and the Mantel-Haenszel method in detecting differential item functioning of verbal ability test of GMMAS according to gender and country variables?

Significance of the Study

The topic covered by the current research is one of the essential topics in psychological fields, especially measurement and evaluation. The study of the differential functioning of the verbal ability test items will provide practical examination for identifying the item with differential functioning and then modifying or deleting them. This, in turn, will lead to an increase in the quality of the test, which will be reflected on the accuracy of the test results, and the fairness of the comparison between individuals taking the test. In addition, this study may lead to results from which it is possible to develop recommendations for developing the verbal ability test. It is noted that the efforts to study differential functioning are rare locally in Arab countries, which makes this study an element of contribution to these efforts, and perhaps it encourages other researchers to give more attention to studying differential functioning and studying bias.

Method

Participants

This study employs quantitative research through following the descriptive approach, as it seeks to describe the statistical characteristics of the verbal ability test in GMMAS using the differential item functioning. The researcher depends on secondary data gathered during the GMMAS standardization by the Arab Office for the Gulf States in 2011. This sample was taken in the fifth and sixth grades, and the ages of the students range from nine years and three months to twelve years and three months. The sample included 2688 with 1269 females and 1419 males. Students' age ranges from nine years and three months to 12 years and three months.

Instrumentation

The study uses the verbal ability test in GMMAS prepared by (Alzayat et al., 2011). It consists of three tests measuring verbal, numerical, and spatial abilities. The verbal ability test consists of three sub-tests, with a total of 30 items, which are of a multiple-choice type. The test assesses three subtests: word synonyms (10 items), word antonyms (10 items), and verbal analogies (10 items). The latest requires students to discover the relationship between a pair of words at the beginning of the question and applying it to new words. the items are scored one mark for a correct answer and zero for a wrong answer. The total score ranges between 0 and 30.

The scale psychometric properties were investigated by (Alzayat et. al., 2011). As for its validity, it was found that the mean raw score for the fifth grade is 19.1 and for the sixth grade is 21.1, while the standard deviation is 6.2 and 5.7 for the fifth and sixth grades, respectively. Descriptive statistics indicated a high level of verbal ability when moving from the fifth to the sixth grade in all Gulf countries. This result indicates that verbal ability increases with the experiences students acquire through the curricula and those they go through at their age. The intercorrelation coefficients of the verbal ability test were positive and medium strength among all subtests of verbal ability and ranged between 0.589 and 0.621. This indicates the interrelationship of the verbal ability subtests, as these subtests measure skills related to verbal ability. Alzayat et. al (2011) showed that the verbal ability correlation coefficients with the Raven successive matrices test are positive and statistically significant as an indicator of the construct validity of the test. also, results showed a presence of positive correlation coefficients between verbal ability and academic achievement in the Arabic language for the fifth and sixth grades. This indicates the predictive validity of the verbal ability test.

As for its reliability, test re-test coefficient of verbal ability was 0.98. The verbal ability showed high internal consistency ability in all grades, where the Cronbach alpha coefficient for verbal ability ranged between 0.850-0.882 for different Gulf countries (Alzayat et. al., 2011).

Study Procedures

Verification the assumptions of item response theory

The assumptions of the item response theory were verified by checking the unidimensionality assumption and the local independence assumption for the verbal ability as follows.

Unidimensionality Assumption

Exploratory factor analysis (EFA)

To verify the assumption of unidimensionality of the test, the adequacy of the sample size was confirmed by the Kaiser-Mayer- Olkin (KMO) and Bartlett's test, and the calculated chi-square value was (13503.534), a function at the level (0.001) and degree of freedom (435), and this result means that the sample size is suitable for conducting exploratory factor analysis. Then, exploratory factor analysis was used according to the method of the principal components of the correlation matrix for the 30 items of verbal ability in the scale. The result showed that there are five latent root factors eigenvalue, each of which is more than one and all together, they explain 37.740% of the variance. The result of dividing the eigenvalue of the first factor (6.345) and the eigenvalue of the second factor (1.572), which equals 4.036 and it is greater than two, is an indication of unidimensionality (Reckase, 1997 cited in Oalla, 2015). The ratio of the explanatory variance of the first factor to the total variance is 56.039. Based on this percentage, it is considered the unidimensional test, which meets the Reckase (1979) criterion of 20% (cited in Lee, 2004). Also, using Cattell's scree plot test (1966) for the 30-item factor analysis. Figure 2 shows the achievement of the unidimensionality of the test by distinguishing the first factor from the rest of the factors.

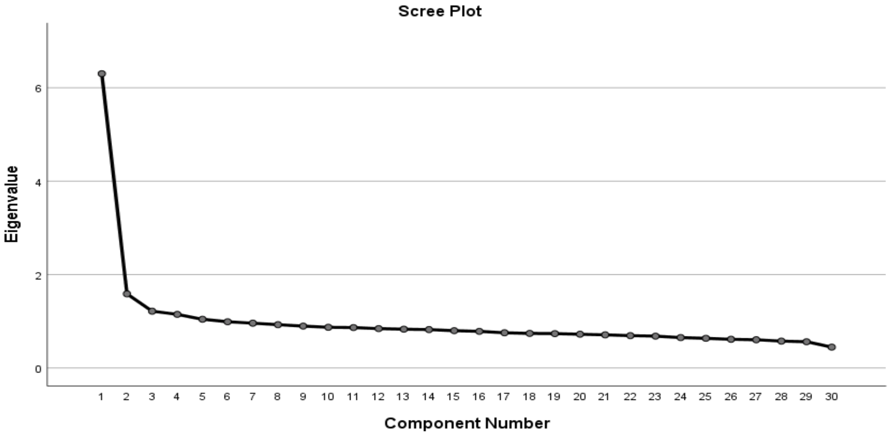

Figure 2. Factor scree plots from principal component analysis of 30 items

Confirmatory factor analysis (CFA)

Another indicator of the fulfilment of the unidimensional assumption of the data, the AMOS program was used to find the value of the Root Mean Square of Residuals (RMSEA) and Tanaka Index (GFI). It is noted from the results that the value of Root Mean Square of Residuals (RMSEA) is equal to .042, which met the Browne and Cudeck (1993) that an RMSEA of .05 or less indicates a good fit. Also, the value of the (GFI) is (0.93), which meets the Tanaka and Huba (1985) criterion.

Local Independence

The second assumption is local independence, and it means, as defined by Hambleton and Swaminathan (1985), that the responses of individuals to the test items of the same ability are statistically independent, and this means that the individual's response to an item should not negatively or positively affect his response to any other item. The assumption of local item independence is equivalent to the assumption of unidimensionality, as shown by (Hambleton and Swaminathan, 1985; Allam, 2005). This means that if the unidimensionality assumption is guaranteed in a scale, then the scale is also assumed to have the assumption of local item independence, but for further verification the researcher has relied on the statistical indicator suggested by Yen (1993), which is the correlation coefficient between the residuals for a pair of items after adjusting the individual's ability (0).

To verify the assumption of the local independence of the verbal ability test, the computer program for Local Dependence Indices for Dichotomous Items (LDID) was used. 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).

The results came to indicate that most of the values of Q3 were less than (0.152), which is a high indication of the fulfilment of the assumption of local independence between the test items. Also, the results indicate that the percentage of the number of independent pairs of items in the verbal ability test was 100%, which is evidence that individual's responses to the test items achieve local independence.

Freedom from Speed

The researcher gave enough time to answer the test item so that the failure of individuals to respond to the test item is due to their low abilities and not to the effect of the speed factor in answering. No sample member during the application objected to the lack of time and its insufficiency during the test application. The researcher also verified the freedom from speed by calculating the percentage of students who were able to finish answering all the test items, as the percentage of students who were able to finish answering all the test items was (100%), which indicates that check borrows freedom from speed in the test.

Psychometric Properties of the Verbal test in According to the IRT

Choosing the Model

To determine which unidimensional logarithmic models are more suitable for the test data, the following indicators were used:

- -2 log Likelihood (−2LL): which tests the hypothesis that adding the discrimination parameter to the 1PL model does not lead to a statistically significant improvement in estimating the parameters and also tests the hypothesis that adding the guessing parameter to the 2PL model does not lead to a statistically significant improvement in estimating the parameters, so that the model is more fit for the data if the difference between the values of (-2LL) for the two models is statistically significant using the chi-squared distribution (x2) with the degrees of freedom of the difference between the parameters of the two models. The lower the value of this indice, the better the model fit; thus, the model with the lowest test-fit indices was chosen for further analysis (Spiegelhalter et al., 1998).
- Index of the values of the information function (Average Information): The best model is the one that provides the most information.
- Root Mean Square Standard Errors of Estimates (RMSE): The best model is the one that gives a lower value for the root mean square standard error of the estimate.
- The number of misfit items for the model: The fewer misfit items, the better the model.

Table 1 shows the values of the model fit indices for choosing the appropriate model for the verbal ability test data.

Table 1
The values of the indicators for choosing the appropriate model for the verbal ability test data

S	Indicators	Model		
		1PL	2PL	3PL
1	-2 log Likelihood	85842.3184	84711.1844	84592.3000
	Model Differences	-	1131.134**	118.8844**
2	Average Test Information	5.679	6.868	7.070
3	RMSE	0.4003	0.3942	0.3766
4	Reliability Index	0.850	0.873	0.876

It is clear from Table 1 that the most suitable model for the test data is the three-parameter logarithmic model (3PL), which considers difficulty, discrimination, and guessing parameters.

Item fit

The suitability of the test items to the three-parameter model was verified using the Excel program, which depends on the Standardized residuals (SRs) index (Wright & Master, 1982).

Standardized residuals, an aspect of this statistical measure, benefit from being less dependent on the size of the sample than chi-square tests.
Standardized residuals are calculated by dividing the ability scale into an equal number of intervals and then computing the difference between the expected and actual performance of examinees in each ability level, as described by (Hambleton et al., 1991). The residual is a term used to describe this dissimilarity. By dividing the residual by the standard error of the predicted performance, we obtain the standardized residual. To judge the fit of the items to the model, the value of the standardized residuals is squared. This indicator follows the chi-squared curve with a degree of freedom equal to one, so we calculate the probability value. The significant value indicates the item is a misfit of the model. The results revealed that all items are fit in the three-parameter model.

Person Fit

The unweighted Almehrizi index was used (Almehrizi, 2010) to check the fit of persons to the three-parameter model. This method aims to collect the squares of the residual difference across all items. It is symbolized by the symbol U$RS$j.

$$SR_j = \sum_{i=1}^{n} SR_{ij} = \sum_{i=1}^{n} (y_{ij} - p_{ij})^2$$

The results revealed 21 cases of misfit for male student response patterns and 28 cases for female student response patterns, which were deleted. The data valid for analysis became 2639.

Reliability of Verbal Ability Test

Three test reliability coefficients were extracted according to the item response theory:
- *Test information function:* It indicates the reliability of the test and the consistency of its item in estimating individuals’ abilities. The higher the test information function's value, the higher the test's accuracy in assessing the measured feature. The maximum value of the verbal ability test information function is 9.38 at the ability level of 0.00, with a standard error of 4.09. This indicates that the test information function provides the largest amount of information at the average ability level.
- *Test Reliability coefficient:* Reliability coefficient refers to the stability of individuals' ability estimates according to the measured trait, which is calculated based on the variation of individuals' estimates and the average function of the test information. The researcher calculated the reliability index for the verbal ability test by using the BILOG MG program. The reliability index of the test reached 0.876, which indicates the test reliability in estimating the individuals’ abilities is high.
-*The empirical reliability of the test:* Empirical reliability refers to the extent to which the ability that was estimated through the models of response theory approaches the real ability of individuals, which is the complement of the ratio of the error variance of individuals' ability estimates to the variance of individuals' ability estimates. The empirical reliability coefficient of the test reached 0.861, which indicates a high-test reliability for estimating individuals’ ability.

Results

1. What are the items in verbal ability test of GMMAS showing differential item functioning according to gender and the country using likelihood ratio test method?

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

Table 2 shows there are three items (10%) have DIF at according to gender; where item 6 in favour of male and item 17 and 20 in favour of female. Also, Table 2 indicated 9 items (30%) showed DIF according to country; where item 2, 4, 10 and 26 showed DIF against Oman, and item 6, 7, 11, 14 and 16 in favour of Oman.

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

	Gender		Country	
ITEM	Estimate	SE	Estimate	SE
1	-0.305	0.245	-0.453	0.274
2	0.064	0.163	-0.515*	0.194
3	-0.111	0.154	0.112	0.199
4	0.105	0.104	-0.58*	0.143
5	0.044	0.404	0.658	0.458
6	-0.32*	0.132	0.377*	0.167
7	0.135	0.102	0.359*	0.121
8	-0.083	0.125	0.061	0.155
9	-0.099	0.099	0.153	0.123
10	0.063	0.095	-0.256*	0.124
11	0.01	0.16	0.543*	0.18
12	-0.011	0.169	-0.041	0.196
13	-0.024	0.119	0.176	0.149
14	0.018	0.145	0.342*	0.159
15	0.134	0.151	-0.085	0.174
16	0.089	0.092	0.253*	0.119
17	0.236*	0.105	0.017	0.142
18	-0.136	0.074	0.142	0.1
19	0.13	0.077	0.087	0.107
20	0.25*	0.077	0.215	0.111
21	-0.185	0.287	-0.629	0.322
22	0.018	0.121	-0.273	0.145
23	-0.185	0.147	-0.188	0.176
24	-0.097	0.211	0.016	0.234
25	-0.17	0.089	-0.111	0.114
26	-0.096	0.141	-0.491*	0.176
27	0.172	0.098	-0.159	0.123
28	0.116	0.089	-0.024	0.12
29	0.117	0.146	0.321	0.18
30	0.121	0.103	-0.03	0.133

* Indicates significant DIF

2. What items in the verbal ability test in the GMMAS scale showing differential functioning according to gender (females vs males) and the country (Oman vs the rest of the Gulf countries) using the Mantel-Haenszel method?

Table 3 shows Chi-squared test values of Mantel and Haenszel, the probability value, the odds ratio, and the D value for the verbal ability test according to gender variable. The values of chi-squared test for Mantel and Haenszel ranged between 0.036 and 13.864. The results indicated that there are DIF for 5 items (16.7%) of the verbal ability test according to gender. Item 6, 18, and 25 showed DIF in favour of males with weak degree of DIF based on D index. In contrast, item 17 and 20 showed DIF in favour of females with weak degree of DIF based on D index.

Also, the values of chi-squared test for Mantel and Haenszel ranged between 0.000 and 32.177 for DIF based on country. The results indicated that 10 items (33.3%) of the verbal ability test showed DIF according to the country of the students. 5 items showed differential functioning against Oman with poor DIF for item 2 and a medium DIF for four items (1, 4, 21 and 26). In contrast, 5 items showed DIF in favour of Oman with poor DIF for item 18, a medium DIF for three items (6, 14 and 16) and a strong DIF for item (11) according to 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 verbal 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	5.808	1.301	-0.618	-	9.137*	1.620	-1.134	MG
2	0.773	0.914	0.2113	-	7.266*	1.508	-0.965	WG
3	0.036	1.020	-0.047	-	0.000	1.009	-0.021	-
4	1.391	0.876	0.3111	-	25.445*	2.263	-1.919	MG
5	0.389	0.936	0.1554	-	6.451	0.737	0.7171	-
6	13.864*	1.416	-0.817	WM	14.348*	0.625	1.1045	MO
7	3.851	0.830	0.4379	-	6.069	0.736	0.7203	-
8	0.184	1.042	-0.097	-	0.116	1.048	-0.11	-
9	1.365	1.120	-0.266	-	0.333	0.972	0.0667	-
10	0.062	0.852	0.3764	-	6.632	1.359	-0.721	-
11	0.193	1.080	-0.181	-	32.177*	0.359	2.4074	SO
12	0.217	1.066	-0.15	-	0.006	1.001	-0.002	-
13	0.263	1.055	-0.126	-	2.565	0.813	0.4865	-
14	0.413	1.110	-0.245	-	11.380*	0.547	1.4178	MO
15	1.550	0.852	0.3764	-	0.013	1.034	-0.079	-
16	1.189	0.885	0.2871	-	11.315*	0.615	1.1424	MO
17	10.114*	0.741	0.7044	WF	0.786	0.884	0.2898	-
18	8.473*	1.351	-0.707	WM	7.229*	0.683	0.896	WO
19	2.483	0..862	0.349	-	0.865	0.878	0.3058	-
20	10.081*	0.753	0.6667	WF	1.655	0.852	0.3764	-
21	4.097	1.340	-0.688	-	9.978*	1.961	-1.583	MG
22	0.287	1.066	-0.15	-	4.836	1.423	-0.829	-

23	6.031	1.332	-0.674	-	2.573	1.301	-0.618	-
24	1.383	1.142	-0.312	-	0.395	0.901	0.245	-
25	9.581*	1.333	-0.675	WM	0.244	1.076	-0.172	-
26	2.805	1.158	-0.345	-	17.128*	1.653	-1.181	MG
27	3.724	0.845	0.3958	-	4.559	1.287	-0.593	-
28	3.516	0.844	0.3986	-	0.012	1.021	-0.049	-
29	1.769	0.887	0.2818	-	0.024	0.976	0.0571	-
30	2.839	0.847	0.3902	-	0.067	1.042	-0.097	-

WM: weak for male; MF: medium for female; WG: weak for gulf; MG: medium for gulf; SO: strong for Oman; MO: 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?

Table 4 presents the differential item functioning of verbal ability test in GMMAS using the likelihood ratio test and Mantel- Haenszel methods according to gender and country variables.

Table 4

Differential item functioning of verbal 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	25	2	LR of Country	No DIF	18	3
	DIF	0	3		DIF	2	7
kappa	0.725			kappa	0.655		
Agreement	93.3%			Agreement		83.3%	

Table 4 shows agreement between the two methods in revealing the existence of DIF for the gender in 28 items, distributed as follows: in 25 items, the two methods agreed that there is no DIF, while two items on the existence of DIF towards the focal group (females), and one item the existence of DIF towards the reference group (males). The remaining items are two items that the two methods do not agree on, of which the Mantel- Haenszel method sees that it has DIF towards males, while the likelihood ratio test method sees that there is no DIF.

To summarize the agreement between the two methods of the Likelihood Ratio Test and Mantel- Haenszel for DIF of gender, the classification stability coefficient kappa was 0.725 which was statistically significant at a significance level α = 0.05 and the percentage of agreement between the two methods was 93.3%. These values indicated a considerable agreement between the two DIF methods for gender (Landis & Koch, 1977).

Also, Table 4 presents the agreement between the two methods in revealing the existence of items with the DIF for the country variable in 25 items. The two methods agreed on 18 items that there is no DIF, the three items on the existence of DIF towards the reference group (The rest of the Gulf countries), and the two methods agreed on the existence of DIF towards the focal group (Oman) for four items. As for the other remaining items, they counted five items that the two methods did not agree on, of which the Mantel- Haenszel method considers that it has a DIF towards Oman, while the Likelihood Ratio Test method sees that there is no DIF for any country. The Mantel- Haenszel method sees that two items have a DIF towards the rest of the countries, while the Likelihood Ratio Test method sees no DIF. On the other hand, there are items that the Likelihood Ratio Test method sees that it has a DIF towards Oman, while the Mantel- Haenszel method sees that there is no DIF for any country, and the Likelihood Ratio Test method sees that there are items that have DIF towards the rest of the countries while the Mantel- Haenszel method sees Hansel states that it has no DIF.

To summarize the agreement between the two methods of the Likelihood Ratio Test and Mantel- Haenszel for DIF based on country, the classification stability coefficient kappa was 0.655 which was statistically significant at the level α = 0.05 and the percentage of agreement between the two methods was 83.3%. These values indicated a considerable agreement between the two DIF methods for country (Landis & Koch, 1977).

Discussion

The study aimed to examine the differential item functioning of verbal ability test in the Gulf Mental Abilities Scale by revealing the items that show differential functioning according to gender and country and to determine the percentage of agreement between the two methods of likelihood ratio test and Mantel-Haenszel.

The study results indicated that there was no strong differential functioning in the verbal ability items of the GMMAS scale according to the gender using the likelihood ratio test and the Mantel-Hansel method, which is evidence of the validity of the verbal ability test on GMMAS. Benito et. al (2018) asserted no differential item functioning (DIF) is recognized as a proof of internal structure validity based on the standards for educational and psychological testing.

These results maybe because the item of this test was prepared based on precise criteria, including their conformity to the curriculum and the levels of students; so that the questions of this test are commensurate with the mental and chronological age of the study sample, and the accuracy in formulating the camouflages was taken into account so that they do not have a clear role in showing the differential functioning of one social type at the expense of another social type. It is possible that cultural differences and differences in context have been considered for both genders.

These results could be justified by the precise development of verbal ability test items, including their conformity to the curriculum and the student's verbal ability levels, so that the test items are commensurate with students' mental and chronological age and the accuracy in formulating the alternatives in multiple choice questions so that they do not have a clear role in showing the differential functioning. It is possible that cultural and contextual differences have been considered for both genders.

The results also showed that only one item from the verbal ability test suffered from a strong differential functioning according to the country variable using the Mantel-Haenszel method, which is evidence of the validity of the items of this test.

These results maybe because the test was conformed to the curricula used in the Arab Gulf states and that it was not affected by the information and goals that the student studied in his country, and the reason for the slight difference in functioning levels may be due to the nature of the test items. Mikyung (2001) showed that the emergence of a differential functioning of an item in terms of language might be due to the use of items unfamiliar to members of a group in the content of the items, and these words have different meanings in some countries.

The Mantel-Haenszel method was more stringent than the likelihood ratio test method in detecting the differential functioning in verbal ability test according to gender and country, as there were 16.7% and 33.3% of items with DIF by Mantel-Haenszel respectively, and 10% and 30% respectively by the likelihood ratio test method. This result is consistent with Mubarak (2006) which showed that the Mantel-Haenszel method revealed 65% of items with DIF according to the country variable, whereas the likelihood ratio test method revealed 50% of items. Further, the agreement between the two methods ranged from 93.3% and 83.3% respectively. The findings of the study are in line with those of Yildirim (2006) who pointed out that percentage of agreement between these methods was 82% in the PISA test, however these percentages was 48% in TIMSS test. This conclusion is especially helpful to psychometricians and applied researchers to use an easier-to-adopt strategy that employs classical test theory to avoid selecting the appropriate model in item response theory.

It is also essential to note that the present study contains several limitations. First, we employed the 3-PL model, which provided an excellent fit to the data, although other potentially well-fitted models could be tested, for instance, applying the 2-PL or Rasch model. Second, the study used two methods to detect differential functioning: the Mantel-Haenszel Method in classical test theory and the likelihood ratio test in item response theory. Other methods can be used, such as Item Characteristic Curve and Lord's chi-square.

In the future, it is necessary to analyse the item content for items with differential functioning of either gender variable or the country, to understand the reasons behind this DIF and to confirm or distribute possible bias. Further, the scope of a DIF analysis should be widened beyond simple demographics like age and race. Also, other assessment-related aspects should be investigated, including test delivery format (i.e., computer-based vs paper-and-pencil) and item response structure (dichotomous vs polytomous).

Conclusions

The current study has shown evidence that few items in the verbal ability test in GMMAS functioned differently across students' gender and country. The results of this study suggest that verbal test in GMMAS is a reliable and valid instrument for investigating cognitive abilities in the future.

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. Int. J. Learn. Man. Sys, 6(2), 61-73.

Allabadi, N. (2008). A comparison between four methods for detecting item function (Assimilation Study). Unpublished doctoral dissertation. Jordan's University.

Almehrizi, R. S. (2010). Comparing among new residual-fit and wright's Indices for dichotomous three -Parameter IRT model with standardized tests. Journal of Educational & Psychological Studies, 4 (2), 14-26.

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.

Alodat, A. M., & Zumberg, M. F. (2018). Standardizing the cognitive abilities screening test (CogAt 7) for identifying gifted and talented children in kindergarten and elementary schools in Jordan. Education of Gifted Young Scientists, 6(2), 1-13. http://doi.org/10.17478/JEGYS.2018.73

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

Alquraan, M., & Alkuwaiti, A. (2017). Differential item functioning in students rating of teaching effectiveness surveys in higher education according to academic disciplines: Data from a Saudi University. Journal of Educational and Psychological Studies [JEPS], 11(4), 770-780.

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. (2014). Standards for educational and psychological testing. American Educational Research Association American Psychological Association National Council on Measurement in Education.

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.org/10.1016/j.stueduc.2018.01.003

Ayala, R. J. (2009). The theory and practice of item response theory. The Guilford press, New York.

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.

Eleje, L., Onah, F., & Abanobi, C. (2018). Comparative study of classical test theory and item response theory using diagnostic quantitative economics skill test item analysis results. European Journal of Educational & Social Sciences 3(1), 71-89.

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. Report, Retrieved in 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.

Hammad, D. (2021). Detecting gender-related differential item functioning in Raven standard progressive matrices and its effect on Saudi sample's cognitive responses. Educational and psychological studies, 36 (111), 1-35.

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:10.1177/0146621616664046.

Kiany, G. R., & Jalali, S. (2009). Theoretical and practical comparison of classical test theory and item response theory. Iranian Journal of Applied Linguistics, 12(1), 167-197. https://www.sid.ir/en/journal/ViewPaper.aspx?id=247630

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

Liu, Q. (2011). Item purification in differential item functioning using generalized linear mixed models. Unpublished doctoral dissertation. Florida State University Libraries.

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

Mubarak, W. (2006). Differential item functioning for science test in (PISA) 2006 international study. An unpublished doctoral thesis. Yarmouk University.

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 English language among the students of Mutah University. Journal of Educational and Psychological Sciences, 19(2), 449- 475.

Ojerinde, D. (2013). Classical Test Theory (CTT) VS Item Response Theory (IRT): An Evaluation of the comparability of item analysis result. Lecture presentation at the institute of education.

Rashwan, R. (2021). Differential item function and its impact on the differential test function using item response theory models and multiple group confirmatory factor analysis. Journal of Educational Sciences and Human Studies, 6(15), 44-93.

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.

Shanmugam, S. (2020). Gender related differential item functioning of mathematics computation items among non-native speakers of English. The Mathematics Enthusiast, 17(1), 108-140. https://doi.org/10.54870/1551-3440.1482

Smith, R. (2011). Investigating the relationship between cognitive ability and academic achievement in elementary reading and mathematics. Retrieved from http://chalkboardproject.org

Spiegelhalter, D. J., Best, N. G., Carlin, B. P., and Van der Linde, A. (1998). Bayesian deviance, the effective number of parameters, and the comparison of arbitrarily complex models. Research Report, 98–009. Available at: http://www.med.ic.ac. uk/divisions/60/biointro.asp (accessed February 2018).

Tinajero, C., Lemos, S., Maria, A., Araujo, M., Ferraces, M., & Paramo, F. (2012). Cognitive style and learning strategies as factors which affect academic achievement of Brazilian university students. Psicologia: Reflexão e Crítica, 25(1), 105-113. https://doi.org/10.1590/S0102-79722012000100013.

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.

Wright, B. D., & Masters, G. N. (1982). Rating scale analysis. Chicago, IL: Mesa.

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.