Journal of Education and Practice
ISSN 2222-1735 (Paper) ISSN 2222-288X (Online)
Vol.11, No.29, 2020

Detection of Sex-Related Differential Item Functioning in Raven's Standard Progressive Matrices Test Using the Mantel-Haenszel Method

Amer J. Almarabheh[1*] Saleh R. Alshammari[2]
1.Department of Family and Community Medicine, Arabian Gulf University, Bahrain
2.Ministry of Education, Science Supervisor, Kuwait

Abstract
This study aimed to determine the presence of sex-related differential item functioning (DIF) in Raven's Standard Progressive Matrices (SPM) Test. The research design was a comparative study, where boys formed the focal group and girls formed the reference group. The software used was SPSS-v26, and binary data of focal and reference groups were analyzed using jMetrik software to detect DIF according to the Mantel-Haenszel method. A total of 1032 students (49.6% boys and 50.4% girls), 570 from intermediate school and 462 from secondary school, were selected from 24 schools using a stratified random-sampling procedure. The statistical analyses showed that the Raven's Standard Progressive Matrices (SPM) Test was one-dimensional. The results showed that, of five moderate DIF items, four items favored the focal group (boys), and one item favored the reference group (girls). The results also showed two large DIF items; the direction of the one item favored the reference group (girls), and the direction of the second item favor the focal group (boys). The findings showed that there were several unbiased items, but some were clearly biased against female performance. According to these findings, we recommend reanalyzing the data using methods depending on the item response theory, such as the logistic regression, simultaneous item bias test (SIB) or IRT-likelihood ratio (IRT-LR) methods to confirm the results seen here.
Keywords: Differential Item Functioning (DIF), Bias Item, Reference Group, Focal Group
DOI: 10.7176/JEP/11-29-10
Publication date: October 31st 2020

1. Introduction

Tests that are used in education and psychology for various purposes should meet specific standards, including validity, reliability, and practicality. These characteristics not only are the fundamental principles of measurement but also are the social values used by decision-makers in addition to measurement. In this regard, items in the test should not provide advantages or disadvantages for any subgroup at the same ability level. Otherwise, the test will be biased toward specific groups (Messick, 1995). The test is used as a data and information collection tool to make educational decisions and judgments. It is a measuring tool that contains a set of stimulants that represent the trait or ability meant to be measured; where researchers focused their efforts, in building tests and developing them, on extracting item Properties in terms of difficulty, distinction, and guesswork. Despite the importance of these features, they are not sufficient for judging the test items' validity for their designated purpose. This is because the response to the test items may be affected by factors other than the ability of the examined individuals, like gender, race, place of residence, language, or socioeconomic status; which may all affect the test results negatively, and subsequently the decisions based on them. Based on that, the items are described as having differential functioning towards a group or a category out of others (Jensen, 1980).

Roever (2005) mentioned that interest in differential item functioning in intelligence tests started in the beginning of the twentieth century, when Binet discovered by chance that the average grades of the economic high class in certain test items were higher than those of the economic low class. He then reviewed the content of these items and discovered that some of them were affected by the socioeconomic status of the examined students (biased towards the higher class). Subsequently, the biased items were removed from the test, and a new amended version was issued.

There are several meta-analyses demonstrating that there are sex differences in some cognitive abilities. The first meta-analysis showed that male students outperform female students in spatial and mathematical ability, but that female students outperform male students in verbal ability" (Francisco et al., 2004. P. 1). Hyde et al. (1990) found a male advantage in quantitative ability, but those researchers noted that many quantitative items were expressed in spatial form. Linn and Petersen (1985) found a male advantage in spatial rotation, spatial relations, and visualization. Voyer et al. (1995) found the same male advantage in spatial ability, finding that it was the most important sex difference in spatial rotation. Feingold (1988) found a male advantage in reasoning ability. Thus, research findings support the idea that the main sex difference may be attributed to overall spatial performance in which male students outperform female students (Neisser et al., 1996). Findings of Colom and Garcia (2002) supported the view that the information content has a role in the estimates of sex differences in general intelligence.

They concluded that researchers must be careful in selecting the markers of central abilities such as fluid intelligence, which is supposed to be the core of intelligent behavior.

There are many studies that focus on differences between male and female students in tests (for instance, Willingham and Cole, 1997; Gallagher et al, 2000). These studies indicate that male students have better spatial ability than female students. This suggests that male students use this spatial ability more often than females when solving problems that can give them advantages when solving certain kinds of problems in geometry. Some studies also indicate that female students are better than their male counterparts in verbal skills which can give them advantages in items where communication is important. Female students also score relatively higher in tests in mathematics that better match course work (Willingham and Cole, 1997). While there are a few studies that treated the sex differences in Raven's Matrices tests according to the DIF approach. Since 2010, some studies have demonstrated slight differential functioning in some items. (Shibaev et al., 2020) except that these studies were conducted on Raven's colored progressive matrices (CPM) and advanced progressive matrices (APM) rather than SPM, which is the subject of this study.

2. Literature Review

The purpose of most standardized achievement tests is to distinguish among ability levels of examinees and thereby rank order individuals on some skill or trait. Ranking examinees accurately requires that all the items in a test discriminate among levels of the valid skill or purported ability. Problems are encountered when a test contains item (s) that also discriminate among levels of abilities other than the valid skill. Unfortunately, because ordering is a unidimensional concept, we cannot order examinees on two or more abilities at the same time unless we base our ranking on (Laveault et al., 1994).

Differential item functioning (DIF) refers to a psychometric difference in how an item functions for two groups. DIF refers to a difference in item performance between two comparable groups of examinees, that is, groups that are matched with respect to the construct being measured by the test. The comparison of matched or comparable groups is critical because it is important to distinguish between differences in item functioning from difference between groups (Dorans, & Holland, 1992). It is important to determine whether items have DIF for at least two reasons. The presence of DIF signals potential bias, and bias has an impact on validity of inferences drawn from group comparisons (Lane, Wang, & Magone, 1996).

Zumbo (1999) mentioned that there are two types of Differential Item Function: the uniform and non-uniform DIF. Uniform DIF appears when the probability of answering the item correctly is consistently higher in one group and across all levels of the ability; thus there is no interaction between the ability level and group membership. Whereas the non-uniform DIF appears when there is an interaction between the level of ability (θ) and group membership; which mean that the pattern of differences in the probability of responding to the item is not the same at all levels of ability, so we find these differences are in favor of one group in a certain ability level, while it is in favor of another group in another ability level. Figure 1. Shows both uniform and non-uniform DIF.

Figure 1. Item characteristic curves demonstrating of uniform and non-uniform DIF

2.1 Detecting DIF Methods

Methods of detecting DIF are basically classified according to classical test theory (CTT) and item response theory (IRT). According to CTT, methods of detecting DIF are analysis of variance (ANOVA), chi-square, converted item index, logistic regression, Mantel-Haenszel (MH). IRT methods are Lord's chi square ($\chi 2$), Raju's area measure, and IRT-likelihood ratio (IRT-LR) and the simultaneous item bias test (SIBTEST) (Camilli & Shepard, 1994; Oshima & Morris, 2008).

2.1.1 Mantel-Haenszel (MH) Method

The mantel-Haenszel (MH) method is a statistically powerful technique for detecting DIF. The MH procedure was first developed by Mantel and Haenszel in 1959 and was used as a method to detect DIF for the first time by Holland and Thayer in 1988 (Holland & Wainer, 1993). The mantel-Haenszel procedure estimates of the common odds ratio α_{MH} across all matched categories. The form at its index is given as follows:

$$\alpha_{MH} = \frac{\sum_i P_{ri} q_{fi} N_{ri} N_{fi/N_i}}{\sum_i q_{ri}\, p_{fi} N_{ri} N_{fi/N_i}} = \frac{\sum_i a_i d_{i/N_i}}{\sum_i b_i c_{i/N_i}} \text{ …….. (1)}$$

Where P_{ri} is the proportion of the reference group in score interval I who answered the item correctly, and q_{ri}=1-P_{ri}. Similarly, P_{fi} is the proportion of the focal group who answered the item correctly, and q_{fi}=1-P_{fi}. Thus, α_i is the ratio of the odds (p/q) that the reference group students have answered the item correctly to the odds that the focal group students have answered the item correctly. If there is no difference in the performance of the two groups on this item within this score interval, then α_i will be equal to 1. If the focal group performs better on the item than the reference group, then α_i<1. If, on the other hands the reference group performs better than the focal group, $\alpha_i > 1$.

α_{MH} is average factor by which the odds that a member of the reference group responds correctly to the item exceeds the odds that a member of the focal group responds correctly to the item. It is observed that the index is weighted by the number of cases in the interval; also, that the interval in which the numbers at cases in the interval; also, that the interval in which the numbers of cases in the two groups are more nearly equal receives the heavier weight. There is a chi-square test associated with the MH approach, namely a test of the null hypothesis, H_0: α_m=1

$$\chi^2{}_{MH} = [|\sum_m R_{rm} - \sum_m E(R_{rm})| - 0.5]/\sum_m Var(R_{rm}) \text{…….. (2)} \quad \text{where}$$

$$E(R_{rm}) = E(R_{rm}|\alpha = 1) = {}^{N_{rm}R_{rm}}/_{N_{tm}} \text{…….. (3)}$$

$$Var(R_{rm}) = Var(R_{rm}|\alpha = 1) = [N_{rm}R_{tm}N_{fm}W_{tm}]/[N_{tm}^2(N_{tm} - 1)] \text{…….. (4)}$$

And where the -0.5 in the expression for $\chi^2{}_{MH}$ serves as a continuity correction to improve the accuracy of chi-square percentage points as approximation to the observed significance levels. The quantity $\chi^2{}_{MH}$ is distributed approximately as a chi-square with one degree of freedom.

For the sake of convenience α_{MH} is transformed to another scale, yielding an index that is referred to as MH D-DIF (Δ_{MH}) by means of the conversion.

$$\Delta_{MH} = -(2.35)\ln\ (\alpha_{MH}) \text{ …….. (5)}$$

This transformation centers the index about the value 0 (which corresponds to the absence of differential item functioning), and puts it on a scale roughly comparable to the Educational Testing Service (ETS) delta scale of item difficulty and reverse the index so that the item positive values of Δ_{MH} indicate that the item favors the focal group; negative values indicate that the item favors the reference group and disfavors the focal group.

To use the Δ_{MH} measure to identify test items that exhibit varying degrees of DIF a classification scheme was developed of ETS for use in test development that puts items into one of three categories: negligible DIF (A), Intermediate DIF (B), and large DIF(C). Items are classified as A for a particular combination of reference and focal groups if either Δ_{MH} is not statistically different from zero or if the magnitude of the Δ_{MH} values is less than one delta unit in absolute value. Items are classified as C if Δ_{MH} both exceeds 1.5 in absolute value and is statistically significantly larger than 1 in absolute value. All other items are classified as category B. In both category A and C statistical significance is at the 0.05 level for a single item. In this study, girls formed the focal group, while boys are formed the reference group (Holland & Wainer, 1993).

2.2 Study Problem and Questions:

Sex differential performance in nonverbal ability tests is a cause for alarm, however in studies that examine sex differential performance in nonverbal ability tests, it is rather challenging to determine whether the significant differences in the nonverbal ability test between boys and girls is due to their true differences in ability or test-related factors such as item-type. Therefore, this study will address the gap in examining this issue of sex difference because of item-type and identify characteristics of content of Raven's Standard Progressive Matrices Test items that cause the differential performance by sex. By detecting DIF according the characteristics of the items content for the Raven's (SPM) test, the issue of the apparently widening sex gap will be explained from a new perspective of item characteristics using the Mantel-Haenszel method. Therefore, the research question is: "What are the items that show differential functioning in Raven's Standard Progressive Matrices Test according of the student's sex?"

2.3 The Significance of the Study:

The main objective of the study is to investigate which items that show DIF, for male and female students on the Raven's Standard Progressive Matrices (SPM) Test using Mantel-Haenszel (MH) which is based on classical test theory (CTT). This study is one of the few studies that shows interest in examining the differential item functioning in (SPM) test to the best of our knowledge. It seeks to find evidence that its items lack differential functioning,

using a statistical method suitable for finding differential functioning (Mantel-Haenszel method). The scientific significance of the study lies in that results may provide evidence of the suitability of the RSPM test and the validity of its results to be used for the functions it was designed for, including but not limited to achieving fairness in screening and accepting students in programs of gifted education despite the difference in sex.

3. Methodology

3.1 Study design:

This is a comparative research study, where the girls form the reference group and boys form the focal group because they form the interest of this study. A total of 12 schools were selected from the six educational districts in the State of Kuwait, where two schools were selected from each educational district chosen randomly (total=24 school).

3.2 Study community and Participants:

The study community consisted of all male and female students in the eighth grade, intermediate school, and the eleventh grade, secondary school, in the public schools of the Kuwaiti Ministry of Education (totaling 23315 students according to the statistics supplied by the Ministry of Education for the academic year 2012/2013. Participants were 1032 students (49.6% boys and 50.4% girls), 570 from intermediate school and 462 from secondary school, ranging in age from 13 to 16 years. Each participant completed the SPM test. The mean SPM score for the total sample was 30.17 (SD=6.46). The mean score for boys was 30.28 (SD=6.05), and for girls it was 30.07 (SD=6.85). All were Kuwaiti citizens and students in the governmental schools in the six districts in Kuwait. Twenty-four schools were selected from the six districts using a stratified random-sampling procedure.

3.3 Instrument:

The Raven's Standard Progressive Matrices (SPM) Test is one of the most widely used measures of cognitive ability. SPM scores are considered among the best estimates of general intelligence. It is a nonverbal test designed to assess ability to reason and solve new problems without relying extensively on declarative knowledge derived from schooling or previous experience. It is one of the most well-known, formal, broad intelligence tests. It was prepared by Raven in 1938 as a tool that measures general intelligence. The test in its original form contained 60 matrices. It is also considered one of the best tests to measure the capacity for abstract (nonverbal) reasoning, and has good psychometric characteristics, upon which a large body of published scientific studies have been built, and it has been accepted and used in the five continents of the world (Abdel-Khalek & Raven, 2006). In Kuwait, the Ministry of Education adjusted the test after studying its standardization, because 12 items were deleted, and the order of items was adjusted according to the difficulty factor.

3.4 Data Analyses:

The classical test theory was used in data analysis. Reliability analyses were carried out using internal consistency reliability using Kuder-Richardson KR-20, which is a reliability measure for a test with binary variables. To evaluate construct validity, we conducted an exploratory factor analysis principal component analysis as the extraction method, and varimax with Kaiser Normalization as the rotation method. A factor was considered important if its eigenvalue exceeded 1.0. The communality represents the percentage of variance of the tool item accounted for all factors. A p-value of < 0.05 was considered as statistically significant for all tests. All statistical analyses were conducted using the Statistical Package for the Social Sciences (SPSS) version 26, and Binary data of focal and reference groups were analyzed using jMetrik 4 software (Meyer, J. P., 2014) to detecting DIF according to the Mantel-Haenszel method, girls formed the focal group, while boys are formed the reference group.

3.5 Validity and Reliability:

The validity of the SPM test has been verified using exploratory factor analysis using the principle component method, then using orthogonal ratio with the varimax method for all items. This was to provide a better explanation of the psychometric properties extracted before rotation. The eigenvalue was used according to Kaiser Criterion where the eigenvalue per factor increases more than one whole, and 0.30 was considered the least significant factor loading of the item according to Guilford Scale. The results showed the Kaiser-Meyer-Olkin Measure to be 0.902 and the Bartlett's Test of Sphericity to be chi-square = 7226.500 and $p < 0.001$, which indicated that the data in this study were suitable for factor analysis. Table 1 shows that the eigenvalue of the first factor was 7.109, and 14.81% of the total variance was explained. The eigenvalue of the second factor was 2.075, and 4.32% of the total variance was explained. When the eigenvalue of the first factor is divided by that of the second factor, the result equals 3.43 (i.e., greater than 2, which is considered an indicator of one-dimensionality (Hattie, 1985)). This means that the SPM test is loaded by one general factor. It appears that this measure of reasoning ability does not require other cognitive abilities to any significant degree. For further explanation, Fig. 2 is the scree plot showing the unidimensionality of the items.

Table 1. Eigenvalues and the deduced degrees of variance for the SPM test

Factor	Eigenvalue	% of Variance	Cumulative %	Factor	Eigenvalue	% of Variance	Cumulative %
1	7.109	14.810	14.810	8	1.143	2.381	35.448
2	2.075	4.323	19.133	9	1.132	2.358	37.806
3	1.626	3.387	22.521	10	1.103	2.298	40.105
4	1.382	2.879	25.400	11	1.08	2.245	42.350
5	1.253	2.611	28.011	12	1.07	2.219	44.569
6	1.239	2.580	30.591	13	1.03	2.155	46.723
7	1.188	2.476	33.067	14	1.01	2.109	48.832
Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO) Index 0.902							
Bartlett's Test of Sphericity Chi-Square = 7226.500; df = 1128; p < 0.001							

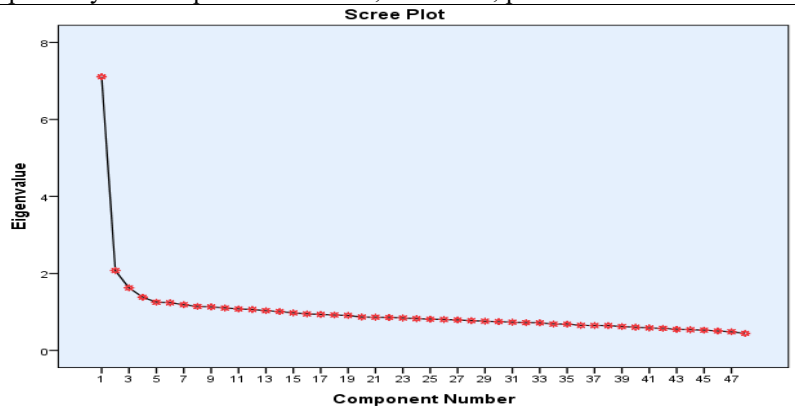

Fig. 2 Scree plot of eigenvalues of factors resulted from psychometric analysis items

The reliability analysis of Kuder-Richardson Formula (KR-20) was used to measure the internal consistency of the SPM test items. The analysis showed that the reliability coefficient for internal consistency of the SPM test was 0.86, which is an acceptable value for proceeding with this study.

4. Results:

4.1 Results of Mantel-Haenszel and Effect Size (odds ratios)

To answer the question of the study, Mantel-Haenszel statistics were calculated, as well as the effect (common odds ratio) for each item in the SPM test using jMetric statistics program, in terms of student sex, where girls were considered the reference group and boys were the focal group. The data in Table 2 shows the summary results (M-H statistics, the significance levels, odds ratios (effect size), and 95% confidence interval) for each of the forty-eight items from the Mantel-Haenszel method to identify Differential Item Functioning on the SPM test. Eight items from the SPM test gave numbers of 9, 10, 17, 18, 31, 42, 44, and 45, suggesting differential functioning according to the sex of the student, where the data showed that the values of Mantel-Haenszel statistic for these items were statistically significant at $\alpha < 0.05$.

Table. 2 Summary Results from the Mantel-Haenszel Method to Identify DIF on the SPM test

Item	M-H Statistics	P-value	Odds Ratio	95% CI		Item	M-H Statistics	P-value	Odds Ratio	95% CI	
1	2.01	0.16	0.35	0.07	1.66	25	1.88	0.17	1.26	0.91	1.75
2	0.15	0.69	0.60	0.04	8.53	26	0.07	0.79	0.96	0.69	1.33
3	2.06	0.15	5.02	0.53	47.84	27	1.50	0.22	0.83	0.61	1.12
4	3.17	0.07	1.94	0.94	3.98	28	0.00	0.97	0.99	0.68	1.45
5	0.94	0.33	0.58	0.19	1.76	29	0.58	0.45	1.12	0.84	1.50
6	1.87	0.17	1.68	0.79	3.59	30	3.18	0.07	0.76	0.57	1.03
7	0.06	0.81	0.90	0.38	2.15	31	5.41	**0.02***	1.41	1.06	1.88
8	1.49	0.22	1.67	0.73	3.81	32	1.56	0.21	0.82	0.60	1.12
9	4.65	**0.03***	0.58	0.36	0.96	33	0.52	0.47	0.88	0.62	1.25
10	4.76	**0.03***	1.99	1.08	3.69	34	0.88	0.35	1.14	0.87	1.50
11	0.55	0.46	0.82	0.49	1.38	35	0.32	0.57	1.08	0.82	1.43
12	0.15	0.69	1.11	0.66	1.86	36	0.01	0.93	0.99	0.73	1.33
13	0.00	0.95	1.01	0.68	1.51	37	2.94	0.09	1.34	0.96	1.87
14	0.02	0.87	0.98	0.73	1.31	38	3.61	0.06	1.35	0.99	1.85
15	0.87	0.35	0.81	0.52	1.26	39	0.03	0.86	0.97	0.73	1.31
16	0.71	0.40	1.13	0.85	1.49	40	2.90	0.09	1.32	0.96	1.83
17	6.66	**0.01***	0.61	0.42	0.89	41	1.17	0.28	1.20	0.86	1.66
18	6.89	**0.01***	0.63	0.44	0.89	42	6.91	**0.01***	0.60	0.41	0.88
19	0.16	0.69	0.92	0.61	1.38	43	0.08	0.77	1.05	0.76	1.45
20	0.82	0.36	0.84	0.58	1.22	44	10.25	**0.00***	1.71	1.23	2.39
21	3.37	0.07	0.71	0.49	1.02	45	8.27	**0.00***	0.47	0.27	0.80
22	0.17	0.68	1.07	0.77	1.48	46	1.43	0.23	0.78	0.51	1.18
23	0.09	0.76	1.05	0.75	1.47	47	1.05	0.30	1.34	0.76	2.37
24	0.07	0.80	0.96	0.71	1.31	48	0.17	0.68	0.89	0.52	1.53

*: Significant at $\alpha < 0.05$

4.2 Results of Delta Mantel-Haenszel and DIF Direction

To determine the amount of differential functioning and its direction, the values of Mantel-Haenszel statistics for the items that showed differential functioning were converted to delta value (Δ_{MH}) according to formula no. (5). Table 3 shows the summary results from the Mantel-Haenszel method to identify Differential Item Functioning on the SPM test. The study results showed five moderate DIF item whose numbers were 9, 17, 18, 42 and 44. The direction of DIF in these items showed that four items 9, 17, 18, and 42 favored the focal group (boys), and one item (44) favored the reference group (girls). The study results showed two large DIF items (10 and 45). The results indicate that the direction of item (10) favors the reference group (girls), and the direction of the second item (45) favors the focal group.

Table 3. Transformation of the Odds Ratio (effect size) to Delta Mantel-Haenszel and Identification of DIF with Results Comparing the Rating Scale

Item	Effect Size (α_{MH})	Transformed value (Δ_{MH})	95% CI		DIF Direction
9	0.58	1.28	2.43	0.11	B^{FG}
10	1.99	-1.62	-0.17	-3.07	C^{RG}
17	0.61	1.16	2.02	0.27	B^{FG}
18	0.63	1.09	1.91	0.28	B^{FG}
31	1.41	-0.81	-0.13	-1.49	A^{RG}
42	0.60	1.20	2.12	0.30	B^{FG}
44	1.71	-1.26	-0.48	-2.05	B^{RG}
45	0.47	1.77	3.06	0.52	C^{FG}

A: Negligible, **B**: Moderate, **C**: Large, RG: Reference Group, FG: Focal Group

Rather than present all the non-parametric Item Characteristic Curves (ICCs) for the items that showed DIF, two ICCs are presented (figure 3) for the purpose of demonstration. These non-parametric ICCs were selected because they demonstrate, first, an item 10 show that the reference group (girls) has a higher chance to answer correctly than the focal group (boys), and then second, an item 45 show that the focal group (boys) has a higher chance to answer correctly than the reference group (girls).

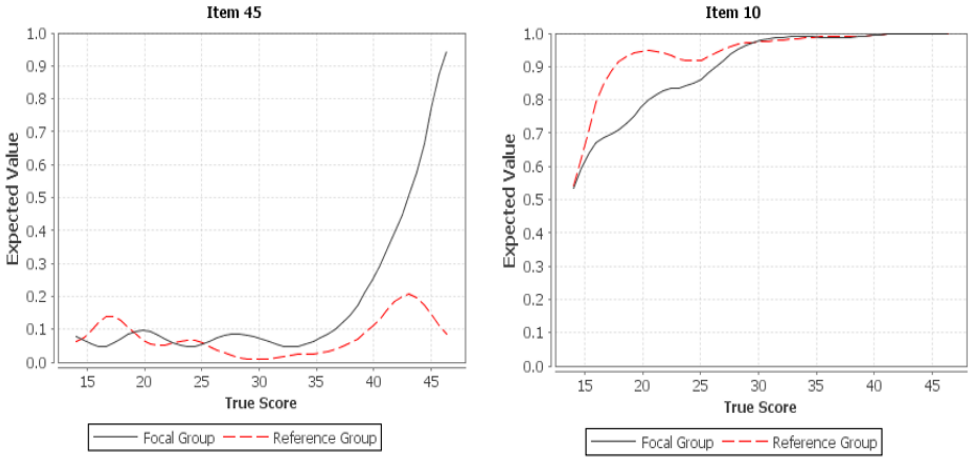

Figure 3. Non-Parametric Item Characteristic Curves

5. Discussion

The statistical analyses showed that the Raven's Standard Progressive Matrices (SPM) Test is one-dimensional. It seems that this measure of reasoning ability does not require other cognitive abilities to a significant degree. The results of the first procedure used in this study indicate that 8 of the 48 items showed differential functioning according to the sex of the student. The results of the second procedure showed that, of five moderate DIF items, four favored the focal group (boys), and one item favored the reference group (girls). The results also showed two large DIF items. The direction of the one item favored the reference group (girls), and the direction of the second item favored the focal group (boys). Finally, the results indicated that there was one item showing negligible DIF favoring the reference group (girls), and it was ignored according to the instructions of the Mantel-Hansel method. Our results support the idea that comparisons between diverse groups show minimal bias when Raven's Standard Progressive Matrices Test is used. Therefore, there is a sex difference in the SPM Test (Colom and Garcia, 2002); however, given that this test is based on abstract figures and that boys have on average a higher spatial ability than girls (Voyer et al., 1995), we predicted that some items may be easier for boys. Thus, boys might solve some items correctly because of their visuo-spatial nature. This could be considered as a threat to bias (Francisco et al., 2004, p. 10).

The results of current study provide evidence that there are sex differences in performance on few test items in SPM test. The differences in cognitive abilities between the two sexes have always been a subject of investigation for researchers (Wechsler et al, 2014). The differences between the sexes in Raven's Matrices tests have always been among the most interesting, most controversial subjects; and yet, the studies have not reached a clear conclusive result (Yang et al, 2014). The results of the studies on the differences in terms of sex are inconsistent and do not follow the same path. Some studies have attributed these differences between the sexes to differences in factors such as the nature of the sample and whether the sample was representative of its community, and the use of statistical methods that fail to identify DIF. A study by Mackintosh & Bennet (2005) showed that boys surpassed girls in some of the items that are similar in a certain pattern. However, their study sample was not large, and the study was conducted using Raven's APM rather than SPM. Therefore, the study emphasized the necessity for researchers to conduct studies that employ methods of qualitative research such as focus groups in order to evaluate the reasons behind the existence of differential item functioning and to verify the sources of variance that affect test scores. This allows determination of whether the subgroups are affected by the same sources of variance and whether any of the sources of variance unfairly favor a subgroup before judging the item's bias. In summary, the authors have investigated the visuo-spatial basis of the SPM test. The male advantage on this test could derive from their visuo-spatial nature.

6. Conclusion and Recommendations

Results of Differential Item Functioning obtained with this study show that boys have on average a higher spatial ability than girls. Based on the description of the results and discussion above, the following conclusions were reached: The study is a comparative study, using DIF data to reveal different performance characteristics of male and female examinees. Therefore, identifying DIF items is as important as determining the underlying source of difficulty across the focal and reference groups. A complete evaluation of test quality must include an evaluation of each question. Therefore, questions should assess only knowledge or skills that are identified as part of the domain being tested and should avoid assessing irrelevant factors and examination items should be fair among examinees from all possible subgroup of the population of the examinees. DIF is an issue that must be properly addressed in examinations and tests designed for heterogeneous groups. Based on these findings, the following

recommendations can be considered for future studies: Conduct further research including additional variables other than sex, especially age and region. It is also necessary for item writers to develop test items and subject them to pilot studies to select items that are free from DIF. Another step in this research would be to include and compare other methods of identifying DIF items, because it was mentioned above that DIF detection methods can vary in their results. Therefore, it may be advantageous to reanalyze the data using a nonmodal-based method (i.e., Item Response Theory) such as the logistic regression, simultaneous item bias test (SIB) and IRT-likelihood ratio (IRT-LR) methods to confirm the results seen here. Future studies can be directed towards examining sex differential performance for intelligence tests, specifically among gifted students. Another important recommendation for future studies is that another method of detecting DIF such as the multidimensional model to detect the presence of differential dimensions (Shelly & Stout, 1993) is used to examine whether both DIF methods flag the same items.

Limitation: The results of this study are limited to using the Mantle-Hansel method to detect the differential item functioning, which is based on the Classic Test Theory (CTT). Another limitation of the study is related to the type of statistical program used in the analysis process, which is jMetrik software.

Acknowledgement: The authors would like to acknowledge all those who contributed to conducting this study, especially the students and cooperative teachers.

References

Abdel-Khalek, A.;Raven, J.(2006).Normative Data from the Standardization of Ravens Standard Progressive Matrices in Kuwait an International Context. *Social Behavior and Personality*, 34(2): 169-179. https://doi.org/10.2224/sbp.2006.34.2.169

Camilli, G., & Shepard, L. A. (1994). Methods for identifying biased test items. California: Sage. https://us.sagepub.com/en-us/nam/methods-for-identifying-biased-test-items/book3416

Colom, R., Garcıa,L. F., Abad,F. J.,& Juan-Espinosa,M. (2002). Null sex differences in general intelligence: evidence from the WAIS-III. *Spanish Journal of Psychology*, 5, 1 29-35.

Laveault, B. D. Zumbo, M. E. Gessaroli, & M. Boss. (1994.). *Modern theories of measurement: Problems & issues*. Ottawa, Ontario: University of Ottawa.

Dorans, N.J., & Holland, P.W. (1992). DIF detection and description: Mantel- Haenszel and standardization. (RR-92-10). Prin. *Educational Testing Service*.

Feingold, A. (1988). Cognitive gender differences are disappearing. *American Psychologist*, 43, 95–103. https://doi.org/10.1037/0003-066X.43.2.95

Francisco J. A., Roberto, C.I., & Rebollo, S., E. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: evidence for bias. *Personality and Individual Differences*, 36, 1459-1470. https://doi.org/10.1016/S0191-8869(03)00241-1.

Jensen, R. (1980). *Bias in mental testing*. New York: Macmillan publishing co. Inc.

Hattie,J.(1985).Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement*. 9, 139-164.

Holland, P. W., & Thayer, D. T. (1988). An alternative definition of the ETS delta scale of item difficulty. *Educational testing Service, Technical report* (85-64)/ *Research Report* (85-43).

Holland, P. W., and Wainer, H. (1993). *Differential Item Functioning*. Hillsdale, NJ: Erlbaum.

Hyde,J.,Fennema,E.,&Lamon,S. (1990). Gender differences in mathematics performance: a meta-analysis. *Psychol Bull*, 107,139–153. https://doi.org/10.1037/0033-2909.107.2.139

Gallagher, A.M. De lesi, R; HoIst, P.C; McGill Cuddy Delsi, A.V, Morley, M. & Cahalan, C. (2000). Gender differences in advanced mathematical problem solving. *Journal of Experimental Child Psychology*, 75, 165-190 https://doi.org/10.1.1.536.2454&rep=rep1&type=pdf

Lane, S., Wang, N. & Magone, M. (1996). Gender-related differential item functioning on a middle-school mathematics performance assessment. *Educational Measurement: Issues and* Practice, 15(4), 21-27. From https://files.eric.ed.gov/fulltext/ED392821.pdf

Linn,M.,& Petersen,A. (1985). Emergence and characterization of sex differences in spatial ability: a meta-analysis. *Child development*, 56, 1479–1498.

Mackintosh N, Bennett E (2005) what do raven's matrices measure? An analysis in terms of sex differences. Intelligence 33: 663–674. https://doi.org/10.1016/j.intell.2005.03.004

Mantel, N., & Haenszel, W. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute*, 22 (4), 719- 748. doi.org/10.1093/jnci/22.4.719

Messick, S. (1995). Validity of psychological assessment: Validation of inferences from persons' responses and performances as scientific inquiry into score meaning. *American Psychologist*, 50(9), 741-749. https://doi.org/10.1037/0003-066X.50.9.741

Meyer, J. P. (2014). *Applied Measurement with jMetrik*. New York: Routledge.

Neisser, U.,Boodoo, G.,Bouchard,T. J., Boyking,W.,Brody,N., Ceci,S.,Halpern,D., Loehlin, J.,Stenberg, R.,&

Urbina,S. (1996). Intelligence: knowns and unknowns. *American Psychologist*, 51,2 77-101. https://doi.org/10.1037/0003-066X.51.2.77

Oshima, T. C., & Morris, S. (2008). Raju's differential functioning of items and tests (DFIT). *Educational Measurement*: *Issues and Practice*, 27(3), 43-50. https://doi.org/10.1111/j.1745-3992.2008.00127.

Roever, C. (2005). That's not fair! Fairness, bias, and differential item functioning in language testing. SLS Brownbag. Retrieved 16/3/2006 from http://www2.hawaii.edu/~roever/brownbag.pdf.

Shibaev V, Grigoriev A, Valueva E, Karlin A. Differential Item Functioning on Raven's SPM+ Amongst Two Convenience Samples of Yakuts and Russians. Psych. 2020 Mar;2(1):44-51.

Voyer, D., Voyer, S., & Bryden, M. P. (1995). Magnitude of sex differences in spatial abilities: A meta-analysis and consideration of critical variables. Psychological Bulletin, 117(2), 250–270. https://doi.org/10.1037/0033-2909.117.2.250

Wechsler, S., Nakano, T., Domingues, S., Rosa, H., Silva, R., Silva-Filho, J. and Minervino, C (2014). Gender differences on tests of crystallized intelligence. *European Journal of Education and Psychology*.7 (1): 59-72. https://doi.org/ 10.1989/ejep.v7i1.152

Willingham, W.W. & Cole, N.S; (1997). Gender and fair assessment. New Jersey, U.S.A: Lawrence Erbaum associate Publishers. https://psycnet.apa.org/record/1997-08628-000

Yang, W., Liu, P., Wei, D., Li, W., Hitchman, G., Li, X.,Qiu,J. & Zhang, Q. (2014).Females and Males Rely on Different Cortical Regions in Raven's Matrices Reasoning Capacity: Evidence from a Voxel-Based Morphometry Study. *PLOS ONE*. 9(3):1-6.

Zumbo, B. D. (1999). *A Handbook on the Theory and Methods of Differential Item Function* (Dif), Ottawa، on Directory of Human Resources research And Evaluation Department of National Defense.