Journal of Psychology in Africa, 2026
Vol. 36, No. 4, 651–658, http://doi.org/10.32604/jpa.2026.070232

WISC-III scores: Evidence of gender-related differential item functioning in Sudan

Ismael Salamah Albursan[1], Mohammad Farhan Al. Qudah[1], Eqbal Z. Darandari[1], Salaheldin Farah Bakhiet[2,*] and Mohammed Ateik Al-Khadher[1]

[1]*Department of Psychology, College of Education, King Saud University, Riyadh, Saudi Arabia*
[2]*Department of Special Education, College of Education, King Saud University, Riyadh, Saudi Arabia*

**Correspondence Author: Salaheldin Farah Bakhiet. Email:* bakhiet@ksu.edu.sa
Received: 11 July 2025; Accepted: 21 January 2026; Published: 31 August 2026

Abstract: The Wechsler Intelligence Scale for Children 3rd version (WISC-III) is a widely used assessment tool for children's intellectual functioning, including in non-Western settings. Yet, the evidence is scarce of possible gender differences among Sudanese students. This study examined possible gender differences in the Wechsler Intelligence Scale for Children 3rd version (WISC-III) scores among Sudanese students by exploring Differential Item Functioning (DIF) across five subtests: picture completion, information, similarities, picture arrangement, and object assembly. The study sample comprised 455 boys and 525 girls aged 6 to 16. The results following a Likelihood-Ratio Test within Item Response Theory, applying the three and 2-parameter logistic models, revealed seventeen items with differential functioning by gender. The girls' scores were higher on eight items across picture completion, information, similarities, and picture arrangement subtests, and the boys' scores were higher on nine items across the subsets of picture completion, information, similarities, picture arrangement, and object assembly. These DIF analysis results suggest a need for actions to ensure fairness and equity in the interpretation of scores from the WISC-111 in the Sudanese context.

Keywords: Testing; WISC-III; DIF; gender differences; IQ tests; Sudan

Introduction

Wechsler Intelligence Scale for Children (WISC). WISC is a widely used assessment tool for measuring intellectual functioning in children. The third edition, known as the WISC-III, was published in 1991, building upon the first edition (Wechsler, 1949), and the second modified edition (WISC-R) was published in 1974 (Wechsler, 1974; Sattler, 2001). With 519 items distributed across 13 subtests, including ten core subtests and three supplemental subtests, the WISC-III is designed for children aged 6 to 16. Use of the WISC-III in the Sudanese settings requires careful study to ensure fairness and equity in the interpretation of scores. The study examined possible gender differences in WISC-III scores among Sudanese students.

WISC-III description. The WISC-III, a western psychological ability test, consists of six verbal scales, each comprising 30 items, assessing skills such as information, vocabulary, comprehension, arithmetic, digit span (supplemental), and similarities. Additionally, there are seven performance scales: picture completion (30 items), picture arrangement (14 items), block design (12 items), mazes (10 items), object assembly (5 items), coding (178 items, with a supplemental part), and symbol search (90 items, with a supplemental part) (Al-Kilani & Al-Rawsan, 2006). The test administration follows a sequence of alternating between performance and verbal scales.

IQ scores derived from the WISC-III are based on ten subtests, five verbal and five performance subtests, while the remaining three are used for clinical diagnostic purposes. The test provides three IQ scores: verbal, performance, and overall intelligence. Additionally, four mental indices are obtained: the Verbal Comprehension Index (VCI), the Perceptual Organization Index (POI), the Freedom from Distractibility Index (FDI), and the Processing Speed Index (PSI). These indices are derived through factor analysis, with each group of subtests representing a different factor (Nicholson & Alcorn, 1993).

Validation studies conducted in the United States on a sample of 2200 respondents representing various demographics revealed strong reliability coefficients for the WISC-III. Split-half reliability coefficients ranged from 0.69 to 0.87 (with an average of 0.78) for different age groups. The reliability coefficients for verbal, performance, and overall intelligence were 0.95, 0.91, and 0.96, respectively. The reliability coefficients for the four mental indices were 0.94, 0.90, 0.87, and 0.85 for verbal comprehension, perceptual organization, freedom from distractibility, and processing speed, respectively. Furthermore, factorial analysis established the construct validity of the scale (Prifitera et al., 1998; Sattler, 2001; Wechsler, 1991).

Of studies that have examined use of the WISC-111 in the Middle Eastern and North Africa region, the study by Özberk and Koc (2017) is the major exception. Özberk and Koc (2017) used several DIF detection techniques to detect gender and socioeconomic status DIF in WISC-IV Turkish Form. The sample included 817 children, and 315 items were used in polytomous and dichotomously scored subtests. Items detected as having DIF were similar across three techniques. DIF results showed that the following items of the picture completion subtest (6, 8, 10, 15, 22, 24, and 33) favored girls, while items (17, 18, 19, 22, and 32) favored the boys. In the similarity subtest, items (5, 9, and 23) favored girls, while items (8 and 21) favored boys.

More recent evidence from research on newer versions of the Wechsler scales further supports the importance of examining whether cognitive test scores operate equivalently across demographic groups. For example, Rodríguez-Cancino et al. (2023) examined measurement invariance in the WISC-V across sex and age groups and emphasized that such analyses are necessary before interpreting group differences in cognitive performance.

Overall, the WISC-III serves as a comprehensive tool for assessing intellectual functioning in children, and its reliability and validity have been established in various populations. The availability of adapted versions allows for its application in different cultural and linguistic contexts, providing valuable insights into children's cognitive abilities.

Score differences by personal and contextual factors. Fair and unbiased testing is crucial because it prevents systematic disadvantages for specific genders or cultures, promotes equitable opportunities, and enhances the validity and fairness of assessments. Unbiased tests identify individuals' actual abilities, rather than cultural or gender-specific knowledge, leading to more accurate and inclusive outcomes in hiring, education, and other high-stakes decisions. This ensures that all groups have an equal chance and that the talent pool is diverse and representative of the population (Tavakol, et al., 2024). Similarly, Holden et al. (2023) argued that modern intelligence assessments must be evaluated not only in terms of reliability and validity, but also in terms of fairness and equity, particularly when used with culturally, linguistically, or demographically diverse groups. Differential Item Functioning (DIF) occurs when test items function differently for individuals from different demographic groups (e.g., gender, ethnicity) who have the same underlying ability or trait. Differential Item Functioning (DIF) can negatively impact educational policy and practices by systematically disadvantaging certain student groups, leading to unfair test scores, discriminatory disciplinary actions, and flawed educational decisions for marginalized groups.

By conducting DIF, test developers can create valid, fair, and culturally appropriate assessments that accurately measure a construct, rather than reflecting cultural or linguistic differences that hinder test-taker performance. DIF analysis could help ensure fairness and equity by identifying test item functioning differently for different groups (e.g., by gender, race, or language proficiency), particularly intelligence tests, even when they have the same ability level. Addressing them could reflect true knowledge, not irrelevant group-specific factors. It is also essential for cultural adaptation of tests to new contexts and ensuring educational equity by detecting and removing construct-irrelevant variance. DIF analysis also informs item revision and development, leading to accurate and equitable measure of all students' abilities (Martinková et al., 2017).

Differential item functioning. Differential Item Functioning (DIF) is a statistically derived function that reveals the difference in item response between two groups with the same ability level. Recent psychometric literature has emphasized that DIF analysis is not merely a statistical procedure for detecting score discrepancies, but also an important approach for evaluating fairness and measurement equivalence across groups. DIF findings may indicate that individuals from different demographic groups interpret, process, or respond to certain test items differently despite possessing comparable levels of the underlying construct, which can threaten the validity of score interpretation in high-stakes assessments (Millsap, 2023; Paek & Cole, 2022). More specifically, DIF is the differences in the likelihood of correct responses to test items in groups having the same ability (Camilli & Shepard, 1994; Hambleton & Rogers, 1995; Hambleton et al., 1991). DIF examines the psychometric characteristics of items regarding how they function differently in different groups of respondents, bias is a systematic error that makes a group's performance on the test better than another's or as a systematic underestimation or overestimation of the population's determiner, and has social implications and indicates the absence of justice and equality (Jensen, 1980; Dorans & Holland, 1994).

Therefore, DIF is considered an essential but insufficient condition to regard an item as biased since DIF is a metric characteristic. In contrast, bias refers to the theoretical explanation for its presence. In other words, if an item reveals a differential functioning for a given group, other procedures shall be taken to judge the item as biased. These other procedures can be arbitration or experimental investigation (Clauser & Mazor, 1998).

DIF approaches. Methods of identifying differential item functioning vary according to the theories underlying them. For example, there are several methods based on the modern test theory. One of these methods is the Likelihood-Ratio Test, a parametric method based on different models when doing the matching. In this method, two models are checked and compared. The first model is called the compact model, in which parameters of all items, including the one examined for differential functioning, are estimated assuming or setting constraints. Hence, the parameters of items are equal in the two groups: the reference group and the focal group. The second model is called the augmented model. It differs from the first model in that the parameters of the item being examined for differential functioning can differ in the reference group and the focal group. The difference in the matching of the two models is examined by computing the likelihood ratio (G^2) according to the following equation:

$$G^2 = -2 \ \log LC - (-2 \ \log LA)$$

where:

log LC represents the log-likelihood of the compact model, and LA log represents the log-likelihood of the augmented model (De Ayalla, 2022).

The present study. We aimed to determine whether certain test items within the WISC-III show differential functioning for Sudanese school boys and girls, potentially impacting the accuracy and fairness of IQ assessments. Our specific research question was: What are the WISC-III picture completion, picture arrangement, object assembly, information, and similarities items with differential functioning by gender?

Method

Participants and setting

A total number of 980 subjects participated in the study, 455 males and 525 females, whose ages ranged from

Table 1. The distribution of subjects according to age and gender

Age	Male	Female	Total
7 years	3	4	7
8 years	163	222	385
9 years	284	294	578
10 years	5	5	10
Total	455	525	980

7 to 10 years. They were all the students whose protocols were taken from candidate students to the National Commission for the Gifted in Sudan who came from all areas of Khartoum (7 Municipalities). Table 1 presents the participant characteristics.

Measures

The present study used the Sudanese version of WISC-III (Al-Housein, 2005). This version was standardized in Khartoum and examined for its characteristics in the Sudanese environment by researchers (e.g., Al-Housein, 2008; Al-Khalifa et al., 2008; Al-Khalifa & Al-Housein, 2012). The version was validated by presenting it to 17 referees. Interrater agreement for verbal, practical, and overall intelligence was 0.95, 0.96, and 0.93, respectively. Moreover, seven percent of the scale was modified based on the referees' recommendations. The scale was then administered to a pilot sample of 110 male and female children from 6 to 16 years. The primary sample consisted of 330 children. The scale showed concurrent validity with age, as the correlation ranged between 0.21 and 0.66, all significant at the 0.01 level. Convergent validity was also established by correlating the scale with WAIS-R in older ages. Al-Housein (2005) computed the internal consistency between subtests and the results of verbal, practical, and overall intelligence to establish its construct validity. Coefficients of internal consistency ranged from 0.66 to 0.89. Reliability for verbal and performance subtests and the three IQs was also established by split-half (reliability coefficients ranging from 0.68 to 0.84), Guttmann equation (from 0.80 to 0.95) and Alpha-Cronbach (from 0.77 to 0.94). The counterpart reliability coefficients for performance subtests ranged from 0.67 to 0.81, 0.73 to 0.93 and 0.69 to 0.91, respectively. The Reliability of performance intelligence by the three methods was 0.76, 0.94, and 0.93, respectively. The counterpart reliability coefficients for verbal intelligence were 0.93, 0.98, and 0.97, respectively. After being modified, the whole scale yielded high-reliability coefficients (0.94, 0.98, and 0.97, respectively) using the Spearman-Brown equation.

Procedures

The Institutional Review Board of King Saud University determined that ethical approval was not required for this study. because it relied exclusively on secondary, de-identified data obtained from the routine assessment protocols of the National Authority for Gifted Education in Sudan. No direct contact, testing, or data collection from children occurred for the purpose of this research, and the dataset contained no personal identifiers. All procedures complied with international ethical guidelines and the principles of the Declaration of Helsinki. The WISC-III had been administered by trained school psychologists under the Commission's standardized assessment protocols. The present study relied exclusively on secondary, de-identified data extracted from these archival records, and no additional testing or direct contact with participants was conducted for research purposes.

Data Analysis. The IRTLRDIF v.2.0b program was used for data analysis. Likelihood-Ratio Test was administered utilized to examine DIF in items based on gender, using Three Parameters Logistic Model (3PL), with dichotomous items; and Samejima's Graded Model (1969, 1997), for polytomous items. The first is the Three Parameters Logistic Model (3PL), used with dichotomous items. The second is Samejima's Graded Model (1969, 1997), used with polytomous items. Both models include the dichotomous model as a particular case.

One group was set as reference group (males) and the other focal group (females). Then goodness of fit was checked by The G^2 test to show if there are differences. LR compares two models for each item: A restricted model, where the item's parameters are the same across the two groups (males and females); and an unrestricted model where they are allowed to differ. If the unrestricted model (with different parameters) fits the data significantly better than the restricted model (with same parameters), you conclude that DIF is present, meaning the item is functioning differently for the two groups. A good-fit for restricted model indicates that DIF does not exist.

The G^2 test has a distribution similar to the Chi-Square test (X2), with degrees of freedom matching the number of parameters. If the value of G^2 is 3.84 or higher, the null hypothesis of no differences in item parameters between the reference group and the focal group is rejected. The critical value in our case is 3.84 using $\alpha = 0.05$.

Results

For the test sections under consideration in the present study: picture completion, picture arrangement, object assembly, information and similarities (all making up 98 items), 17 (17.4% of the total number of items) items showed differential functioning; 8 items were in favor of girls and belong to picture completion, information, similarities, and picture arrangement; and 9 were in favor of boys from the subsets: picture completion, information, similarities, picture arrangement, and object assembly. The Likelihood-Ratio Test (G^2) for all items showing differential functioning are presented below.

Picture completion

Out of the 30 items composing this subtest, items 12, 14, and 21 showed differential functioning (see Table 2) in favor of girls. As to parameters through which differential functioning emerged, items 12 and 14 showed differential functioning through the guessing parameter, as the value of G^2 for the guessing parameter for item 12 was 11.8 with a p-value of 0.0081. Regarding item 14, the value of G^2 for the guessing parameter was 24.4 ($p < 0.0001$). The parameter resulting in differential functioning for item

Table 2. G^2 values for the picture completion subtest items showing DIF

Item	Determiner	G^2	*p*-value	a_b	b_b	a_g	b_g
12	All	11.8	0.0081	0.53	0.06	0.43	1.21
14	All	24.4	0.0000	0.27	−2.24	0.72	−0.04
21	All	6.1	0.0135	0.84	1.21	0.71	2.08

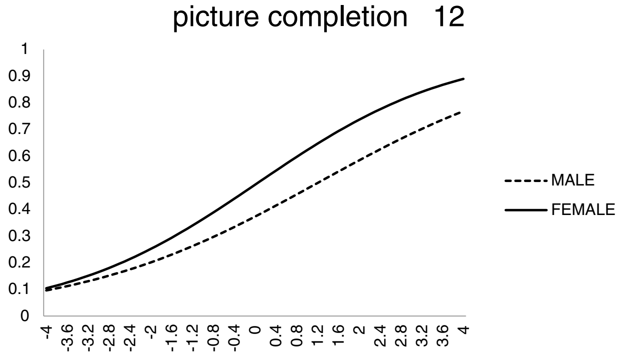

Figure 1. ICC for item 12 in picture completion for the two groups

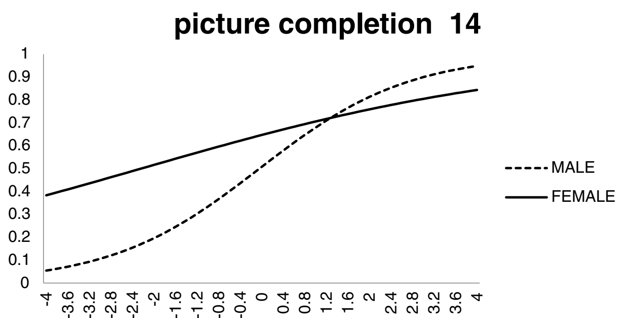

Figure 2. ICC for item 14 in picture completion for the two groups

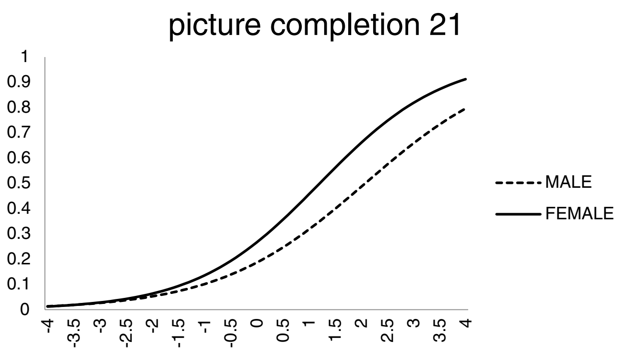

Figure 3. ICC for item 14 in picture completion for the two groups

21 was the difficulty parameter. The value of G^2 for this parameter was 6.1 ($p = 0.0135$) (See the Figs. 1–3).

Information

Only one item (item 10) out of the 30 items in this subtest showed differential functioning in favor of boys. The parameter responsible for this difference is the difficulty parameter. The value of G^2 of this parameter for item 10 was 4.1 ($p = 0.0429$; see Table 3)

Table 3. G^2 values for item 10 of the information subtest that showed DIF

Item	Determiner	G^2	*p*-value	ab	bb	ag	bg
10	All	4.1	0.0429	0.53	0.06	0.43	1.21

Table 4. G^2 values for the similarities subtest items showing DIF

Item	Determiner	G^2	*p*-value	a_b	b_b	a_g	b_g
2	All	4.1	0.043	1.46	−4.15	0.24	−2.97
14	All	9.5	0.0233	1.59	0.96	1.34	1.43
16	All	5.2	0.0226	2.73	2.01	1.5	2.4

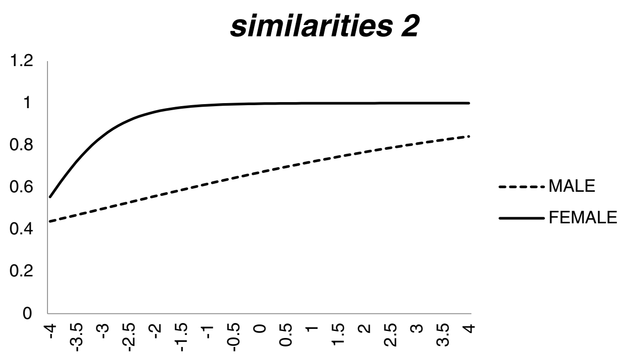

Figure 4. ICC for item 2 in similarities for the two groups

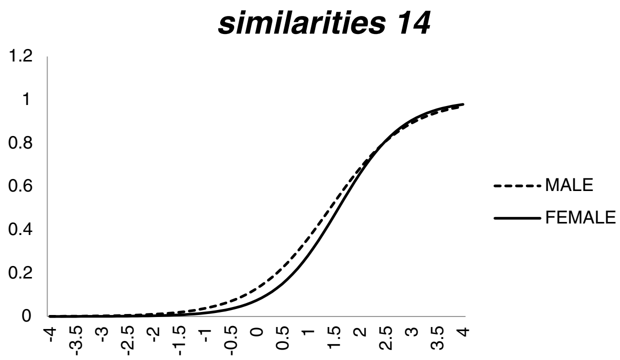

Figure 5. ICC for item 12 in similarities for the two groups

Similarities

Of the 19 items included in this subtest, 3 showed differential functioning. These are items 2, 14, and 16 (see Table 4). The value of G^2 for item 2 was 4.1 ($p = 0.043$) in favor of boys. The G^2 value for item 14 was 9.5 ($p = 0.0233$) in favor of girls. The G^2 value for item 16 was 5.2 ($p = 0.0226$) in favor of girls (See the Figs. 4–6).

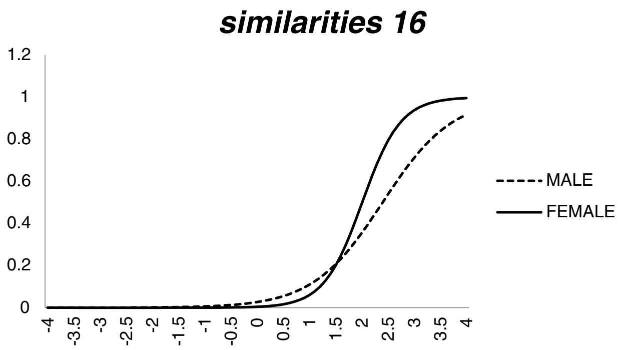

Figure 6. ICC for item 16 in similarities for the two groups

Table 5. G^2 values for the picture arrangement subtest items showing DIF

Item	Determiner	G^2	*p*-value	DIF
1	all	9.4	0.0244	Yes
	a	1.3		
	b	8.1	0.0044	Yes
6	all	4.7	0.030	Yes
	a	0		
	b	4.7	0.0302	Yes
7	all	9.1	0.027	Yes
	a	0.8		
	b	8.3	0.0040	Yes
8	all	5.2	0.0226	Yes
	a	0		
	b	5.2	0.0226	Yes
9	all	10.5	0.0012	Yes
	a	3.2		
	b	7.3	0.0069	Yes

Picture arrangement

Of the 11 items in the picture arrangement subtest, 5 showed differential functioning. These are items 1, 6, 7, 8, and 9 (see Table 5). Item 9 showed the highest differential functioning with a G^2 value of 10.5 ($p = 0.0012$) in favor of girls. Then came item 1 with a G^2 value of 9.4 ($p = 0.0022$) in favor of girls. The G^2 values for items 7, 8, and 6 were 9.1 ($p = 0.027$) in favor of boys, 5.2 ($p = 0226$) and 4.7 ($p = 0.030$) in favor of girls, respectively.

Pictures assembly

All five items in this subtest showed differential functioning with G^2 values ranging from 7.2 ($p = 0.0226$) for item 4 to 12.3 ($p = 0.0004$) for item 1. All differences were in favor of boys (see Table 6).

Discussion

The five WISC-III subtests included items showing differential functioning. In dichotomous items, the parameters responsible for differential functioning were difficulty, discrimination, and guessing. In the picture completion subtest, the guessing parameter induced differential functioning of items 12 and 14, whereas the difficulty parameter induced differential functioning in item 21. The differential functioning of item 10 was attributed to the difficulty parameter.

Table 6. G^2 values for the object assembly subtest items showing DIF

Item	Determiner	G^2	*p*-value	DIF
1	all	12.3	0.0004	Yes
	a	0.1		
	b	12.2	0.0005	Yes
2	all	11.7	0.0006	Yes
	a	0.8		
	b	10.9	0.0010	Yes
3	all	7.7	0.027	Yes
	a	0.1		
	b	7.5	0.0040	Yes
4	all	7.2	0.0226	Yes
	a	0.2		
	b	7	0.0082	Yes
5	all	10.3	0.0013	Yes
	a	0.8		
	b	9.6	0.0020	Yes

Eight items showed differential functioning in favor of girls (8% of the 98 items in the five subtests), while nine showed differential functioning in favor of boys (9% of the 98 items in the five subtests). The percentage of DIF items displayed in this study is 17% which is far less than the other studies (Maller's study was 34.4%, and Gokiert & Ricker's study was 43.3%). The proportion of items that showed significant gender DIF varied from one subtest to another, where the information subtest has the lowest (1 item), followed by similarities and picture completion subtests (3 items each), and picture arrangement and picture assembly subtests (5 items each).

For the Information subtest, this study agreed with the results of the Canadian study on item 10, which showed DIF, with more discrimination for girls and more difficulty for boys. No other items showed DIF in that subtest. In the Canadian study, Gokiert and Ricker (2004) reported DIF in nine items (4, 7, 9, 10, 15, 17, 22, 24, and 25) for the b-parameter (17 and 24 favored boys, and the remaining items favored girls, including item 10). No items displayed DIF on the discrimination (a) parameter.

The result of this study is consistent—to some extent-with the American study (Maller, 2001) that reported DIF in item 10 (in addition to other items: 7, 8, 15, and 16) with science content, and it was more difficult for girls, which is in contrast with this study.

Although the information subtest is of the verbal type in which girls usually outperform boys (Anastasi, 1958; Feingold, 1993; Maccoby, 1966; Maccoby & Jacklin, 1974; Tyler, 1965), the item content and cultural differences between countries could play a significant role.

Regarding the similarities subtest, which is of the verbal type, this study concurred with the Canadian and the American samples in items 2 and 16, which showed differential functioning. However, this study did not concur with the Canadian and American samples in the differential functioning of other items.

This study showed DIF in item 14, favoring girls, which did not appear in the other studies. Conversely, item 2 was more discriminating and difficult for boys, while item 16 was more discriminating for boys and girls. DIF was nonuniform, where item 2 favored females with low ability, and item 16 favored females with high ability. These results are consistent with the results of the American study. Maller (2001) reported DIF in fourteen items, including 2 and 16 (2, 3, 4, 6, 7, 8, 9, 10, 12, 13, 15, 16, 17, 19). Some (7, 8, and 13) exhibited nonuniform DIF. Item 2 was more difficult for boys, and item 16 was more discriminating for boys and girls.

The results contradict the Canadian study. Gokiert and Ricker (2004) reported DIF in eleven items, including 2 and 16 (1, 2, 3, 5, 6, 7, 10,15, 16, 17, and 18). Items 2 and 5 were more discriminating for boys, while items 10, 15, and 16 were more discriminating for girls. In addition, items 5, 10, and 18 were easier for boys, while items 6, 7, 16, and 17 were easier for girls.

The study conducted on the Turkish version of WISC-IV (Özberk & Koc, 2017) confirms that some items in the subtest still have DIF; some favored girls (3 items) while others (2 items) favored boys.

Other studies also reported these inconsistent results, particularly at ages 10–12 (Teresi & Fleishman, 2008) These results could be explained through the content of items (Maller, 2001), the distribution of the sample's age, and the nonuniform DIF that could give different results if not detected.

As to the picture completion subtest (a performance subtest), this study concurred with the Canadian and the American samples to some extent on the items that showed differential functioning (12, 14, 21). The results concurred with Maller's study (2001) on items 14 and 21 and Gokiert & Ricker,'s study (2004) on item 12, favoring girls.

In this study, item 12 was more discriminating for boys and more difficult for girls, item 14 was more discriminating and difficult for girls, and item 21 was more discriminating for boys and more difficult for girls. All were nonuniform DIF (where items 12 and 21 favored females with high ability, and item 14 favored females with low ability and males with high ability).

The results of this study support the conclusion that gender DIF items were inconsistent across countries (Bakhiet & Lynn, 2015; Lin & Steedle, 2020). This highlights that cultural factors and national contexts influence how test items perform and how they are interpreted by different genders.

For future directions, other DIF approaches may need to be considered (Li & Becker, 2021; Lim, et al., 2021), with nonuniform, in addition to uniform DIF and DIF patterns (i.e., balanced, shift-low, or shift-high and other patterns of DIF) with Differential Step Functioning framework (Penfield, 2007).

Future research should explore DIF in subtests beyond the five examined in this study, to provide a complete picture of potential biases across different cognitive domains. Investigating DIF in the other subtests could enhance the understanding of gender-related biases in the WISC.

Conclusion

In conclusion, this study investigated gender-related biases in the Wechsler Intelligence Scale for Children 3rd version (WISC-III) by examining Differential Item Functioning (DIF) in five specific subtests. The results showed that a significant proportion of items demonstrated differential functioning, with some favoring girls and others favoring boys. The percentage of items displaying DIF in this study was 17%, which is lower than in previous studies. The presence of DIF varied across subtests, with the information subtest having the lowest number of items showing DIF.

The findings of this study partially aligned with previous research conducted in Canada and the United States while also showing some inconsistencies. Factors such as item content, the age distribution of the sample, and nonuniform DIF may contribute to the varied results observed across different studies and countries.

Acknowledgement: The authors extend their appreciation to the Ongoing Research Funding Program, King Saud University, Riyadh, Saudi Arabia, for funding this work.

Funding Statement: This Research was funded by Ongoing Research Funding Program, number (ORF-2026-1533), King Saud University, Riyadh, Saudi Arabia.

Author Contributions: Ismael Salamah Albursan conceptualized the study, supervised the research design, and critically revised the manuscript for its significant scientific and intellectual content. Mohammad Farhan Al. Qudah contributed to the development of the methodology, planned the statistical analyses, participated in interpreting the Differential Item Functioning (DIF) results, and assisted in reviewing and editing the manuscript. Eqbal Z. Darandari contributed to the psychometric interpretation of the WISC-III constructs, participated in data validation, and assisted in refining the manuscript. Salaheldin Farah Bakhiet served as the principal investigator in the Sudan context, supervised data acquisition from the National Commission for the Gifted in Sudan, conducted advanced psychometric analyses (including DIF procedures), drafted major sections of the manuscript, and led the theoretical integration and final scientific review. Mohammed Ateik Al-Khadher managed and coded the data, supported the implementation of the statistical analyses, prepared the tables and presentation of results, and contributed to the technical editing and formatting of the manuscript. All authors reviewed and approved the final version of the manuscript.

Availability of Data and Materials: Data available on request from the authors.

Ethics Approval: The Institutional Review Board of King Saud University determined that ethical approval was not required for this study, because it relied exclusively on secondary, de-identified data obtained from the routine assessment protocols of the National Authority for Gifted Education in Sudan. No direct contact, testing, or data collection from children occurred for the purpose of this research, and the dataset contained no personal identifiers. All procedures complied with international ethical guidelines and the principles of the Declaration of Helsinki.

Conflicts of Interest: The authors declare no conflicts of interest.

References

Al-Khalifa, O. H., & Al-Hussain, A. T. (2012). Performance rates on WISC-III in Sudanese States. *Journal of Educational and Psychological Studies*, *6*(2), 121–134. (In Arabic). https://doi.org/10.53543/jeps.vol6iss2pp121-134.

Al-Housein, A. (2005). Adapting and standardizing WISC-III in Khartoum. Unpublished MA Thesis, Khartoum, Sudan: Alneelain University. (In Arabic).

Al-Housein, A. (2008). Adapting and standardizing WISC-III in the Northern States. Unpublished Ph.D. Dissertation, Khartoum, Sudan: Alneelain University. (In Arabic).

Al-Khalifa, O., Taha, A., & Al-Housein, A. (2008). Adapting WISC-III in Sudan and Japan: A cross-cultural study. *The Arab Journal For Special Education*, *12*, 171–194. (In Arabic).

Al-Kilani, A., & Al-Rawsan, F. (2006). *Evaluation in Special Education*. 1st ed. Amman, Jordan: Dar Al-Maysarah Publisher. (In Arabic).

Anastasi, A. (1958). *Differential Psychology: Individual and Group Differences in Behavior*. 3rd ed. New York, NY, USA: Macmillan.

Bakhiet, S. F., & Lynn, R. (2015). Gender differences on the wechsler intelligence scale for children-III in Bahrain and the United States. *Psychological Reports*, *117*(3), 794–798. https://doi.org/10.2466/03.17.pr0.117c26z9

Camilli, G., & Shepard, L. (1994). *Methods for Identifying Biased Test Items*. London, UK: SAGE publications.

Clauser, B. E., & Mazor, K. M. (1998). Using statistical procedures to identify differentially functioning test items. *Educational Measurement: Issues and Practice*, *17*(1), 31–44. https://doi.org/10.1111/j.1745-3992.1998.tb00619.x

De Ayalla, R. (2022). *The Theory and Practice for Item Response Theory*. New York, NY, USA: Guilford Press.

Dorans, J., & Holland, W. (1994). *DIF Detection and Description Mantel-Haenszel and Standardization*. Princeton, NJ, USA: Educational Testing Service.

Feingold, A. (1993). Gender differences in personality: A meta-analysis. *Psychological Bulletin*, *116*(3), 429–456. https://doi.org/10.1037/0033-2909.116.3.429

Gokiert, R. J., & Ricker, K. L. (2004). Gender differential item functioning on the WISC-II: Analysis of the Canadian standardization sample. In *Centre for research in applied measurement and evaluation*. Edmonton, Alberta: University of Alberta.

Hambleton, R., & Rogers, J. (1995). Item bias review practical assessment, research and evaluation. [cited 2026 Jan 20]. http://edresearch.Org/pare/get.ASP=4&n=6.

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

Holden, L. T. R., Hart, S. A., & Lakin, J. M. (2023). Modern assessments of intelligence must be fair and equitable. *Journal of Intelligence*, *11*(7), 126. https://doi.org/10.3390/jintelligence11070126.

Jensen, A. (1980). *Bias in Mental Testing*. New York, NY, USA: A Division of Macmillan Publishing Co., Inc.

Li, L., & Becker, B. J. (2021). Assessing differential bundle functioning using meta-analysis. *Journal of Educational Measurement*, *58*(4), 492–514. https://doi.org/10.1111/jedm.12303

Lim, H., Choe, E., & Han, K. (2021). A residual-based differential item functioning detection framework in item response theory. Paper presented at the Annual Meeting of the National Council on Measurement in Education (NCME), Virtual.

Lin, Q., & Steedle, J. (2020). Investigating assessment conditions potentially associated with differential item functioning. *ACT Research & Policy, Technical Brief (R1852). ACT, Inc*. [cited 2026 Jan 20]. https://www.act.org/content/dam/act/unsecured/documents/R1852-assessment-conditions-DIF-2020-11.pdf.

Maccoby, E. E. (1966). *The Development of Sex Differences*. Stanford, CA, USA: Stanford University Press.

Maccoby, E. E., & Jacklin, C. N. (1974). *The Psychology of Sex Differences*. Stanford, CA, USA: Stanford University Press.

Maller, S. (2001). Differential item functioning in the WISC-III item parameters for boys and girls in the national standardization sample. *Educational and Psychological Measurement*, *61*(5), 793–817. https://doi.org/10.1177/00131640121971527

Martinková, P., Drabinová, A., Liaw, Y. -L., Sanders, E. A., McFarland, J. L., & Price, RM. (2017). Checking equity: Why differential item functioning analysis should be a routine part of developing conceptual assessments. *CBE—Life Sciences Education*, *16*(2), 1–13. https://doi.org/10.1187/cbe.16-10-0307

Millsap, R. E. (2023). *Statistical Approaches to Measurement Invariance*. Abingdon, Oxfordshire, UK: Routledge, https://doi.org/10.4324/9781003260242.

Nicholson, C., & Alcorn, C. (1993). Interpretation of the WISC-III and its subtests. Paper presented at the Annual Meeting of the National Association of School Psychologists (25th, Washington, DC, USA, April 17, 1993).

Özberk, U. E., & Koc, K. (2017). Gender and socioeconomic status DIF on the WISC-IV Turkish form: A comparison of DIF detection techniques. *Journal of Measurement and Evaluation in Education and Psychology*, *8*(1), 112–127.

Paek, I., & Cole, K. (2022). *Using R for Item Response Theory Model Applications*. Abingdon, Oxfordshire, UK: Routledge.

Penfield, R. D. (2007). An approach for categorizing DIF in polytomous items. *Applied Measurement in Education*, *20*, 335–355. https://doi.org/10.1080/08957340701431435

Prifitera, A., Weiss, L., & Saklofske, D. (1998). The WISC-III in context. In: Prifitera, A., Saklofske, D. (Eds.), *WISC-III clinical use and interpretation (1–35) (Scientist Practitioner Perspectives)*. San Diego, CA, USA: Academic Press.

Rodríguez-Cancino, M., Salas, G., & Valenzuela, M. (2023). WISC-V measurement invariance according to sex and age. *Journal of Intelligence*, *11*(9), 180. https://doi.org/10.3390/jintelligence11090180

Sattler, J. (2001). *Assessment of Children: Cognitive Applications*. 4th ed. San Diego, CA, USA: Academic Press.

Tavakol, M., Stewart, C., & Sharpe, C. (2024). Ensuring fairness in assessment in health professions education: Rapid analysis tools to detect differential item functioning across groups. *International Journal of Medical Education*, *15*, 80–83. https://doi.org/10.5116/ijme.6694.de69

Teresi, J. A., & Fleishman, J. A. (2008). Occurrences and sources of Differential Item Functioning (DIF) in patient-reported outcome measures. *Psychology Science Quarterly*, *49*(4), 428–443.

Tyler, L. (1965). *The Psychology of Human Differences*. 3rd ed. New York, NY, USA: Appleton.

Wechsler, D. (1949). *Wechsler Intelligence Scale for Children*. New York, NY, USA: The Psychological Corporation.

Wechsler, D. (1974). *Manual for the Wechsler Intelligence Scale for Children—Revised*. New York, NY, USA: Psychological Corporation.

Wechsler, D. (1991). *Manual for the Wechsler Intelligence Scale for Children*. 3rd ed. (WISC-III). San Antonio, TX, USA: The Psychological Corporation.