Journal of the Association of Arab Universities for Research in Higher Education (مجلة اتحاد الجامعات العربية (للبحوث في التعليم العالي

Volume 43 | Issue 2 Article 19

2023

Constructing a criterion-referenced test in psychometric subjects according to item response theory

Sami Al-Massarweh
Isra University- Jordan, msarwahsami@gmail.com

Fatima Al-Qeyam
Isra University- Jordan, fatima.al-qeyam@iu.edu.jo

Ahmad Hani Al-qatawneh
Isra University - Jordan, ahmad.qatawneh@iu.edu.jo

Contents lists available at: Digital Commons

مجلة اتحاد الجامعات العربية للبحوث في التعليم العالي

Journal of the Association of Arab Universities for Higher Education Research (JAARU-RHE)

Journal homepage: https://digitalcommons.aaru.edu.jo/jaaru_rhe/

Constructing a criterion-referenced test in psychometric subjects according to item response theory

بناء اختبار معياري مرجعي في موضوع القياس النفسي وفق نظرية استجابة البند

Sami Al-Massarweh*
Assistant Professor
Psychology, Faculty of Arts
Isra University, Jordan
Sami.almassarweh@iu.edu.jo

سامي مصاروة
أستاذ مساعد
علم نفس ، كلية الآداب
جامعة الإسراء- الأردن

Fatima Al- Qeyam
Assistant Professor
English Language and Linguistics, Faculty of Arts
Isra University - Jordan
fatima.al-qeyam@iu.edu.jo

فاطمة القيام
أستاذ مساعد
أدب انجليزي ولغويات، كلية الآداب
جامعة الإسراء- الأردن

Ahmad Hani Al-qatawneh.
Associate professor
History of International Relations, Faculty of Arts
Isra University - Jordan
ahmad.qatawneh@iu.edu.jo

أحمد هاني القطاونة
أستاذ مشارك
تاريخ العلاقات الدولية، كلية الآداب
جامعة الإسراء- الأردن

Received: 31/10/ 2022 Accepted: 28/11/ 2023 Published: 15/06/ 2023

Abstract

The article aimed to construct criterion-referenced in the psychometric subject based on item response theory. The sample consisted of (121) participants (54 male & 67 female) selected from the Department of Psychology at Isra University (Jordan) during the second semester of the academic year 2021/2022. The criterion-referenced test consisted of (36) items following a multiple-choice shape in which each item has 4 options. The results showed the assumptions of items response theory in the study data and matched the responses to (34) items. The results found that there were two items that did not match the model of item response theory that were deleted. Finally, the results of the parameter assessments of the items (discrimination, difficulty, and estimation) indicated that they were agreeable with the test criteria mentioned in psychometric literature and educational. In light of the results obtained, the study recommended using the test, which was developed by the researchers to assess student's achievement in the subject of psychometrics for students of Psychology, because it demonstrates acceptable validity and reliability and complies with the requirements of the logarithmic three-parameter model, and the possibility of using the same methods as assessments for other courses after ensuring their psychometric qualities.

Keywords: Psychometric subject, item response theory, criterion-referenced test.

المستخلص

هدف البحث بناء اختبار معياري مرجعي في موضوع القياس النفسي بناءً على نظرية استجابة البند. وقد تكونت العينة من (121) مشاركا (54 ذكر و67 أنثى) تم اختيارهم من قسم علم النفس بجامعة الإسراء (الأردن) خلال الفصل الدراسي الثاني من العام الجامعي 2022/2021. وقد تكوّن الاختبار المرجعي المعياري من (36) عنصرًا يتبع شكل الاختيار من متعدد حيث يكون لكل عنصر 4 خيارات. وقد أظهرت النتائج فرضيات نظرية استجابة المفردات في بيانات الدراسة ومطابقتها مع (34) فقرة، وكشفت النتائج أن هناك عنصرين لا يتطابقان مع نموذج نظرية استجابة البند، فتم حذفهما. وأشارت نتائج تقييمات العناصر (التمييز، الصعوبة، التقدير) إلى أنها كانت مقبولة ضمن معايير الاختبار المذكورة في الأدبيات السيكومترية والتعليمية، وفي ضوء ذلك، تمت التوصية باستخدام الاختبار الذي طوره الباحثون لتقييم تحصيل الطلاب في مادة القياس النفسي لطلاب قسم علم النفس، لأنه يوضح مصداقية وموثوقية مقبولة ويتوافق مع متطلبات النموذج اللوغاريتمي ثلاثي المتغيرات، وإمكانية استخدام نفس الأساليب المستخدمة في التقييمات لمقررات أخرى بعد التأكد من صفاتها السيكومترية.

كلمات مفتاحية: موضوع القياس النفسي، نظرية استجابة البند، اختبار مرجعي للمعيار.

Introduction

University education aspires to develop a set of students who possess knowledge, skills, values, and experiences, and who will use "these skills and knowledge to employ and discover new applications that touch on" society's natural reality, particularly students in the Department of Psychology, who would have been a new boom that transforms the local labour market. The educational psychometric process is based on instruments and assessments that provide quantitative data that enable educators to precisely understand educational phenomena. In order to ascertain the extent to which educational objectives are met and students' achievement levels, performance tests were developed as a necessary and integral part of the educational process. Through the numerous sorts and formats of examinations, we amass the database upon which we base our judgments and disclose pupils' strengths and shortcomings in order to classify them.

A person's performance is judged by how well he or she does on benchmark tests, which do not need to be compared to how well his or her colleagues do. These tests measure levels that can be viewed in terms of levels that require an accurate understanding of the behavioural domain that they measure, not just looking at the outcomes achieved compared to the performance of his or her colleagues (Allam, 2011). "Standard reference tests compare the performance of an individual with his peers, while benchmark tests compare performance to a specific level without taking into account the performance of other people". The goal of standard reference tests is to make it easier to compare people in the field that the test measures. As with criterion-referenced tests, the goal is to look at how well the person did on a specific set of skills (AL, Dosari, 1999).

Tests like this one have been linked to a new way of assessing things like educational and psychological tests called the Item “Repose Theory” IRT. It tried to fix the problems with the classical theory. It can be done to "get rid of the criticisms that were made about the analogy classical theory" posited by (Hambleton & Swaminathan,1985) which are: The psychometric characteristics of the test are calculated for a group of people, and the characteristics of individuals can only be calculated using a particular test and not in absolute terms (Rogers, Hambleton & Swaminathan, 1999).

The Study Problem

Recently, researchers have demonstrated a strong interest in item response theory since it attracted the attention of test administrators and was used to develop a variety of tests, mental and accomplishment assessments, and attitude scales. Because the researchers taught undergraduates at Al-Isra University, it became clear that there aren't any standard criteria for evaluating students' work that is made up of correct objective scientific methods.

Fan (2009) emphasized that the Rasch model (single-parameter) is more accurate in estimating both ability and difficulty compared to the other two- and three-parameter models, while Courvill (2004) proved that the binary model was the best of the three models and the lowest in terms of error rates, as Al-Akalia (2007) proved that the three-parameter model is the most accurate model in selecting data compared to the other two models.

In light of the abovementioned, the researchers developed a special measure in light of the item response theory to assess the students of the Department of Psychology at Al-Israa University.

Questions of the study

This study will attempt to address the following questions:

1. To what extent do the three-parameter model's assumptions about item response theory hold true in the current study's data?
2. To what extent do individual answers correspond to the three-parameter model?
3. How were the three parameters of the test items (discrimination, guessing and difficulty) estimated using the item response theory's three-parameter logistic model?

Significance of the Study

1. Examine the item response theory's assumptions in connection to elucidating the influence of the subjects' characteristics and the items on the accuracy of the subjects' parameters and the assessment of the item.
2. This study used a logarithmic three-parameter model to produce criterion-referenced test in order to demonstrate the widespread interest in this sort of exam across many individuals.

Objectives of the study

This article will attempt to verify the following:

1. The extent to which the three-parameter model's assumptions about item response theory hold true in the current study's data.
2. The extent to individual answers corresponds to the three-parameter model.
3. The extent to which the three parameters of the test items (discrimination, guessing and difficulty) were estimated using the item response theory's three-parameter logistic model.

Terminology

Criterion-referenced test: This sort of exam is used to evaluate the performance of individuals in regard to a criterion (performance absolute level) without requiring comparison to other people's performance (Allam, 2011).

Al-Ajili (2005) describes it as "The test in which the student's level is determined in relation to a fixed criterion (level) without reference to the performance of others, and this level is usually related to the behavioral goals of the subject." It is described procedurally as a collection of multiple-choice questions with four choices and one right response produced according to the principles of item response theory and analyzed using the taught triple model to ensure the correctness and objectivity of assessing student success.

Item Response Theory: This is a modern theory of psychological and educational measurement in which the connection between the respondent's performance and the underpinning characteristic of the subject of the measurement is decided by a specific mathematical function. Latent trait models are used to explain how well people do on the vertebrae (Hambleton & Swaminathan, 1985). These models say that how well people do on the vertebrae is linked to how good they are.

Achievement test: A organized technique that assesses students' knowledge and ability to apply subject-matter information and abilities by having them respond to a series of paragraphs that cover the academic subject's material (Odeh, 2001).

Al-Abadi (2006) describes it as "a set of vocabulary (questions) that are given to the student to answer them verbally or in writing, and it may be objective, essay, drawings or forms used for comparison and measurement."

Literature review

Advantages of the criterion-referenced tests:

1. The aim of the contextual reference tests is not just to check how well a student is doing on his or her own, but also how well the school is doing and how well students are chosen for graduate school, for example (Wikstrom, 2005).
2. Criterion-referenced test rely mostly on figuring out a list of specific academic performance and "their specific levels of achievement. This type of test is useful when different educational institutions offer different content to their students because it tells the institutions that students must reach" certain levels of mastery in learning specific knowledge and skills (Wiberg, 2004).
3. The criterion-referenced test can help teachers and psychologists figure out why students do not know how to do certain things and what to do when they don't. They can also help teachers figure out what to do when students do not meet some of the objectives on which these tests are rooted and what to do when they do not meet the goals (Al-Anzi, 2004).
4. It does not have to be at the top of the characteristic continuum to use the term "criterion." Instead, the criterion-referenced test levels can be outlined at any point, and these levels will be used to explain tasks that the student should be able to do. The results were compared in light of these tiers, and

these levels can change from period to period as the student grows (Abdul Salam, 1996).

5. The Item Repose Theory, or I R T, was one of the first theories in educational measurement and evaluation. This theory was able to fix a lot of the problems with the classical theory, and it came up with good ways to deal with things like calibrating items, making question banks, and making criterion referenced test (Hambleton & Swaminathan, 1985).

So, the item response theory helps to solve "problems that the classical theory could not solve, like adaptive measurement, question banks", and independent capacity estimates for people who were exposed to a sample of items. Lord (1980) says that the independence of measurement is the main difference between both the "classical theory and the item response theory" (Rup & Zumbo, 2006). The following are some of the fundamental assumptions upon which paragraph response theory is founded, as defined by (Hambleton & Swaminathan, 1985; Hambleton & Jones, 1993) and alluded to in Al-Sherifin 2006: This suggests that there is a single capability that explains uniformity:

1. Assumption of one dimension: one-dimensional models are used to describe an individual's performance on a test (Local Self-Sufficiency).

2. Assuming that your position is not important: it means that the person's response to one item does not affect how he or she responds to other items. This means that the responses of people to the test items are statistically independent.

3. Assuming Item Characteristic Curve: The item characteristics curve is a mathematical function that links the probability of a correct answer for paragraph P (∅) and the subject capacity (∅) measured by a set of items in the test that was constructed.

4. Speediness: that is, participants who fail to respond to test objects do so because to their inadequate ability, not due to a lack of time to reach and respond to the item. Some of the most important models in the item response theory are shown here:

1. "A single-parameter logistic model: the one-parameter model is one of the most basic and most often used models of response theory to items". It is stated mathematically as the relationship between the likelihood of a responder giving the correct answer to an item and the chance of the respondent giving the incorrect answer. Hambleton and Swaminathan (1985) came up with a model that says that all items have the same power, and that the parameter is assumed to be zero for all items. This model also says that the numerical function that reflects probability is shown in this equation:

$$P_i(\theta) = \frac{e^{D(\theta - bi)}}{1 + e^{D(\theta - bi)}}, i = 1,2,3,\ldots\ldots,n$$

"i when (∅): the probability of a correct answer for an individual whose ability pi (∅)"

"(D) Scaling Factor equals (1.7)".

"∅ = Individual capacity".

"bi: the difficulty parameter of paragraph I".

"ai: items highlighting parameter".

This model is based on the difference between the responder's ability, () as indicated by the letter "S" in the characteristic to be evaluated (the ability underlying the subject's response) and the degree of the item "I" to which the respondent wishes to reply (represented by the symbol (bi)) (Al Taqi, 2009).

Figure 1: presented of the three items curves, which represent its significant logarithm, include one parameter for represents the differences between the items, namely the item difficulty, (bi) item difficulty parameter, it showed the individual ability when the probability of the item correct answer = (0.5) i.e. (P (θ) = 0.5). it means that the item difficulty value of answering any items probability = 0.5 (Crocker & Aljina, 2009).

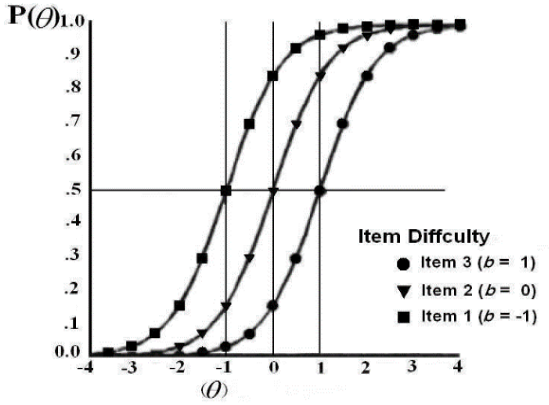


Figure 1: showed the three items' curves in single parameter model

Source: (Crocker & Aljina,2009)

2. "A two-parameter logistic model": this model suggests that both the discrimination and difficulty parameters are changeable, "that the guess for all items is zero, and that its calculations are more complex than those of the parameterized single model". The following equation expresses this model:

$$P_i(\theta_j) = \frac{e^{Da_i(\theta_j - bi)}}{1 + e^{Da_i(\theta_j - bi)}} \quad , \quad i = 1,2,3,\ldots\ldots,n$$

The incline of the curve grows as (ai)); The curves' tendency is not identical and overlap because the discrimination parameter is changeable "for each item in this model (Lord, 1980: 125)", as shown in Figure (2).

ai: items highlight parameter

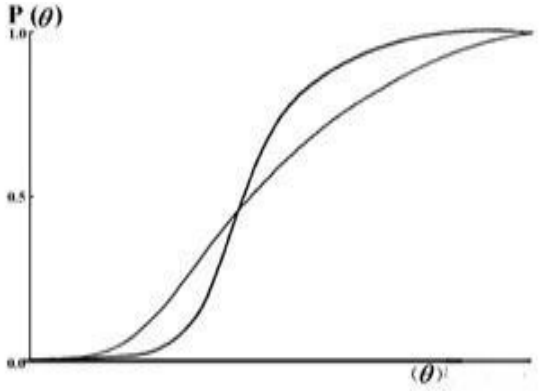


Figure 2: Two curves of Two two-items signify in the two-parameter model.

Source: (Lord,1980)

3. "Logistic model with three parameters: It is graded according to three criteria: Difficulty, Discrimination, and Guess. What separates this model out from the two-parameter model is the addition of the parameter of guessing, as expressed in the following equation":

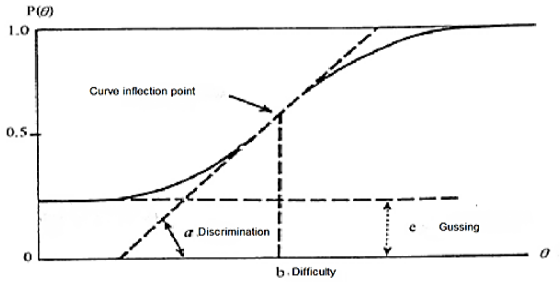
"$P_i(\theta_j) = c_i + (1 - c_i)\frac{e^{Da_i(\theta_j - bi)}}{1 + e^{Da_i(\theta_j - bi)}}, i = 1,2,3,\ldots\ldots,n$ "

"The characteristic curve of the three-feature model function" is depicted in Figure (3), along with three parameters: the estimation parameter (c), which indicates the height of the "asymptote below the curve, and the parameter b)), which indicates the complexity of the item at the point of inflection of the curve, that is, when the probability" of receiving the correct answer equals c + 1/2;. The item discrimination strength parameter (a) is the angle of the curve at its inflection point, which shows the variation of the response level on the item in relation to the power level. (Lord, 1980; Toland, 2008; (11).

Figure 3: presented the "curve item in the three-parameter model

Source: (Lord, 1980: 124).

As a result of the foregoing, it is clear how far modern measurement theory has



advanced in its liberation from the effect of individual characteristics on item parameters and from the effect of item features on individual characteristics when developing tests, as well as the numerous studies that have been conducted on this subject. Bani Atta (2019) published a paper titled "The Effect of Item Characteristics and Individuals' Ability to Assume Positional Independence in Item Response Theory." It sought to investigate the relationship between the item's characteristics and an individual's ability to assume local independence.

To accomplish this, responses from (1000) "individuals were generated on (100) items, and after analysing them using a two-parameter logistic model to determine the item's parameters and the ability of individuals, four

models of tests were formed according to the item's level of difficulty, discrimination, and inability of individuals". Al-Balawi's & Abd al-Aal's (2018) study objective was to determine the influence of test time on item response theory and classical theory estimations of the item's parameters. A multiple-choice achievement exam was constructed using 28 questions from the mathematics course, and three different reaction time images were created for the test. It was administered to a sample of (451) students, and then the results were analyzed using the (ITEMAN, (BILOG-MG3) software. "The findings indicated agreement between item response theory and classical theory". "There were statistically significant differences (-0.05) in the averages of estimating discrimination features due to time in favor of the appropriate time (30 minutes), as well as significant differences in the accuracy of estimating difficulty and discrimination parameters in favor of the second image (the appropriate time 30 d), which is the most accurate in estimating these parameters according to item response theory" and the third image has a percentage of 68 percent.

According to Al-Nasraween (2019), "the purpose of his study was to demonstrate the effect of the number of alternatives in a multiple-choice test on the item information function and the test using a triple model parameterized within the context of item response theory. The purpose of the study was verified by developing a multiple-choice achievement test in the second part of the mathematics subject for tenth grade students in government schools in the capital governorate". The study sampled (1530) tenth-grade students and discovered "statistically significant differences in reliability between the five- and four-alternate models, as there were no statistically significant differences in the arithmetic means of the information function due to the variable number" of alternative items. Among the studies examining the criterion referenced test, Al-Mutairi's (2018) study used item response theory to develop a criteria reference test for assessing the research efficiency and competence of students at the College of Education. Where the reference's narrated news was constructed, its items focused on several areas related to scientific research, including types of research according to methodology, components of research, types of data for variables, academic research, educational research instruments, types of samples, psychometric characteristics of educational research, and statistics. The test, which consisted of (40) items, was administered to a sample of (362) male and female students. The study developed a test consisting of (37) questions to assess research abilities using the single model, since it has validity due to its conformance to the Rush model's assumptions and am excellent stability score for both the item's difficulty and the individual's ability.

Tokhi (2017) calibrated the Quality of Life scale using Item response theory, and developing norms that help interpreting the examinee levels estimated by the total scale. The sample of the study consisted of (600) male students from Umm Al Qura University. The study tool represented in the Quality Of Life Scale for university students designed by (Muncie, and Kazim, 2006) which consists of 60 items measuring 6 dimensions. The study used the Winsteps (version 3.67) program for calibrating the scale according to Item response theory. The study results to calibrator the items of the Quality of Life scale on one linear metric. The process of calibration involved eliminating (14) items of the scale that showed statistical misfit to Item response theory, so the calibrated scale consisted of (46) items, and possible to calculate the person measures corresponding to each possible total score on the scale in its final form, and was able to study the verification of the reliability and validity of the scale finalized after calibrate using Item response theory, make the measurement standards of scale that can be used to interpret the estimated measures of examinees on the total scale.

Al-Shehri (2018) used the "one-parameter logistic model (Rush model)" to develop a criterion referenced test to assess "mathematical competencies in engineering thinking. An achievement test measuring the

levels and abilities of engineering thinking was developed, consisting of (38) multiple choice items". The study sampled (480) students in the middle third grade in governorate of Al-Jawf. Additionally, the results demonstrated that items met several requirements in the theoretical literature for measurement, and the results shown for the research sample individuals' estimations, removing (3) persons who do not fit the model. The study sample's ability characteristics varied from (-1.389) to (1.674).

Afolabi's (2015) study attempted to examine whether it is possible to use the eighth-grade criterion referenced test -language mathematical aptitude test to "predict student behaviour in higher school. The study employed a quantitative approach with a correlative design to ascertain whether there are relationships between selected variables and the students' semantic test grades". The variables examined were successful in the eighth-grade mathematics course. The study's results were examined statistically. The study's findings indicated a highly positive correlation between students' scores on a criterion-referenced test mathematical aptitude exam and their achievement in a ninth-grade mathematics course or topic. Additionally, the study's findings indicated a positive correlation between students' scores on a criterion-referenced test mathematical aptitude test and their success on the Georgia test for high school graduation. These findings emphasize the importance of developing appropriate predictions or indicators to forecast students' future behaviour. The study's findings indicated that this should be done immediately to lower the chance of children dropping out of high school.

Methodology

The study relied on the analytical descriptive approach and a number of statistical methods using Statistical Package Program for social sciences (SPSS).

The population

The study population consists of the students of psychology department at University of Isra - Jordan, during the second semester of 2021/2022.

Sample of the study

Sample of the study made up 121 participants (54 male & 67 female) was selected from different academic level. The participants are Arabic language native speakers. Also, there are homogenous in age rating between 18- 30 years. Moreover, the participants represent the almost of social classes, i.e., middle, low and high class.

The Study Tool

To accomplish the study's objective, the following procedures were taken during the psychometric course:

1. "Determining the content: the content was determined" in order to identify the fundamental concepts and characteristics of the psychometric course, and the vocabulary includes the following: the notion of measurement and its significance, grades of measurement and their types of psychological tests and measures characteristics, statistics in educational and psychological measurement, central tendency and dispersion measures, types of criteria test, and the tests psychometric characteristics.
2. Defining the behavioural objectives that were addressed in the topics and that students were expected to master.
3. A specification table for the test has been created to connect content to target levels.
4. (41) Multiple-choice questions with four options and one right answer were developed, taking into account the prerequisites for developing this type of test item and their content and cognitive level compliance with the behavioural aim they assess.

The scale of psychometric characters

"To ensure content validity, the researcher presented the 41 test items, along with a list of behavioural objectives, content, and a table of specifications, to several experts and specialists in measurement and evaluation at the university. To ensure scientific accuracy and to accomplish the research objectives, the researcher relied on the opinions of experts and referees" (Al-Batsh and Abu Zina, 2007). Global validity was determined for the test data by doing

principal component analysis and obtaining the Eigen value for each factor, as well as noting the Explained Variance for each factor. Related to the reliability of the scale, the researcher has examined the scale reliability based on Cronbach Alpha and Split-half analysis method as shown in the following table:

Table 1: "results of Cronbach Alpha and Split-half analysis"

Variable	Split-half	Cronbach Alpha
Scale	0.93	0.90

As illustrated in table (1) the results showed that the "scale enjoying an excellent reliability grade, which means the stability of the scale for purpose of the study".

Results of the study

First question: "To what extent are the assumptions of item response theory fulfilled in the current study data according to the three-parameter model?"

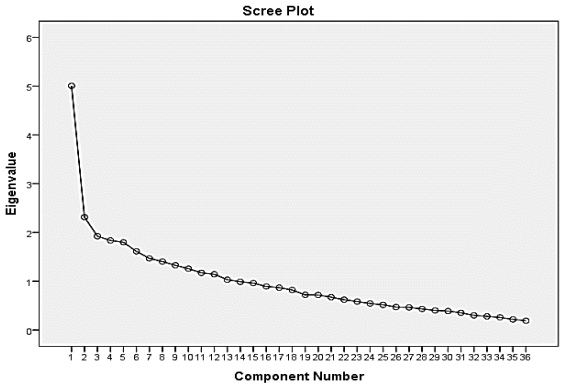
1. The assumption of unidimensionality: The factor analysis was performed on the final test data for (115) male and female students on (36) items using spss statistical programming; the analysis produced (13) factors, of which the first (13.908 per cent) was explained by the variance, and the remaining factors (64.740 per cent) explained the remaining variance. The values of the entire root and the explained variance ratio for each factor, as well as the cumulative interpreted variance ratio, are shown in Table (2)

Table 2: Results of factor analysis of participants response on (36) items

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings
	Total	**Variance%**	**Cumulative%**	**Total**	**Variance%**	**Cumulative %**	**Total**
1	4.01	12.90	12.90	4.00	12.90	12.90	2.24
2	3.32	5.42	19.33	1.31	5.42	19.33	1.00
3	2.91	4.34	24.68	2.92	4.34	24.68	2.85
4	2.82	4.10	31.78	2.83	4.10	31.78	2.84
5	1.79	5.00	35.78	1.80	5.01	35.78	1.72
6	1.60	4.48	40.26	1.61	4.48	40.26	1.70
7	1.46	4.08	44.34	1.47	4.08	44.34	1.66
8	1.41	3.90	48.25	1.40	3.90	48.25	1.65
9	1.33	3.68	51.93	1.32	3.68	51.93	1.63
10	1.26	3.49	55.43	1.25	3.49	55.43	1.60
11	1.18	3.25	58.69	1.17	3.25	58.69	1.54
12	1.15	3.18	61.87	1.14	3.18	61.87	1.49
13	1.04	2.86	64.74	1.03	2.86	64.74	1.35
14	0.99	2.75	67.49				
15	0.96	2.67	70.16				
16	0.89	2.48	72.65				
17	0.87	2.41	75.06				
18	0.82	2.28	77.35				
19	0.72	2.00	79.35				
20	0.71	1.99	81.35				
21	0.67	1.87	83.22				
22	0.62	1.72	84.95				
23	0.58	1.62	86.57				
24	0.54	1.51	88.09				
25	0.51	1.43	89.52				
26	0.47	1.31	90.83				
27	0.46	1.29	92.13				

Component	Initial Eigenvalues			Extraction Sums of Squared Loadings			Rotation Sums of Squared Loadings
	Total	**Variance%**	**Cumulative%**	**Total**	**Variance%**	**Cumulative %**	**Total**
28	0.43	1.20	93.33				
29	0.40	1.12	94.45				
30	0.39	1.08	95.53				
31	0.35	0.98	96.52				
32	0.30	0.83	97.35				
33	0.28	0.78	98.13				
34	0.25	0.71	98.85				
35	0.21	0.61	99.46				
36	0.19	0.53	100.01				

Table (2) showed that the first factor latent root = (5.00), which explains 13.90% of total variance. the result of (first latent root /second latent root) = (2.16). Since this value is more than (2.00), it provides that the first factor is a control factor, which is confirms the unidimensional assumption. The results in above table Achieve a one-dimensional assumption the final form of the test, furthermore the test measures a single characteristic, and the one-dimensional assumption can be enhanced by representation of latent roots using Scree Plot as illustrated in following Graph.



B- "Assumption of Local Independence: because the one-dimensional assumption is" satisfied, this implies that the local independence assumption is satisfied, as the one-dimensional assumption is identical to the local independence assumption.

C- "Speediness assumption: The students took sufficient time to" respond to the test items, and the researcher received no notes about a lack of time during the test's administration, indicating and demonstrating that the student's inability to respond to the test items is due to their low abilities, not to the effect of the speed factor.

Second question: "To what extent do the individuals' responses correspond to the three-parameter model?"

The participants' data were input for the final exam's (36) items using the Bilog-MG3 application. The study revealed that all individual responses were similar to the model except for (5) students, for whom the "Chi-Square values were statistically significant" at level 0.05. The analysis was re-run using the same tool to determine the extent to which the model's test items were fulfilled. The Chi-Square test at a significance level of 0.05 revealed that two items with the numbers (14 and 23) were not similar, despite the probability value being less than 0.05. Table (3) Conduct an evaluation of these indicators.

*Table 3***:** Chi-square results and "significance level for the three-Parameters Logistic Model"

No.	Chi-square for good matching	Sig	No.	Chi-square for good matching	Sig
1	12.41	0.088	19	5.32	0.502
2	21.14	0.140	20	19.69	0.265
3	6.82	0.454	21	3.83	0.208
4	18.57	0.365	22	16.78	0.075
5	4.78	0.582	23	13.04	0.001
6	9.32	0.122	24	12.87	0.077
7	12.17	0.094	25	12.34	0.419
8	11.73	0.106	26	24.45	0.096
9	9.78	0.288	27	11.26	0.293
10	4.19	0.764	28	14.33	0.166
11	5.32	0.502	29	14.02	0.065
12	11.68	0.754	30	5.94	0.751

13	12.77	0.895	31	6.38	0.503
14	17.86	0.011	32	4.66	0.331
15	7.85	0.446	33	10.33	0.155
16	3.88	0.960	34	14.19	0.903
17	10.94	0.526	35	7.60	0.185
18	8.71	0.187	36	7.31	0.423

Through Table (3), "the Chi-Square value and its statistical significance are shown at the significance level $\alpha \leq 0.05$, and it is found that (34) test items out of the total items (36) are not statistically significant, and this indicates their conformity to the logarithmic three-parameter model".

Third question: "What are the values of the estimates of the parameters of the test items (difficulty, discrimination, guessing) for the test item according to the 3-parameter logistic model of item response theory?"

To address this question, we used the Bilog-MG3 program to determine the paragraph's properties (difficulty, discrimination, and estimate) and standard error. The evaluation of these estimates for the test items in their final form (34) is shown in Table No. (4). (After deleting the non-conforming items).

*Table 4***:** Values of "parameters of the items and the standard error of the test in its final form according to the three-parameter model".

N	Difficulty parameter	Standard error of the difficulty parameter	Discrimination parameter	Standard error of the discrimination parameter	Guess parameter	Standard error of the guessing parameter
1	-0.131	0.032	1.653	0.225	0.153	0.047
2	0.423	0.024	0.768	0.121	0.226	0.036
3	1.331	0.103	1.183	0.232	0.205	0.023
4	0.168	0.256	0.654	0.088	0.132	0.061
5	1.411	0.120	0.892	0.244	0.240	0.025
6	-0.821	0.217	0.609	0.145	0.212	0.040
7	0.860	0.191	0.878	0.171	0.207	0.022
8	0.667	0.165	1.882	0.293	0.187	0.042
9	0.681	0.254	0.932	0.236	0.258	0.046
10	-0.600	0.073	1.634	0.160	0.242	0.059
11	-0.981	0.268	0.485	0.073	0.244	0.030
12	0.186	0.080	1.275	0.434	0.411	0.027
13	-0.971	0.096	1.364	0.190	0.163	0.048
15	0.751	0.419	1.550	0.186	0.173	0.051
16	0.645	0.206	1.828	0.174	0.313	0.052
17	1.033	0.108	0.961	0.116	0.216	0.061
18	1.005	0.152	0.898	0.136	0.420	0.060
19	0.782	0.146	0.481	0.059	0.242	0.081
20	1.692	0.565	0.886	0.144	0.321	0.041
21	0.680	0.115	1.051	0.166	0.202	0.034
22	0.362	0.278	0.757	0.083	0.212	0.075
24	0.976	0.111	0.890	0.135	0.222	0.034
25	-0.882	0.176	1.183	0.103	0.255	0.085
26	-1.291	0.125	1.212	0.415	0.206	0.066
27	-0.718	0.090	0.891	0.148	0.240	0.037
28	1.517	0.123	1.252	0.270	0.252	0.028
29	-0.153	0.084	0.891	0.205	0.214	0.023
30	0.648	0.240	0.783	0.192	0.310	0.047
31	-0.835	0.157	1.118	0.181	0.196	0.032
32	1.145	0.241	0.794	0.148	0.203	0.040
33	-1.361	0.071	0.840	0.165	0.186	0.071
34	1.513	0.124	1.258	0.279	0.253	0.028

As shown in Table (4), the difficulty parameter values were between (-1.31 and 1.692), with a mean (0.304), the difficulty parameter's standard error value varied between (0.025 and 0.565), the discrimination parameter value varied between (0.485 and 1.882), with mean of 1.054, and the discrimination parameter's standard error value was between (0.059 and 0.42), and for (0.2359), The standard error of the estimated parameter was between (0.020 and 0.083), and Ree (1983) verified that the discrimination parameter normal values in the model of logarithmic three-parameter were between (0.5 and 2.5), with an arithmetic mean of (0.5). (1). Hambleton (1989) asserts that the logarithmic three-parameter model is consistent when the difficulty parameter is between (-2 and 2), the discrimination parameter is between (0.4 and 2), and the guessing parameter is between (0.4 and 2). (0 and 0.25).

Recommendations:

Considering the results obtained, the researchers recommend the following:

1. Using the test, which was developed by the researchers to assess student's achievement in the subject of psychometrics for students in the Department of Psychology, because it demonstrates acceptable validity and reliability and complies with the requirements of the logarithmic three-parameter model.
2. Using the same methods as assessments for other courses and ensuring their psychometric qualities.
3. Conducting more studies on student evaluation in the light of item response theory

References

Abdel Salam, N. (1996). *Building a two-reference test and a reference group test (a comparative study) in Anwar Al-Sharqawi and his colleagues.* Contemporary Trends in Measurement, Psychological and Educational Assessment, Cairo, The Anglo Library (86-120).

Afolabi. A. (2015). *Using the criteria referenced competency tests to predict student success in high school*, Unpublished Doctoral Dissertation, Capella University, USA

Al-Abadi, R. (2006). *School tests*. Arab Society Library, Amman, 1st floor

Al-Ajili, S. (2005). *Educational measurement and evaluation*. Education Centre for Printing and Publishing, College of Education, University of Sanaa, Yemen, 3rd Edition.

Al-Akaila, A. Nasser, (2007), *A comparative study between the classical theory and the modern theory of measurement in detecting and estimating the standard error in cognitive abilities tests for basic stage students in Jordan*, Unpublished PHD dissertation, Cairo University, Egypt.

Al-Anzi, H. (2004). *Building an achievement test to measure the degree of mastery of basic skills in mathematics for upper grades of elementary school students*, an unpublished master's thesis, King Saud University. Faculty of Education.

Al-Balawi, F. & Abdel-Aal, M. (2018). The effect of test time on estimating the parameters of the items and the stability of the test. Comparative study: item response theory and the traditional theory, *International Journal of Educational and Psychological Studies* / Volume 4-Issue 2 - 2018, pp. 252-pg 276.

Al-Batsh, M. & Abu Zina, F. (2007). *Scientific Research Methods*, Research Design and Statistical Analysis, 1st Edition, Dar Al-Masirah for Publishing, Amman.

Al-Dosari, I. (1999). A reference frame in the educational assessment of Arabic Gulf states. *Journal of Education, 9(31)*. 138 - 158.

Allam, S. (2011). *Educational and psychological measurement and evaluation*. Cairo: Arab Thought House, 5th Edition.

Al-Mutairi, T. (2018). Building a reflexology test for the research methods for students of the College of Education using the item response theory: The Rush model. *Journal of the College of Education in Psychological Sciences*, Volume 42, Issue 2, pp. 212 - p. 250.

Al-Nasraween, M. (2019). The function of item information, test, and stability when using three models of the multiple-choice test within the framework of the item response theory, *International Journal of Educational Research* / United Arab Emirates University - Volume (43) Issue (3) October 2019.

Al-Sharifain, N. (2006). Psychometric properties of the reference spoken test in measurement and evaluation according to the modern theory of educational and psychological measurement, *Journal of Educational and Psychological* **Sciences**, Volume (7). Issue (4).

Al-Shehri, A. (2018). Building a spoken-reference test to measure mathematical competencies in engineering thinking among middle school students according to the one-teacher logistic model. *Educational Sciences Studies* / University of Jordan, Volume 45, Issue 4, Appendix 2, pp. 63 - pg. 77.

Al-Taqi, A. (2009). *The Modern Theory of Measurement*, Dar Al Masirah for Publishing, Distribution and Printing: Amman, Jordan.

Bani Atta, Z. (2019). The effect of paragraph characteristics and the level of individuals' ability to assume positional independence in the item response theory, *The Jordanian Journal of Educational Sciences*, Vol. 15, No. 1, p. 99, p. 111.

Corville, T., (2004), *An empirical comparison of item response theory and classical test theory item –person statistics***,** Unpublished PHD dissertation, Texas University.

Crocker, L. & Aljina, J. (2009). *An Introduction to the Traditional and Contemporary*

Theory of Measurement (Zinat Dana's translation) Amman: Dar Al-Fikr (No year of publication of the original book is mentioned).

Fan, A., (2009), Item Response Theory Model an Empirical Comparison of Thir Item Statistics, *Educational and Psychological measurement*, Vol.44, No.2.

Hambleton, K & Swaminathan, H. & Rogers, H. (1999). *Fundamentals of item response theory*. Newbury park, CA: Sage Publication.

Hambleton, R. & Swaminathan, H. (1985). *Item Response Theory: Principles and applications.* Boston MA:Kluwer-Nyjhoff.

Hambleton, R. (1989). *Principles and selected applications of item response theory.* New York: Macmillan Publishing Company.

Lord, F. (1980). *Applications of Item Response Theory to Practical Testing Problems.* New Jersey: Lawrence Erlbaum's associates.

Odeh, A. (2001). *Measurement and evaluation in the teaching process*. Dar Al-Amal for Publishing and Distribution: Irbid, Jordan.

Ree, M. & Jensen, H. (1983). Effects of sample size on linear equating of item characteristic curve parameters. In D.J. weiss (Ed.), *New horizons in testing*. pp135-146. NewYork.

Rup, A.A. &. Zumbo, B.D. (2006). Understanding parameter invariance in unidimensional IRT models. *Educational and Psychological Measuremen*t, 66, 63-84.

Tokhi, Laila, (2017), Using Item response theory for calibrating Quality of Life Scale of the male students in Umm Al Qura University in Holly Makkah, *Journal of Arab Studies in Education and Psychology***,** No.90, Arab Educators Association.

Toland, M. (2008). *Determining the Accuracy of Item Parameter Standard Error of Estimates in BILOG-MG 3.* Unpublished doctoral dissertation, The University of Nebraska - Lincoln, AAT 3317288.

Wiberg, M. (2004): *Classical Test Theory vs. Item Response Theory an Evaluation of The Theory Test in the Swedish Driving-License Test*, Umea University, Department of Educational measurement.

Wikstrom, C. (2005). *Criterion-Referenced Measurement for Educational Evaluation and Selection*, Umea University.