Journal of the College of Education for Women

مجلة كلية التربية للبنات

June 30, 2024 [Vol. 35(2)]

P-ISSN: 1680-8738; E-ISSN: 2663-547X

A Differential Performance in the Ability Difference to Employ Test Wiseness Strategies According to Contemporary Measurement Theory

Waleed Khalid Abdulkareem Baban

Department of Psychological and Educational Counseling, College of Education, Salahaddin University/ Erbil, Iraq

waleed.baban@su.edu.krd

https://doi.org/10.36231/coedw.v35i2.1744

Received: April 22, 2024; **Accepted**: May 17, 2024; **Published**: June 30, 2024

Abstract

The current research aims to utilize the Andrich Model, described as one of the Polytomous models, within the framework of Contemporary Measurement Theory. It focuses on differential performance based on the ability to employ test-wiseness strategies. To achieve this goal, the researcher relied on the Test-Wiseness Scale prepared by (Hamad 2010). A stratified random sample of (447) male and female students from the tenth, eleventh, and twelfth grades was selected. The assumptions of the Item Response Theory (IRT) were verified, including the one-dimensionality assumption. This was done through factor analysis of the test items using Principal Components Analysis (PCA) for individuals' responses to the test items. The Eigenvalue, explained variance, and cumulative explained variance for each factor were calculated. It was found that there was one factor with a meaningful interpretation for the scale, and through this assumption, the local independence assumption was also confirmed. To analyze the data of the scale items, the researcher used the Andrich Model and employed computer software (ConstructMap-4.6). The estimated item location values on the latent trait indicated that they ranged from (2.55) to (2.71) logits, with an average of (0.027) logits. This suggests that the scale covers a wide range of the measured trait, from low to high levels of ability. Furthermore, the standard error for the mean of the item difficulty estimates was (0.032), which is a low value close to zero. This indicates the accuracy of the item location estimates on the latent trait of wisdom.

Keywords: Differential Performance, Test Wiseness Strategies, Andrich Model, Contemporary Measurement Theory.

الأداء التفاضلي لاختلاف القدرة على توظيف استراتيجيات الحكمة الاختبارية وفق نظرية القياس المعاصرة

وليد خالد عبد الكريم بابان

قسم الارشاد النفسي والتربوي، كلية التربية، جامعة صلاح الدين/ أربيل، العراق

waleed.baban@su.edu.krd

https://doi.org/10.36231/coedw.v35i2.1744

تاريخ الاستلام: ٢٠٢٤/٤/٢٢، **تاريخ القبول:** ٢٠٢٤/٥/١٧، **تاريخ النشر الالكتروني:** ٢٠٢٤/٦/٣٠

المستخلص

يهدف البحث الحالي إلى الاستفادة من نموذج أندريش، بوصفه أحد نماذج الاستجابة المتعددة التدريج لنظرية القياس المعاصرة. حيث ركز الباحث على قياس الأداء التفاضلي استنادا للقدرة على استخدام استراتيجيات الحكمة الاختبارية. ولتحقيق هذا الهدف اعتمدت الباحث على مقياس الحكمة الاختبارية والمعد من قبل (حمد، ٢٠١٠). لذلك تم اختيار عينة عشوائية طبقية قوامها (٤٤٧) طالباً وطالبة من الصفوف العاشر والحادي عشر والثاني عشر الإعدادي، وتم التحقق من فرضيات نظرية الاستجابة للفقرة(IRT) ، بما في ذلك فرضية البعد الواحد، وذلك من خلال التحليل العاملي. لفقرات الاختبار باستخدام طريقة تحليل المكونات الرئيسية (PCA) لاستجابات الأفراد لفقرات الاختبار، وذلك بحساب قيمة الجذر الكامن ونسبة التباين المفسر،وكذلك التبـاين المفسر التراكمي لكل عامل من العوامل، ومن خلال هذا الافتراض تم تأكيد فرضية الاستقلال المحلي أيضاً. ولتحليل بيانات فقرات المقياس استخدم الباحث نموذج أندريش، وباستخدام برنامج الحاسوب (ConstructMap-4.6) حيث أشارت قيم موقع الفقرة المقدرة على السمة الكامنة إلى أنها تراوحت من (٢,٥٥) إلى (٢,٧١) لوغاريتم، بمتوسط (٠,٠٢٧) لوغاريتم. وهذا يشير إلى أن المقياس يغطي نطاقًا واسعًا من السمة المقاسة، من الأقل إلى الأعلى كما بلغ الخطأ المعياري لمتوسط تقديرات صعوبة الفقرة (٠,٠٣٢)، وهي قيمة منخفضة قريبة من الصفر، مما يدل على دقة تقديرات موقع الفقرة على سمة الحكمة الكامنة خلف الاستجابة للاختبار.

الكلمات المفتاحية: الأداء التفاضلي، استراتيجيات الحكمة الاختبارية، نموذج أندريش، نظرية القياس المعاصرة

1. Introduction

The introduction Measuring students' learning outcomes is highly critical and demands meticulous attention, given its substantial influence on subsequent assessment-related decisions. It directly affects choices concerning advancing to higher levels of education or the next academic phase and also relates to the student's requirement for additional preparation and learning. Consequently, it is closely intertwined with the evaluation procedure. If the measurement process contains flaws or errors, decisions founded on such inaccurate assessments will also be flawed.

Research Problem

One of the main concealed challenges that cause inaccuracies in measuring learning outcomes across different educational curricula and levels is the issue of random error arising from the adoption of test-wiseness tactics. Test-wiseness has been shown to contribute to the success of numerous students, even if they lack sufficient knowledge of the subject matter tailored for the test context. Conversely, the absence of test-wiseness results in lower levels of achievement among high-performing students who do not possess such strategies. Consequently, it becomes vital for many students to be acquainted with test-wiseness techniques in order to effectively employ their intellectual abilities during testing situations (Almaliki, 2010: 9).

The impact of this performance differential error among students within the same grade or academic level becomes even more pronounced, especially when the measurement process relies on comparing students against each other to select the best or most capable ones. This practice leads to the exclusion of highly competent students from the curriculum, in favor of those who achieve a lower level, not due to any deficiency in their academic abilities, but because of their lack of proficiency in employing test-wiseness strategies.

Certain students have been noticed to express their frustration over being unable to attain high exam grades despite their thorough preparation, whereas others manage to achieve high scores despite being less prepared. This indicates that the group with higher scores, but inadequate preparation, likely utilized test-wiseness strategies, resulting in an inherent random error in assessing students' performance levels. While this error is unavoidable and cannot be completely eliminated, it can be estimated (Howard, 2003: 62-63).

This clarifies the recent and noticeable surge in students' fascination with test-wiseness skills. Test-wiseness is seen as a suggested model to explain how individuals' exam scores serve as a factor in students' performance, even when their abilities are similar. Furthermore, the increasing volume of information and knowledge incorporated into textbooks to keep pace with the knowledge explosion has driven many students to embrace these strategies. By adopting test-wiseness techniques, they aim to manage the overwhelming amount of study material and extensive curriculum content, enabling them to navigate exams successfully and attain high grades.

Based on the foregoing, the researcher finds that we are facing an urgent problem that requires thorough investigation and exploration. Addressing this problem falls within the researcher's sphere of interest, prompting them to delve into it. Consequently, the researcher will conduct this study to answer the following two questions:

- What is the impact of the ability to employ test-wiseness strategies on differential performance, according to contemporary measurement theory and in accordance with Andrich's Graded Response Model?

- Investigating the differences in the use of test-wiseness strategies among individuals with low and high abilities.

Research Significance

In recent times, certain educators are promoting the adoption of test-wiseness strategies among students as a means of self-help for psychological and educational support, aiming to mitigate the prevalent issue of test anxiety experienced by many students. These strategies empower students to enhance their memory retention, mentally prepare for test scenarios, acquaint themselves with various question formats, practice effective answering techniques, and emphasize the significance of carefully reading and adhering to test instructions. Consequently, these approaches aid in reducing students' anxiety levels, ultimately leading to improved performance in examinations (Alzahrani, 2015: 221).

This highlights the extensive impact of test-wiseness strategies and their application by students during test situations. Students effectively utilize these strategies to identify the correct answers or formulate responses skillfully, even when they are uncertain about the correct answer. Consequently, these strategies assist them in achieving higher grades (Saleh & Obaid, 2020: 123).

Despite the encouragement of test-wiseness strategies, the crucial aspect of accuracy in the measurement process, which aims to evaluate students' knowledge of the curriculum content rather than their deductive thinking abilities or guessing skills, has been overlooked. Therefore, the current research holds significance in examining how the utilization of test-wiseness strategies impacts the success of a broad group of students, especially those in secondary school. These students are being prepared mentally and educationally for the transition to higher education, as the challenges in the education system are becoming more numerous and complex. This motivation drives the researcher to address the obstacles that hinder quality education and to take responsibility in diagnosing these challenges, aiming to contribute to the continual progress of the educational process in the right direction.

To ensure the objectivity of the assessment, researchers must not discriminate or show bias based on factors such as gender, race, or other characteristics. They should not favor any specific group over others in the evaluation process. All measurements should be impartial and based solely on the merits of the test items, considering the context of the assessment. (Kim & Cohen, 1994).

Consequently, it is essential to employ contemporary measurement approaches as experimental studies have demonstrated their capability to achieve the precision and impartiality sought in psychological and educational sciences. Identifying differential performance on psychological assessment items is vital as it tackles psychometric concerns. This identification can unveil potential biases in test items, which can impact the fairness, validity, and reliability of the tests (Salubayba, 2013).

The significance of the current research lies in its utilization of the contemporary measurement theory, specifically the Andrich Model, to unveil the dangers of certain non-educational practices, such as employing test-wiseness strategies and investigating their impact on learning outcomes. These practices can lead to differential performance among students, with some demonstrating the ability to employ these strategies effectively, while others exhibit lower proficiency in their use. This research aims to raise awareness among educators and stakeholders about the importance of cautious measurement of learning outcomes in educational assessments. By employing the Andrich Model as one of the polytomous models in contemporary measurement theory, this research aims to achieve this objective.

Research Objectives

1. Measuring differential performance in the ability to employ test-wiseness strategies, based on the graded response model of the contemporary measurement theory, Andrich Model.

2. Detecting the extent of variation in the utilization of test-wiseness strategies between individuals with lower and higher abilities.

2. Theoretical Framework

2.1 Key Words

2.1.1Differential Performance

Differential performance refers to the variations in test scores between two distinct groups, and it indicates the potential impact of test-wiseness strategies on the outcomes, showing how test-wiseness individuals might perform differently from non- test-wiseness individuals (Wood, 2009, p42).

Procedurally, it indicates that there are disparities in the utilization of test-wiseness strategies between individuals with high and low abilities.

2.1.2 Ability

The underlying capacity to manipulate and process the raw data upon which the performance tasks that measure the degree of skills and knowledge depend (Kazem, 1988, p56).

Procedural definition: It refers to the practical application of test-wiseness strategies.

2.1.3 Test-wiseness strategies

Hammad, (2010): Cognitive ability acquired through a set of skills to utilize the characteristics of the testing situation, which the examinee practices during the test to enhance their score (Hammad, 2010, p303).

Procedural definition**:** It is the threshold that distinguishes between students who possess a high level of this ability and those who have a low level of it. These students are identified using a scale (Hammad, 2010), which is employed by the researcher in the current study.

2.2 Item Differential Functioning (DIF)

Hambleton and Rogers (1995) indicate that an item is biased when the difference in the area under the Item Characteristic Curve (ICC) between different equivalent groups in terms of ability is statistically significant. This could include differences between genders or different ethnic groups. In other words, the likelihood of a correct response to the item varies for individuals within subgroups who possess the same level of ability. Crocker and Algina (Croker & Algina, 1986) argue that an item is biased if it remains invariant across different sources of variance at the same level of ability, despite variations in the groups to which individuals belong.

On the other hand, Embretson and Reise (Embretson & Reise 2000) suggest that an item is biased if it operates differently for one group compared to another. Camilli and Shephard propose that an item is biased if it is more difficult for one group compared to another group at the same ability level for the trait being measured (Shephard& Camilli 1994, p321).

The concept of differential item performance can be succinctly as the disparity in performance between two groups of individuals possessing equivalent levels of abilities but exhibiting contrasting responses to the same item. This discrepancy is observed intrinsically in factors such as ethnicity, culture, language, or gender within these two groups of individuals. The initial group is termed the "Reference Group," serving as a comparative baseline, while the subsequent group is denoted as the "Focal Group," undergoing experimental evaluation. (Hidalgo & Gomez-Benito, 2010).

2.3 Models Specific to Polytomous Responses

In these models, the responses are ordered, such as those obtained from surveys, assessment scales, and personality measures. They are divided into three models: the Graded Response Model (GRM), the Partial Credit Model (PCM), and the Rating Scale Model developed by Andrich. Among these, the Andrich model is most aligned with the current study's instrument, making it the researcher's choice. This model is suitable for data derived from rating scales (Masters & Wright, 1984, p536).

The researcher will adopt the Andrich Model due to its compatibility with the nature of data based on Likert scaling. It's worth mentioning that the Andrich Model is one of the Polytomous IRT Models and was developed by Andrich in 1988 to suit data derived from multi-response Likert scaling. The concept behind this model is that each item in the scale carries an overall affective load, and the model estimates this load for each item based on the probabilistic mathematical function that the model employs (Gruijter & Kamp, 2005, p101). In many cases, especially with Likert scales or similar types of rating scale formats, individuals are asked to respond to an item using a predetermined set of responses. The same set of response alternatives is then applied to all items in the test (Altaqi, 2005, p50).

This model dissects the difficulty level of the item (i.e., the threshold between two consecutive values, x and x-1) into two components: the first component represents the item difficulty (βi), and the second component represents the distance (j) from the difficulty level, denoted as tau (τj). This value remains constant across all items composing the scale (Embretson & Reise, 2000, p115). It also assumes an equal number of response categories for all items in the scale. When the number of response categories varies among different items, the estimation process is carried out for each group of items with the same number of categories. When comparing difficulty levels among various items becomes challenging, the derived ability does not get affected (Altaqi, 2013, p50).

The scaling of the Likert scale is accomplished by modeling the thresholds for different levels in the items, based on ordered response categories (such as: Poor, Fair, Good, Excellent), according to (m = 1) the number of categories (0, ..., m). Thus, the probability of selecting category (k) for item (i) can be expressed as shown in the following equation:

$$(\theta) = \frac{EXP[k(\theta - \beta_i) - \sum_{j=0}^{k} \tau_{i(m+1)}]}{\sum_{k=0}^{m} EXP[h(\theta - \beta_i) - \sum_{j=0}^{k} \tau_{i(m+1)}]} p_{ix}$$

τj(m+1) represents the thresholds parameter for all the items.
(m = 1) = a common number of response categories.
The following figure represents five curves with hypothetical values (0, 1, 2, 3, 4) for a question with a difficulty level of (0.30). The values (-1.8, 0.3, -0.2, 2.0) indicate the difficulties of the categories across all the question curves,

including this one.

The following figure represents five curves with hypothetical values (0, 1, 2, 3, 4) for a question with a difficulty Item of (0.30). The values (-1.8, 0.3, -0.2, 2.0) indicate the difficulties of the Items across all question curves, including this question.

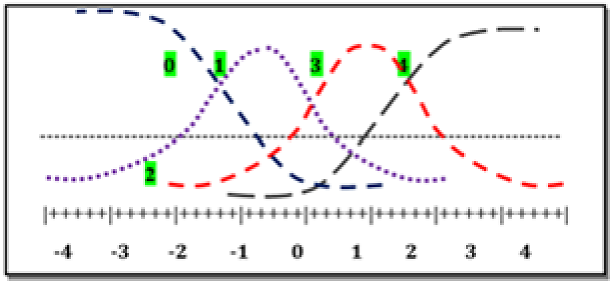

Figure (1)

Represents response probability curves for a four-level item of a scale according to Note: From Altaqi, 2013, p50

2.4 Related Works

Zakri's study (2020) entitled: “Identifying differential item functioning of the "EMRU" test of parental rearing styles among a sample of secondary school students.”

The research aimed to identify differential items functioning (by using the Mantel-Hanszel method) of the "EMBU" fest of parental rearing styles; according to the gender variable, of secondary school students. By using cluster randomized method, the research sample comprised (274) second-grade students, 134 males and 140 females of the academic year 2018/2019 (Department of Education in Sabya Province). In order to achieve the objectives of the research, the research sample responded to the "EMBU" test of parental rearing styles, which consists of (74) items. The "EMBU" test is translated and standardized by Abd Al-Rahman & Al-Mughrabi, 1990. Findings indicated that: (14) items of the "EMBU" test (father-image) showed differential functioning according to the gender variable; that included (8) items related to male students, and (6) items related to female students. There is no statistical significant effect of the internal validity indicators (RMSEA, NCP, AIC, SRMR, CFI) of the "EMBU" test (father-image) as a result of excluding the differential items functioning from the test. (17) items of the "EMBU" test (mother-image) showed differential functioning according to the gender variable; that included (9) items related to female students, and (8) items related to male students. There is no statistical significant effect of the internal validity indicators (RMSEA, NCP, AIC, SRMR, CFI) of the "EMBU" test (mother-image) as a result of excluding the differential items functioning from the test.

Abdullah's study (2022) entitled: “The Effect of Sample Size on the Item Differential Functioning in the Context of Item Response Theory.”

The study examined the effect of different sample sizes to detect the Item differential functioning (DIF). The study has used three different sizes of the samples (300, 500, 1000), as well as to test a component of twenty polytomous items, where each item has five categories. They were used Graded Response Model as a single polytomous item response theory model to estimate items and individuals' parameters. The study has used the Mantel-Haenszel (MH) way to detect (DIF) through each case for the different samples. The results of the study showed the inverse relationship between the sample size and the number of items, which showed a differential performer.

3. The Analytical Part

3.1 Methodology of the Study

The current study adopted a descriptive survey methodology to achieve its research objective.

3.2 Research Population

The current research population comprises secondary school students in grades 10, 11, and 12, who are enrolled in the scientific and literary tracks, within the educational districts of western Sulaymaniyah city. The total number of students is 10,875, with 7,259 students in the scientific track and 3,616 students in the literary track. The researcher will further elucidate and present the research population through the following table (Table 1).

Table (1)

Illustrates the research population

Total	Literary specialization		Science Specialization		Class
	Females	Males	Females	Males	
3699	503	539	1465	1192	10 Grade
2631	510	368	1107	646	11 Grade
4545	947	749	1697	1152	12 Grade
10875	1960	1656	4269	2990	Total

3.3 Research Sample

A representative sample was selected from the research population in order to understand the alignment of the responses of the sample individuals to the items of the test-wiseness Scale and to assess the extent of the impact of using test-taking wisdom strategies on differential performance, based on the assumptions of the Andrich model. The goal was also to determine the psychometric characteristics of the scale items. The sample was chosen using a stratified random sampling method, and the sample size was determined according to the guidelines provided by (Alam, 2005: 99), which states that, according to the item response theory, particularly the Andrich model, the minimum required number of individuals should not be less than 200. The research sample consisted of 530 male and female students from the tenth, eleventh, and twelfth grades. However, due to the fact that 83 students had responses falling between the high and low levels in employing test-wiseness strategies – individuals whose total scores on the test-wiseness Scale ranged from 105 to 207 – their data were excluded. Therefore, the final sample comprised 447 male and female students from the preparatory stage, accounting for 4.87% of the research population. Among them, 207 students belonged to the low-level group, while 240 students belonged to the high-level group in employing these strategies. They were selected using the stratified random sampling method. Table (2)illustrates the size and distribution of the sample used for the current research purposes:

Table (2)

Illustrates the research sample

Percentage	Total	Employment of test-wiseness Strategies		Level
		Females	Males	
%45.2	240	128	112	High
%15.6	83	32	51	Moderate
%39.2	207	71	136	Low
%100	530	231	299	Total

3.4 Research Instrument

The researcher adopted the test-wiseness Scale developed by (Hammad, 2010). The test-wiseness Scale consists of 52 items, with 50 negatively worded items. However, items 10 and 28 were rephrased in a positive and effective manner, indicating the use of the four test-wiseness strategies: time management, error avoidance, guessing, and utilizing test construction features. Each item presents five response options: "Always Applicable," "Mostly Applicable," "Sometimes Applicable," "Rarely Applicable," and "Never Applicable." Therefore, respondents' scores on the scale range from 52 as the lowest value to 260 as the highest value.

3.5 Face Validity

In this analysis, the scale or test items are presented to a group of specialized experts to assess their suitability in measuring the intended construct (Alam, 2000, p227). The scale is examined to determine the extent to which its items represent the facets of the trait it is supposed to measure (AbdulRahman, 1998, p185).

Ebel (1972, p522) suggests that consulting experts regarding the measurement of the intended trait is the best way to ensure the face validity of the scale. Following Ebel's viewpoint, the preliminary version of the scale items was presented to 18 experts specializing in educational and psychological sciences. They were asked to provide their opinions on the appropriateness of the items in the test-wiseness scale. These opinions were analyzed using percentages and the Chi-Square test of goodness-of-fit (χ^2). A item was considered valid when the calculated χ^2 value was significant at the 0.05 level, corresponding to 3.83 or higher, which is equivalent to an agreement rate of 83% among experts and reviewers. Table (3) illustrates this process:

Table (3)

Presents the results of the face validity of the test-wiseness scale

Item Sequence	Agreed	Disagreed	Chi-Square Value
1, 2, 3, 5, 9, 10, 11, 13, 14, 15, 16, 18, 21, 23, 24, 28, 30, 32, 33, 34, 37, 38, 39, 40, 41, 42, 44, 47, 50, 52	18	100%	0
4, 6, 8, 20, 22, 26, 31, 45, 48, 49, 51	17	49.4%	1
7, 17, 19, 25, 27, 35, 43, 46	16	88.8%	2
12, 29, 36	15	83.3%	3

3.6 The relationship between the item score and the total score of the dimension it belongs to

Anastassi (1976, p206) pointed out that the correlation of an item with an external or internal criterion is an indicator of its validity. When a suitable external criterion is not available, the total score of the respondent serves as the best internal criterion in assessing this relationship. The relationship of items with the total score means that the scale measures a single trait. The coefficient of the correlation between the item score and the total test score can be computed using Pearson's correlation coefficient. It is assumed that this relationship should be positive to indicate construct validity, contributing partially to establishing construct validity as an empirical validation.

Based on this, the researcher relied on the relationship between the scores of each item and the total score of the dimension (domain) to which it belongs in the scale, in order to assess the item's validity. To achieve this, the researcher used Pearson's correlation coefficient. It was found that all items were statistically significant at a significance level of 0.001 with degrees of freedom of 445. Table (4), illustrate the correlation values between the item score and the total score for the four dimensions of the scale, respectively.

Table (4)

Illustrates the correlation coefficient between the item score and the totalscoreTest

Item	coefficient of correlation value	Item	coefficient of correlation value	Item	coefficient of correlation value	Item	coefficient of correlation value
1	$.322^{***}$	14	$.526^{***}$	27	$.650^{***}$	40	$.346^{***}$
2	$.293^{***}$	15	$.598^{***}$	28	$.615^{***}$	41	$.346^{***}$
3	$.530^{***}$	16	$.537^{***}$	29	$.459^{***}$	42	$.445^{***}$

4	.542***	17	.546***	30	.430***	43	.578***
5	.545***	18	.497***	31	.444***	44	.398***
6	.608***	19	.532***	32	.425***	45	.427***
7	.449***	20	.515***	33	.504***	46	.461***
8	.533***	21	.576***	34	.438***	47	.375***
9	.439***	22	.644***	35	.398***	48	.513***
10	.380***	23	.604***	36	.441***	49	.453***
11	.517***	24	.487***	37	.446***	50	.557***
12	.519***	25	.549***	38	.456***	51	.612***
13	.503***	26	.526***	39	.459***	52	.693***

# The critical value for the correlation coefficient with degrees of freedom (445) at a significance level.

֍*0.05 = (0.088)
֍**0.01 = (0.128)
֍***0.001 = (0.169)

It is evident from the values presented in the above table that all the items were statistically significant at the significance levels adopted by the researcher.

3.7 Reliability

Reliability, in the context of scientific research, refers to the consistency of the scores obtained from the items of a measurement instrument that is intended to measure what it is designed to measure (Marshall, 1972, p104). Reliability can be assessed using various methods, including the split-half method and variance analysis (Thorne and others, 2001, p140). To examine the reliability of the scale used in this study, the researcher applied the Rollon equation as a corrective equation for the split-half method. Additionally, the researcher employed the variance analysis method and utilized the alpha-Cronbach equation to suit the nature of the scale used. Both Thorndike and Hagen (1977) emphasized that establishing reliability using this method depends on the consistency in individuals' responses to each item of the scale (Thorndike & Hagen, 1977, p82).

After using the two equations to calculate the reliability of the Test wiseness scale, it was found that the Rollon reliability coefficient ranged from (0.847 to 0.776). This indicates that these values are consistent with the acceptable reliability values for the purposes of this study. Therefore, the Test wiseness scale demonstrates a high level of reliability. Table number (5) illustrates these results.

Table (5)

Illustrates the coefficient of reliability and the standard error of measurement for the test-wiseness scale using both Rollon and Cronbach's Alpha equations

Alpha-Cronbach		Rollon		Standard Deviation	items		Dimension	The sequence
Standard Error	Reliability	Standard Error	Reliability		To	from		
3.731	0.784	3.231	0.838	8.029	14	1	Test time management	1
2.713	0.759	2.377	0.815	5.527	23	15	Error Avoidance	2
1.566	0.737	1.445	0.776	3.055	32	24	Guessing	3
5.130	0.793	4.454	0.847	11.387	52	33	Utilizing Test Structure Characteristics	4

3.8 Derivation of Criteria for Classifying the Ability to employ test-wiseness Strategies

Criteria are one of the primary objectives used to classify individuals who possess a certain trait or those who lack it. They are based on the raw score, which is the result derived from test application before being subjected to statistical treatment (Hassanein, 2001, p29). Accordingly, the researcher extracted the raw scores and then established five levels for the scale, based on the number of answer alternatives used in the scale's gradation. The range was calculated by subtracting the highest scale score, which is 260, from the lowest score, representing (52), and then dividing the result by (5), representing the number of levels defined by the researcher. Individuals whose raw scores fell between (52) and (104) were classified as having low ability to employ test-wiseness strategies, while those with scores between (208)and (260)were considered to have high ability. Individuals whose scores ranged from (105) to (207) were excluded, as their scores represent an average or neutral level of classification based on the number of answer alternatives in the scale. These individuals cannot be classified as having either low or high ability to employ test-wiseness strategies.

3.9 Verifying Assumptions of the Andrich Model

The Item Response Theory is based on three main assumptions: the one-dimensionality assumption, local independence of items, and fit to the item characteristic curve. Verifying these assumptions is essential prior to employing the model in statistical analysis. This was done as follows:

Firstly: Investigating One-dimensionality Assumption: This assumption was verified in the current study by adopting certain indicators based on a widely used method, which is factor analysis. Factor analysis is a statistical method employed to handle interconnected data with varying degrees of correlation, summarizing them into independent classifications based on qualitative criteria. (Hattie, 1985, p146).

3.10 Factor Analysis

This is accomplished through testing a set of external criteria in addition to the test for which the validity coefficient is to be determined. The intercorrelations between the external criteria and the test are calculated, and these correlations are then analyzed to determine the extent to which each item is saturated with the common factor, as well as any other common factors that might be present. The degree of saturation of an item with the common factor indicates its validity in measuring that factor (Abdul Rahman, 1999, p192).

Before embarking on the use of the factor analysis method, the researcher verified the necessary conditions in the correlation matrix for the factor analysis. It is essential that the determinant of the correlation matrix is not equal to zero, which means ($R\neq 0$). The researcher found that the determinant of the correlation matrix equals (1.812×10^{5}) and this value is greater than zero.

3.11 Measurement of sample homogeneity in relation to sample size

This is done using the chi-square (χ^2) value for the Bartlett's test, and Table (6) illustrates this.

Sample adequacy and sufficiency: This is achieved by calculating the Kaiser-Meyer-Olkin (KMO) measure, which should not be less than 0.5 according to Kaiser's criterion. As shown in Table (6), the KMO value is 0.564, which is greater than 0.50. This means that it indicates.

Table (6)

Kaiser-Meyer-Olkin (K.M.O) Value and Bartlett's Test

0,564	Sample adequacy test Kaiser-Meyer-Olkin Measure of Sampling Adequacy. (KMO)
11124.678	Bartlett's Test of Sphericity
1770	Degrees of freedom (df)
.000	Statistical significance (sig)

3.12 The values of prevalence

The adequacy of sample is assessed by calculating the level of each variable using the Measures of Sampling Adequacy (MSA) test. This test indicates whether the correlation level between each variable and the other variables in the correlation matrix is sufficient for conducting factor analysis.

The values of prevalence (MSA) found in the diagonal elements of the inverse correlation matrix (Anti-image Matrices) all exceeded (0.50), ranging between (0.507) and (0.800). According to Kaiser's criterion, these values are considered acceptable, as Kaiser sets a minimum threshold of (0.50), for considering the correlation level of each variable with the other variables suitable for continuing with factor analysis. If a variable does not exceed the critical value, it is removed due to its lack of interdependence with the other variables or its independence from the structure of the other variables (Tigza, 2012, p90). This is illustrated in Table (7).

3.13 Factor saturation of items

The extent to which each item in the scale is associated with the common factor and any other shared factors (if present) is represented by the degree of factor saturation. The magnitude of factor saturation for each item onto the common factor and other factors is indicative of the reliability of measuring that common factor (Abdulrahman, 1998, p192). This can be identified through the adoption of a criterion such as the Guttman criterion, which is the same criterion the researcher adopted in the current study. Items are accepted if their saturation onto the common factor is equal to or exceeds (30%). This criterion was met with values ranging from (0.557) to (0.653), indicating that all items of the scale were valid. As a result, no item was excluded. These results are presented in Table (7).

Table (7)

illustrates the eigenvalues results for the Measures of Sampling Adequacy (MSA) test

Items	The values of prevalence (MSA)	Factor Loadings	Items	The values of prevalence (MSA)	Factor Loadings	Items	The values of prevalence (MSA)	**Factor Loadings**
1	.578	.457	**19**	.628	.473	**36**	.708	.412
2	.580	.319	**20**	.559	.467	**37**	.670	.392
3	.702	.365	**21**	.718	.466	**38**	.509	.390
4	.775	.328	**22**	.676	.459	**39**	.599	.388
5	.598	.371	**23**	.792	.454	**40**	.528	.383
6	.642	.444	**24**	.652	.453	**41**	.603	.412
7	.626	.454	**25**	.594	.450	**42**	.698	.346
8	.544	.531	**26**	.583	.450	**43**	.589	.301
9	.619	.525	**27**	.726	.445	**44**	.545	.348
10	.524	.524	**28**	.635	.438	**45**	.527	.322
11	.670	.509	**29**	.681	.437	**46**	.557	.339
12	.737	.506	**30**	.568	.430	**47**	.507	.481
13	.533	.490	**31**	.539	.427	**48**	.800	.448
14	.675	.484	**32**	.603	.426	**49**	.574	.331
15	.666	.481	**33**	.564	.426	**50**	.508	.465
16	.663	.479	**34**	.577	.419	**51**	.589	.425
17	.650	.477	**35**	.596	.417	**52**	.586	.496
18	.580	.476						

Furthermore, exploratory factor analysis was conducted to identify the underlying factors within the test. The researcher employed the Principal Components analysis method for individuals' responses to test items. This involved calculating the Eigen Value, which represents the latent root, as well as the Explained Variance for each factor. Additionally, the cumulative explained variance for each extracted factor was calculated. Table (8) illustrates the Eigen Values, Explained Variance, and Cumulative Explained Variance for the extracted factors.

Table (8)

presents the Eigen Values, Explained Variance, and Cumulative Explained Variance values

Cumulative Explained Variance Ratio	Explained Variance Ratio	Eigenvalue	Factor	S
60.410	60.410	42.093	Test Time Management Strategy	1
81.316	20.906	18.161	Error Avoidance Strategy	2
92.743	11.427	7.981	Guessing	3
100.000	7.257	5.381	Test Construction Features Utilization Strategy	4

From Table (8), it is evident that the final eigenvalues for the four factors are greater than one, as per Kaiser's criterion. This indicates that the scale is distributed across four factors. Additionally, it is also evident that the first factor explains the largest proportion of variance in the scores of the examined individuals compared to the other factors. The eigenvalue for the first factor was (42.093), accounting for a variance of (60.410%). This implies that this factor dominates in explaining the overall variance in the scale scores.

The test is considered valid if all the coefficients of the saturation matrix for the factor loadings are greater than (0.30), and if the absolute value of the determinant of the correlation matrix is greater than (0.00001). Additionally, the Kaiser-Meyer-Olkin (KMO) measure and the Measures of Sampling Adequacy (MSA) test for each variable should not be less than (0.5). Based on these results, all of these conditions have been met in the scale. This indicates that the scale is unidimensional, meaning there is a single underlying latent trait, namely the test wisdom, that the scale measures. This latent trait is responsible for explaining the variance in the scale scores.

3.14 Local Independence

Local independence and one-dimensionality are similar concepts, but they are not equivalent in meaning. When verifying the one-dimensionality assumption, local independence is also verified, but the reverse is not true. In other words, the local independence assumption can be met without having a unidimensional structure, as long as all factors influencing test results are considered (Erguven, 2014, p26). Moreover, local independence serves as an indicator of one-dimensionality when the employed model estimates an individual's capacity as unidimensional. Based on the foregoing, after confirming the one-dimensionality assumption, the researcher has implicitly confirmed the local independence assumption as well.

3.15 Thirdly: Nature of the Item Characteristic Curve

Where the continuous increase of the item characteristic curve indicates an increasing probability of success for individuals with higher scores in responding to this item, with a higher likelihood than individuals with lower scores on the trait (Erguven, 2014, p26). This assumption refers to the nature of the distinctive curve or function for each item, which describes the relationship between the ability and performance on the item. The shape of the item characteristic curve depends on the item's difficulty, discrimination, and individuals' ability. The distinctive curves for the items in the Andrich model are parallel. The researcher utilized the program "ConstructMap 4.6" to draw the item characteristic curves for the items, illustrating the distinctive curves for some items.

3.16 Checking Data Fit for the Andrich Model

In the realm of measurement, assessment, and statistical research, conducting item fit analysis serves as a systematic verification of how test or scale items operate in measuring the trait. Item fit analysis has been proposed as a method to identify extraneous factors that influence item responses. Similarly, individual fit analysis aims to diagnose individuals with response patterns that deviate from the specified model.(Embretson & Reise, 2000, p127-128).

The computer program (ConstructMap 4.6) was employed to conduct the statistical analysis for estimating the model parameters. This analysis encompassed various components, including the statistical calibration of scale items and the estimation of their parameters, individual ability estimation, statistical fit assessment for both items and individuals, and the standard errors associated with these estimations. Furthermore, the analysis involved identifying the measurement properties of the entire scale.

The matching of individuals' responses has been computed by calculating the ability of each individual (the location of the individual on the measured trait) in addition to the standard error in measuring the ability. Moreover, the overall statistical fit values have been calculated, characterized by two indices: the **Infit index**, also expressed as the mean square fit (MNSQ) convergence statistic, is a statistical indicator of unexpected behaviors that affect individuals' responses to items that are closely aligned with their ability levels. The **Outfit index**, also referred to as the mean square fit for outliers, is an alternative statistical indicator, exhibiting similar or parallel characteristics to the former. However, it is more sensitive to unexpected behaviors of individuals. Both indices are associated with each ability estimate. Table (9) illustrates the mean locations of individuals and the standard error in the estimates, along with the in-fit and out-fit statistics for both convergent and divergent matching values.

Table (9)

Displays the mean locations of individuals, the standard error in estimates, the convergent and divergent fit squares, and the t-statistic

T-Statistic for External Fit	External Fit (OUTFIT)	T-Statistic for Internal Fit	Internal Fit (INFIT)	Standard Error	Mean Locations of Individuals
0.08 (Negative)	1.04	0.12 (Negative)	1.00	0.162	0.434

When examining the values of the weighted person fit statistics and the t-statistic for each individual in the sample, it was found that there were 29 individuals whose observed responses deviated from the expected responses based on their abilities. This means that the values of the outfit mean squares (MnSq) corresponding to their abilities exceeded the range of (0.75 to 1.33) suggested by both "Adams and Khoo" (Adams & Khoo, 1996), which is the same range adopted by the program. Alternatively, the corresponding t-statistic values for their abilities were either greater than (+2) or less than (-2).

As pointed out by Alastair & Hutchinson (1987), if the value of this statistic exceeds (+2), then the individual's ability is considered mismatched with the abilities of the group of individuals. Therefore, these individuals are not fitting the model, and they should be excluded to proceed with the analysis (Wilson, 2005, p15). The t-statistic is a transformation of the outfit mean square (MnSq) into the standard normal distribution. Values above (+2) or below (–2) are generally considered large. Hence, it is recommended to use both the mean square values and the t-statistic together (Hamadneh, 2013, p94). When both of them indicate significant misfit, further investigation should be conducted within the item to understand the reason.

After excluding the non-fitting individuals (29 individuals), the analysis for testing the fit of the (52) items to the Andrich model were repeated. The fit data for the items are presented in Table (10):

Table (10)

Displays the mean person locations and the standard error of measurement for each item, as well as the mean square values for convergent and divergent fit, and the t-statistic for each item

Item Locations	Standard Error	Internal Fit (INFIT)	t-Statistic	External Fit (OUTFIT)	t-Statistic
Mean	0.024	0.032	1.00	0.1-	0.98

When examining the values of the convergent and internal fit statistics for items, which indicate the stability of relative difficulty levels of items across different ability levels, it became evident that all items were found to be in good fit. Consequently, no items were excluded from the assessment. The analysis was then repeated to assess the fit of both individuals and items to the Andrich model. The fit indices demonstrated excellent fit for both individuals and items, as shown in Table (11) below:

Table (11)

Presents the estimation of difficulty, internal and external fit, and the t-values for the items of the test-wiseness Scale

Outfit		Infit		S. E	Item Location	Item
T	Mnsq	T	Mnsq			
1.8	1.16	1.8	1.14	0.032	1.03	1
0.0	1.00	0.5-	1.08	0.031	0.95-	2
0.9-	1.06	1.2-	0.95	0.031	2.07	3
0.3	1.01	0.4-	1,07	0.032	1.45-	4
1.6-	0.88	1.8-	0.86	0.031	0.47-	5
2.0-	0.88	1.6-	0.86	0.032	0.71	6
1.0-	0.95	1.3-	0.94	0.032	2.71	7
0.3-	0.98	0.8-	0.96	0.031	1.58	8
1.9-	0.91	2.0-	0.90	0.031	1.81	9
1.3	1.33	1.9	1.29	0.033	0.67-	10
1.2-	0.94	2.0-	0.91	0.032	2.19-	11
1.2	1.21	1.2	1.21	0.032	0.22-	12
1.8-	0.77	1.8-	0.76	0.032	1.75-	13
0.1	1.00	1.1-	0.95	0.032	1.17-	14
1.8-	0.91	1.9-	0.90	0.032	1.36	15
1.0-	0.95	1.7-	0.93	0.034	2.33	16
1.3	1.06	1.0	1.05	0.033	1.20	17
2.0-	0.89	2.0-	0.89	0.033	1.58	18
1.9-	0.75	1.9-	0.76	0.032	0.26-	19
1.1	1.10	1.3	1.09	0.031	0.09	20
0.9-	0.96	1.2-	0.94	0.031	0.76	21
0.8	1.04	0.6	1.03	0.031	1.79	22
2.0	1.12	1.9	1.09	0.031	0.38-	23
0.4	1.02	0.4-	0.98	0.033	2.28-	24
1.5-	0.93	2.0-	0.89	0.032	1.14-	25
1.8-	0.83	1.5-	0.80	0.031	0.48-	26
0.0	1.00	0.5-	0.98	0.031	0.85-	27
0.2-	0.99	0.7-	0.97	0.032	1.02-	28
1.9	1.14	1.8	1.13	0.032	1.59	29
2.0	1.10	2.0	1.10	0.033	2.28	30

1.7	1.08	1.7	1.08	0.032	2.08	31
1.8-	0.87	2.0-	0.86	0.032	0.76-	32
2.0-	0.89	1.9-	0.87	0.031	0.32-	33
0.9-	0.91	1.2-	0.85	0.034	2.07	34
1.0-	0.95	1.7-	0.93	0.034	2.32	٣٥
1.5-	0.93	2.0-	0.89	0.032	1.10-	36
1.7-	0.92	1.9-	0.87	0.031	1.61-	37
1.8-	0.87	1.1-	0.86	0.031	1.49-	38
1.7	1.19	1.6	1.17	0.033	1.92	39
2.0	1.10	1.7	1.08	0.032	1.31	40
1.9	1.26	1.9	1.25	0.032	1.21	41
0.4-	0.98	1.6-	0.93	0.032	1.89-	42
0.6	1.03	0.1	1.00	0.031	0.76	43
0.2	1.01	1.2-	0.95	0.032	1.73-	44
1.8	1.16	1.8	1.14	0.032	1.13	45
1.2	1.06	0.9	1.04	0.031	0.68	٤٦
1.9	1.21	1.9	1.20	0.031	0.88	47
1.1	1.05	1.1	1.05	0.031	0.53	48
1.4-	0.93	1.3-	0.89	0.032	2.18-	49
1.4-	0.80	1.9-	0.79	0.032	0.02-	50
0.8-	0.96	1.2-	0.94	0.031	1.10	51
0.4-	0.98	0.8-	0.96	0.031	0.63	52

Outfit Fit		Infit Fit		Standard Error	Estimates of Ability Levels	Average
MnSq	T	MnSq	T			
0.99	0.26-	0.96	0.30-	0.151	0.201-	

From the above table, it is observed that the free estimates of item locations on the latent trait for the model fit assumptions ranged between (2.55-) and (2.71) logits, with an average of (0.027). This indicates a broad range of the measured trait covered by the scale, spanning from low to high levels of ability. The standard error of the mean difficulty estimates was (0.032), a low value close to zero. Similarly, for the values of individual ability estimates, along with their means, standard errors, infit and outfit fit statistics, as shown in the lower part of the table, they reflect the accuracy of the estimates of item locations on the measured latent trait (the test-wiseness).

3.17 Indicators of the Validity of the test-wiseness Scale

The fundamental principles of test and measurement validity remain consistent, whether the test is criterion-referenced or reference-narrative. The concept of validity in reference-narrative tests is no different from that in criterion-referenced tests, despite the evidence relying on the nature of the tool, reflecting variations in their purposes. Despite the nearly identical concepts, some experts differentiate between types of validity, mentioning descriptive validity as an alternative to content (face) validity, and functional validity as an alternative to experimental validity. Selection validity of the behavioral domain is offered as a substitute for conceptual or theoretical construct validity (Ababneh, 2009, p151-152).

3.18.1 Descriptive Validity

A scale is considered valid if it can be used to accurately describe an individual's performance within the behavioral domain that the scale or test measures. Descriptive validity is the first step along this path, and sometimes this type of validity is referred to as content validity. The reason for choosing the concept of descriptive validity is that it is more general than content validity (Alam, 1986, p82). Descriptive validity can be estimated for a scale by involving a group of experts and reviewers in the field of specialization to assess the content validity of the scale's items. This aspect has been confirmed in the literature that has addressed this scale, and it was previously mentioned in the stages of testing the items of the tool. This form of validity has been pre-verified through the procedures followed by the researcher in calculating apparent validity.

3.18.2 Domain Sampling Validity (DSV)

Internal consistency is used to confirm the validity of domain sampling. It is derived from the statistical program used in the current research and is primarily an indicator of homogeneity. Because the degree of homogeneity helps describe and define the characteristics of the measured domain or trait, the consistency of the scale is related to the extent of the hypothetical construct's validity. As a method to establish internal relationships among item scores, the theory of psychological measurement considers the strength of the correlation between the items designed to measure the trait as a statistical indicator of construct validity. Therefore, indicators of consistency can be preliminary indicators of validity (Odeh, 1998, p387). For example, the Rollon coefficient for the four domains of the scale were (0.847, 0.776, 0.815, 0.838) respectively, while the values of Cronbach's alpha coefficient were (0.784, 0.759, 0.737, 0.793). Considering that consistency is one of the indicators of instrument validity, these values are good and indicate the consistency of the items in measuring what they were intended to measure. This relationship serves as a statistical indicator of the instrument's validity.

3.18.3 Model Fit Validity

To verify the objectivity conditions, evidence must be provided to confirm the assumptions of the Rating Scale Model for the test-wiseness scale. Previous references have indicated the fulfillment of a fundamental assumption of the model, backed by a set of indicators that signify it, namely one-dimensionality. Additionally, it is necessary to highlight other indicators of meeting other assumptions. Among these, the measurement invariance across the characteristics of the ability distribution for the research sample is crucial. This implies that the relative difficulty values of the items do not significantly differ for most individuals across various levels of the trait. There are two key indicators for this:

The first indicator is the average value of the statistical fit indices, both the Proximal Fit (Infit) and the Distal Fit (Outfit), also known as the fit statistic or static fit. It measures the proximity of the observed data to the ideal data patterns as assumed by the model. This value typically falls within the range of (0.75 - 1.33), reflecting the optimal fit of the data to the model's expectations.

The second indicator involves utilizing the statistical value (t) for the fit indices of both Proximal Fit (Infit) and Distal Fit (Outfit) for items and individuals in the sample. It includes excluding items and individuals that do not meet the suitability criteria of this statistical value. Specifically, items and individuals with fit indices (t) exceeding the boundaries of (-2) and (2) are removed, as mentioned earlier. These indicators contribute to the assessment of construct validity. These indicators serve to assess the degree to which the observed data aligns with the expected model patterns. They demonstrate whether the observed item characteristic curves closely resemble the expected curves of the model. Additionally, they indicate whether the item characteristic curves exhibit a similar slope or curvature. When the item characteristics are independent of the sample, the items possess relatively equal discriminatory power.

3.19 Reliability Indicators for the test-wiseness Scale

The reliability coefficients for the test-wiseness Scale were calculated using two methods. The first method involved assessing the internal consistency of the entire scale using the Cronbach's alpha coefficient, which was computed using the Construct map 4.6 software. The obtained value was 0.82, indicating a high level of internal consistency and suggesting that the instrument has strong reliability.

The second method involved using Item Response Theory (IRT) to calculate reliability coefficients for the test-wiseness Scale. After obtaining the estimated values for both item difficulties and individual abilities, two types of reliability coefficients were computed using the Construct map 4.6 software: Person Reliability and Test Reliability.

The test reliability is computed in the software according to the following procedure as indicated by "Mislevy and colleagues" (Mislevy, Beaton, Kaplan, and Sheehan, 1992).

$$\text{Reliability} = \text{Var}(\hat{\theta}_{\text{EAP}})/\hat{\sigma}^2$$

When the latent trait is normally distributed, estimating the variance of the population is obtained from the Maximum Marginal Likelihood (MML) estimation, and the distribution becomes nearly normal.

From the aforementioned information, one can directly benefit from the reliability values provided by the (ConstructMap.4.6) software. Table (15) presents the reliability coefficients as computed in the statistical program for the Table (12): Reliability Coefficients for Individual and Test Wisdom Scale Stability Scale:

Table (12)

Reliability Coefficients for Individual and Test Wisdom Scale Stability Scale Stability

Scale Reliability Coefficient	Person Reliability Coefficient	Reliability from the Information Function
0.74	0.76	0.775

From the above table, it is evident that the value of the (Individuals' Reliability Coefficient) is of good quality, indicating the reliability of discriminating among individuals, sample adequacy, and item reliability. Thus, it contributes to defining the latent trait measured by these items. It should be noted in this context that the reliability coefficient values in this method are equivalent to the values of reliability coefficients obtained using Cronbach's alpha method in classical theory, which represents the minimum threshold of reliability. It's worth mentioning that the reliability value of the instrument was determined using the approach proposed by Mislevy, Beaton, Kaplan, and Sheehan (1992), extracted from the statistical program. Following a series of steps, including the removal of individuals who were not suitable for the model, the largest change in reliability value after removing unsuitable individuals was (0.034), with a minimal decrease. This strongly indicates the cohesion and consistency among the scale items, confirming the good reliability value.

3.20 Statistical Methods: In order to achieve the research objectives, the researcher utilized several statistical programs and methods as follows

Statistical Package for the Social Sciences (SPSS): This was used for extracting: (Principal Component) factor analysis and Guttman's factor analysis to confirm one-dimensionality and construct the scale.

Statistical software (ConstructMap.4.6): This was used to extract the following: estimates and map the item and person locations on the latent trait continuum. It also aided in determining the convergent fit (Infit) and divergent fit (Outfit) for the scale items, using fit statistics such as t-test, and mean square (MnSq) for item likelihood ratios, based on the Andrich model.

- Rulon Equation for Reliability Calculation.
- Cronbach's Alpha Equation for Reliability Calculation.
- Information Function and Standard Error Function.
- Test Reliability Coefficient according to the method proposed by (Mislevy, Beaton, Kaplan, and Sheehan, 1992), calculated from the variance ratio, using the estimated marginal expectation method based on the equation.

- Person Reliability Coefficient based on the Separation Coefficient among sample individuals.

4. Presentation and Interpretation of Results

4.1 First Aim: Measuring differential performance in the ability to employ test-wiseness strategies, based on the graded response model of the contemporary measurement theory, Andrich Model

The estimated ability of an individual derived from the scale is adjusted for the abilities of the remaining individuals who respond to the same scale. If the estimated abilities corresponding to each possible score on the scale are statistically equivalent to the score resulting from the performance analysis of individuals from one of the two samples on this scale, while taking into account the standard error of these estimates along with those estimates derived from the performance analysis of the entire sample, it means that the estimated ability of the individual obtaining a specific total score on this scale is not affected by the variation in the performance level of the analysis sample.

Therefore, the researcher divided the statistical analysis sample based on their utilization of the test-wiseness strategies. This was done by calculating and splitting the total sample into a high-ability group and a low-ability group in employing the test-wiseness strategies according to a criterion of score mediation, relying on the Score File. Subsequently, the results of each sample's responses on the scale were individually analyzed using the (ConstructMap.4.6) program to calculate ability estimates and their standard errors. Following that, a comparison of the ability criterion was conducted. Table (13) illustrates this process.

Table (13)

Illustrates the ability to employ test-wiseness strategies (in logits) and corresponding standard errors for each potential total score, derived from the overall, low-ability, and high-ability samples

High-Ability Sample			Low-Ability Sample			Overall Sample		
Standard Error	Difference Between Estimates	Ability Estimation	Standard Error	Difference Between Estimates	Ability Estimation	Standard Error	Ability Estimation	Item
*****	*****	*****	*****	*****	*****	*****	*****	1
0.741	-0.02	0.79	0.074	0.09	0.64	0.321	0.73	2
0.583	0.02	-0.56	0.048	-0.07	-0.41	0.272	-0.34	3
0.259	0.03	0.62	0.017	0.04	0.43	0.257	0.59	4
0.696	0.03	-1.53	0.003	0.00	-1.50	0.192	-1.50	5
0.727	0.04	-0.84	0.019	-0.06	-0.72	0.224	-0.76	6
0.698	0.02	1.53	0.094	0.01	1.51	0.295	1.50	7
0.541	0.01	0.57	0.074	0.09	0.37	0.146	0.46	8

0.641	0.03	1.09	0.008	0.01	1.07	0.136	1.06	9
0.610	0.04	-0.43	0.074	-0.09	-0.28	0.114	-0.37	10
0.655	0.04	-1.08	0.002	-0.02	-1.00	0.374	-1.02	11
1.011	-0.03	1.45	0.021	0.11	1.31	0.217	1.42	12
0.804	0.02	-2.09	0.008	0.01	-2.06	0.302	-2.05	13
0.707	0.04	-1.59	0.028	-0.06	-1.45	0.224	-1.51	14
0.627	0.00-	0.11	0.018	0.08-	0.09	0.208	0.01	15
0.597	0.01-	0.47	0.037	0.10-	0.28	0.213	0.38	16
0.809	0.01	-2.23	0.074	-0.09	-2.01	0.323	-2.10	17
0.806	0.03	2.09	0.003	0.00	1.66	0.304	2.06	18
0.605	0.03	0.19	0.020	0.05	0.17	0.031	0.22	19
0.295	0.04	1.37	0.048	0.07	1.01	0.208	1.33	20
0.824	-0.01	1.07	0.017	-0.04	0.88	0.230	1.06	21
0.809	0.04	-1.23	0.074	-0.09	-0.97	0.223	-1.10	22
0.591	0.01	0.34	0.008	0.01	0.24	0.184	0.33	23
0.658	-0.03	0.79	0.054	0.16	0.46	0.129	0.76	24
0.658	-0.03	0.79	0.054	0.16	0.66	0.129	0.76	25
1.068	0.03	-2.92	0.010	0.02	-1.91	0.369	-2.89	26
0.809	0.13	-2.23	0.074	-0.09	-2.01	0.323	-2.10	27
0.627	0.03	-0.62	0.003	0.00	-0.59	0.147	-0.59	28
0.346	0.03	1.97	٠,018	0.08	1.02	0.253	1.94	29
0.605	0.01	0.19	0.020	0.05	0.17	0.031	0.22	30
0.251	0.01	-0.39	0.003	0.00	-0.40	0.149	-0.40	31
0.627	0.10-	0.11	0.018	0.08-	0.09	0.008	0.01	32
0.585	0.00	-0.00	0.003	0.00	0.00	0.078	-0.00	33
0.639	0.02	-1.09	0.008	0.01	-1.06	0.233	-1.07	34
0.622	-0.04	1.64	0.028	0.06	1.45	0.309	1.51	35
0.590	0.01	-0.35	0.008	0.01	-0.33	0.083	-0.34	36
0.607	0.01	-0.70	0.008	0.01	-0.68	0.100	-0.69	37
0.359	0.03	1.27	0.048	0.07	1.10	0.364	2.07	38
0.257	0.03	0.55	0.017	0.04	0.26	0.155	0.52	39
0.645	-0.04	1.29	0.028	0.06	1.11	0.659	1.17	40
0.608	0.00	0.70	0.008	0.01	0.69	0.602	0.68	41
0.627	0.03	-0.62	0.003	0.00	-0.59	0.647	-0.59	42
0.696	0.03	-1.53	0.003	0.00	-1.50	0.692	-1.50	43
0.806	0.03	2.09	0.003	0.00	2.06	0.804	2.06	44
0.296	0.00	-1.46	0.011	0.03	-1.19	0.249	-1.46	45
0.290	0.01	-1.38	0.011	0.03	-1.11	0.288	-1.37	46
0.707	0.04	-1.59	0.028	-0.06	-1.45	0.724	-1.51	47
0.295	0.04	1.37	0.048	0.07	1.20	0.295	1.33	48

0.768	0.03	-1.26	0.020	-0.05	-1.04	0.787	-1.19	49
0.655	0.03	-1.08	0.008	-0.02	-1.00	0.674	-1.02	50
0.698	0.04	1.57	0.076	-0.09	0.73	0.712	1.64	51
*****	*****	*****	*****	*****	*****	*****	*****	52

From the previous table, we observe that the ability to employ test-wiseness strategies may lead to differential performance. This is indicated by the differences between the responses of individuals with high-level proficiency in using test-wiseness strategies compared to their overall score, and individuals with low-level proficiency in using these strategies. This suggests that students' utilization of test-wiseness strategies results in differential performance that affects the amplification of their ability scores, as well as the impact on their measurement properties such as reliability. This is further evident from the differences in the values of the standard error.

Thus, the researcher has answered the First Aim, which states (Measuring differential performance in the ability to employ test-wiseness strategies, based on the graded response model of the contemporary measurement theory, Andrich Model).

4.2 The Second objective: Detecting the extent of variation in the utilization of test-wiseness strategies between individuals with lower and higher abilities

In order to achieve this objective, the researcher calculated the mean for individuals with low abilities, which was (91.64), with a variance of (3.44). As for individuals with high abilities, their calculated mean was (227.81), with a variance of (2.34). The researcher utilized these results to compute the differences between both groups by employing an independent samples t-test. The results indicated that the calculated t-value was (2.29), with degrees of freedom (445) and a significance level of (0.05). Upon comparing the calculated t-value with the critical t-value (1.96), the results suggested that there were statistically significant differences in favor of those with high abilities. Table (14) illustrates these findings:

Table (14)

Illustrates the level of differences in the utilization of test-wiseness strategies based on individuals' low and high abilities

Significance Level at (٠,٠٥)	*Degrees of Freedom*	*t-Value*		*Variance*	*Mean*	*Sample Size*	*Group*
		Tabular	*calculated*				
Signific	*445*	*1.96*	*2.29*	*3.44*	*91.64*	*207*	*lower*
				2.34	*227.81*	*240*	*higher*

The results above indicate that there are differences in the utilization of test-wiseness strategies between individuals with low and high abilities, in favor of those with higher abilities. This suggests that students resorting to using test-wiseness strategies result in differential performance that inflates the scores, leading to the generation of random errors. This, in turn, affects the measurement accuracy and widens the gap between observed (raw) scores and true scores.

5. Recommendations of the Study

Based on the results obtained, the researchers can recommend the following:

1. Using the Andrich Model for the Purpose of Developing Psychological and Educational Scales**: Given its statistical capabilities, the Andrich Model should be employed for the construction and enhancement of psychological and educational scales. This model possesses the potential to produce precise indicators for each individual in the sample and for each item within the scale.

2. Utilizing the Statistical Software (ConstructMap-4.6) for Analyzing Data from Psychological and Educational Scales: The statistical software (ConstructMap-4.6)should be employed for the analysis of data from psychological and educational scales. This software offers statistical capabilities that enable the production of accurate information regarding the sample and the instrument. This is especially beneficial in the diagnostic domain for each individual within the sample .These recommendations underscore the importance of employing standardized measurement tools, models, and advanced statistical software in psychological and educational research. Such practices can lead to precise results and enable a more informed use of data to improve educational and assessment processes.

6. Suggestions of the Study

In light of the results obtained in the current research, the researcher proposes the following:

Replicating the Same Study Using Different Models from Item Response Theory (IRT): Conducting a similar study using different models from Item Response Theory (IRT) that deal with graded response would be beneficial. This could involve exploring how individuals respond to different types of items or assessing different aspects of the measurement process using alternative (IRT) models.

Replicating the Current Study with Different Personality Variables: Replicating the current study with other personality variables, such as using the Social Desirability Scale with the Andrich Model or any other (IRT) model, would be valuable. This would allow for a broader examination of how various personality traits relate to the measurement process.

These suggestions emphasize the importance of conducting further research to explore the applicability and generalizability of the findings, both in terms of different (IRT)models and in relation to different psychological constructs. Replicating and extending the study can provide a more comprehensive understanding of the relationships between psychological variables and measurement processes.

References

Ababneh, E. Kh. (2006). Empirical verification of stocking equations in determining levels of parallel ability for maximum likelihood estimation of item parameters in item response theory. *Jordanian Journal of Educational Sciences*, *2*(2), 53-63.

Abdullah, Z. (2022). The effect of sample size on the item differential functioning in the context of item response theory. *Educational and Psychological Journal*, *19*(72), 119-123.

Abdul-Rahman, S. (1998). *Psychological measurement: Theory and application*. Kuwait: Alfalah Library.

Alam, S. M. (1986). *Contemporary developments in psychological measurement*. Kuwait University.

Alam, S. M. (2000). *Educational and psychological measurement and evaluation: Fundamentals, applications, and contemporary approaches* (1st ed.). Cairo: Dar Alfikr Alarabi.

Alam, S. M. (2005). *Models of unidimensional and multidimensional item response and their applications in educational and psychological measurement*. Cairo: Dar Alfikr Alarabi.

Albashabsheh, Kh. M. (2016). *The effect of test performance on the cognitive development in mathematics in the context of PISA 2012: A study from Jordan*. (Master's Thesis), Jordan.

Allabadi, N. (2008). *Comparison of four methods for detecting differential item functioning*. (PhD dissertation), Yarmouk University, Jordan.

Almaliki, Dh. B. A. (2010). *The relationship between test anxiety and test wiseness among a sample of secondary school students in Allaith educational province*. [Unpublished master's thesis], College of Education, Umm Alqura University.

Altaqi, A. M. (2013). *Modern theory of measurement* (2nd ed.), Amman: Dar Almaseera for Publishing and Distribution.

Alzaher, Z. M., Jacqueline, T., & Judat, A. A. (1999). *Principles of measurement and assessment in education*. Amman: Dar Althaqafa for Publishing and Distribution.

Alzahrani, M. R. (2015). Psychometric characteristics of the test wiseness scale among university students in the Saudi environment. *Scientific Journal of the College of Education,3*(4), 217-266.

Anastassi, A. (1976). *Psychological testing* (4thed). New York, Macmillan, 206.

Awda, A. (1998). *Measurement and evaluation in the teaching process* (2nded). Irbid: Dar Alamal for Publishing and Distribution.

Camilli, G. & Shepared, L. (1994). *Methods for identifying bias test item*. USA: Stage publication.

Chung, W. & Huisu, Y. (2004). Effects of average signed area between two item characteristic curves and purification procedures on the DIF detection via the mantel-haenszel method. *Applied Measurement in Education, 17*(2), 113-144.

Crocker, L. & Algina, J. (1986). *Introduction to classical and modern test theory*. New York.

Dawood, H. (2005). Teaching test presentation strategies. *Education Magazine, Qatar National Committee for Education, Culture and Science, 34* (125), 102.

DeGruijter, D. N. M., & Van Der Kamp, L. J. Th. (2005). *Statistical test theory for education and psychology*. © D. N. M. de Gruijter & L. J. Th. Van der Kamp.

Ebel, R. L. (1972). *Essentials of educational measurement* (1sted.). New Jersey; prentice Hall Inc: 522.

Embretson, S. E. & Reise, S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ, US: Lawrence Erlbaum Associates Publishers.

Erguven, M. (2014). Two approaches to psychometric process: Classical test theory and item response theory. *Journal of Education*; 26.

Finch, H. W., Immerkus, C.J., & Frensh, F.B. (2016). *Applied psychometrics using SPSS and AMOS*. Information Age publishing, INC.

Góomez - Benito, J., & Navas - Ara, M. (2000). A comparison of x2, RFA and IRT based procedures in the detection of DIF. *Quality & Quantity, 34* (1), 17 - 31.

Hamad, D, F, A. (2010). The relationship of test wiseness to achievement test performance with a multiple-choice test constructed according to the Rasch model among female college of education students in the literary departments at Umm Alqura university. *Journal of Arab Studies in Education and Psychology,4*(4) Umm Alqura University, 297-338.

Hamadneh, I. M., & Bani Kh. M. S. (2013). Constructing a scale of attitudes towards cyberbullying among a sample of social media users at Al albayt university. *Almanarah Journal*, *19*(3).

Hambleton, R., & Rogers, J. (1995) Item bias review.*Research & Evaluation,13*(7). http://pareonline.net/getvn.asp?v=13&n=7

Hassanein, M. S. (2001). *Measurement and evaluation in physical education*. (Part 1), (4thed.). Cairo: Arab Thought House.

Hattie, J. (1985). Methodology review: Assessing unidimensionality of tests and items. *Applied Psychological Measurement, 9*(2), 139-164.

Hidalgo, M. D., & Gomez-Benito, J. (2010). Education measurement: Differential item functioning. In P. Peterson, E. Baker, & B. McGaw (Eds.) *International encyclopedia of education* (3rd ed.). USA: Elsevier – Science & Technology.

Howard, B. L. (2003). *Test scores and what they mean*. Englewood Cliffs, NJ: New York, Prentice–Hall. 62-63.

Hulin, C., Drasgow, F., & Parsons, C. (1983). *Item response theory: Application to psychological measurement*. Homewood IL: Dow Jones - Irwin.

Kaafarani, A. (1988*). Dress norms in nadhriyyah discourse: A sociolinguistic analysis of the modesty of Saudi women's clothing*. (Master's thesis), Sultan Qaboos University, Oman.

Kim, S., Cohen, AS., & Kim, H. (1994). An investigation of Lord's procedure for the detection of differential item functioning. *Applied Psychological Measurement, 18*(3), 217-228

Marshall, C. J. (1972). *Essentials testing Addison Wesley* (1st ed.). California, U. S. A, 104.

Masters, G. N., & Wright, B. D. (1984). The essential process in a family of measurement models. *Psychometrika*, *49*, 529−544.

Michaelides, M. P. (2008). An illustration of a mantel-haenszel procedure to flag misbehaving common items in test equating. practical assessment.*Research & Evaluation,* *13*(7). https://pareonline.net/getvn.asp?v=13&n=7

Osterlind, S. (1983). *Test item bias*. Beverly Hills; Sage publications.

Rebecca, Z., Dorothy, T. & John, M. (1997). Descriptive and inferential procedures for assessing differential item functioning in polytomous items. *Applied Measurement in Education*, *10*(4),321-344.

Saleh, A. A., & Obaid, M. A. (2020). Test wiseness and its relationship with attentional control among graduate students. *Alqadisiyah Journal for Humanities*, *23*(4), 121-148.

Salubayba, T. M. (2013). Differential item functioning detection in reading comprehension test using mantel-haenszel, item response theory, and logical data analysis.*The International Journal of Social Sciences, 14*(1), 76-82.

Swaminathan, H. & Rogers, J. (1990). Detecting differential item functioning using logistic regression procedures. *Journal of Educational Measurement, 127*(4).

Thorndike, L. R. & Elizabeth P. H. (1977). *Measurement and evaluation in psychology and education*. (4thed.), New York: John Wiley &amp; Sons: 82.

Tigza, A. B. (2012). *Exploratory and confirmatory factor analysis: Concepts and methodology using SPSS and LISREL*. Amman: Dar Almaseera for Publishing, Distribution, and Printing, p. 90.

Uiterwijk, H. & Vallen, T. (2005). Linguistic sources of item bias for second generation immigrants in Dutch tests. *Language Testing*, *22*(2), 211 - 234.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Erlbaum.

Woods, C. (2009). Empirical selection of anchors for tests of differential item functioning. *Applied Psychological Measurement*, *33*(1), 42-57.

YAN, S. (2005). *Gender-related differential item functioning in mathematics assessment on the third international mathematics and science study-repeat (TIMSS-R)*. The University of Toledo, ProQuest Dissertations Publishing, 2005. 3177610.

Zakri, A. M. (2020). Identifying differential item functioning of the "EMBU" test of parental rearing styles among a sample of secondary school students. *Journal of the Faculty of Education, Al-Azhar University*, *39*(3), 677-720.

Zumbo, D. B. (1999). *A handbook on the theory and methods of differential item functioning (DIF): Logistic regression modeling as a unitary framework for binary and Likert- type item scores*. Ottawa:Directorate of Human Resources Research and Evaluation, Department of National Defense.