MANKIND QUARTERLY Mankind Quarterly, (), 1-13 Edition
ISSN:

Gender Differential Item Functioning in Raven's Coloured Progressive Matrices Test (CPM)

Ismael Salamah Albursan* Intisar Abonagma† Salaheldin Farah Attallah Bakhiet‡
Madina Hussien Dawsa§ Ikhlas Taher Haj Adam¶

Abstract

This study aimed to reveal differential item functioning (DIF) by gender for the items of Raven's Coloured Progressive Matrices (CPM) test, separately for age groups of 4-6 years, 7-9 years, and 10-12 years. The number of students in the study sample was 2624 including 1190 males and 1434 females, selected from the kindergarten and primary school levels from the states of Khartoum and Gadaref in Sudan. The 36-item Coloured Progressive Matrices (CPM) test was applied, and differential item functioning was analysed on the total sample and for the three age groups using the Mantel-Haenszel method. The number of items that show differential item functioning by gender rose from 8% in the 4-6 years age group (preschool) to 11% for ages 7-9 (primary school) and 32% for the 10-12 years age group. This study concludes with a number of recommendations calling for studies of DIF in intelligence tests of the Raven type in relation to gender and also age and ethnic groups both in Sudan and elsewhere.

Keywords: IQ tests, Differential item functioning, Mantel-Haenszel method, Coloured Progressive Matrices, CPM, Sudan.

1 Introduction

Historically, the study of differential item functioning (DIF) of intelligence tests dates back to 1911. When Binet reviewed the first results on his test, which he had developed to measure mental ability, he noted that the average performance of children from the higher economic classes was much higher than the average performance of children from the lower economic classes (Roever, 2005). He was convinced that there were items in his test that were affected by the social and economic levels of the examined students. Some items of the test were deleted accordingly. The human rights movement in the late sixties and early seventies played a role in highlighting the issue of bias in mental abilities tests. The aim was to achieve justice and equality between individuals in educational and employment opportunities by making tests as free of bias as possible (Conoley, 2003).

The term differential item functioning (DIF) has been used since the beginning of the 1980s to denote the methods and statistical treatments that are used to detect the bias of test items. Differential item functioning is a statistically derived function to express the difference in response to an individual item between two groups at the same ability level. There are those who use the term differential item functioning as a synonym for bias. This, however, is not justified. The term differential functioning in items of a test is

*King Saud University, Department of Psychology, College of Education, Riyadh, Saudi Arabia; email: ibursan@ksu.edu.sa

†University of Khartoum, Department of Psychology, College of Arts, Khartoum, Sudan; Email: Intisar.abunagma@gmail.com

‡Gifted Education Program, Department of Special Education, College of Education, Administrative and Technical Sciences, Arabian Gulf University, Manama, Kingdom of Bahrain; Email: salaheldinbakhiet@gmail.com

§University of El Fasher, Faculty of Education, Department of Psychology, El Fasher, Sudan; Email: algesses@gmail.com

¶Magadeem Private Schools, Gedaref, Sudan; Email: Simber.org@gmail.com

used to identify items in which the probability of a correct answer is different between two groups *with the same level of ability*, where ability is defined as the ability construct that the test is designed to measure: in technical terms, the average performance on the complete test.

Items with differential functioning can be described as those biased in favour of one group over the other for reasons not related to the ability of individuals only (Camilli & Shepard, 1994). Hambleton and Rogers (1995) described differential item functioning as difference in the probability of the correct answer on an item in different groups of equal ability. Hambleton et al. (1991) argue that differential item functioning is present when the response functions are different in different subgroups. Jensen (1980) defines bias as a systematic error that makes performance on a test better for one group than for another.

Dorans and Holland (1994) distinguished between the concept of bias and the concept of differential item functioning in this way: Differential item functioning in psychometrics is present when the item works differently in one group from the other group, while bias of an item carries social meaning, inequality and equality. Hence, differential item functioning is a prerequisite for considering an item biased, but it is not sufficient for it, in the sense that when an item shows differential performance for a group, then additional procedures are required to decide that the item is biased. Such procedures may include arbitration of that item or an empirical evaluation.

One example of a study of differential item functioning by gender is the study of Gamer and Engelhard (1999) on mathematical performance which confirmed that males score higher in geometry and measurement, while females excel in algebra. A study of Mendes-Barnett & Ercikan (2006) identified sources of DIF and differential bundle functioning for males and females on a Principles of Mathematics exam, showing that males perform better on items that require higher cognitive processes while females were superior in items of mathematical relationships and nominal descriptions. Drina (2007) studied differential item functioning on a ninth-grade mathematics proficiency exam according to gender and region. This study found four items out of 36 with differential item functioning linked to gender, while no region-related differences appeared.

A study of Ryan and Chiu (2001) on the sources of differential item functioning in a mid-western mathematics placement exam indicated that males performed better than females on difficult vocabulary items that need high thinking skills and on geometry, as well as on items that include drawings and pictures. Among studies conducted in the Arab world, Al-Bursan (2013) examined gender-based differential item functioning and found that DIF in favour of males increased as the cognitive level of test items decreased — that is, gender-related differences became more pronounced on lower-level items. This suggests that the complexity and cognitive demands of test items may influence the degree of gender-related DIF observed.

Studies of gender differences in measured general and special abilities (Hyde & Linn, 1988; Hyde et al., 1990; Linn & Petersen,1985; Maccoby & Jacklin, 1974; Voyer et al., 1995) found that boys outperform girls in spatial and mathematical reasoning and on tests measuring quantitative abilities, spatial rotation, spatial relations and visualization, whereas females outperform males in verbal ability. They concluded that there are gender differences in some cognitive abilities. This contrasts with investigations by other researchers (Colom et al., 2000, 2002; Jensen, 1998a), who found a negligible sex difference in general intelligence. This indicates that cognitive gender differences result from differences in specific cognitive abilities, but not from differences in the core of intelligence (Abad et al., 2004). This conclusion was further reinforced by Jensen (1998b), McLaurin et al. (1973) and Paul (1985), who found that there is no gender difference on the total score of the Progressive Matrices (PM) test (Raven et al., 1998), one of the most widely used measures of cognitive ability (g). Males and females obtain similar scores in the PM Test.

Up to date, few studies have applied DIF analysis to interethnic comparisons using Raven's Matrices. Summarizing the results of these studies, there are very few or no significant differences in item functioning related to ethnicity. Rushton et al. (2004) concluded that Matrices tests are ethnically unbiased. The study of Al-Qati (2016) of gender bias of the Wechsler test items for measuring children's intelligence (Saudi version) showed that the ratio of gender-biased items was not high, distributed almost equally, and that the effect of bias on the overall test score was negligible. The study of Al-Bustanji (2004) comparing four procedures for detecting gender-based differential item functioning on a test of specific mental abilities in Jordan showed that differential functioning in general was in favour of males in math and spatial ability, and in favour of females in verbal ability.

When researchers compared the correlation coefficients, reliability and logistic of the Mantel-Haenszel (MH) method with other models of DIF (Baghi & Ferrara, 1990; Dodeen, 2004; Harris & Carlton, 1995; Kim & Cohen, 1992; Penfield, 2001; Raju et al., 1993), they found that the MH method was the most accurate, easiest to use and most powerful method for detecting differential performance on an item.

The current study is based on these previous studies. The basic conclusion of these studies was: DIF describes differences in the statistical properties of an item between groups of equal ability, determined by factors specific to group membership such as differential opportunities to learn or differences in socialization. While hormonal and other biological differences have been suggested as explanations of gender differences in specific abilities, they are not typically considered to explain gender differences on specific items. The presence of DIF can have serious consequences for the interpretation of test scores for both groups and individuals. One of our observations is that gender differences in item performance reported by various studies were few. However, some indicated that males perform better on mathematics tests than females and there are significant gender differences in special abilities. Such a study has never been performed in Sudan. Our study addresses the following questions:

1) Is there differential item functioning by gender of Coloured Progressive Matrices (CPM) test items in the total sample of the study?

2) Is there differential item functioning related to age in CPM test items?

Methods of detecting differential item functioning depend on the theory on which it is based. There are several methods based on traditional test theory. *Analysis of variance* has been one of the most used methods until the end of the eighties of the last century. It depends on the statistical significance of the interaction of the item score with the group score (Labadi, 2008). The *coefficient of discrimination* is another method used to assess item performance. It involves calculating the correlation between individual item scores and the total test score, then ordering these correlations — either ascending or descending — to compare item performance between high- and low-performing groups (Berk, 1982). Another approach is the *goodness-of-fit method*, which compares the proportion of correct responses across subgroups within the same total test score category, allowing for the identification of items that may function differently among examinees of similar overall ability (Crocker & Algina, 1986). *Logistic regression* treats the item response (correct or incorrect) as the dependent variable and includes the examinee's estimated ability level, group membership, and their interaction as independent variables. The *likelihood-ratio method*, a parametric technique, compares two models to detect differential item functioning (DIF). The first, known as the compact model, includes parameter constraints that assume item characteristics are equal across groups. This is then compared to an augmented model that does not impose those constraints, allowing the parameters to vary between groups. If the comparison between the two models reveals statistically significant differences, this is considered evidence of differential item functioning (DIF), indicating that the item may be biased in favour of one group over another despite equivalent overall ability among individuals (Hambleton et all., 1991).

2 Methods

Table 1: Table 1. Descriptive statistics of the sample

Age	Age range (years)	Females	Males
4–6	Preschool (Kindergarten)	342	372
7–9	Primary stage (Grades 1–3)	701	563
10–12	Primary stage (Grades 4–6)	391	255
	Total	**1434**	**1190**

2.1 Cognitive test

The study utilized the Coloured Progressive Matrices (CPM), a non-verbal intelligence test developed by John C. Raven. The CPM is designed for individuals aged 5 to 11 years, the elderly, and those with moderate to severe learning difficulties. It consists of 36 items divided into three sets (A, AB, B), each containing 12 items. The items are presented on a coloured background to enhance visual appeal and maintain the participant's attention. The CPM is widely used in research and clinical settings to assess non-verbal reasoning abilities.

In Sudan, an adapted version of the CPM was administered to primary school students in Khartoum State by (Al-Khatib et al., 2006a,b, 2021). The sample included 1,683 students, with 57% males and 43% females. The test demonstrated high internal consistency, with stability coefficients ranging from 0.72 to 0.91. The Spearman-Brown reliability coefficient ranged from 0.63 to 0.81, indicating good reliability across subgroups and the total test score. Regarding the consistency of scores, the study found that the scores of different age groups had high correlations between the item score and set score, similarly between the set score and the total score, suggesting that the CPM effectively measures cognitive abilities across various age groups.

The study showed a positive correlation of the test scores with the chronological age of the participants, indicating that older students tended to score higher on the CPM as expected. Discrimination analysis revealed that most items in sets A and AB, as well as the first seven items in set B, effectively differentiated between higher and lower ability groups. The study also established percentile norms for males and females for each item, providing a reference for interpreting individual scores. Based on these findings, the study recommended the use of the CPM for various purposes, including classification and diagnostic assessments.

3 Results

Table 2 shows the Mantel-Haenszel statistic, odds ratio, and p-value (significance) of the Mantel-Hanzel statistic for the thirty-six test items.

Table 2: Table 2. Mantel–Haenszel's method applied to gender for the thirty-six test items, complete sample

Mantel–Haenszel statistic	Significance (p)	Odds ratio	Difference favouring	Is there differential performance?	Item
6.696	0.010	1.270	females	Yes	1
0.226	0.635	0.502	–	No	2
0.603	0.438	0.606	–	No	3
31.465	0.000	1.619	females	Yes	4
0.352	0.553	0.841	–	No	5
0.240	0.624	0.865	–	No	6
1.059	0.303	0.926	–	No	7
1.903	0.168	0.746	–	No	8
2.407	0.121	0.967	–	No	9
0.003	0.956	0.832	–	No	10
1.195	0.274	0.929	–	No	11
4.847	0.028	0.668	males	Yes	12
1.260	0.281	0.883	–	No	13
0.254	0.614	0.720	–	No	14
1.429	0.232	0.924	–	No	15
0.701	0.403	0.900	–	No	16
4.939	0.026	1.029	females	Yes	17

Table 2 (continued)

Mantel–Haenszel statistic	Significance (p)	Odds ratio	Difference favouring	Is there differential performance?	Item
0.000	0.991	0.823	–	No	18
1.038	0.308	0728	–	No	19
3.568	0.059	0.671	–	No	20
1.773	0.183	0.700	–	No	21
1.183	0.277	0.741	–	No	22
0.001	0.970	0.836	–	No	23
10.536	0.001	0.604	males	Yes	24
0.321	0.571	0.865	–	No	25
1.446	0.229	0.993	–	No	26
0.218	0.640	0.865	–	No	27
3.782	0.052	1.003	–	No	28
0.012	0.914	0.834	–	No	29
7.484	0.006	0.627	males	Yes	30
3.145	0.076	0.699	–	No	31
0.102	0.749	0.757	–	No	32
0.137	0.711	0.771	–	No	33
4.222	0.040	0.661	males	Yes	34
7.025	0.008	1.008	females	Yes	35
1.684	0.194	0.679	–	No	36

Table 2 reveals that 8 of the 36 CPM items (22%) showed DIF at a significance level of $\underline{p} < .05$, equally divided in 4 items (1, 4, 17, 35) favouring females and 4 items (12, 24, 30, 34) favouring males. The effect of DIF on score differences between males and females is negligible.

Applying the method to the three age groups, we obtained the results summarized in Tables 3-5. The number of DIF items increased from 3 items (8%) at age 4-6 years to 4 items (11%) at age 7-9 years and 12 items (33.3%) at age 10-12 years.

Table 3: Table 3. Mantel–Haenszel's method applied to gender for the thirty-six test items, age 4–6 years

Mantel–Haenszel statistic	Signifi-cance (p)	Odds ratio	Difference favouring	Is there differential performance?	Item
1.510	0.219	0.605	–	No	1
0.219	0.640	0.000	–	No	2
0.213	0.645	0.507	–	No	3
1.545	0.214	0.788	–	No	4
7.453	0.006	1.225	females	Yes	5
0.001	0.970	0.733	–	No	6
7.841	0.005	1.162	females	Yes	7
0.007	0.933	0.672	–	No	8
0.002	0.965	0.751	–	No	9
2.082	0.149	0.932	–	No	10
0.320	0.572	0.805	–	No	11

Table 3 (continued)

Mantel–Haenszel statistic	Significance (p)	Odds ratio	Difference favouring	Is there differential performance?	Item
2.507	0.113	0.953	–	No	12
0.002	0.969	0.622	–	No	13
0.330	0.565	0.579	–	No	14
0.123	0.725	0.765	–	No	15
0.471	0.492	0.804	–	No	16
1.774	0.183	0.913	–	No	17
1.210	0.271	0.863	–	No	18
0.060	0.806	0.637	–	No	19
0.112	0.737	0.743	–	No	20
1.641	0.200	0.444	–	No	21
1.046	0.306	0.573	–	No	22
2.477	0.115	0.522	–	No	23
0.198	0.656	0.772	–	No	24
3.541	0.060	0.479	–	No	25
2.653	0.103	0.539	–	No	26
1.920	0.166	0.547	–	No	27
0.026	0.871	0.703	–	No	28
0.317	0.573	0.782	–	No	29
1.989	0.158	0.502	–	No	30
0.002	0.963	0.697	–	No	31
1.162	0.281	0.481	–	No	32
4.056	0.044	0.440	males	Yes	33
3.347	0.067	0.458	–	No	34
1.009	0.315	0.831	–	No	35
1.788	0.181	0.507	–	No	36

Table 4: Table 4. Mantel–Haenszel's method applied to gender for the thirty-six test items, age 7–9 years

Mantel–Haenszel statistic	Significance (p)	Odds ratio	Difference favouring	Is there differential performance?	Item
3.623	0.057	0.000	–	No	1
1.207	0.272	0.280	–	No	2
0.408	0.523	0.533	–	No	3
34.957	0.000	1.883	females	Yes	4
0.814	0.367	0.518	–	No	5
0.095	0.757	0.758	–	No	6
2.291	0.130	0.950	–	No	7
0.003	0.955	0.782	–	No	8
0.374	0.541	0.846	–	No	9
0.808	0.369	0.675	–	No	10
2.704	0.100	0.967	–	No	11

Table 4 (continued)

Mantel–Haenszel statistic	Signifi-cance (p)	Odds ratio	Difference favouring	Is there differential performance?	Item
1.480	0.224	0.585	–	No	12
0.002	0.961	0.606	–	No	13
0.265	0.606	0.569	–	No	14
0.005	0.945	0.671	–	No	15
1.176	0.278	0.632	–	No	16
2.443	0.118	0.955	–	No	17
0.062	0.803	0.774	–	No	18
4.829	0.028	0.514	males	Yes	19
2.028	0.154	0.579	–	No	20
0.585	0.444	0.838	–	No	21
0.484	0.486	0.837	–	No	22
1.968	0.161	0.930	–	No	23
14.132	0.000	0.434	males	Yes	24
1.250	0.264	0.884	–	No	25
0.748	0.387	0.856	–	No	26
1.690	0.194	0.914	–	No	27
1.691	0.193	0.914	–	No	28
0.000	0.991	0.742	–	No	29
0.042	0.837	0.717	–	No	30
1.477	0.224	0.622	–	No	31
0.007	0.936	0.710	–	No	32
1.693	0.193	0.571	–	No	33
7.335	0.007	0.483	males	Yes	34
0.007	0.931	0.726	–	No	35
2.907	0.088	0.462	–	No	36

Table 5: Table 5. Mantel–Haenszel's method applied to gender for the thirty-six test items, age 10–12 years

Mantel–Haenszel statistic	Signifi-cance (p)	Odds ratio	Difference favouring	Is there differential performance?	Item
5.481	0.019	1.324	females	Yes	1
0.005	0.943	0.456	–	No	2
0.382	0.536	0.414	–	No	3
1.618	0.203	0.841	–	No	4
0.322	0.571	0.483	–	No	5
0.043	0.835	0.594	–	No	6
4.170	0.041	0.444	males	Yes	7
1.675	0.196	0.536	–	No	8
4.081	0.043	1.028	females	Yes	9
0.132	0.717	0.606	–	No	10
0.223	0.637	0.742	–	No	11

Table 5 (continued)

Mantel–Haenszel statistic	Significance (p)	Odds ratio	Difference favouring	Is there differential performance?	Item
4.745	0.029	0.332	males	Yes	12
1.858	0.173	0.848	–	No	13
0.038	0.846	0.468	–	No	14
0.885	0.347	0.740	–	No	15
1.895	0.169	0.903	–	No	16
0.496	0.481	0.785	–	No	17
8.418	0.004	0.313	males	Yes	18
0.000	0.988	0.649	–	No	19
3.856	0.049	0.428	males	Yes	20
7.330	0.007	0.381	males	Yes	21
4.098	0.043	0.436	males	Yes	22
0.138	0.710	0.742	–	No	23
0.079	0.778	0.693	–	No	24
3.294	0.070	0.983	–	No	25
1.283	0.257	0.821	–	No	26
0.001	0.977	0.592	–	No	27
1.111	0.292	0.834	–	No	28
1.602	0.206	0.515	–	No	29
9.434	0.002	0.368	males	Yes	30
2.348	0.125	0.475	–	No	31
0.364	0.546	0.718	–	No	32
13.073	0.000	1.568	females	Yes	33
1.624	0.202	0.881	–	No	34
16.261	0.000	1.697	females	Yes	35
10.814	0.001	1.545	females	Yes	36

4 Discussion

The first question addressed in our study was: *Is there differential item functioning (DIF) by gender in the CPM test items in the total sample?* Table 2 provides the relevant data and indicates that 8 out of 36 items (22%) displayed DIF. These were evenly split, with four items favouring females and four favouring males. This balanced outcome suggests that, across the 4–12-year age span, there is no consistent item bias favouring one gender over the other.

However, it is important to clarify that these findings do not reflect differences in general intelligence between boys and girls. Rather, they highlight that some individual test items function differently across genders, even when ability levels are held constant. This distinction is crucial: We are examining potential sources of measurement bias, not true gender differences in cognitive ability. As such, our findings align with prior research (e.g., Colom et al., 2000, 2002; Jensen, 1998) that found negligible gender differences in general intelligence. We only assess whether specific items within the CPM may distort the measurement of ability across genders, not whether males or females are inherently more intelligent.

This observation supports the idea that while overall cognitive ability (g) may be equivalent between genders, the manifestation of certain abilities—and consequently, performance on specific test items—may vary. Abad et al. (2004) made a similar point, suggesting that observed gender differences are more likely

to emerge in specific cognitive domains rather than in general intelligence. Our findings are consistent with Al-Qati (2016), who found a relatively low and evenly distributed presence of gender-biased items in the Wechsler test. Likewise, Colom et al. (2000, 2002) found negligible sex differences in intelligence using large adult samples and Spanish versions of WAIS-III. Importantly, these studies and ours do not claim that certain genders perform better overall, but that a few items may favour one gender due to factors unrelated to the underlying construct of general intelligence.

This analysis assumes that the Raven's test measures a unidimensional construct—general intelligence (g). We recognize that this assumption may not fully hold if the test inadvertently taps into multiple cognitive domains with differing gender strengths. In such cases, observed DIF may reflect genuine domain-specific differences rather than bias per se.

Some prior studies, such as Mends & Ercikan (2006), found that males outperform females on items demanding higher cognitive processes. This may be partly explained by gender differences in variability: When standard deviation in test performance is higher among males, a larger proportion of males will be found at the upper tail of the distribution, potentially explaining why they succeed more often on the most difficult items. Gamer & Engelhard (1999) similarly reported that males outperformed females in geometry and measurement tasks, a finding echoed by Rushton et al. (2004). Al-Bustanji (2004) found that DIF in mental ability tests tended to favour males in math/spatial domains and females in verbal domains.

For example, Ryan and Chiu (2001) reported that males outperform females on vocabulary items requiring higher-order thinking, and on visual-spatial tasks. But without knowing whether they controlled for total test performance or latent ability, it is unclear if their findings represent DIF or actual ability differences.

Our findings suggest that the observed DIF could be related to item characteristics. It is plausible that some matrix items require skills more aligned with visualization (which tends to favour males), while others might involve pattern recognition or strategies that may not be strictly spatial, potentially benefiting females. However, quantifying these item features (e.g., verbal versus visual load) is complex and was not within the scope of our study. Future research could explore these dimensions to better understand the sources of DIF in matrix-type tests.

The second research question concerned DIF in relation to age. While this is partially descriptive, a notable trend emerged: The number of DIF items increased with age, particularly in the oldest age group (10–12 years). This may indicate that as children grow older, cognitive abilities become more specialized and differentiated, possibly contributing to increased item-level bias across genders. It may also reflect the onset of puberty and associated hormonal changes that influence cognitive development and gender-differentiated experiences.

Bors & Forrin (1995), Hertzog (1989), and Salthouse (1996) discussed how age-related changes affect the structure of cognitive abilities. However, those works largely focused on aging populations, and their relevance to children must be interpreted cautiously. Nonetheless, our results suggest that gender differences in item performance become more pronounced with age, indicating an age-related increase in item-specific variance rather than in general ability.

This contrasts with findings that suggest increasing general ability with age reduces item-specific variance (e.g., Verhaeghen & Salthouse, 1997), but again, most of those studies focused on aging adults rather than child development. In our study, the developmental trajectory seems to point to increased differentiation and potentially more gender-specific learning experiences or strategies, which in turn may create or amplify DIF.

Our findings are also consistent with Al-Bursan (2003), who reported increases in gender-related DIF at higher developmental levels. This trend may be driven by the interaction of maturing cognitive systems and socially mediated gender-specific experiences, which could influence how children approach and solve certain types of test items.

5 Conclusion

In summary, our study finds no systematic gender bias in the overall CPM scores, but does detect item-level DIF that increases with age. These results underscore the importance of using DIF analysis in test construction and validation, particularly when the goal is to ensure fairness across demographic groups. Future studies should investigate whether item-level DIF in matrix reasoning tasks can be predicted based on verbal, spatial, or strategy-related content, and how such characteristics interact with age and gender during cognitive development

6 Recommendations

1. Conduct item-level cognitive analysis. Future studies should analyse the cognitive demands of individual CPM items (e.g., spatial visualization, pattern recognition, or logical reasoning) to determine whether specific item features predict gender-based DIF. This can help identify whether certain cognitive processes systematically favour one gender over the other.

2. Develop a framework for item content classification. Researchers should develop or adopt a validated framework to categorize CPM items based on their visual, verbal, or abstract reasoning load. This would help quantify and predict sources of DIF and provide clearer guidance for item construction in gender-fair assessments.

3. Apply multiple DIF detection methods. To improve robustness and comparability, future research should use a combination of DIF detection techniques (e.g., Mantel-Haenszel, logistic regression, item response theory-based methods) across different populations and age groups. This would help confirm the presence of DIF and its consistency across methods.

4. Expand research across regions and cultures. Given the context-specific nature of test performance, it is recommended to replicate this study in other regions of Sudan and in other Arab or African countries. This would help determine whether the patterns of DIF observed are culturally generalizable or regionally specific.

5. Investigate age-related cognitive differentiation. The increase in DIF with age suggests the need to study how gender differences in cognitive strategies or experiences develop with age. Longitudinal studies could track when and how these differences emerge and whether they are stable or shift during key developmental periods (e.g., around puberty).

6. Design gender-inclusive test items. Test developers should use findings from DIF analyses to design or modify items to minimize gender bias. Pilot testing items for DIF before they are included in final versions of cognitive ability tests is essential for fairness and validity.

7 Implications of the study

1. Incorporate gender-sensitive educational interventions. If certain cognitive domains show consistent gender differences (e.g., males in spatial tasks, females in pattern-based tasks), educational programs should aim to strengthen underrepresented skills in each gender from an early age to reduce long-term disparities.

2. Collaborate with psychometricians and educational psychologists. Interdisciplinary collaboration can improve the development of more equitable assessments by integrating cognitive theory, psychometric modelling, and classroom-based evidence.

8 Limitations of the study

1. Limited generalizability. The study sample was drawn from a specific geographic and cultural context within Sudan, which may limit the generalizability of the findings to other regions, cultures, or educational systems. DIF patterns may vary in different sociocultural environments.

2. Cross-sectional design. The research design was cross-sectional, which limits the ability to draw conclusions about developmental changes over time. A longitudinal approach would provide stronger evidence for how DIF patterns evolve with age and cognitive development.

3. Absence of Socioeconomic and Educational Background Data. Important contextual variables such as socioeconomic status, parental education, or school quality were not controlled for. These factors might influence test performance and interact with gender or age effects.

References

Abad, F. J., Colom, R., Rebollo, I., & Escorial, S. (2004). Sex differential item functioning in the Raven's Advanced Progressive Matrices: Evidence for bias. *Personality and Individual Differences, 36*(6), 1459-1470. https://doi.org/10.1016/S0191-8869(03)00244-9

Al-Bursan, I. (2013). Gender related differential item functioning for Jordanian national test for mathematics learning quality control for tenth grade. *Educational & Psychological Studies, 79,* 229-270.

Albustanji, M. M. (2004). *Comparing four procedures for detecting gender-based differential functioning of items of a specific mental abilities test for the 15–16-year age group in Jordan*. Doctoral dissertation, Amman Arab University.

Al-Khatib, M., Al-Mutawakkil, M., Bakhiet, S., & Abd Al-Rahim, N. (2021). *Standardization of the Colored Progressive Matrices Test in the Sudanese environment for children aged 4–6 years*.

Al-Khatib, M., Al-Mutawakkil, M., & Hussein, A. (2006a). Validity and reliability implications of the Colored Progressive Matrices Test for pre-school and primary school children in Khartoum State. *Educational Studies, 14,* 138-163.

Al-Khatib, M., Al-Mutawakkil, M., & Hussein, A. (2006b). *Standardization of the Colored Progressive Matrices Test for first-cycle primary school students in Khartoum State*. Khartoum: Sudan Currency Printing Press Company Ltd.

Al-Qati, A. A. (2016). Bias of Wechsler Modified Children's Intelligence Test items (Saudi image) by gender. *Journal of Educational Sciences, 5*(2).

Albustanji, M. M. (2004). *Comparing four procedures for detecting gender-based differential functioning of items of a specific mental abilities test for the 15–16-year age group in Jordan*. PhD thesis, Amman Arab University.

Baghi, H., & Ferrara, S. F. (1990). Detecting differential item functioning using IRT and Mantel–Haenszel techniques: Implementing procedures and comparing results (ERIC Document No. ED325479). ERIC.

Berk, R. A. (1982). *Handbook of Methods for Detecting Test Bias*. Baltimore MD: Johns Hopkins Univ. Press.

Bors, D. A., & Forrin, B. (1995). Age, speed of information processing, recall, and fluid intelligence. *Intelligence, 20*(3), 229-248. https://doi.org/

Camilli, G., & Shepard, L. A. (1994). *Methods for Identifying Biased Test Items*. London: Sage.

Colom, R., Juan-Espinosa, M., Abad, F. J., & García, L. F. (2000). Negligible sex differences in general intelligence. *Intelligence, 28*(1), 57-68. https://doi.org/10.1016/S0160-2896(99)00032-7

Colom, R., Contreras, M. J., Botella, J., & Santacreu, J. (2002). Vehicles of spatial ability. *Personality and Individual Differences, 32*, 903-912. https://doi.org/10.1016/S0191-8869(01)00099-0

Conoley, A. C. (2003). *Differential item function in the Peabody Picture Vocabulary Test–Third Edition:*

Partial correlation versus expert judgment. Doctoral dissertation, Texas A&M University.

Crocker, L., & Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. New York: CBS College Publishing.

Dodeen, H. (2004). Stability of differential item functioning over a single population in survey data. *Journal of Experimental Education, 72*(3), 181-193. https://doi.org/10.3200/JEXE.72.3.181-193

Dorans, N. J., & Holland, P. W. (1994). *DIF detection and description: Mantel-Haenszel and standardization*. Princeton, NJ: Educational Testing Service.

Drina, E. (2007). *Gender differential item functioning on a ninth-grade mathematics proficiency test in Ohio.* Doctoral dissertation, Ohio University.

Fennema, E., & Lamon, S. J. (1990). Teachers' attributions and beliefs about girls, boys, and mathematics. *Educational Studies in Mathematics, 21*(1), 55-69. https://doi.org/10.1007/BF00311015

Gamer, M., & Engelhard, G. Jr. (1999). Gender differences in performance on multiple-choice and constructed response mathematics items: A many-faceted Rasch model approach to differential item functioning. *Applied Measurement in Education, 12*(1), 29-51. https://doi.org/10.1207/s15324818ame1201_2

Hambleton, R. K., & Rogers, J. (1995). Item bias review. *Practical Assessment, Research, and Evaluation, 4*(6).

Hambleton, R. K., Swaminathan, H., & Rogers, H. J. (1991). *Fundamentals of Item Response Theory*. Newbury Park CA: Sage.

Harris, A. M., & Carlton, S. T. (1995). Patterns of gender differences on mathematics items on the Scholastic Aptitude Test. *Applied Measurement in Education, 6*(2), 137-151.

Hertzog, C. (1989). The influence of cognitive slowing on age differences in intelligence. *Developmental Psychology, 25*(4), 636-651. https://doi.org/10.1037/0012-1649.25.4.636

Hyde, J. S., & Linn, M. C. (1988). Gender differences in verbal ability. *Psychological Bulletin, 104*(1), 53-69.

Hyde, J. S., Fennema, E., & Lamon, S. J. (1990). Gender differences in mathematics performance: A meta-analysis. *Psychological Bulletin, 107*(2), 139-155.

Jensen, A. R. (1980). *Bias in Mental Testing*. New York: Macmillan.

Jensen, A. R. (1998a). *The g Factor: The Science of Mental Ability*. Praeger.

Jensen, A. R. (1998b). The g factor and the design of education. In R. J. Sternberg & W. M. Williams (Eds.), *Intelligence, Instruction, and Assessment: Theory into Practice* (pp. 111-131). Lawrence Erlbaum.

Kim, S.-H., & Cohen, A. S. (1992). Effects of linking methods on detection of DIF. *Journal of Educational Measurement, 29*(1), 51-66. https://doi.org/10.1111/j.1745-3984.1992.tb00367.x

Labadi, N. (2008). *Comparing four methods for detecting differential paragraph functioning (A simulated study)*. Unpublished PhD thesis, University of Jordan, Amman.

Linn, M. C., & Petersen, A. C. (1985). Emergence and characterization of sex differences in spatial ability: A meta-analysis. *Child Development, 56*(6), 1479-1498. https://doi.org/10.2307/1130467

Maccoby, E. E., & Jacklin, C. N. (1974). *The Psychology of Sex Differences*. Stanford Univ. Press.

McLaurin, W. A., Jenkins, J. F., Farrar, W. E., & Rumore, M. C. (1973). Correlations of IQs on verbal and nonverbal tests of intelligence. *Psychological Reports, 33*(3), 821-822. https://doi.org/10.2466/

pr0.1973.33.3.821

Mendes-Barnett, S., & Ercikan, K. (2006). Examining sources of gender DIF in mathematics assessments using a confirmatory multidimensional model approach. *Applied Measurement in Education*, *19*(4), 289-304.

Paul, A. M. (1985). Sex differences in the Raven Progressive Matrices. *Journal of Clinical Psychology, 41*(5), 637-641. https://doi.org/10.1002/1097-4679(198509)41:5<637::AID-JCLP2270410516>3.0.CO;2-1

Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: A comparison of three Mantel–Haenszel procedures. *Applied Measurement in Education, 14*(3), 235-259. https://doi.org/10.1207/S15324818AME1403_3

Raju, N. S., Drasgow, F., & Slinde, J. A. (1993). An empirical comparison of the area methods, Lord's chi-square test, and the Mantel–Haenszel technique for assessing differential item functioning. *Educational and Psychological Measurement, 53*(2), 301-314. https://doi.org/10.1177/0013164493053002001

Raven, J., Raven, J. C., & Court, J. H. (1998). *Coloured Progressive Matrices*. Oxford, UK: Oxford Psychologists Press.

Roever, C. (2005). "That's not fair": Fairness, bias, and differential item functioning in language testing. Retrieved February 6, 2010, from http://www.hawaii.edu/2Roeve/brounbag.doc

Rushton, J. P., Skuy, M., & Bons, T. A. (2004). Construct validity of Raven's Advanced Progressive Matrices for African and non-African engineering students in South Africa. *International Journal of Selection and Assessment, 12*(3), 220-229. https://doi.org/10.1111/j.0965-075X.2004.00276.x

Ryan, K. E., & Chiu, S. (2001). An examination of item context effects, DIF, and gender DIF. *Applied Measurement in Education, 14*(1), 73-90.

Salthouse, T. A. (1996). The processing-speed theory of adult age differences in cognition. *Psychological Review, 103*(3), 403-428. https://doi.org/10.1037/0033-295X.103.3.403

Verhaeghen, P., & Salthouse, T. A. (1997). Meta-analyses of age–cognition relations in adulthood: Estimates of linear and nonlinear age effects and structural models. *Psychological Bulletin, 122*(3), 231-249. https://doi.org/10.1037/0033-2909.122.3.231

Voyer, D., Voyer, S. D., & 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