International Journal of Evaluation and Research in Education (IJERE)
Vol. 15, No. 6, December 2026, pp. 4897~4910
ISSN: 2252-8822, DOI: 10.11591/ijere.v15i6.39685

Sex-based differential item functioning in the Wechsler Intelligence Scale for Children–fifth edition: implications for educational assessment in Jordan

Nizar Allabadi[1], Mohammad Al-Karamneh[2], Jehad M. Al-Anati[1], Laith Jumah[1], Omar Alorani[3]
[1]Department of Educational Psychology, The University of Jordan, Amman, Jordan
[2]Department Administration Curricula and Teaching Methods, Amman Arab University, Amman, Jordan
[3]Department of Counseling and Special Education, The University of Jordan, Amman, Jordan

Article Info

Article history:

Received Mar 26, 2026
Revised Oct 12, 2026
Accepted Oct 19, 2026

Keywords:

Differential item functioning
Fairness
Item bias
Psychometric analysis test
WISC-V

ABSTRACT

This study aimed to investigate differential item functioning in the Arabic version of the Wechsler Intelligence Scale for Children–fifth edition (WISC-V) in the Jordanian context by sex, using logistic regression and probability ratio testing. A descriptive approach was employed, and 15 subtests of the WISC-V scale were administered to a sample of 468 children, including normal children, gifted, learning-disabled, and mildly intellectually disabled children, selected using stratified random sampling. The children's ages ranged from 6 to 6.11 years, and the age groups were defined as 11 age categories, each encompassing one calendar year. The logistic regression analysis revealed 14 items with sex-based differential performance in the cube design, memory capacity, information, visual memory, arithmetic, and shape weighting tests. The differential performance was uniform in 10 items and non-uniform in 4. The probability ratio test results showed sex-based differential performance in 10 items of the information test, including arithmetic and shape weighting. Differential performance was uniform in 3 items and non-uniform in 7 items. The overall agreement rate between the detection methods (likelihood-ratio test, logistic regression) was 97.4%. The study recommends employing multiple methods for detecting differential item functioning to enhance the accuracy of interpretation, as well as training psychologists to use test results with caution and to consider potential sex-related differences in specific items.

This is an open access article under the CC BY-SA license.

Corresponding Author:

Mohammad Al-Karamneh
Department Administration Curricula and Teaching Methods, Amman Arab University
Al-Urdon St., Amman, Jordan
Email: m.karamneh@aau.edu.jo

1. INTRODUCTION

Psychological and educational measurements are fundamental cornerstones in assessing individuals' abilities, behaviors, and learning outcomes. They aim to design standardized tools for measuring cognitive, emotional, and behavioral characteristics across diverse environments [1]. The availability of accurate measurement tools enables objective assessment of individuals' cognitive abilities and promotes fairness by minimizing systematic bias that may affect individuals or groups due to cultural, linguistic, social, or economic factors [2]. From this perspective, intelligence tests occupy a pivotal role as systematic and standardized instruments for evaluating cognitive performance and predicting academic achievement and life outcomes [3]. Furthermore, such tests contribute to enhancing fairness in educational assessment, which

remains a critical educational issue [4], [5]. These considerations highlight the importance of ensuring that psychological measurements are both valid and equitable across different populations.

In line with this perspective, the Wechsler Intelligence Scale for Children (WISC-V) is one of the most widely used standardized intelligence tests for assessing the cognitive abilities of children and adolescents aged 6 to 16 years. Developed by Wechsler and published in 2014, it represents the latest development in the Wechsler series of tests, the most recent theories of intelligence with psychometric advancements [6]. Unlike previous forms, the WISC-V offers a multidimensional assessment structure, encompassing five main cognitive domains: verbal comprehension, visual-spatial, fluid reasoning, working memory, and processing speed [7]. Sub-index scores are derived from these domains, in addition to the full-scale IQ (FSIQ), which serves as an international indicator of overall intellectual functioning for population groups. These features enhance the clinical and educational utility of the scale by providing a more comprehensive profile of cognitive strengths and weaknesses.

Building on this structure, the WISC-V scale was developed based on a large, representative sample, giving it strong psychometric properties and applicability to diverse population groups [8]. Its effectiveness extends beyond cognitive description to include supporting clinical diagnosis, educational decision-making, and the identification of developmental strengths and weaknesses, making it a needed instrument in psychological and educational assessment [9]. In addition to its core subtests, the WISC-V contains a range of secondary and ancillary subtests that enhance the flexibility and diagnostic precision of the assessment. Subtests allow examinees to discover specific cognitive processes, cross-validate results, or substitute when certain core tasks are not appropriate. For example, the information and comprehension subtests can provide deeper insight into verbal knowledge and social reasoning, while image concepts taps into categorical reasoning and abstract thinking [6]. The arithmetic subtest offers an applied measure of working memory and numerical reasoning, and cancellation extends the evaluation of processing speed with a more complex visual scanning task [8]. Together, these additional measures allow practitioners to generate expanded index scores, such as the general ability index (GAI) or cognitive proficiency index (CPI), which provide alternative interpretations when the FSIQ may not fully represent a child's abilities [10]. This multidimensional structure strengthens the interpretive value of the scale in both clinical and educational contexts.

Beyond its structural and psychometric strengths, the WISC-V is an internationally recognized and widely used instrument, demonstrating its validity across multiple cultural settings. A study conducted in Greece [11] indicated the importance of using the test in assessing learning disabilities and identifying specific cognitive deficits. In the same vein, studies conducted in Sweden and Finland explored the cross-cultural applicability of the WISC-V, confirming its effectiveness in assessing cognitive abilities in these populations [12]. The WISC-V was first published in 2014 and has since undergone extensive translation and revision in some countries, including Canada, Australia, Spain, and some Scandinavian countries. These revisions were made to ensure its cultural relevance and psychometric validity across different target populations [13]. In the Arabic context, the WISC-IV scale has been adapted to meet the linguistic and cultural needs of Arabic-speaking children. This version has undergone substantial modifications to align with the cultural context of Arabic-speaking children, ensuring its relevance and accuracy in assessing cognitive abilities [14]. In the Jordanian context, there is currently no empirical evidence examining the psychometric properties or item-level fairness of the WISC-V in Jordan.

Despite its widespread use, Holden and Tanenbaum [5] pointed out that intelligence testing has historically been associated with concerns regarding bias, discrimination, and inequality. Among these are the sex differences in cognitive abilities, as proposed by Geary [15], which remains a subject of debate in the literature. Geary [15] conducted an analysis of sex differences using the WISC scale. The results showed small differences in general cognitive performance (FSIQ). In addition, females performed better on tasks that required high processing speed and maintaining focus on the task for a short period, such as programming, while males performed better on tasks that required visual manipulation of stimuli, such as block design.

In this context, this change becomes even more important when one takes into account the differences between males and females, which are often influenced by social, cultural, and educational contexts rather than merely biological factors. For example, parental expectations, sex-stereotypical educational practices, and the availability of educational resources all contribute to differences in the development of verbal, visual, and problem-solving skills between males and females [16], [17]. Socioeconomic status, school quality, and cultural norms regarding sex roles all play a significant role in intelligence test scores, highlighting the importance of interpreting these differences within their social and cultural context [18]. Therefore, analyzing sex differences in cognitive performance requires a comprehensive understanding of the interaction between biological and environmental factors to ensure a fair and culturally appropriate scientific explanation. This trend is supported by the study [18], which showed small and statistically insignificant differences on the WISC scale in most cognitive dimensions, such as

verbal comprehension and working memory, which means that sex alone does not explain these differences, and other surrounding factors must be taken into account to understand the full picture.

In light of these findings, issues of fairness and justice are increasingly important in the design and interpretation of intelligence tests, as they are essential components of the framework of validity evidence needed to justify the use of test scores for a specific purpose. Accordingly, relying on standardized tests such as the WISC-V provides a precise framework for analyzing these differences and enables researchers and educators to identify patterns of strengths and weaknesses, which can help design personalized educational interventions, reduce assessment bias, and support equal learning opportunities [19]. In this context, the concept of differential performance emerges as an indicator of systematic variability in test scores across individuals or groups when using standardized assessment tools, this is where differential item functioning (DIF) individuals from different groups (such as sex or race) despite having the same baseline level of cognitive ability, in other words, an item exhibits DIF if it gives an advantage to one group over another despite equal mental ability, indicating potential within-item bias [20].

Given its importance, ensuring assessment fairness requires the detection of DIF. Researchers can prevent any group from being unfairly disadvantaged or marginalized by removing or altering items that perform differently [21]. Equal opportunity in research comparisons and instructional guidance is ensured by the use of DIF-free instruments. DIF does not always imply bias, even though it shows a difference in item performance between groups. According to Zumbo [22], bias arises when DIF causes a group's ability to be systematically overestimated or underestimated. Therefore, DIF is a statistical phenomenon, while bias has interpretive and practical implications for fairness. Understanding this distinction is essential for assessing the validity and integrity of tests. DIF often manifests as differences across cognitive domains, in subtest scores, or between specific groups such as sex, age, socioeconomic status, or cultural background [9]. It is important to emphasize that differential performance is not necessarily an indicator of superiority or deficiency, but rather reflects a difference in the pattern of cognitive performance, which may have educational, clinical, or developmental implications [23]. In this context, several studies have shown increasing interest in analyzing differential performance on the WISC. Adepoju *et al.* [24] noted that 145 items on the WISC perform differently, while 125 items show no bias. Ozberk and Koc [25] concluded that some items show differences based on sex and others based on socioeconomic status. Giofrè *et al.* [26] showed that the WISC exhibits sex variance and that there are some very unequal subtests, such as the "block design" and "coding" tasks, which showed performance differences between boys and girls.

To further explain this, there are many reasons that contribute to DIF on intelligence tests. On the one hand, individual, cognitive, and neurodevelopmental factors, such as individual differences in reasoning, memory, attention, and processing speed, contribute to determining mental performance [10]. On the other hand, environmental factors play an equally important role, such as the quality of the educational environment, parental support, exposure to rich educational experiences, and socioeconomic conditions [18]. Language and cultural differences also directly influence performance, particularly on verbal tasks, where test items may include vocabulary or concepts from specific cultural contexts [27]. Additionally, test construction features like item difficulty phrasing and scoring systems may unintentionally favor some groups, underscoring the necessity of creating assessment instruments that are both psychometrically sound and culturally equitable. To evaluate test fairness and detect potential item bias, psychometric research has identified some effective statistical methods for detecting DIF. Logistic regression is a main example, used to detect uniform differential item functioning (UDIF) and non-uniform differential item functioning (NUDIF), it features a statistical significance scale and an effectiveness size, requires a small sample size, and utilizes the logit function to transform probabilities from the range (0, 1) to the range $(-\infty, \infty)$ in linear prediction. It can be used for multiple-response items [28]. To assess the magnitude of differential performance in this method, Zumbo and Thomas, cited in Hidalgo and López-Pina [29], suggested using $\Delta R2$, where the size is small if $\Delta R2<0.13$, medium if $0.13\leq\Delta R2\leq0.26$, and high if $\Delta R2>0.26$. Another method for detecting DIF is the likelihood-ratio test, which is used to test the null hypothesis that there are no differences in item parameters between groups. If the likelihood-ratio test is statistically significant, it indicates the presence of DIF in the item; this method is used to detect DIF in two-response items or multi-response items [30].

Despite these methodological advances, there remains a critical lack of empirical research investigating the functioning of WISC-V items within the Jordanian context, particularly with respect to fairness across groups. To ensure fair evaluation and equal educational opportunities in the Jordanian context, it is essential to understand how items function across various categories. Jordanian children are exposed to particular social, cultural, and educational elements, such as bilingual education, socioeconomic inequality, and sex-related educational practices, that may have an impact on their cognitive performance. However, no prior study has systematically examined whether these contextual factors are reflected in DIF within the WISC-V. Teachers, psychologists, and legislators may be able to design culturally appropriate interventions and accurately identify learning difficulties or giftedness by learning about potential biases in cognitive assessment instruments through studies of DIF in the WISC-V [27], [31].

The results of this study can help improve the interpretation of test scores and reduce bias in decision-making. For schools, the study supports more equitable placement of students in special education and gifted programs. For policymakers, the results provide evidence to support the need for culturally validated and fair assessment tools.

Despite the widespread international use of the WISC-V, no empirical study to date has examined DIF across sex within the Jordanian context. Existing research in Arab settings has primarily focused on earlier intelligence measures such as (TOMA-3, TONI-4), leaving a significant gap in the literature regarding the psychometric fairness of the WISC-V [32]–[39]. Moreover, although previous studies have investigated DIF in intelligence tests, there is limited evidence on whether the WISC functions equivalently across sexes in Jordan, particularly in high-stakes educational decision-making contexts. This gap is particularly important given the high-stakes use of intelligence test scores in Jordan for educational placement, including identification for special education and gifted programs. Without evidence of item-level fairness, the validity of these decisions remains uncertain.

This study offers several novel contributions to the literature on psychometric fairness and intelligence assessment. First, it represents the first empirical investigation of sex-based DIF in the Arabic version of the WISC-V within the Jordanian context, thereby addressing a significant gap in Arab psychometric research. Second, unlike previous studies that primarily focused on earlier intelligence measures or non-Arab populations, the present study evaluates the item-level fairness of a contemporary and internationally recognized cognitive assessment instrument in an underrepresented cultural setting. Third, the study contributes methodologically by employing two complementary DIF detection techniques, ordinal logistic regression and the likelihood ratio test, and systematically examining the level of agreement between them. This dual-method approach enhances the robustness and credibility of DIF detection and provides additional evidence regarding the consistency of the statistical procedures used in psychometric fairness research. Collectively, these contributions advance understanding of culturally responsive cognitive assessment and support the development of fairer intelligence-testing practices in educational and psychological contexts.

2. METHOD

2.1. Research design

The descriptive research design was adopted in this study due to its suitability for the research objective. This design is commonly used to examine phenomena as they naturally occur, without any experimental manipulation. It allows researchers to analyze the characteristics of a phenomenon and to understand the relationships among its variables in a systematic and objective manner. In addition, this design facilitated the comparison of item performance between male and female participants using appropriate statistical techniques.

2.2. Population

The study population consisted of all school students spread across the regions of the Hashemite Kingdom of Jordan during the 2025-2026 academic year. The participants' ages ranged from 6 to 16.11 years, which corresponds to the age range targeted by the WISC-V scale. According to official educational statistics, the total population included approximately 2,244,751 male and female students.

2.3. Participants

To cover the age range that is appropriate for the scale, schools were selected from different regions of Jordan, including Amman, Salt, Zarqa, and Madaba in the central region, Irbid, Ajloun, and Jerash in the northern region, and Karak in the southern region. This geographic distribution was intended to enhance the representativeness of the sample across the Kingdom. A total of 468 children were selected, with ages from 6-16.11 years, including normal, gifted, learning disabled, and mildly mentally disabled children, who were selected using a stratified random sampling method.

The sample was organized into 11 groups, each representing a one-year age interval to ensure developmental comparability across groups. The confidentiality of the collected information was guaranteed. The children were given the freedom to participate voluntarily, with their right to withdraw at any stage of the study being respected; the findings will not be shared elsewhere but will be used solely for educational purposes, particularly for research endeavors. Therefore, several ethical principles of research ethics were meticulously adhered to, including integrity, honesty, caution, respect, and transparency. Table 1 shows the distribution of the target sample from each age group according to sex.

Table 1. Distribution of study sample members according to student sex and age group

Age	Male	Female	N
6–6.11	20	13	33
7–7.11	23	16	39
8–8.11	25	17	42
9–9.11	27	28	55
10.11–10	14	25	39
11.11–11	33	24	57
12–12.11	18	15	33
13–13.11	28	26	54
14–14.11	26	20	46
15–15.11	17	15	32
16–16.11	19	19	38
Sum	250	218	468
%	53.4%	46.6%	100%

2.4. Scale WISC-V in its original form

The WISC-V consists of 21 subtests designed for children aged 6 to 16.11 years. The validity and reliability of the scale in its initial form were verified through its application to a sample of 2,200 children. Convergent and discriminant validity were demonstrated through correlations among all subtests.

Further psychometric analyses confirmed that the WISC-V measures five interrelated cognitive abilities using confirmatory factor analysis. Criterion-related validity was calculated through the correlation coefficient between the composite score and the subtest on the WISC-V and the Kaufman Assessment for Children (KABC-II). The correlations ranged from 0.77 to 0.81, while the corrected correlations between the corresponding subtests ranged from 0.5 to 0.74. The split-half reliability value for the total test score was (0.96), the reliability coefficients for the results of the main indicators ranged from (0.88-0.93), the reliability values for the subtests ranged from (0.80-0.94), the average retest reliability coefficient for the subtests at different ages ranged from (0.71-0.90), the corrected reliability coefficients for the results of the main indicators ranged from (0.75-0.94), and the reliability coefficient for the total test score was (0.92) [6].

2.5. Arabic version of the WISC-V for the Jordanian context

The Arabic version of the WISC-V for the Jordanian environment consisted of 15 tests. Six subtests were deleted from its original version, namely: picture concepts, speed of naming knowledge, quantity of naming speed, immediate interpretation of symbols, delayed interpretation of symbols, and knowledge of interpreting symbols. The Higher Council for the Affairs of the Disabled in Jordan explained that these tests are used more in clinical psychological and clinical assessment than in assessment related to learning disabilities and intelligence testing. This center diagnoses individuals with disabilities and ensures their access to services and social inclusion. The 15 subtests measure five main factors: verbal comprehension, spatial perception, abstract thinking, working memory, and processing speed. Table 2 shows the distribution of the subtests among these factors.

The validity and reliability of the scale were verified in the Jordanian environment through logical validity procedures carried out by Princess Tharwat College by obtaining the original (American) version, translating it into Arabic, and presenting it to experts to verify the translation's conformity to the original version. In light of their suggestions, items 7 and 17 of the similarities test and item 35 of the information test were modified, and the reliability coefficient by the test-retest method reached 0.82-0.99.

Table 2. Distribution of the Jordanian version of the WISC-V subtests across the five factors

Main factors	Subtests
Verbal comprehension	Similarities vocabulary information and comprehension
Spatial perception	Visual puzzles, cube design
Abstract thinking	Intellectual matrices, shape weights
Working memory	Number memory capacity, image memory, letter number sequence
Processing speed	Coding, symbol's, deletion

3. RESULTS AND DISCUSSION

3.1. RQ1: Which items of the WISC-V test, Arabic version for the Jordanian environment, showed differential performance according to the sex variable using logistic regression?

To answer this question, logistic regression was used, and the results showed differential performance between males and females across the following subtests. Table 3 shows differential performance for item 7 in the cube design test, items 6 and 7 in the memory capacity test, items 6, 9, 11, 12, 13, and 18 in the shape and weights test, item 7 in the image memory test, items 17 and 18 in the information

test, and items 11 and 29 in the arithmetic computation test, where the Δ3 values for all items were statistically significant at (α=0.01). The table also shows that differential performance was non-uniform in the memory capacity, cube design, and arithmetic computation test items, and in items 6, 9, 11, 12, and 13 of the Shape and weights test, because the corresponding probability value Δ1 for these items was not statistically significant at the (α=0.01) level. However, differential performance appeared uniform in item 18 of the shape and weights test, favoring males (b=-1.192), and in item 7 of the image memory test. The performance was favorable for females, with values of (b=0.813) and items 17 and 18 of the information test, and favorable for males, with values of (b=1.072) and (=-0.775) respectively. The effect of the differential performance for item 7 of the cube design test, items 6 and 7 of the memory capacity test, items 9, 11, 12, 13, and 18 of the shape and weights test, item 7 of the image memory test, items 17 and 18 of the information test, and item 11 of the arithmetic computation test was small, with ΔR^2 being less than 0.13, while it was moderate for item 6 of the shape and weights test and item 29 of the arithmetic computation test, where $0.13 \leq \Delta R^2 \leq 0.26$.

Table 3. Differential performance items according to sex variable using logistic regression

Subtests	Item No	Δ1 (chi2-ch1)	sig	Δ2 (chi3-ch2)	sig	Δ3 (chi3-ch1)	sig	b	ΔR^2 ($R^2$3-$R^2$1)
Cube design	CD7	1.67	0.196	8.993	0.003*	10.663	0.005*	NUDIF	0.0344
Memory	MC 6	0.017	0.896	9.979	0.002*	9.996	0.007*	NUDIF	0.0347
capacity	MC 7	2.466	0.116	6.921	0.009*	9.387	0.009*	NUDIF	0.0455
Shape weights	SW_6	1.511	0.219	9.66	0.002*	11.171	0.004*	NUDIF	0.183
	SW_9	4.195	0.041	9.32	0.002*	13.515	0.001*	NUDIF	0.069
	SW_11	2.105	0.147	8.63	0.003*	10.735	0.005*	NUDIF	0.036
	SW_12	6.457	0.011	5.966	0.015*	12.423	0.002*	NUDIF	0.028
	SW_13	6.019	0.014	6.207	0.013*	12.226	0.002*	NUDIF	0.021
	SW_18	9.305	0.002*	0.09	0.764*	9.395	0.009*	-1.192	0.011
Image memory	IM_7	12.715	0.000*	0.411	0.521*	13.126	0.001*	0.813	0.033
Information	INF_17	11.109	0.001*	0.518	0.472*	11.627	0.003*	1.072	0.015
	INF_18	7.517	0.006*	2.661	0.103*	10.178	0.006*	-0.77	0.034
Arithmetic	AC_11	4.652	0.031	4.672	0.031*	9.324	0.009*	NUDIF	0.046
computation	AC_29	0.113	0.737	13.464	0.000*	13.577	0.001*	NUDIF	0.157

Statistically significant at the significance level (α=0.01); Non-uniform differential item functioning: NUDIF**
Δ1, Δ2, and Δ3 represent Chi-square differences between nested logistic regression models.

3.2. RQ2: Which items of the WISC-V test, Arabic version for the Jordanian environment, showed differential performance according to the sex variable using the likelihood-ratio test method?

To answer this question, the likelihood-ratio test method was used, which is based on item response theory (IRT). The assumption of unidimensionality for the subtest was verified using an exploratory factor analysis (EFA). Table 4 shows the eigenvalues of the factors explaining performance on the WISC-V subtests. Kaiser's criterion was used to determine the number of factors, where a factor is considered significant if its (Eigenvalue>1). It is also noted that the sum of the variance explained by factors on the subtests was all greater than 50% [40]. The results indicate that the ratio of the eigenvalues first factor to the eigenvalues second factor was greater than 2 for all subtests except for the memory capacity and letter number sequence tests, where the ratio was 1.9, close to 2. This confirms the unidimensional assumption [41], meaning that the scale items measure a single trait: intelligence. The results also showed differential performance between males and females on the following subtest items according to the likelihood-ratio test method.

Table 5 shows that items 2, 6, 9, 11, 12, and 13 exhibited differential performance in the shape weights test, with the G^2 value for all items being greater than 3.84. The table also shows that items 8, 17, and 26 exhibited differential performance in the information test, with the G^2 values for all items exceeding 3.84. The table also shows that item 11 exhibited differential performance in the arithmetic computation test, with its G^2 value being greater than 3.84. All items in the subtest were statistically significant at the (α=0.05). To determine whether the differential performance was uniform or non-uniform, the shape curves of the item characteristics in the tests that exhibited differential performance (shape weights, information, and arithmetic computation) were extracted, as shown in Figures 1 to 3.

Figure 1 shows that the curve of the characteristics of item 2 reveals a uniform differential performance in favor of females, and the differential performance was non- uniform for items 2, 6, 9, 11, 12, 13. For students with low abilities, the differential performance was in favor of males, but as the students' abilities increased, the differential performance was reversed to become in favor of females. Furthermore, Figure 2 shows that the curve of the characteristics of item 2 reveals a uniform differential performance in favor of females, and the differential performance was non-uniform for items 8, 17, 26. For students with low abilities, the differential performance was in favor of males, but as the students' abilities increased, the

differential performance was reversed to become in favor of females. Lastly, Figure 3 shows that the curve of the characteristics of item 11 reveals non-uniform differential performance. For students with low abilities, the differential performance favored males, but as students' abilities increased, the differential performance reversed to favor females.

Table 4. Eigenvalues of WISC-V subtests according to EFA

Subtest	Component	Initial eigenvalues			Extraction sums of squared loadings		
		Total	% of variance	Cumulative (%)	Total	% of variance	Cumulative (%)
Cube design	1	11.56	31.24	31.24	11.56	31.24	31.24
	2	3.46	9.36	40.60	3.46	9.36	40.60
	3	1.96	5.29	45.89	1.96	5.29	45.89
	4	1.53	4.13	50.02	1.53	4.13	50.02
	5	1.21	3.27	53.29	1.21	3.27	53.29
Similarities	1	8.56	37.22	37.22	8.56	37.22	37.22
	2	2.75	11.97	49.19	2.75	11.97	49.19
	3	1.34	5.82	55.01	1.34	5.82	55.01
Intellectual matrices matrix reasoning	1	9.58	29.93	29.93	9.58	29.93	29.93
	2	4.48	14.00	43.93	4.48	14.00	43.93
	3	2.08	6.50	50.43	2.08	6.50	50.43
	4	1.54	4.81	55.24	1.54	4.81	55.24
	5	1.17	3.65	58.89	1.17	3.65	58.89
	6	1.04	3.26	62.15	1.04	3.26	62.15
Memory capacity	1	8.68	32.16	32.16	8.68	32.16	32.16
	2	4.57	16.92	49.08	4.57	16.92	49.08
	3	2.24	8.29	57.37	2.24	8.29	57.37
	4	1.56	5.80	63.17	1.56	5.80	63.17
	5	1.02	3.77	66.93	1.02	3.77	66.93
Vocabulary	1	8.56	37.22	37.22	8.56	37.22	37.22
	2	2.75	11.97	49.19	2.75	11.97	49.19
	3	1.34	5.82	55.01	1.34	5.82	55.01
Shape weights	1	10.39	30.56	30.56	10.39	30.56	30.56
	2	4.76	14.00	44.55	4.76	14.00	44.55
	3	2.72	8.00	52.55	2.72	8.00	52.55
	4	2.05	6.03	58.59	2.05	6.03	58.59
	5	1.43	4.22	62.80	1.43	4.22	62.80
Visual puzzles	1	8.49	29.28	29.28	8.49	29.28	29.28
	2	3.88	13.39	42.67	3.88	13.39	42.67
	3	2.07	7.12	49.79	2.07	7.12	49.79
	4	1.28	4.42	54.21	1.28	4.42	54.21
	5	1.07	3.70	57.91	1.07	3.70	57.91
Image memory	1	7.74	29.76	29.76	7.74	29.76	29.76
	2	3.02	11.61	41.37	3.02	11.61	41.37
	3	2.37	9.10	50.47	2.37	9.10	50.47
	4	1.74	6.68	57.15	1.74	6.68	57.15
	5	1.33	5.12	62.27	1.33	5.12	62.27
Information	1	11.56	31.24	31.24	11.56	31.24	31.24
	2	3.46	9.36	40.60	3.46	9.36	40.60
	3	1.96	5.29	45.89	1.96	5.29	45.89
	4	1.53	4.13	50.02	1.53	4.13	50.02
	5	1.21	3.27	53.29	1.21	3.27	53.29
Letter number sequence	1	3.75	41.66	41.66	3.75	41.66	41.66
	2	1.99	22.13	63.79	1.99	22.13	63.79
	3	1.15	12.73	76.52	1.15	12.73	76.52
Comprehension	1	7.80	35.47	35.47	7.80	35.47	35.47
	2	2.32	10.55	46.03	2.32	10.55	46.03
	3	1.01	4.59	50.61	1.01	4.59	50.61
Arithmetic Computation	1	10.73	31.55	31.55	10.73	31.55	31.55
	2	5.19	15.26	46.81	5.19	15.26	46.81
	3	2.77	8.14	54.95	2.77	8.14	54.95
	4	1.75	5.14	60.09	1.75	5.14	60.09
	5	1.23	3.62	63.71	1.23	3.62	63.71

3.3. RQ3: What is the degree of agreement between the logistic regression method and the likelihood-ratio test in detecting differential performance on items of the WISC-V test (Arabic version for the Jordanian environment)?

To answer this question, we calculated the percentage of agreement between the two methods on the subtests that showed differential performance, namely arithmetic computation, information, and shape weights. Table 6 shows a high degree of agreement between the results of differential performance detection

using the logistic regression and likelihood-ratio test, with an overall agreement rate of 94.7%. The agreement rate in the subtests ranged from 92.6% to 97%. The results indicate that for items where both methods agreed on the presence of differential performance, the agreement extended not only to the detection of differential performance but also to its type, whether uniform or non-uniform.

Table 5. Differential performance items according to sex variable using the likelihood-ratio test

Subtest	Item No	G^2	d.f	Male			Female		
				a	b	c	a	b	c
Shape weights	2	4.0	3	0.50	-5.03	0.25	1.90	-3.60	0.25
	6	8.1	3	2.07	-2.11	0.21	0.80	-2.46	0.23
	9	10.4	3	2.88	-1.19	0.17	1.22	-1.13	0.17
	11	4.1	3	2.86	-0.86	0.14	1.46	-0.86	0.13
	12	5.8	3	2.84	-0.89	0.13	1.56	-0.76	0.12
	13	6.0	3	3.74	-0.81	0.12	1.76	-0.69	0.12
Information	8	4.5	3	3.99	-1.64	0.23	2.41	-1.47	0.22
	17	4.9	3	2.2	0.45	0.07	2.54	0.76	0.07
	26	4.0	3	1.86	0.76	0.09	3.05	0.96	0.09
Arithmetic computation	11	8.8	3	1.66	-0.58	0.13	3.26	-0.84	0.12

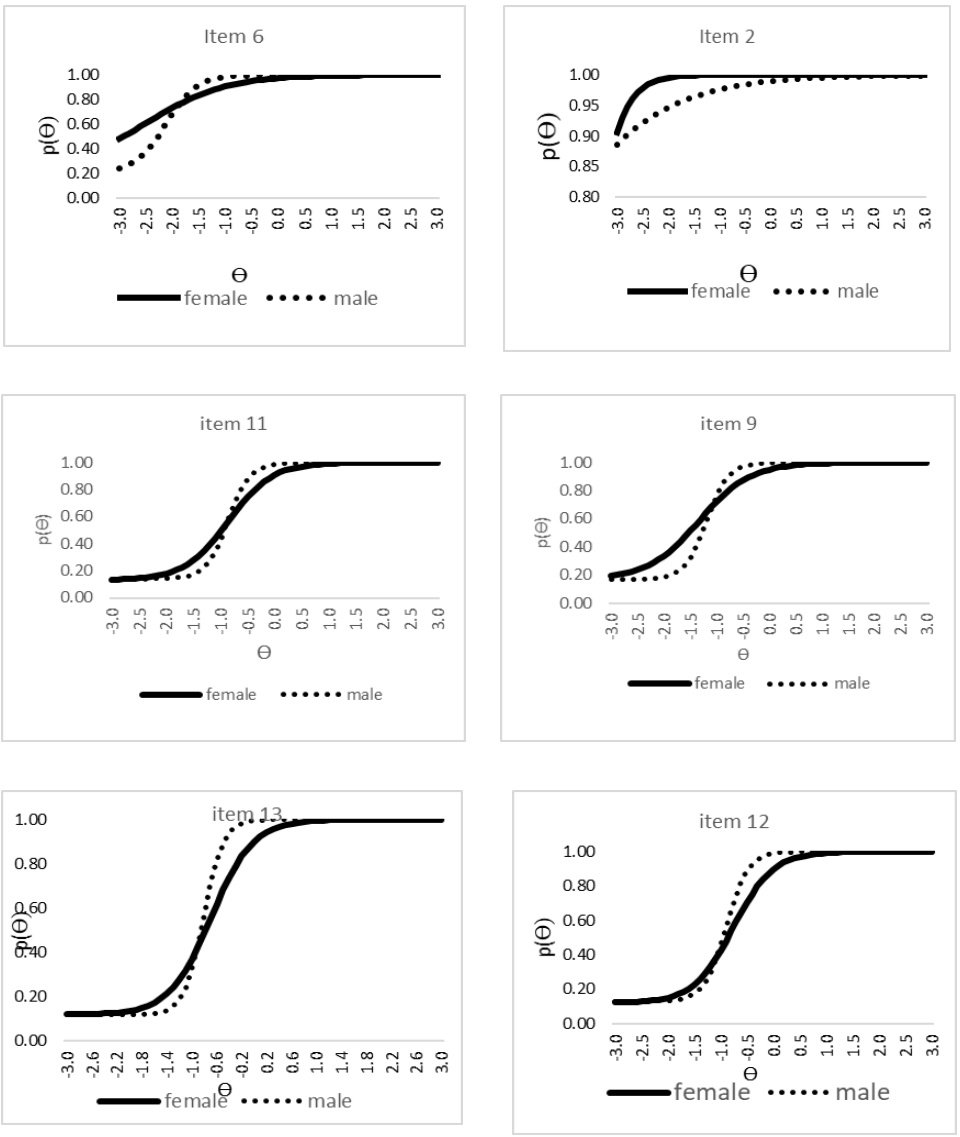


Figure 1. Curves of the characteristics of items 2, 6, 9, 11, 12, 13 in the shape weights test

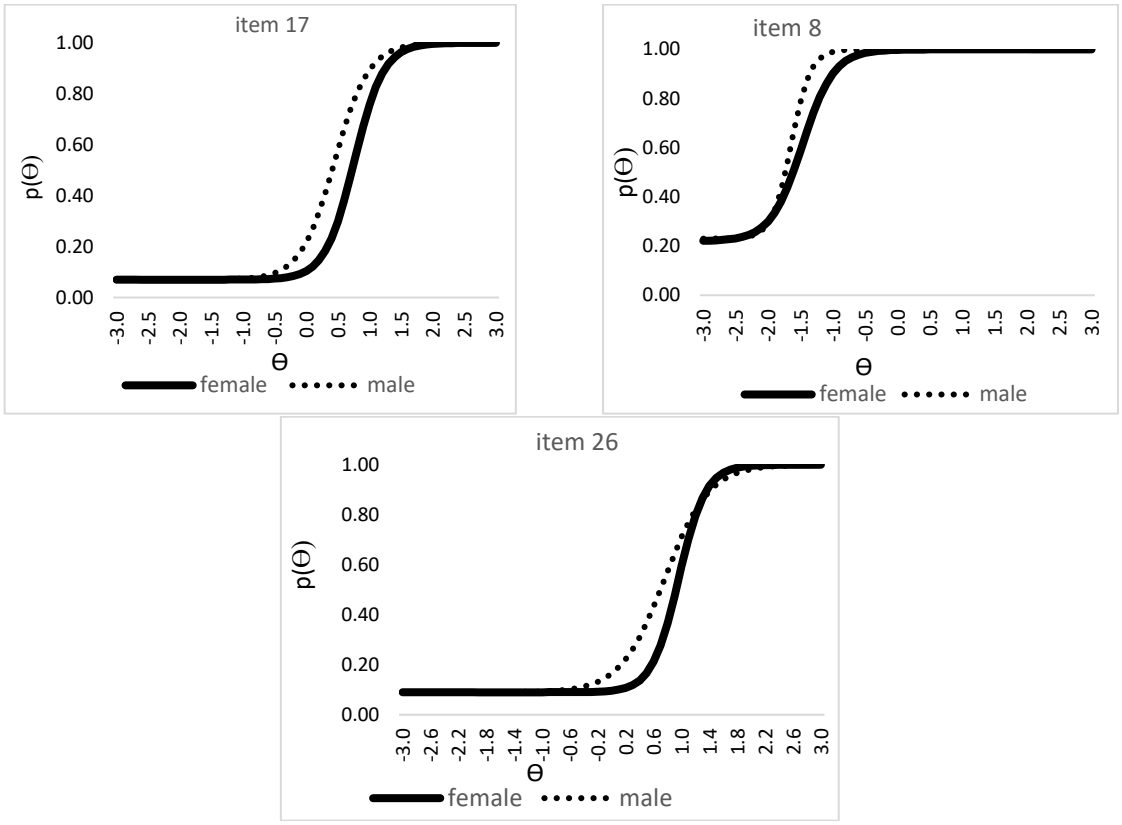


Figure 2. Curves of the characteristics of items 8, 17, 26 in the information test

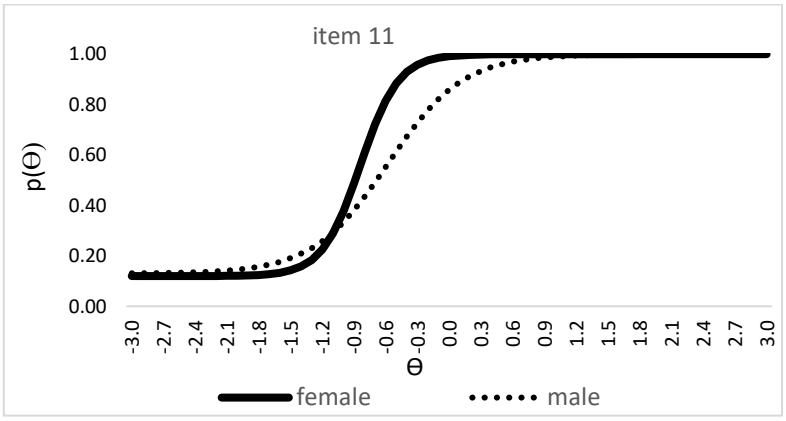


Figure 3. Curves of the characteristics of Item 11 in the arithmetic computation test

Table 6. Degree of agreement between the logistic regression and the Likelihood-ratio test in detecting the differential performance of items in the WISC-V test, Arabic version for the Jordanian environment

Test	Method	DIF	Differential performance uniform, non-uniform	Impact of differential performance	Compatibility ratio (%)
Shape weights	Logistic regression	6, 9, 11, 12, 13, 18	Non- uniform for 6, 9, 11, 12, 13	Small for 9, 11, 12, 13 Medium for 6	92.6
	Likelihood-ratio test	2, 6, 9, 11, 12, 13			
Information	Logistic regression	17, 18	Uniform for 17 and in favor of males	Small for 17	94.6
	Likelihood-ratio test	8, 17, 26			
Arithmetic computation	Logistic regression	11, 29	Non- uniform for 11	Small for 17	97
	Likelihood-ratio test	11			

Overall agreement: 94.7%

3.4. Discussion

The results of the study related to the first question showed that 14 items out of 307 items in the WISC-V scale, the Arabic version for the Jordanian environment, performed differently according to the sex variable when applying the logistic regression method. It was found that the items were concentrated in the cube design, memory capacity, arithmetic computation, information, image memory, and shape weights subscales, representing 4.56%, which is very weak. DIF did not appear in the remaining subscales. Total 12 items that showed weak (negligible) differential performance were classified according to the Zumbo and Thomas classification referred to in Hidalgo and López-Pina [29], representing 85.7%. Only 2 items were classified as performing with moderate differential performance, representing 14.3%. Differential performance was irregular in 10 items, representing 71.4%, and in 4 items, representing 28.6%, it was performed with regular differential performance. The results of the second question showed the presence of 10 items that exhibited differential performance based on sex when the likelihood-ratio test was applied. The items were found to be concentrated in the Shape weights, Arithmetic Computation, and Information subscales, while DIF was not observed in the other subscales. Differential performance was consistent in 3 items (30%) and inconsistent in 7 items (70%). The DIF observed in these items did not follow a single pattern, sometimes favoring males and other times females, indicating the absence of sex bias in the Arabic version of the WISC-V scale adapted for the Jordanian context. The results for question three showed an overall agreement rate of 97.4% with the detection method (logistic regression likelihood-ratio test), a high percentage indicating the reliability of the results and the validity of the analytical procedures.

The presence of differential performance in subtests can be explained by the difference in item performance across groups, but this does not always imply bias. Bias occurs when differential performance (DIF) systematically leads to an overestimation or underestimation of ability in a particular group, which the results did not demonstrate [22]. Therefore, the presence of DIF in this limited number of items of the WISC-V Arabic version for the Jordanian context is a statistical phenomenon [9]. Differential performance does not necessarily indicate superiority or deficiency, but rather reflects a difference in cognitive performance patterns, which may have educational, clinical, or developmental implications [23]. This result is consistent with Adepoju *et al.* [24], who indicated that 145 items in the WISC scale worked differently, while 125 items showed no bias. It also aligns with Ozberk and Koc [25], who showed that some items exhibit DIF based on sex and others based on socioeconomic status, while Giofrè *et al.* [26] clarified that the scale WISC shows sex variation and the presence of some unequal tasks (subtests), such as “block design” and “coding” tasks, which showed a difference in performance between males and females. The findings can be explained by the presence of several factors contributing to differential performance, such as cultural, biological, and environmental factors related to sex, which play a significant role in performance on the scale [42].

Nisbett *et al.* [43] showed only slight differences between males and females result on the WISC scale in most cognitive dimensions, such as verbal comprehension and working memory, indicating that sex alone does not explain these differences and that other surrounding factors must be considered to understand the full picture. Giofrè *et al.* [44] indicated that females outperformed males in tasks requiring attentional control, while males outperformed females in some tasks requiring visuospatial processing.

The results of the study indicate important practical implications for educational decision-making processes in Jordanian schools, as differences in the performance of certain items across groups may affect the accuracy of estimating students’ abilities in intelligence tests, which in turn is reflected in educational classification decisions. In special education programs, this type of variability may lead to an underestimation of some students’ abilities and an increased likelihood of diagnosing them with learning difficulties or placing them in programs that do not reflect their true abilities. Conversely, in gifted programs, it may contribute to either an unfair inflation or reduction in performance estimates, which may affect the accuracy of student nomination and eligibility. Accordingly, teachers may benefit from the results of DIF analysis by becoming more aware that certain test items may not measure the same ability with equal fairness across all groups. It is therefore recommended not to rely on a single test or a single item in student assessment, but rather to use multiple sources of evaluation, such as classroom observation, practical performance, and continuous assessment. This also supports a more cautious interpretation of scores in important educational decisions, thereby reducing the likelihood of bias.

Similarly, school psychologists can use these findings to improve the accuracy of the clinical interpretation of test results by avoiding exclusive reliance on standardized scores and instead integrating multiple sources of information, such as behavioral interviews and clinical observations. This contributes to improving educational classification decisions and reducing errors in placing students in special education or gifted programs, thereby enhancing fairness and accuracy in decision-making. Furthermore, these findings contribute to the development of assessment standards in Jordanian schools by supporting the review and standardization of tests to detect biased items and by strengthening procedures for validating reliability and fairness. In addition, incorporating the concepts of DIF analysis into school psychologists’ training programs

enhances their competence in interpreting tests and making more objective educational decisions. At the school practice level, reliance on multi-source assessment promotes educational fairness and reduces classification errors, ensuring equal opportunities for students.

4. CONCLUSION

This study aimed to examine DIF in the Arabic version of the WISC-V within the Jordanian context. The findings indicated that the majority of WISC-V items functioned equivalently across males and females, demonstrating a high level of fairness and consistency. With the exception of a small number of items that showed weak or inconsistent differential performance, most subtests exhibited no evidence of DIF. These results suggest that observed differences in performance between sexes are more likely attributable to variations in confidence levels or problem-solving strategies rather than bias in test design. Accordingly, the WISC-V can be considered a fair and unbiased measure of intelligence for both males and females. To further enhance the validity and fairness of the scale, the study recommends the use of multiple statistical methods for detecting DIF. Additionally, it emphasizes the importance of training psychologists to interpret test results cautiously, taking into account the potential presence of sex-related differences in specific items.

FUNDING INFORMATION

Authors state no funding involved.

AUTHOR CONTRIBUTIONS STATEMENT

This journal uses the Contributor Roles Taxonomy (CRediT) to recognize individual author contributions, reduce authorship disputes, and facilitate collaboration.

Name of Author	C	M	So	Va	Fo	I	R	D	O	E	Vi	Su	P	Fu
Nizar Allabadi	✓		✓	✓	✓	✓	✓	✓		✓				
Mohammad Al-Karamneh	✓	✓	✓	✓		✓	✓		✓	✓	✓			
Jehad M. Al-Anati		✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓		
Laith Jumah		✓	✓		✓		✓		✓	✓				
Omar Alorani					✓	✓		✓				✓		

C : **C**onceptualization
M : **M**ethodology
So : **So**ftware
Va : **Va**lidation
Fo : **Fo**rmal analysis
I : **I**nvestigation
R : **R**esources
D : **D**ata Curation
O : Writing - **O**riginal Draft
E : Writing - Review & **E**diting
Vi : **Vi**sualization
Su : **Su**pervision
P : **P**roject administration
Fu : **Fu**nding acquisition

CONFLICT OF INTEREST STATEMENT

Authors state no conflict of interest.

INFORMED CONSENT

We have obtained informed consent from all individuals included in this study.

ETHICAL APPROVAL

The research related to human use has been complied with all the relevant national regulations and institutional policies in accordance with the tenets of the Helsinki Declaration and has been approved by the authors' institutional review board.

DATA AVAILABILITY

The information that backs the conclusions of this research can be obtained by contacting the corresponding author, [MA-K]. This data includes details that may jeopardize the confidentiality of the research subjects, which is why it is not accessible to the public because of specific limitations.

REFERENCES

[1] S. Urbina, *Essentials of psychological testing*, 2nd ed. Hoboken, NJ, USA: Wiley, 2014.

[2] American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, *Standards for educational and psychological testing*. Washington, DC, USA: American Educational Research Association, 2014.

[3] R. J. Sternberg, *Rethinking intelligence: Schools that develop the mind*. Thousand Oaks, CA, USA: Corwin Press, 2020.

[4] S. K. Benson, "Contextualizing inclusion policy: Views from Jordanian special education teachers," *Educational Research for Policy and Practice*, vol. 22, pp. 499–515, 2023, doi: 10.1007/s10671-023-09344-5.

[5] L. R. Holden and G. J. Tanenbaum, "Modern assessments of intelligence must be fair and equitable," *Journal of Intelligence*, vol. 11, no. 6, p. 126, 2023, doi: 10.3390/jintelligence11060126.

[6] D. Wechsler, *Wechsler Intelligence Scale for Children-Fifth Edition (WISC-V)*. San Antonio, TX, USA: Pearson, 2014.

[7] G. L. Canivez, M. W. Watkins, and S. C. Dombrowski, "Structural validity of the Wechsler Intelligence Scale for Children–Fifth Edition: Confirmatory factor analyses with 16 primary and secondary subtests," *Psychological Assessment*, vol. 29, no. 4, pp. 458–472, 2017, doi: 10.1037/pas0000358.

[8] J. M. Sattler, *Assessment of children: Cognitive foundations and applications*, 6th ed. San Diego, USA: Sattler Publisher, 2016.

[9] D. P. Flanagan and E. M. McDonough, *Contemporary intellectual assessment*: *Theories, tests, and issues*. New York, NY, USA: The Guilford Press, 2018.

[10] A. S. Kaufman, S. E. Raiford, and D. L. Coalson, *Intelligent testing with the WISC-V*. Hoboken, NJ, USA: John Wiley & Sons, 2020.

[11] S. Lioliou *et al.*, "Exploring the Prevalence of Learning Disabilities in a Community Sample of Children Using the Greek Wechsler Intelligence Scale for Children (WISC-V GR)," *International Journal of Environmental Research and Public Health*, vol. 22, no. 3, p. 377, 2025, doi: 10.3390/ijerph22030377.

[12] J. Rosenqvist, S. Slama, and A. Haavisto, "Generalizability of the Swedish WISC-V to the Finland-Swedish minority–the FinSwed study," *Child Neuropsychology*, vol. 31, no. 1, pp. 53–79, 2024, doi: 10.1080/09297049.2024.2331277.

[13] Pearson, *WISC-V Efficacy Research Report*. 2017. Accessed: May 14, 2026. [Online]. Available: https://www.pearson.com/content/dam/one-dot-com/one-dot-com/global/Files/efficacy-and-research/reports/WISC-V-Research-Report-v5.pdf

[14] N. El-Zahhar and Y. Meleka, *Wechsler Intelligence Scale for Children – Fourth Edition (Arabic Adaptation and Standardization Manual)*. Cairo: Anglo-Egyptian Bookshop,2010.

[15] D. C. Geary, "Now you see them, and now you don't: An evolutionarily informed model of environmental influences on human sex differences," *Neuroscience & Biobehavioral Reviews*, vol. 125, pp. 26–32, 2021, doi: 10.1016/j.neubiorev.2021.02.020.

[16] N. M. Else-Quest, J. S. Hyde, and M. C. Linn, "Cross-national patterns of gender differences in mathematics: A meta-analysis," *Psychological Bulletin*, vol. 136, no. 1, pp. 103–127, 2010, doi: 10.1037/a0018053.

[17] D. Voyer and S. Voyer, "Gender differences in scholastic achievement: A meta-analysis," *Psychological Bulletin*, vol. 140, no. 4, pp. 1174–1204, 2014, doi: 10.1037/a0036620.

[18] D. C. Geary, M. K. Hoard, and L. Nugent, "Boys' visuospatial abilities compensate for their relatively poor in-class attentive behavior in learning mathematics," *Journal of Experimental Child Psychology*, vol. 211, p. 105222, 2021, doi: 10.1016/j.jecp.2021.105222.

[19] R. Lynn and P. Irwing, "Sex differences in means and variability on the progressive matrices in university students: A meta-analysis," *British Journal of Psychology*, vol. 96, no. 4, pp. 505–524, 2005, doi: 10.1348/000712605X53542.

[20] P. W. Holland and H. Wainer, *Differential item functioning*. New York, NY: Routledge, 2018.

[21] G. Camilli and L. A. Shepard, *Methods for identifying biased test items*, 3rd ed. Thousand Oaks, CA, USA: SAGE Publications, 2017.

[22] B. D. Zumbo, "*Validity and fairness in testing: Differential item functioning and differential test functioning*," in *Handbook of Test Development*, 2nd ed., S. M. Downing and T. M. Haladyna, Eds., New York, NY, USA: Routledge, 2016, pp. 490–512.

[23] K. S. McGrew, "*The Cattell-Horn-Carroll theory of cognitive abilities*," in Contemporary intellectual assessment, 4th ed., D. P. Flanagan and P. L. Harrison, Eds. New York, NY, USA: The Guilford Press, 2019, pp. 61–100.

[24] O. A. Adepoju, I. K. Dauda, and A. K. Lawal, "Differential Item Functioning (DIF) Of The Wechsler Preschool and Primary Scale of Intelligence Administered on Nigerian Children," *Iconic Research and Engineering Journals*, vol. 8, no. 11, pp. 1511–1520, 2025.

[25] E. U. Ozberk and N. Koc, "Gender and Socioeconomic Status DIF on The WISC-IV Turkish Form Items: A Comparison of DIF Detection Techniques," (in Turkish), *Journal of Measurement and Evaluation in Education and Psychology*, vol. 8, no. 1, pp. 112–117, 2017, doi: 10.21031/epod.287577.

[26] D. Giofrè, E. Toffalini, L. Esposito, and C. Cornoldi, "Gender differences in the Wechsler Intelligence Scale for Children in a large group of Italian children with attention deficit hyperactivity disorder," *Journal of Intelligence*, vol. 11, no. 9, p. 178, 2023, doi: 10.3390/jintelligence11090178.

[27] J. E. Helms, "Culture, race, and ethnicity in psychological assessment," in *Cambridge Handbook of Intelligence*, 2nd ed., R. J. Sternberg and S. B. Kaufman, Eds., Cambridge, UK: Cambridge University Press, 2018, pp. 558–583.

[28] S. Chaimongkol, "Modeling Differential Item Functioning (DIF) Using Multilevel Logistic Regression Models: A Bayesian Perspective," Ph.D. dissertation, University of Florida, Florida, United States, 2005.

[29] M. D. Hidalgo and J. A. López-Pina, "Differential Item Functioning Detection and Effect Size: A Comparison Between Logistic Regression and Mantel-Haenszel Procedures," *Educational and Psychological Measurement*, vol. 64, no. 6, pp. 903–915, 2004, doi: 10.1177/0013164403261769.

[30] J. Al-Anati, "The Differential Performance of the Toni-4 Non-Verbal Intelligence Test Scores According to the Sex of the Student, the Family Income, and The Educational Level Of The Mother," (in Arabic), *Dirasat*, no. 60, pp. 60–72, 2017, doi: 10.34118/0136-000-060-005.

[31] D. P. Flanagan, S. O. Ortiz, and V. C. Alfonso, *Essentials of cross-battery assessment*, 3rd ed. Hoboken, NJ, USA: Wiley, 2017.

[32] S. Abu-Drei, "Psychometric Properties of the Jordanian Version of the Wechsler Intelligence Scale–Fourth Edition in Sign Language for Deaf Individuals Aged 6–16.11 Years," (in Arabic), *Journal of Applied Science - Humanities Series*, vol. 27, no. 2, pp. 1–31, 2021, [Online]. Available: http://search.mandumah.com/Record/1106768.

[33] A. Al-Hroub, "Use of the Jordanian WISC-III for twice exceptional identification," *International Journal for Talent Development and Creativity*, vol. 8, no. 1, pp. 121–144, 2021, doi: 10.7202/1076752ar.

[34] E. G. Ababneh, E. K. Duku, C. Reid-Westoby, A. Gaskin, and M. Janus, "Psychometric properties and factor structure of the Early Development Instrument in a sample of Jordanian children," *BMC Psychology*, vol. 10, no. 1, p. 316, 2022, doi: 10.1186/s40359-022-01014-0.
[35] S. Al-Anzi, "The Differential Functioning of Raven's Progressive Matrices Test in Light of Item Response Theory According to the Variables of Gender and Specialization," (in Arabic), *North Journal of Humanities*, vol. 9, no. 1, pp. 415–434, 2024, [Online]. Available: http://search.mandumah.com/Record/1525273.
[36] D. Hammad, "Detecting Gender Related Differential Item Functioning Raven Standard Progressive Matrices and its Effect on Saudi Samples Cognitive Responses," (in Arabic), *Journal of the Faculty of Education*, vol. 36, no. 111, pp. 1–35, 2021, doi: 10.21608/sec.2021.167210.
[37] W. Al-Wabel and S. Abd Rabbuh, "Differential Item Functioning of the Triarchic Intelligence Based on Gender and Academic Major among Qassim University Students," (in Arabic), *Journal of Arts for Psychological and Educational Studies*, vol. 5, no. 1, pp. 59–102, 2023, [Online]. Available: http://search.mandumah.com/Record/1366936.
[38] T. Al-Khursheh, "Differential Item Functioning of the Test of Mathematical Abilities "TOMA-3" According to Gender and its Effect on the Accuracy of Estimating the Person's Ability and Information Function According to Item Response Theory "IRT"," (in Arabic), *Journal of Qualitative Education Research*, no. 77, pp. 403–423, 2023, [Online]. Available: http://search.mandumah.com/Record/1492560.
[39] N. Hemedan and L. Maharmeh, "The psychometric properties of the Jordanian version of the Childhood Autism Rating Scale Second Edition (CARS2-HF)," *Dirasat: Human and Social Sciences*, vol. 49, no. 2, pp. 561–572, doi: 10.35516/hum.v49i2.1814.
[40] P. F. Merenda, "A guide to the proper use of factor analysis in the conduct and reporting of research: Pitfalls to avoid," *Measurement and Evaluation in Counseling and Development*, vol. 30, no. 3, pp. 156–164, 1997, doi: 10.1080/07481756.1997.12068936.
[41] M. D. Reckase, "Unifactor latent trait models applied to multifactor tests: Results and implications," *Journal of Educational and Behavioral Statistics*, vol. 4, no. 3, pp. 207–230, 1979, doi: 10.3102/10769986004003207.
[42] L. Esposito and D. Giofrè, "Sex differences in intelligence on the WISC: A meta-analysis on children with specific learning disabilities," *Journal of Intelligence*, vol. 13, no. 2, pp. 1–17, 2025, doi: 10.3390/jintelligence13020018.
[43] R. E. Nisbett *et al.*, "Intelligence: New findings and theoretical developments," *American Psychologist*, vol. 67, no. 2, pp. 130–159, 2012, doi: 10.1037/a0026699.
[44] D. Giofrè, K. Allen, E. Toffalini, and S. Caviola, "The impasse on gender differences in intelligence: A meta-analysis on WISC batteries," *Educational Psychology Review*, vol. 34, no. 4, pp. 2543–2568, 2022, doi: 10.1007/s10648-022-09705-1.

BIOGRAPHIES OF AUTHORS



Nizar Allabadi is an academic and researcher in Jordan specializing in Educational Psychology. He is an assistant professor at Department of Educational Psychology, Faculty of Educational Sciences, The University of Jordan. His research interests include: item response theory (IRT) and the psychological aspects of technology use. He can be contacted at email: n.labadi@ju.edu.jo.



Mohammad Al-Karamneh is a Ph.D. holder in Educational Psychology/ Measurement and Evaluation. He is an assistant professor at Amman Arab University since 2023, in the Department of Administration, Curriculum and Instruction. His research interests include: educational psychology, measurement and evaluation, and the construction of psychological and educational tests. He can be contacted at email: m.karamneh@aau.edu.jo.



Jehad M. Al-Anati is a Ph.D. holder in Educational Psychology/Measurement and Evaluation. He is an associated professor at Department of Educational Psychology, The University of Jordan. His research interests include: measurement and evaluation, item banks and e-testing, classroom assessment, program evaluation, and impact studies. He can be contacted at email: j.alanati@ju.edu.jo.



Laith Jumah is an assistant professor in Educational Psychology at Department of Educational Psychology, Faculty of Educational Sciences, The University of Jordan. He earned his Ph.D. in Educational Psychology and Learning Systems from Florida State University. His research interests include: epistemic beliefs and cognition, self-regulated learning, educational psychology, and learning systems. He can be contacted at email: l.jumah@ju.edu.jo.



Omar Alorani is an assistant professor specializing in Psychological Counseling at Department of Counseling and Special Education, Faculty of Educational Sciences, The University of Jordan. His research interests include: psychological well-being, counseling interventions, psychotherapy, educational sciences, and social factors affecting students and specific populations in Jordan. He can be contacted at email: O.alorani@ju.edu.jo.