Library Progress International
Vol.44 No. 3, Jul-Dec 2024: P. 18913-18928

Print version ISSN 0970 1052
Online version ISSN 2320 317X

Original Article

Available online at www.bpasjournals.com

Performance of the Generalized Mantel-Haenszel Method in Graded-Response Items Using Empirical Data

Majed M. Aljodeh

Department of Education and Psychology, University of Tabuk, Tabuk, Saudi Arabia
Majed Mahmoud Aljodeh https://orcid.org/0009-0003-1530-930X
majed_jodeh@hotmail.com

How to cite this article: Majed M. Aljodeh (2024) Performance of the Generalized Mantel-Haenszel Method in Graded-Response Items Using Empirical Data. *Library Progress International,* 44(3), 18913-18928.

ABSTRACT

The current study aimed to examine the performance behavior of the Generalized Mantel-Haenszel (GMHDIF) method in detecting differential item functioning (DIF) in graded response based on the gender variable while altering the sample size. It used real data obtained from the responses of a sample of Tabuk University students on a scale to evaluate the quality of academic advising. Six sample size levels were used: 250, 500, 1000, 1500, 2000, and 2500. The study concluded that the differential items detected by the method in small sample sizes may not appear as such in larger samples. Conversely, items that do not seem differential in small samples may show differential functioning in larger samples. Some items appeared to be different across all sample sizes, including the smaller ones. Therefore, the effectiveness and ability of the method to detect DIF items increases with larger sample sizes. Items expected to have a high level of differential functioning are easier for the method to detect, even in smaller sample sizes.

***Keywords**:* Performance, Graded-Response, GMHM, Empirical Data

Introduction

Traditionally, researchers study the differential behavior of test and scale items, which is referred to as Differential Item Functioning (DIF). Differential Item Functioning (DIF) is defined as the difference in the probability of individuals responding to one of responses for an item in particular (e.g., the correct answer) based on different groups of the same ability respondents in the trait being measured. (Holland & Thyer, 1988).

In the context of dichotomous response items, differential item functioning can be easily interpreted as the difference in the probability of the same ability individuals in the measured trait answering correctly across different respondent groups. However, the situation becomes more complex when interpreting differential item functioning for graded-response items or items with graded responses. Researchers have distinguished between two approaches to interpreting DIF for such items. Some view DIF as a directional difference in responses to the item among the same ability individuals in the traits being measured across all response options, varying between different groups of respondents. Others view DIF in graded response items as the difference in the probability of choosing one of the response options for the item, particularly among individuals with the same ability in the traits being measured, varying based on the respondent group (Penfield et al., 2009; Penfield, 2010).

Regardless of the type of items and the mechanisms for interpreting their differential functioning, the existence of DIF itself is concerning and negatively impacts the properties of measurement tools. It may favor one group over another in performance despite both having the same ability in the traits being measured when DIF is present. It also affects measurement equivalence across different respondent groups. Detecting differential functioning in scale items is essential for ensuring the fairness of measurement tools and enhancing the accuracy of predictions regarding the traits being measured (Jafari et al., 2013; Thissen, 2001).

Differential Item Functioning (DIF) can negatively impact scale conclusions and individual classification. Several studies have investigated this in achievement tests, like the TOEFL (Test of English

Language and Scholastic Assessment Test) and the MELAB (Michigan English Language Assessment Battery) (Park, 2008; Wagner,2004; Eom, 2008; Vahid et al., 2011).

In the field of test construction and the development of measurement tools, the importance of providing evidence for the validity of these tools has increased, including indicators of Differential Item Functioning (DIF). The latest version of the standards for developing educational and psychological tests has considered DIF an important indicator of the measure's construct validity. (American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 2014)

Researchers have distinguished between two types of differential item functioning: uniform differential item functioning (UNDIF) and non-uniform differential item functioning (NUNDIF), focusing on the relationship between group membership and level of ability. According to Ackerman (1992), the item characteristic curves (ICCs) for the scale are parallel for UNDIF and non-parallel for NUNDIF (Ackerman, 1992; Mellenbergh, 1989; Millsap & Everson, 1993; Narayanon & Swaminathan, 1996).

Evaluating Differential Item Functioning (DIF) of items in Graded Response Using the Generalized Mantel-Haenszel Method

In the context of graded response items, which are the focus of this study, one of the common and proven methods for detecting differential item functioning (DIF) in such items is the General Mantel-Haenszel (GMH) method. The Mantel-Haenszel method is widely used to examine DIF in dichotomous response items and was subsequently applied to graded response items, proving to be an effective tool for this purpose (Penfield, 2001).

The original Mantel-Haenszel method for detecting differential item functioning (DIF) is an extension of the Chi-Square (χ^2) test for significance. This method compares the responses of individuals between two groups: the first is called the Reference Group, and the second is referred to as the Focal Group. The comparison is made across different levels of individuals' abilities. When testing the null hypothesis regarding the absence of DIF, the MH χ^2 value appears in equations 1 and 2.

$$\chi^2 = \frac{\left\{\left|\sum_{j=1}^{k}[A_j - E(A_j)]\right| - 0.5\right\}^2}{\sum_{j=1}^{k} Var(A_j)} \quad (1)$$

$$Var(A_j) = \frac{nR_j nF_j m_{1j} m_{0j}}{T_j^2 (T_j - 1)} \quad (2)$$

Where:

K: The number of ability levels for the trait being measured.

J: Ability level number j.

A_j: The number of correct responses observed in the reference group for the item under differential functioning study.

$E(A_j)$: The expected number of correct responses in the reference group for the item under differential functioning study.

Var (A_j): The variance of the correct responses observed in the reference group for the item under differential functioning study.

nR_j: The number of respondents to the item under differential functioning study in the reference group at ability level j.

nF_j: The number of respondents to the item under differential functioning study in the focal group at ability level j.

Tj: The total sample size at ability level j.

m_{1j}: The number of correct responses to the item under differential functioning study at ability level j.

m_{0j}: The number of incorrect responses to the item under differential functioning study at ability level j. (Penfield, 2001)

The ability levels are typically divided based on the total responses of individuals to all items on the scale.

Equations 1 & 2 can be extended to include graded-response items. The data obtained from individuals' responses to the graded-response item are arranged in a contingency table (matrix) of rank 2xTxK. Where T represents the number of response categories for the graded-response item, and K represents the number of ability levels of the respondents, with ability typically represented by the total scores of individuals on the scale. At each ability level K, there is a 2xT contingency table.

Let's assume that m_1, m_2, m_3, … m_T represent the scores assigned to each item response. The contingency

table for ability level k would look as shown in Table 1 bel

Table 1

Contingency Table for Ability Level k for an Item in a Scale

	The scores assigned to the item responses				
	m_1	m_2	m_3	m_T	Total
Reference group	nR1k	nR2k	nR3k	nR_Tk	nR+k
Focal group	nF1k	nF2k	nF3k	nF_Tk	nF+k
Total	n+1k	n+2k	n+3k	n+Tk	n++k

Where:

nR_Tk: Represents the number of individuals in the reference group at ability level k who received the score assigned to the item response, which is m_t

nF_Tk: Represents the number of individuals in the focal group at ability level k who received the score assigned to the item response, which is m_t.

"+": Represents the sum across a row or column in the contingency matrix in Table 1.

"++": Represents the total sum of the rows or columns in the contingency matrix in Table 1.

The General Mantel-Haenszel method treats the response categories of an item as nominal-level data. Given the multiple responses and categories at different ability levels for each item, the data are handled in the form of matrices. The generalized form of equation 1 for multiple-response items is expanded and given as follows:

$$Q_{GMH}=[\sum D_k - \sum E(D_k)]^{\mathbf{I}}[\sum V(D_k)]^{-1}[\sum D_k - \sum E(D_k)] \qquad (3)$$

Where:

$$D_k = \begin{bmatrix} nR1k \\ \text{nR2k} \\ \text{nR3k} \\ \dots \\ \text{nR(T}-1)\text{k} \end{bmatrix} N_k = \begin{bmatrix} \text{n} + 1\text{k} \\ n + 2k \\ \text{n} + 3\text{k} \\ \dots \\ n + (T-1)k \end{bmatrix} E(D_k) = (nR + k)N_k\,/(n + +k)$$

$$d_{Nk} = \begin{bmatrix} nR1k & 0 & 0 & 0 \\ 0 & nR2k & 0 & \dots \\ \dots & 0 & \dots & 0 \\ 0 & 0 & 0 & nR(T-1)k \end{bmatrix}$$

$$V(D_k) = (nR + k)(nF + k)\left[\frac{(n + +k)d_{NK} - N_k N'_k}{(n + +k)^2 - \big((n + +k) - 1\big)}\right]$$

Researchers believe that the General Mantel-Haenszel method is an effective approach for detecting differential item functioning (DIF) of items in multiple-response items. It can also be applied to dichotomous response items and is more advanced and comprehensive than the traditional Mantel-Haenszel method (Fidalgo & Madeira, 2008).

Detecting differential item functioning (DIF) is an issue related to scale items, such as a test. Certain variables, such as sample size, test length, individual characteristics, item format, and others may influence it. Hence, this study was conducted to examine the ability of the General Mantel-Haenszel method to detect differential item functioning based on gender under varying sample sizes. The study uses real data, specifically the responses of a sample of students from the University of Tabuk, on a scale designed to evaluate the university's academic advising quality.

Upon reviewing the theoretical literature related to the impact of various conditions and factors on methods for detecting differential item functioning (DIF), specifically the General Mantel-Haenszel method for graded-response items, we find a scarcity of such studies(Daoud et al., 2024)(Allahham, Sharabati, Al-Sager, et al., 2024). Most of the research has focused on the Mantel-Haenszel method for detecting DIF in dichotomous items. In this regard, the results of the study by Ya-Hui & Wen-Chung (2005) compared three methods for detecting differential item functioning (DIF) of items in polytomous response items using simulated data under various conditions. These conditions included the proportion and magnitude of DIF in the items, as well as the effect on test power and the identification of Type I errors. The value of DIF has a higher and more significant effect than the proportion of its presence in the items. It also showed that the General Mantel-Haenszel method had the weakest test power for detecting DIF compared to the other methods used in the study.

The impact of sample size on methods for detecting DIF, most of the research has focused on dichotomous response items. Notably, the majority of these studies used simulated data rather than real experimental data(Almustafa et al., 2023). This might explain the lack of research on the effect of sample sizes on DIF detection methods for polytomous response items, as generating data that accurately simulates real data for this type of item is challenging(Al Mawahreh et al., 2024). In this context, we mention some studies that have examined the effectiveness of DIF detection methods under varying sample size conditions(Allahham et al., 2023).

In a simulation study conducted by Kabasakala et al. (2014), three methods were compared: the Mantel-Haenszel method, the item response theory likelihood ratio (IRT-LR) method, and the SIBTEST method. The comparison was made under the influence of sample sizes, ability distribution, and test length, examining the impact on Type I error and test power(Alkhazaleh et al., 2023). The study revealed a clear effect of sample sizes, particularly on the Mantel-Haenszel method. Significant changes were observed in the Type I error rate and test power. As the sample size increased, the Type I error rate decreased, and the test power reached its highest levels(Awawdeh et al., 2024).

The study by Aljoudeh (2021), which used simulated data for dichotomous response items, examined the performance of the IRT-LR (Item Response Theory Likelihood Ratio) method under various conditions of sample size and the magnitude of DIF in the items. Four sample size levels were generated: 250, 500, 750, and 1000, representing responses to 40 dichotomous response items. Some items were intentionally designed to exhibit DIF at different levels, under two conditions: uniform DIF and non-uniform DIF. The study concluded that with a sample size of 1000 individuals, the IRT-LR method demonstrated a high level of performance in detecting items with uniform DIF across all levels(A. A. et al., 2024). However, its performance decreased for items with non-uniform DIF across all levels and sample sizes examined.

Finch (2005) conducted a comparison among several methods for detecting differential item functioning (DIF), including the Multiple Indicators Multiple Causes (MIMIC) model, the Mantel-Haenszel (MH) method, SIBTEST, and the Item Response Theory Likelihood Ratio (IRT-LR) method. This comparison was made in light of test length, sample size, ability distribution of the examinees, and the level and magnitude of DIF in the items. The results indicated that methods for detecting differential item functioning (DIF) were more effective with longer tests, larger sample sizes, or in cases of dichotomous-parameter items(Demirbag et al., 2006). The effectiveness decreased when the number of items was lower, especially when there were only 20 polytomous-parameter items(Atieh Ali, Sharabati, Allahham, et al., 2024).

The study by Woods (2009) demonstrated the superiority of the Multiple Indicators Multiple Causes (MIMIC) model over Item Response Theory models in detecting differential item functioning (DIF), particularly when using small sample sizes and dichotomous items. This conclusion was drawn from a study of simulated data with varying sample sizes and test lengths(Shehadeh et al., 2024).

In a study conducted by Ugurlu and Atar (2020), the researchers compared two methods for detecting differential item functioning (DIF) in dichotomous items. The first method was the Multiple Indicators Multiple Causes (MIMIC) model, and the second was the logistic regression method. This comparison was based on generated data under varying sample size conditions, the impact on Type I error, and the proportion of items exhibiting DIF(Allahham, , et al., 2024).

In a study conducted by Ugurlu and Atar (2020), the researchers compared two methods for detecting differential item functioning (DIF) of items in dichotomous items. The first method was the Multiple Indicators Multiple Causes (MIMIC) model, and the second was the logistic regression method. This comparison was based on simulated data under varying sample size conditions, the impact on Type I error, and the proportion of items exhibiting DIF(Atieh Ali, Sharabati, Alqurashi, et al., 2024). The study concluded that the proportion of items exhibiting differential item functioning (DIF) changed from 20% to 40% when the sample size was increased from 2000 to 4000. The impact of sample size had a clear effect on the effectiveness of both methods in detecting DIF, as evidenced by the decreased rates of Type I error.

Arikan et al. (2016) conducted a study comparing four methods for detecting differential item functioning (DIF) in items: MIMIC, SIBTEST, logistic regression (LR), and Mantel-Haenszel (MH). This comparison was carried out under varying sample size conditions, with subsamples of 300, 600, 1000, 1200, and 2000 selected from a total dataset of 340,000. The study found that with larger sample sizes, such as 2000 or more, the results were more effective in detecting differential item functioning (DIF), and the methods were more consistent. Unlike

smaller sample sizes, All four methods could identify the same items exhibiting DIF, which showed discrepancies in the items detected as exhibiting DIF across each method.

Recently, Elyan and Al jodeh (2024) conducted a study aimed at examining the effectiveness of the likelihood ratio method for detecting differential item functioning (DIF) in dichotomous response items based on gender under varying sample sizes and test lengths(Alrjoub et al., 2021). This was done using real data obtained from the results of tenth-grade students in Jordan who participated in the 2018 PISA international assessment. The researchers selected three sample size levels: 342, 200, and 100, as well as three test length levels: 30, 20, and 10 items. The researchers found that the effectiveness of the likelihood ratio method of Mantel-Haenszel in detecting differential item functioning (DIF) increases with larger sample sizes while keeping the test length at a constant level. Conversely, its effectiveness decreases with longer test lengths when the sample size is held constant at a certain level(A. A. A. Sharabati, Awawdeh, et al., 2024). The study concluded that using a large sample size along with a short test length maximizes the effectiveness of the method(Atta et al., 2023).

It is noted from the review of some studies addressing differential item functioning (DIF) that most of the studies presented utilized simulated data, with the exception of the study by Arikan et al. (2016) and the study by Elyan and Al jodeh (2024). Overall, most of the studies addressed dichotomous response items(Morshed et al., 2024). Through examining the conditions of varying sample sizes and their impact on the effectiveness of the methods in detecting differential item functioning (DIF of items), it was observed that increasing the sample size plays a significant role and has a clear effect on the effectiveness of those methods. In line with these studies, the current research aims to evaluate the general Mantel-Haenszel method's ability to detect differential item functioning, specifically in graded response items. This focus sets this study apart from others, especially since it utilizes real data, enhancing its findings' realism(*Bataineh,etl AI.*, 2023).

Study Problem and its Importance

The tools of measurement, psychological and personality scales, and tests vary according to the different purposes for which they were developed(Sharabati et al., 2023). They play a significant role in various fields, such as classifying individuals, their admission to universities and educational institutions, distributing them across different specializations, as well as assessing their satisfaction with various services and activities, their quality, and their inclinations and attitudes(A. ali et al., 2024). This information informs important decisions that aid in diagnosing areas of strength and weakness, ultimately contributing to enhancing academic quality and the functioning of educational institutions(Aljabari et al., 2024).

In order for these scales to fulfill their intended purpose, they must possess the necessary characteristics that ensure validity, reliability, accuracy, and fairness among respondents. One of the key and prominent issues that has recently become a focus of attention for researchers and those involved in scale development is item bias and differential item functioning (DIF). Furthermore, the latest edition of the Standards for Educational and Psychological Testing has recognized the study of differential item functioning (DIF) as evidence and an indicator of the construct validity of test items, as previously mentioned.

Differential item functioning (DIF) based on variables such as gender, race, specialization, language, culture, and others in measurement scales may be influenced by various factors, just as other characteristics like validity, reliability, and item parameters are affected. One of the factors that may play a significant role in influencing differential item functioning (DIF) is the variation in sample sizes. This has been observed in the studies discussed earlier. As an appropriate sample size may be necessary to enhance the ability of methods used to detect DIF. The presence of differential item functioning (DIF) in the items of a scale based on a certain variable negatively impacts the scale and its characteristics. Bias in favor of one group over another in the scale's items raises doubts about the validity of the results. It makes it difficult to trust the decisions based on these results. Hence, studying the factors that influence the performance of methods for detecting differential item functioning (DIF) in scale items based on a certain variable—especially sample size—is important. The findings of such a study can provide valuable recommendations and guidelines for researchers regarding the optimal sample size that suits the method being used for accurate results. Notably, since this study used real experimental data, it adds a realistic dimension to its outcomes.

Given that researchers widely use the Generalized Mantel-Haenszel (GMH) method to detect differential item functioning (DIF), especially for polytomous response items, studying this method under varying sample sizes holds significant value in the theoretical literature related to detecting DIF in scale items.

Study Objective

This study aims to investigate the effectiveness of the Generalized Mantel-Haenszel (GMH) method in detecting items with differential item functioning (DIF) based on the gender variable at different sample sizes. Specifically, the study will address the following main question:

- What is the effectiveness of the Generalized Mantel-Haenszel (GMH) method in detecting differential item functioning (DIF) in Graded-response scale items based on gender when varying the sample size?

Methodology and Procedures

• **Study Tool:** This study used a scale designed to evaluate the quality of academic advising, which was developed by the researchers in the 2024 study by Alshehri and Aljoudeh (2024). The final version of the scale consists of 25 Graded-response items, based on a five-point Likert scale, distributed across five main axes. The original version of the scale is attached in Appendix 1.

• **Study Data:** A sample of responses was obtained from 2,760 students, consisting of 1,687 female and 1,073 male students, from all undergraduate students at the University of Tabuk in 2023.

Given that the primary goal of the current study is to evaluate the effectiveness of the Generalized Mantel-Haenszel (GMH) method in detecting differential item functioning (DIF) for polytomous response items under varying sample sizes, rather than assessing the quality of academic advising, the sample was divided into six levels randomly, taking into account that the numbers were close based on the variable of gender, so that it would serve this purpose. The following table 2 illustrates these levels:

Table 2

Sample Size Levels in the Study and their Distribution by Gender Variable

Level	Sample size	Male No.	Female No.
First	250	118	132
Second	500	221	279
Third	1000	475	525
Forth	1500	837	837
Fifth	2000	870	1130
Sixth	2500	1073	1427

It is noted that all male students were included in the sixth level of the sample size, representing the largest sample size of males who responded to the study tool.

Using the SPSS program, a master file containing all student responses was created, which was then divided into sub-files based on the different sample size levels. These files were converted into Comma Delimited (*.csv) data files to be used in the GMHDIF program developed by Fidalgo (2010). The GMHDIF program operates in two stages, and in both stages, the respondents' scores on all items of the scale are collected to estimate the students' levels on the scale or their abilities. The number of levels or categories is equal to the distribution range of total scores. Then, in the first stage, the program estimates the (Q_{GMH}) value for each item along with its statistical significance to identify items that demonstrate initial differential item functioning (DIF). Then, in the second stage, the program excludes the scores of students on the items that demonstrated initial differential item functioning (DIF) from their total scores (i.e., the total scores for determining students' levels are not calculated by adding the students' scores on these items) from the first stage. The QGMH value is then re-estimated for all items (including those that demonstrated initial differential functioning) to identify the items that exhibit final differential functioning.

Study Results

The GMHDIF program was run six times, which corresponds to the six sample size levels adopted in this study. Below is a presentation of the results for each level, showing the QGMH values, their statistical significance, and item numbers for both the first and second stages.

Results for the First Sample Size Level (250):

Table 3 presents the results of DIF items analysis, and the direction of DIF for these items at the first sample size level, which is 250.

Table 3

Results of DIF Items at Sample Size Level 250

Item No.	Stage	QMH	Sig.	Mean for male	Mean for female	Direction of DIF
1	1	QMH = 0.5591	p = 0.4546	3.23	3.64	female
	2	QMH = 5.1033	p = 0.0239*			
3	1	QMH = 7.4134	p = 0.0065*	3.39	3.75	female
	2	QMH = 5.9526	p = 0.0147*			

From Table 3, it is evident that two items exhibited final differential item functioning (DIF) based on gender: Items 1 and 3.

Results for the Second Sample Size Level of 500

Table 4 presents the results of DIF items analysis, and the direction of DIF for these items at the second sample size level of 500.

Table 4

Results of DIF Items Analysis at a Sample Size of 500

Item No	Stage	QMH	Sig.	Mean for male	Mean for female	Direction of DIF
1	1	QMH = 5.8608	p = 0.0155*	3.26	3.52	female
	2	QMH = 9.7865	p = 0.0018*			
3	1	QMH = 3.9848	p = 0.0459*	3.56	3.62	\ female
	2	QMH = 4.4321	p = 0.0353*			

It can be observed from Table 4 that items 1 and 3 demonstrated differential functioning based on gender. These are the same two items that exhibited differential functioning at the sample size level of 250.

Results for the Third Sample Size Level of 1000

Table 5 presents the results of DIF items analysis, and the direction of DIF for these items at the third sample size level of 1000.

Table 5

Results of DIF Analysis at a Sample Size of 1000

Item No	Stage	QMH	Sig.	Mean for male	Mean for female	Direction of DIF
1	1	QMH = 8.7419	p = 0.0031*	3.35	3.58	female
	2	QMH = 8.7419	p = 0.0031*			
3	1	QMH = 1.1377	p = 0.0286*	3.53	3.65	female
	2	QMH = 1.1436	p = 0.0284*			
4	1	QMH = 3.0404	p = 0.0812	3.31	3.42	female
	2	QMH = 5.8757	p = 0.0154*			
8	1	QMH = 2.1207	p = 0.0453*	2.96	2.76	male
	2	QMH = 2.8154	p = 0.0134*			

It is evident from Table 5 that four items demonstrated differential functioning based on gender. In addition to items 1 and 3, which were identified in the first and second levels, items 4 and 8 were detected, which the method did not identify in the initial two sample sizes.

Results for the Fourth Sample Size Level of 1500

Table 6 presents the results of DIF items analysis, and the direction of DIF for these items at the fourth sample size level of 1500.

Table 6

Results of DIF Items Analysis at a Sample Size of 1500

Item No	Stage	QMH	Sig.	Mean for male	Mean for female	Direction of DIF
1	1	QMH = 28.571	p = 0.0000*	3.26	3.52	female
	2	QMH = 35.360	p = 0.0000*			

4	1	QMH = 7.8844	p = 0.0050*	3.26	3.35	female
	2	QMH = 6.6029	p = 0.0102*			
8	1	QMH = 5.0085	p = 0.0252*	2.92	2.71	male
	2	QMH = 8.5116	p = 0.0035*			
9	1	QMH = 2.6673	p = 0.1024	3.12	2.99	male
	2	QMH = 5.2036	p = 0.0225*			
12	1	QMH = 1.2531	p = 0.2630	3.03	2.88	male
	2	QMH = 6.2048	p = 0.0127*			
17	1	QMH = 7.5229	p = 0.0061*	3.10	3.25	female
	2	QMH = 7.2319	p = 0.0072*			
24	1	QMH = 8.4386	p = 0.0037*	3.26	3.1	male
	2	QMH = 7.0643	p = 0.0079*			
25	1	QMH = 9.1131	p = 0.0025*	3.26	3.13	male
	2	QMH = 9.9469	p = 0.0016*			

Table 6 clearly shows that eight items demonstrated differential functioning based on gender: 1, 4, 8, 9, 12, 17, 24, and 25.

Results for the Fifth Sample Size Level of 2000

Table 7 presents the results of DIF items analysis, and the direction of DIF for these items at the fifth sample size of 2000.

Table 7

Results of DIF Items Analysis at a Sample Size of 2000

Item No	Stage	QMH	Sig.	Mean for male	Mean for female	Direction of DIF
1	1	QMH = 32.694	p = 0.0000*	3.34	3.51	female
	2	QMH = 29.127	p = 0.0000*			
4	1	QMH = 11.804	p = 0.0006*	3.34	3.42	female
	2	QMH = 10.167	p = 0.0014*			
8	1	QMH = 6.2938	p = 0.0121*	3.00	3.24	female
	2	QMH = 9.1548	p = 0.0025*			
9	1	QMH = 4.3014	p = 0.0381*	3.21	3.00	male
	2	QMH = 4.5477	p = 0.0330*			
12	1	QMH = 6.3810	p = 0.0115*	3.11	2.88	male
	2	QMH = 7.7994	p = 0.0052*			
17	1	QMH = 7.7129	p = 0.0055*	3.27	3.13	male
	2	QMH = 8.6495	p = 0.0033*			
20	1	QMH = 7.3852	p = 0.0066*	3.47	3.32	male
	2	QMH = 5.2298	p = 0.0222*			
24	1	QMH = 4.7667	p = 0.0290*	3.32	3.15	male
	2	QMH = 4.5752	p = 0.0324*			
25	1	QMH = 10.045	p = 0.0015*	3.36	3.15	male
	2	QMH = 10.108	p = 0.0015*			

It can be seen from Table 7 that nine items also demonstrated differential functioning based on gender. However, there is a difference with one item: item 5 did not appear at this level, while item 20 appeared at it, but did not appear in the fourth level. The items are 1, 4, 8, 9, 12, 17, 20, 24, and 25.

Results for the Sixth Sample Size Level of 2500

Table 8 presents the results of DIF items analysis, and the direction of DIF for these items at the sixth sample size level of 2500.

Table 8

Results of DIF Items Analysis at a Sample Size of 2500

Item No	Stage	QMH	Sig.	Mean for male	Mean for female	Direction of DIF
1	1	QMH = 39.862	p = 0.0000*	3.34	3.54	female
	2	QMH = 43.808	p = 0.0000*			
4	1	QMH = 10.788	p = 0.0010*	3.23	3.36	female
	2	QMH = 9.3219	p = 0.0023*			
5	1	QMH = 4.8197	p = 0.0281*	3.38	3.42	female
	2	QMH = 4.8291	p = 0.0280*			
8	1	QMH = 5.9645	p = 0.0146*	2.98	3.18	female
	2	QMH = 7.1078	p = 0.0077*			
9	1	QMH = 6.1145	p = 0.0134*	3.20	3.03	male
	2	QMH = 4.8640	p = 0.0274*			
12	1	QMH = 5.5035	p = 0.0190*	3.09	2.92	male
	2	QMH = 6.1238	p = 0.0133*			
17	1	QMH = 5.2028	p = 0.0226*	3.28	3.15	male
	2	QMH = 6.6153	p = 0.0101*			
20	1	QMH = 5.0948	p = 0.0240*	3.45	3.34	male
	2	QMH = 4.2028	p = 0.0404*			
24	1	QMH = 5.5042	p = 0.0190*	3.31	3.18	male
	2	QMH = 4.5284	p = 0.0333*			
25	1	QMH = 9.4753	p = 0.0021*	3.34	3.18	male
	2	QMH = 7.8102	p = 0.0052*			

It can be observed from Table 8 that ten items demonstrated differential functioning based on gender, with the reappearance of item 5 and the continued presence of item 20 at this level. The items are: 1, 4, 5, 8, 9, 12, 17, 20, 24, and 25.

To better understand, discuss, and draw conclusions from the results, a summary table was created to consolidate the study's findings during the program's six implementations, representing six sample size levels. This summary is presented in Table 9.

Table 9

Summary of DIF Item Numbers and Counts at Each Sample Size Level

Level	Sample size	Items numbers with DIF	The number of items with DIF
First	250	1,3	2
Second	500	1,3	2
Third	1000	1,3,4,8	4
Fourth	1500	1, 8, 9, 12, 17, 24, 25	8
Fifth	2000	1, 4, 8, 9, 12, 17, 20, 24, 25	9
Sixth	2500	1, 4, 5, 8, 9, 12, 17, 20, 24, 25	10

Discussion

The current study aimed to examine the performance of the Generalized Mantel-Haenszel method in detecting differential item functioning for polytomous response items based on gender across various sample size levels. By closely examining the results of the differential item functioning analysis in the previous tables, especially focusing on Table 9, which summarizes the results of the program's execution across all selected sample size conditions. We can observe that increasing the sample size enhances the method's effectiveness in detecting items that exhibit differential functioning based on gender. We notice that the number of items with differential functioning detected ranged from 2 to 10. This means that some items that did not show differential functioning in smaller sample sizes were identified as such when the sample size increased. This result aligns with the findings

of studies on differential item functioning under varying sample size conditions for dichotomous response items (Aljodudeh, 2021; Arikan et al., 2016; Kabasakala et al., 2014; Elyan & Al jodeh, 2024; Ugurlu & Atar, 2020).

By carefully examining the differential item numbers in Table 9, we find, for instance, that item 1 appeared as a differential item across all sample size levels, even in the smaller ones. This suggests that such items may have a high level of differential functioning, causing them to be identified as differential even in small sample sizes. When examining item 3, it appeared in the first three sample size levels but disappeared in the three larger ones.

This result may be misleading, as items that appear to demonstrate differential functioning in smaller sample sizes may not exhibit the same behavior when the sample size increases. Conversely, some items that did not show differential functioning in smaller sample sizes were identified as having differential functioning in larger sample sizes, such as items 4, 9, 13, 17, 24, and 25. What best explains this phenomenon is that such items may have a low level of differential functioning, making it difficult for the method to detect them in smaller sample sizes. They require an increased sample size for the method to identify them. For example, item 20 only appeared in the two largest sample size levels. Likewise, item 5 needed 2,500 responses to be recognized as a differential item, which may also be attributed to these items' very low level of differential functioning. These items only appear differential when using large sample sizes.

The presence of certain items with differential functioning across all sample size levels, such as item 1, prompts the researcher to reconsider such items—either by removing them from the scale or by modifying them to align with the variable under study regarding differential functioning. This result seems reasonable and is generally agreed upon, as it is considered a differential item even in lower sample size levels.

As for the items that appear differential in smaller sample sizes but not in larger ones. As well as those that do not show differential functioning in small sample sizes but do in larger ones. This opens the door for a scientific and logical discussion of this phenomenon.

What we can rely on, whether in studies of differential functioning or others, is that increasing the sample size enhances the validity and accuracy of the results. It also increases the power of statistical tests and reduces errors, particularly Type I errors. This has been indicated in some studies, such as the study by Kabasakala et al. (2014). From this, we see that increasing confidence in the ability and effectiveness of the Generalized Mantel-Haenszel method to detect DIF of the polytomous response type is achieved using larger sample sizes, as using smaller sizes may lead to misleading results. This could be seen as a negative aspect of this method, as increasing the sample size places a greater burden and effort on researchers. This reinforces the findings of the study by Ya-Hui and Wen-Chung (2005), which concluded that the Generalized Mantel-Haenszel method is the weakest among the other methods compared in the study.

Conclusions

The findings of this study lead us to conclude that items that appear to have differential functioning in small sample sizes may not exhibit the same behavior in large sample sizes when using the Generalized Mantel-Haenszel method. Conversely, items that do not show differential functioning in small sample sizes may be identified as differential when larger sample sizes are used. The Generalized Mantel-Haenszel method is significantly influenced by sample size in detecting DIF for polytomous response items. Its ability to do so increases with larger sample sizes, and heavily depends on the magnitude of DIF. The higher the magnitude of DIF, the easier for the method to detect it, even under small sample conditions. To further clarify the characteristics of this method, it is essential to study other factors that may influence it, such as the length of the scale items, the magnitude of DIF, and its type.

References

- Ackerman, T. A. (1992). A didactic explanation of item bias, item impact, and item validity from a multidimensional perspective. *Journal of Educational Measurement*, *29*(1), 67–91. https://doi.org/10.1111/j.1745-3984.1992.tb00368.x
- Aljoudeh, M. (2021). Item response theory likelihood ratio test performance for deducting DIF items in different levels in samples sizes and different levels of DIF items. *Vidyabharati International Interdisciplinary Research Journal*, *13*(1), 392–399.
- Alshehri, Y., & Aljoudeh, M. (2024). Quality of academic advising from the perspective of students at the University of Tabuk for the 2022 academic year. *Humanities and Educational Sciences Journal, (37)*,

422–451,. https://doi.org/10.55074/hesj.vi37.992

- American Educational Research Association, American Psychological Association, and& National Council on Measurement in Education. (2014). *Standards for Eeducational and Ppsychological Ttesting*. Washington, DC: American Educational Research Association.
- Vahid A., Christine C. & Lee O.Aryadoust, V., Goh, C. C. M., & Kim, L. O. (2011). An Iinvestigation of Ddifferential Iitem Ffunctioning in the MELAB Llistening Ttest. *Language Assessment Quarterly*, *8*(4), 361–385. DOI:https://doi.org/10.1080/15434303.2011.628632
- Elyan, R. M., & Al jodehAl jodeh, M. M., M. M. (2024). The effectiveness of Mantel –Haenszel log odds ratio method in detecting differential item functioning across different sample sizes and test lengths using real data analysis. *Dirasat: Educational Sciences*, *51*(3), 37–46. https://doi.org/10.35516/edu.v51i3.6755
- Eom, M. (2008). Underlying factors of MELAB listening construct. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *6*, 77–94.
- Fidalgo, A. M. (2010). *GMHDIF: User''s Mmanual*. Oviedo, Spain: Universidad de Oviedo.
- Fidalgo, A. M., & Madeira, J. M. (2008). Generalized Mantel-–Haenszel Mmethods for Ddifferential Iitem Ffunctioning Ddetection. *Educational and Psychological Measurement*, *68*(6), 940-958. https://doi.org/10.1177/0013164408315265
- Finch, H. (2005). The MIMIC Mmodel as a Mmethod for Ddetecting DIF: Comparison Wwith Mantel-–Haenszel, SIBTEST, and the IRT Llikelihood Rratio. *Applied Psychological Measurement*, *29*(4), 278–295. https://doi.org/10.1177/0146621605275728
- Jafari, P., Bagheri, Z., Hashemi, S. Z., & Shalileh, K. (2013). Assessing whether parents and children perceive the meaning of the items in the PedsQLTM 4.0 quality of life instrument consistently: aA differential item functioning analysis. *Global Journal of Health Science*, *5*(5), 80 – 88. https://doi.org/10.5539/gjhs.v5n5p80.
- Kabasakala, K., Arsan, N., Gok, B., & Kelecooglu, H. (2014). Comparing Performances (Type I error and Power) of IRT Likelihood Ratio SIBTEST and Mantel-–Haenszel Methods in the Determination of Differential Item Functioning. *Educational Sciences: Theory &and Practice*, *14*(6), 2186–2193. DOI: https://doi.org/10.12738/estp.2014.6.2165
- Mellenbergh, G. J. (1989). Item bias and item response theory. *International jJournal of eEducational rResearch*, *13*(2), 127–143. https://doi.org/10.1016/0883-0355(89)90002-5.
- Millsap, R. E., & Everson, H. T. (1993). Methodology review: Statistical approaches for assessing measurement bias. *Applied pPsychological mMeasurement*, *17*(4), 297–334. https://doi.org/10.1177/014662169301700401
- Narayanon, P., & Swaminathan, H. (1996). Identification of items that show nonuniform DIF. *Applied pPsychological mMeasurement*, *20*(3), 257-–274. https://doi.org/10.1177/014662169602000306
- Park, G.-P. (2008). Differential Iitem Ffunctioning on an English Llistening Ttest across Ggender. *TESOL Quarterly*, *42*(1), 115–123. https://doi.org/10.1002/j.1545-7249.2008.tb00212.x
- Penfield, R. D. (2001). Assessing differential item functioning among multiple groups: aA comparison of three Mantel-–Haenszel procedures. *Appl.ied Meas.urement in Educ.ation*, *14*(3), 235–259. doi: https://doi.org/10.1207/ S15324818AME1403_3
- Su, Y. -H., & Wang, W. -C. (2005). Efficiency of the Mmantel, Ggeneralized Mantel–Haenszel, and Llogistic Ddiscriminant Ffunction Aanalysis Mmethods in Ddetecting Ddifferential Iitem Ffunctioning for Ppolytomous Iitems. *Applied Measurement in Education*, *18*(4), 313–350. https://doi.org/10.1207/s15324818ame1804_1
- Thissen, D. (2001). *IRTLRDIF v.2.0b: Software for the computation of the statistics involved in Item Response Theory Likelihood-Ratio tests for Differential Item Functioning. L.L*. Thurstone Psychometric Laboratory, University of North Carolina, Chapel Hill, NC.
- Uğurlu, S., & Atar, B. (2020). Performances of MIMIC and logistic regression procedures in detecting DIF. Journal of Measurement and Evaluation in Education and Psychology. *Eğitimde ve Psikolojide Ölçme ve Değerlendirme Dergisi*, *11*(1), 480-4911–12. https://doi.org/10.21031/epod.531509.
- Uğurlu, S., Atar, B., & Akin Arikan, Ç. (2015). MIMIC, SIBTEST, Lojistik Regresyon ve Mantel-

Haenszel Yöntemleriyle Gerçekleştirilen DMF ve Yanlılık Çalışması. *Hacettepe University Journal of Education,* 1–1. https://doi.org/10.16986/HUJE.2015014226

- Wagner, A. (2004). A construct validation study of the extended listening sections of the ECRE and MELAB. *Spaan Fellow Working Papers in Second or Foreign Language Assessment*, *2*, 1–23.
- Woods, C. M. (2009). Evaluation of MIMIC-model methods for DIF testing with comparison to two-group analysis. *Multivariate Behavioral Research*, *44*(1), 1–27. DOI:https://doi.org/10.1080/00273170802620121
- Al Mawahreh, M. A. L., Awawdeh, H. Z., Bani Ahmad, A. Y. A., Almajali, W. I., Ali, A. A. A., & Allahham, M. (2024). How Does Digital Marketing Influence Consumer Behavior? Examining the Mediating Role of Digital Entrepreneurship in the Healthcare and Pharmaceuticals Sector. *Library Progress International*, *44*(3), 5858–5877.
- Aljabari, M., Althuwaini, S., Bouguerra, A., Sharabati, A. A. A., Allahham, M., & Allan, M. (2024). The impact of digital marketing strategies on innovation: The mediating role of AI: A critical study of SMEs in the KSA market. *International Journal of Data and Network Science*, *8*(4), 2029–2036. https://doi.org/10.5267/j.ijdns.2024.7.006
- Alkhazaleh, A., Assaf, A., Shehada, M., Almustafa, E., & Allahham, M. (2023). Analysis of the Impact of Fintech Firms' Lending on the Expansion of Service Base Companies in Jordan. *Information Sciences Letters*, *12*(8), 2891–2902. https://doi.org/10.18576/ISL/120837
- Allahham, M., Sharabati, A. A. A., Al-Sager, M., Sabra, S., Awartani, L., & Khraim, A. S. L. (2024). Supply chain risks in the age of big data and artificial intelligence: The role of risk alert tools and managerial apprehensions. *Uncertain Supply Chain Management*, *12*(1), 399–406. https://doi.org/10.5267/j.uscm.2023.9.012
- Allahham, M., Sharabati, A. A. A., Almazaydeh, L., Sha-Latony, Q. M., Frangieh, R. H., & Al-Anati, G. M. (2024). The impact of fintech-based eco-friendly incentives in improving sustainable environmental performance: A mediating-moderating model. *International Journal of Data and Network Science*, *8*(1), 415–430. https://doi.org/10.5267/j.ijdns.2023.9.013
- Allahham, M., Sharabati, A. A. A., Hatamlah, H., Ahmad, A. Y. B., Sabra, S., & Daoud, M. K. (2023). Big Data Analytics and AI for Green Supply Chain Integration and Sustainability in Hospitals. *WSEAS Transactions on Environment and Development*, *19*, 1218–1230. https://doi.org/10.37394/232015.2023.19.111
- Almustafa, E., Assaf, A., & Allahham, M. (2023). IMPLEMENTATION OF ARTIFICIAL INTELLIGENCE FOR FINANCIAL 1 INTRODUCTION Artificial intelligence (AI) is the intelligence of robots , not humans . Most academic textbooks characterize artificial intelligence as studying " intelligent agents ." These agent. *RGSA – Revista de Gestão Social e Ambiental*, *17*(9), 1–17.
- Alrjoub, A. M. S., Almomani, S. N., Al-Hosban, A. A., & Allahham, M. I. (2021). the Impact of Financial Performance on Earnings Management Practice Behavior (an Empirical Study on Financial Companies in Jordan). *Academy of Strategic Management Journal*, *20*(Special Issue 2), 1–15.
- Atieh Ali, A. A., Sharabati, A. A. A., Allahham, M., & Nasereddin, A. Y. (2024). The Relationship between Supply Chain Resilience and Digital Supply Chain and the Impact on Sustainability: Supply Chain Dynamism as a Moderator. *Sustainability (Switzerland)* , *16*(7), 1–20. https://doi.org/10.3390/su16073082
- Atieh Ali, A. A., Sharabati, A. A., Alqurashi, D. R., Shkeer, A. S., & Allahham, M. (2024). The impact of artificial intelligence and supply chain collaboration on supply chain resilience: Mediating the effects of information sharing. *Uncertain Supply Chain Management*, *12*, 1801–1812. https://doi.org/10.5267/j.uscm.2024.3.002
- Atta, A. A. B., Ahmad, A. Y. A. B., Allahham, M. I., Sisodia, D. R., Singh, R. R., & Maginmani, U. H. (2023). Application of Machine Learning and Blockchain Technology in Improving Supply Chain Financial Risk Management. *Proceedings of International Conference on Contemporary Computing and Informatics, IC3I 2023*, 2199–2205. https://doi.org/10.1109/IC3I59117.2023.10397935
- Awawdeh, H. Z., Al Mawahreh, M. A. L., Allahham, M., Almajali, W. I., Ali, A. A. A., & Bani Ahmad, A. Y. A. (2024). The Impact of Digital Marketing on Building Consumer Confidence the Role Mediating

of Information sharing and AI: An Empirical Study of the Telecommunications Sector in Jordan. Library Progress International, 44(3), 5844–5857.

- *Y. A. B. Ahmad, N. Verma, N. M. Sarhan, E. M. Awwad, A. Arora and V. O. Nyangaresi, "An IoT and Blockchain-Based Secure and Transparent Supply Chain Management Framework in Smart Cities Using Optimal Queue Model," in IEEE Access, vol. 12, pp. 51752-51771, 2024, doi:10.1109/ACCESS.2024.3376605*
- *Bani Ahmad, A. Y., Fraihat, B. A. M., Hamdan, M. N., Ayasrah, F. T. M., Alhawamdeh, M. M., & Al-Shakri, K. S. (2024). Examining the mediating role of organizational trust in the relationship between organizational learning and innovation performance: A study of information systems and computer science service firms.*
- *Bataineh, A. Q., Abu-AlSondos, I. A., Almazaydeh, L., El Mokdad, S. S., & Allahham, M. (2023). Enhancing natural language processing with machine learning for conversational AI. (2023). 2023.*
- *Daoud, M. K., Sharabati, A. A., Samarah, T., Alqurashi, D., & Alfityani, A. (2024). Optimizing online visibility : A comprehensive study on effective SEO strategies and their impact on website ranking. 8(7).*
- *Demirbag, M., Koh, S. C. L., Tatoglu, E., & Zaim, S. (2006). TQM and market orientation's impact on SMEs' performance. Industrial Management and Data Systems, 106(8), 1206–1228. https://doi.org/10.1108/02635570610710836*
- *Morshed, A., Maali, B., Ramadan, A., Ashal, N., Zoubi, M., & Allahham, M. (2024). The impact of supply chain finance on financial sustainability in Jordanian SMEs. Uncertain Supply Chain Management, 12(4), 2767–2776. https://doi.org/10.5267/j.uscm.2024.4.025*
- *Sharabati, A. A. A., Awawdeh, H. Z., Sabra, S., Shehadeh, H. K., Allahham, M., & Ali, A. (2024). The role of artificial intelligence on digital supply chain in industrial companies mediating effect of operational efficiency. Uncertain Supply Chain Management, 12(3), 1867–1878. https://doi.org/10.5267/j.uscm.2024.2.016*
- *Sharabati, A. A. A., Rehman, S. U., Malik, M. H., Sabra, S., Al-Sager, M., & Allahham, M. (2024). Is AI biased? evidence from FinTech-based innovation in supply chain management companies? International Journal of Data and Network Science, 8(3), 1839–1852. https://doi.org/10.5267/j.ijdns.2024.2.005*
- *Selvasundaram, K., Jayaraman, S., Chinthamani, S. A. M., Nethravathi, K., Ahmad, A. Y. B., & Ravichand, M. (2024). Evaluating the Use of Blockchain in Property Management for Security and Transparency. In Recent Technological Advances in Engineering and Management (pp. 193-197). CRC Press.*
- *Ramadan, A., Maali, B., Morshed, A., Baker, A. A. R., Dahbour, S., & Ahmad, A. B. (2024). Optimizing working capital management strategies for enhanced profitability in the UK furniture industry: Evidence and implications. Journal of Infrastructure, Policy and Development, 8(9), 6302.*
- *Fouzdar, A. S., Yamini, S., Biswas, R., Jindal, G., Ahmad, A. Y. B., & Dawar, R. (2024). Considering the Use of Blockchain for Supply Chain Authentication Management in a Secure and Transparent Way. In Recent Technological Advances in Engineering and Management (pp. 259-264). CRC Press.*
- *Yahiya, A., & Ahmad, B. (2024). Automated debt recovery systems: Harnessing AI for enhanced performance. Journal of Infrastructure, Policy and Development, 8(7), 4893.*
- *Feng, Y., Ahmad, S. F., Chen, W., Al-Razgan, M., Awwad, E. M., Ayassrah, A. Y. B. A., & Chi, F. (2024). Design, analysis, and environmental assessment of an innovative municipal solid waste-based multigeneration system integrating LNG cold utilization and seawater desalination. Desalination, 117848.*
- *Zhang, L., Ahmad, S. F., Cui, Z., Al Razgan, M., Awwad, E. M., Ayassrah, A. Y. B. A., & Shi, K. (2024). Energy, exergy, hermoeconomic analysis of a novel multi-generation system based on geothermal, kalina, double effect absorption chiller, and LNG regasification. Desalination, 117830.*
- *Iqbal, S., Tian, H., Muneer, S., Tripathi, A., & Ahmad, A. Y. B. (2024). Mineral resource rents, fintech technological innovation, digital transformation, and environmental quality in BRI countries: An insight using panel NL-ARDL. Resources Policy, 93, 105074.*
- *Wu, J., Ahmad, S. F., Ali, Y. A., Al-Razgan, M., Awwad, E. M., & Ayassrah, A. Y. B. A. (2024). Investigating the role of green behavior and perceived benefits in shaping green car buying behavior*

with environmental awareness as a moderator. Heliyon, 10(9).

- *Zhao, T., Ahmad, S. F., Agrawal, M. K., Ahmad, A. Y. A. B., Ghfar, A. A., Valsalan, P., ... & Gao, X. (2024). Design and thermo-enviro-economic analyses of a novel thermal design process for a CCHP-desalination application using LNG regasification integrated with a gas turbine power plant. Energy, 295, 131003.*
- *Geetha, B. T., Gnanaprasuna, E., Ahmad, A. Y. B., Rai, S. K., Rana, P., & Kapila, N. (2024, March). Novel Metrics Introduced to Quantify the Level of Circularity in Business Models Enabled by Open Innovation. In 2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies (pp. 1-6). IEEE.*
- *Geetha, B. T., Kafila, K., Ram, S. T., Narkhede, A. P., Ahmad, A. Y. B., & Tiwari, M. (2024, March). Creating Resilient Digital Asset Management Frameworks in Financial Operations Using Blockchain Technology. In 2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies (pp. 1-7). IEEE.*

- *Naved, M., Kole, I. B., Bhope, A., Gautam, C. S., Ahmad, A. Y. B., & Lourens, M. (2024, March). Managing Financial Operations in the Blockchain Revolution to Enhance Precision and Safety. In 2024 International Conference on Trends in Quantum Computing and Emerging Business Technologies (pp. 1-6). IEEE.*
- *Sharabati, A. A., Ali, A., Ali, A., Allahham, M. I., Hussein, A. A., Alheet, A. F., & Mohammad, A. S. (2024). The Impact of Digital Marketing on the Performance of SMEs : An Analytical Study in Light of Modern Digital Transformations. 1–25.*
- *Sharabati, A. A., Allahham, M., Yahiya, A., Ahmad, B., & Sabra, S. (2023). EFFECTS OF ARTIFICIAL INTEGRATION AND BIG DATA ANALYSIS ON ECONOMIC VIABILITY OF SOLAR MICROGRIDS : MEDIATING ROLE OF COST BENEFIT ANALYSIS. 6(3), 360–379.*
- *Peiran Liang, Yulu Guo, Sohaib Tahir Chauhdary, Manoj Kumar Agrawal, Sayed Fayaz Ahmad, Ahmad ,Yahiya Ahmad Bani Ahmad, Ahmad A. Ifseisi, Tiancheng Ji,2024" Sustainable development and multi-aspect analysis of a novel polygeneration system using biogas upgrading and LNG regasification ,processes, producing power, heating, fresh water and liquid CO2" Process Safety and Environmental ,Protection*
- *,Peiran Liang, Yulu Guo, Tirumala Uday Kumar Nutakki, Manoj Kumar Agrawal, Taseer Muhammad ,Sayed Fayaz Ahmad, Ahmad Yahiya Ahmad Bani Ahmad, Muxing Qin 2024. " Comprehensive assessment and sustainability improvement of a natural gas power plant utilizing an environmentally friendly combined cooling heating and power-desalination arrangement" Journal of Cleaner, ,Production,Volume 436,,140387*
- *Rumman, G., Alkhazali, A., Barnat, S., Alzoubi, S., AlZagheer, H., Dalbouh, M., ... & Darawsheh, S. (2024). The contemporary management accounting practices adoption in the public industry: Evidence from Jordan. International Journal of Data and Network Science, 8(2), 1237-1246.*
- *William, P., Ahmad, A. Y. B., Deepak, A., Gupta, R., Bajaj, K. K., & Deshmukh, R. (2024). Sustainable Implementation of Artificial Intelligence Based Decision Support System for Irrigation Projects in the Development of Rural Settlements. International Journal of Intelligent Systems and Applications in Engineering, 12(3s), 48-56.*
- *Yahiya Ahmad Bani Ahmad (Ayassrah), Ahmad; Ahmad Mahmoud Bani Atta, Anas; Ali Alawawdeh, Hanan; Abdallah Aljundi, Nawaf; Morshed, Amer; and Amin Dahbour, Saleh (2023) "The Effect of System Quality and User Quality of Information Technology on Internal Audit Effectiveness in Jordan, And the Moderating Effect of Management Support," Applied Mathematics & Information Sciences: Vol. 17: Iss. 5, Article 12. DOI: https://dx.doi.org/10.18576/amis/170512*
- *Yahiya, A., & Ahmad, B. (2024). Automated debt recovery systems: Harnessing AI for enhanced performance. Journal of Infrastructure, Policy and Development, 8(7), 4893.*
- *Zhan, Y., Ahmad, S. F., Irshad, M., Al-Razgan, M., Awwad, E. M., Ali, Y. A., & Ayassrah, A. Y. B. A. (2024). Investigating the role of Cybersecurity's perceived threats in the adoption of health information systems. Heliyon, 10(1).*

- Appendix 1

Academic Advising Quality Evaluation Scale Used in the Study

Item No. in the scale	Domaine	Item
1	**Academic Policies and Regulations**	The college directs the student to the academic advisor at the beginning of the semester.
2		The college assigns an academic advisor to me and provides all the necessary information
3		The academic advisor is available during office hours and is designated for academic advising when needed.
4		My academic advisor provides me with useful information about the courses and their relationships.
5		My academic advisor helps me understand the academic policies and regulations.
6	**Developing the Student' s Academic and Professional Skills**	My academic advisor discusses my career plan with me for the future after graduation.
7		My academic advisor guides me toward developing the necessary skills and competencies for success in the job market after graduation.
8		My academic advisor discusses with me how to choose suitable job locations after graduation.
9		My academic advisor is concerned with my education and ensuring I gain a distinguished learning experience.
10		The academic advisor monitors my progress throughout the semester.
11	**Developing the Student' s Personal and Leadership Skills**	My academic advisor is concerned with the development of my personality as a student, leader, and member of the community
12		My academic advisor helps me build a plan to achieve my educational goals.
13		My academic advisor assists me in solving academic problems (such as falling behind or underperforming in courses(
14		My academic advisor helps me find appropriate solutions when needed.
16		My academic advisor guides my thinking and encourages me to succeed in my academic field.
18		My academic advisor helps me organize my study schedule appropriately.
15	**The Relationship Between the Academic Advisor and the Student**	My relationship with my academic advisor is friendly and good.
19		I can reach my academic advisor even without a prior appointment.
21		After each meeting with my academic advisor, I feel motivated to achieve my educational goals and overcome the challenges I face.
22		My academic advisor gives me the freedom to express and discuss my feelings.
17	**Skills and Competencies of the Academic Advisor**	My academic advisor helps me choose the appropriate courses and schedules.
20		My academic advisor gives me enough time during my designated session.
23		The academic advisor answers all my questions, and if they cannot, they direct me to the appropriate person or office.
24		My academic advisor guides me to various resource locations within the university.

25	My academic advisor is sufficiently knowledgeable about my study plan and graduation requirements.