Statistical indicators between the traditional and modern measurement theories in the compatibility of an achievement test of measurement and evaluation for students of the Faculty of Physical Education, Sadat City University

Dr/ Ahmed Rabie Mahmoud Saad*

Introduction :

Accordingly, the owners of the contemporary trend in educational and psychological measurement and evaluation focused their efforts to reach the highest levels of accuracy and objectivity in measurement so that the most accurate relationship between the measurement tool and the trait to be measured is achieved.

This type of test has been associated with a contemporary approach to educational and psychological measurement called paradigm response theory models, as it came with premises that address many of the negative aspects of the classical theory of measurement.

Measurement scholars distinguish between two main approaches to analyzing and grading items: the traditional theory of measurement, which is called the theory of true and false degrees. Criticisms of traditional measurement theory.

The traditional measurement theory provided solutions to some of the problems facing researchers in building and developing tests, but it failed to solve other problems, as it assumes that the standard error in measurement is equal for all subjects, and this assumption lacks accuracy, and an individual's the expression of ability is based on Through the real degree that is evident through his performance on the test as a whole, and not at the level of the paragraph, and therefore the status of the ability of the individual will change according to the change in the level of the test. And ease, and that theory is not suitable for building criterion reference tests

And since the traditional theory of measurement leads to the construction of inflexible tests, the measurement specialists directed their efforts to a more objective measurement system that focuses on selecting the test items better, and allows adding or deleting items to the test without affecting the test as a whole. The modern theory of measurement, as this theory is considered a revolution in psychological and educational measurement.

Modern measurement theory has helped provide many solutions to problems related to building and developing tests, especially with regard to the equivalence of tests and their equivalence, building referenced tests, building question banks, detecting paragraph bias, and so on. The theory also played an important role in analyzing test items, and thus Contribute to the evaluation of the

* Assistant Professor, Department of Fundamentals of Physical Education, Sadat City University

quality of these tests, and the great advantage of the paragraph response theory, is that it leads to paragraphs whose parameters do not change when the sample is changed, in addition to the mathematical complexities in the methods used to obtain these estimates. The current research focused on the two-parameter logistic model, as this model is based on the assumption that the paragraphs differ in their difficulty and distinction, and the absence of the guessing factor. The model is more realistic than the one-parameter logistic model. Because it is difficult to find a group of vertebrae with the same discriminatory ability at different levels of ability, and the mathematical equation for this model is:

$Pi(\Theta) = [1 + e^{-Dai(\Theta - bi)}]^{-1}$

Where:

Pi(Θ): The probability that the subject with the ability (Θ) will answer paragraph (i) correctly.

D: scaling factor, which is a constant for ce : is the natural logarithmic base and is equal to (2.718).

The two-parameter logistic model was relied upon in this research because the statistical indicators extracted from this model are more consistent with traditional indicators compared to other models (Hernandez, 2009), in addition to the preference of this model in selecting items (Pelton, 2002).

The issue of distinguishing between standard-referenced tests and standard-referenced tests is not easy, as the same test can be used in both cases, and the main difference between them lies in the issue of interpretation of the results. And in this regard That the similarities between them may have been clearer than the differences, and this was shown through several points, the most important of which are: that both types were prepared to measure specific teaching objectives, but the referenced test requires a high degree of harmony between the objective and the paragraph that measures it.

The reality of education is no longer limited to simply distinguishing between students in the measured characteristic, but rather it must focus on their acquisition of certain skills and achievement of specific goals, and even mastering those skills and information, and thus comparing the individual’s performance with a specific level of performance in a field of behavior, and this is what the tests aim at. Spoken reference.

In the criterion-referenced tests, the student's performance is determined in the light of a statistic known as the cutoff score, which represents the degree that the individual should obtain in order to be considered proficient in the measured trait. In view of the awareness of education specialists of the importance of determining cut-off degrees; A large number of them worked hard to develop methods and models (Jager, 1989; Sizmur, 1997). Most of the methods in determining the degree of cut-off are based on the estimates of the arbitrators. Therefore, according to Jager (1989), these arbitrators must be characterized by experience and knowledge in the field of testing, and that experts understand wide patterns of practices in the intended field.

• Research problem :

Many researchers were interested in comparing the statistical indicators extracted according to the traditional and modern theories of measurement, in order to obtain a test with good psychometric properties. The results of previous studies have varied in this field. Some studies have indicated that the indicators of the modern theory of measurement are superior to the indicators extracted according to the traditional theory of measurement, such as

Previous studies also differed in the type of logistic model used. Although there is a clear difference between the modern and traditional theories of measurement with regard to the methods of selecting the test items, there are not enough justifications for preferring one theory over the other, and therefore the idea of this research came by using one of the modern models in measurement, which is the two-parameter logistic model, which is one of The most consistent models with the statistical indicators of the traditional theory of measurement compared to other models, as well as the preference of this model in selecting the test items, in order to reveal the compatibility between the modern and traditional theories of measurement in matching the referenced test items.

research aims :

The current research aimed to determine the extent of compatibility between the traditional theory of measurement and the two-parameter logistic model in selecting the referenced test items, by identifying the similarities or differences between the two approaches in selecting the items, in terms of the number of items, their level of difficulty and distinction, as well as the validity and reliability of the test.

Research hypotheses

What is the extent of the conformity of the items of the referenced test with the traditional theory of measurement and the two-parameter logistic model?

• What is the extent of compatibility between the traditional theory of measurement and the two-parameter logistic model in selecting the referenced test items?

• What are the psychometric characteristics of the referenced test according to the traditional theory of measurement and the two-parameter logistic model?

• Is there a statistically significant difference at the level of statistical significance ($\alpha = 0.05$) between the stability coefficients estimated using the traditional theory of measurement and the two-parameter logistic model?

•Is there a statistically significant difference at the level of statistical significance ($\alpha = 0.05$) between the two validity coefficients estimated using the traditional theory of measurement and the two-parameter logistic model؟

search terms:

Psychometric characteristics of the test:It means the coefficients of validity and reliability of the test based on the reference prepared in this research.

Cutting degree:

It is the degree that the student must pass in order to be considered

proficient (Allam, 2001 AD). In this research, it refers to the degree that the student must pass in order to be considered proficient in measurement and evaluation, and this degree has been reached according to the Nadelsky method (15).

Research Methodology :

The current research used the descriptive approach in the research.

- research community :

The research community consisted of all (1048) fourth class students distributed over the first semester of the academic year 2020/2021 AD.

- The research sample :

The sample of students was chosen randomly, and the research sample was (140) students. The researcher also selected (7) arbitrators in a simple random way from the community of teachers and educational supervisors in order to determine the cut-off degree.

Search tool:

In order to achieve the objectives of the research, the researcher built a referential test in measurement and evaluation, as the researcher took into account the scientific foundations in constructing referential tests (Allam, 2000). The test consists of (30) items of multiple choice type. Each item has four alternatives, one of which represents the correct answer. The researcher made sure that each paragraph measures a specific goal, according to the list of goals prepared by the researcher for this unit. The content of the unit was analyzed carefully and in detail, and a list of thirty comprehensive detailed objectives for the test content was prepared.

In order to ensure that the goals are included in the study unit and that they represent the three cognitive levels (remembering, understanding, and application), the researcher presented the list of goals to a group of specialists in teaching measurement and evaluation. This is in order to express an opinion on the extent to which the goals cover the subject of the study unit and the way it is formulated. After approving all the detailed objectives of the study unit concerned, the researcher wrote a set of possible questions on each objective separately. The number of possible paragraphs for each objective ranged between five to six, and then the researcher randomly selected one for each of the objectives.

The effectiveness of the paragraphs in the prepared test was verified by collecting the opinions of the arbitrators, in order to indicate the degree of compatibility between the paragraph and the goal that it measures.

The researcher presented a graded assessment scale from (1-5) to a group of arbitrators and specialists in the field of teaching measurement and evaluation. This is to take their views on the validity of the test paragraph in measuring the specific goal, and to suggest what they see as an amendment in the light of their response to the items of the scale in terms of: the text of the paragraph, the attractiveness of the subtractors, the independence of the paragraphs, their freedom from errors,

T for two independent samples. Table No. (1) shows the results of the analysis:

Table No. (1)
The results of the t-test to test the difference between the two groups and the statistical significance.

probability value	Degrees of freedom	)t(	standard deviation	Arithmetic mean	the number	the group
**٠,٠٠	١٧٧	١٨,٥٤	٧,٣٨	١١,٨٧	٣٩	before teaching
			٣,٢٧	٢٦,٩٢	١٤٠	after teaching

It means: statistically significant at the level of statistical significance (0.01 = α)

It is noted from the results of Table No. (1) that there is a statistically significant difference between the performance of the two groups, as the calculated value of (T) was (18.54) with a probability of (0.00), and this value is statistically significant at the level of significance (α = 0.01), which indicates that the test distinguishes between groups Characteristically differentiated, that is, the test measures performance related to a specific behavioral field of the knowledge and skills that students have learned in measurement and evaluation.

The validity of the research tool was verified through the validity associated with the achievement criterion in measurement and evaluation, with a value of (0.87). In order to estimate the stability of the internal consistency of the test, the researcher applied the test to an exploratory sample consisting of (27) tenth grade students after they had learned the unit. The value of Cronbach's alpha stability coefficient for the test was (0.89). It should be noted in this regard that the method of estimating Cranbach's alpha stability is suitable for standard-referenced tests, and when used with referenced tests, it gives a low estimate of stability, which calls us to trust the stability of the current test used in this research. The values of the difficulty coefficients for the items of the test in its survey form ranged between (0.33 - 0.78), and the values of the discrimination coefficients ranged between (0.27 - 0.57), and therefore all (30) items of the test were accepted.

In the test based on the reference prepared in the current research, the student's performance was determined in the light of a statistic known as the cutoff score, which represents the degree that the student should obtain in order to be considered proficient in the measured trait. Where the researcher determined the cut-off degree by selecting a simple random sample of (7) arbitrators; (5) teachers studying measurement and evaluation for the tenth grade, (2) measurement and evaluation supervisors specializing in curricula and teaching methods, where each arbitrator was asked to identify among the alternatives for each of the multiple-choice paragraphs those that could be excluded by the examinees with low ability ; Because in their view, it does not represent the correct answer to the paragraph, and the minimum level for the student to pass

the test paragraph is the inverse of the number of remaining alternatives, which represents the probability of the correct answer to the paragraph. Appendix No. (1) shows the opinions of the arbitrators to determine the degree of cutting according to the Nadelsky method. After completing the arbitration, the values were collected for all the test items and all the arbitrators, and then the result was divided by the number of arbitrators. The resulting value represents the average minimum level of passing in the prepared test, which represents the cut-off degree.

After coming up with the final version of the test, it was finally applied to the members of the research sample consisting of (140) students from the tenth grade students, each in his school, and collectively inside the classroom, where the students were informed of the date of the test, and the researcher sought the help of the teachers of the subject of measurement and evaluation in those Schools to apply the test.

Statistical treatments:

The data were entered into the computer's memory, and the statistical program (SPSS) and the program (BILOG-MG3) were used to perform the necessary statistical analyzes to answer the research questions.

Research results and discussion:

The results related to the first question and their discussion: To what extent do the items of the referenced test match the traditional theory of measurement and the two-parameter logistic model?

The researcher first verified the assumptions of the paragraph response theory in the total test items amounting to (30) items, where the assumption of Unidimensionality was verified using indicators that adopted the factor analysis of the main components through the use of factor analysis, where the analysis produced (10) The factors of the latent root value for each of them are greater than one. Table (2) shows the values of the latent roots and the percentage of explained variance for each factor, as well as the cumulative explained variance percentage.

Table No. (2)

The results of the factorial analysis of the total test

Cumulative Explained Variance %	variance	Explained	Latent root factor
٣٠,٣٨٤	٣٠,٣٨٤	٩,١١٥	١
٣٨,٨٣٤	٨,٤٥٠	٢,٥٣٥	٢
٤٥,٩٢٢	٧,٠٨٨	٢,١٢٧	٣
٥٢,٧٦٢	٦,٨٤٠	٢,٠٥٢	٤
٥٨,٢٣٠	٥,٤٦٨	١,٦٤٠	٥
٦٣,٢٣٤	٥,٠٠٤	١,٥٠١	٦
٦٧,٨٤٣	٤,٦٠٩	١,٣٨٣	٧
٧٢,١٠١	٤,٢٥٨	١,٢٧٧	٨
٧٥,٨٧٦	٣,٧٧٥	١,١٣٣	٩
٧٩,٣٨٢	٣,٥٠٦	١,٠٥٢	١٠

It is noted from the results of Table (2) that the value of the latent root of the first factor is (9.115), and it explains (30.384%) of the total variance. 207 - 230). It is also adopted in the one-dimensional factorial analysis through the ratio of the potential root of the first factor to the potential root of the second factor, and it is a large ratio of not less than (2) (Hambleton and Swaminathan, 1985: 157). The value of the latent root of the first factor over the value of the latent root of the second factor is equal to (3.60), which is a percentage that exceeds the criterion (2), and when looking at the ratio of the difference between the latent root of the first factor and the latent root of the second factor to the difference between the latent root of the second factor and the latent root of the third factor, It turns out that the percentage is large and equal to (16.13), and this indicates that the prepared test achieves a one-dimensional assumption, and therefore it is possible to rely on its paragraphs to reach sound decisions. The value of Cronbach's stability coefficient was alpha (0.903), and this value is high. The researcher contented himself with verifying the assumption of one-dimensionality to indicate the realization of the local independence assumption (Hambleton and Swaminathan, 1985: 22-25). With regard to the assumption of the Item Characteristic Curve (ICC) characteristic of the item or item, it is known that the student's probability of answering the test item with a correct answer increases with the increase of his ability. It was verified that the speed factor did not play a role in the students' response to the test items, as the students were given enough time to answer the test items.

After verifying the previous assumptions, the (BILOG-MG3) program was used to check the good conformity of the data according to the two-parameter logistic model. This is to detect individuals who do not conform to the two-parameter logistic model. The results of the analysis using the chi-square test at the level of significance (= 0.01 μ) showed that (4) individuals did not conform to the model, and the probability value was less than (0.01). The analysis was re-analyzed to choose the items that match the logistic model used, and the results showed that (28) items matched the model used, and the two items (10, 17) in the test form did not match.). The researcher also analyzed the data using the (SPSS) program in order to obtain the coefficients of difficulty and discrimination for the test items from the perspective of the classical theory of measurement. Table No. (3) shows the statistical indicators of the test items before deleting the non-conforming items according to the modern (IRT) and traditional (CTT) theories of measurement:

Table No. (3)
Difficulty and discrimination coefficients for the test items according to the modern and traditional theories of measurement

٢PLM		١CTT		n	٢PLM		١CTT		n
(a)	(b)	rpbi	p		(a)	(b)	rpbi	p	
١,٣١	٠,٠٩	٠,٤٠	٠,٥٩	١٦	٠,٨٤٩	٠,٥٤	٠,٣٨	٠,٤٧	١
٠,٣٠	٠,١٢	٠,٦٠	٠,٥٥	١٧	٠,٧٦	١,٢١-	٠,٦٢	٠,٥٢	٢
١,٢٩	٠,٥٣	٠,٤٨	٠,٥٢	١٨	١,٩٧	١,٢٤-	٠,٤٠	٠,٦٩	٣
١,٥٠	١,٤٩-	٠,٦٩	٠,٤٠	١٩	١,٣١	١,٤٠-	٠,٣٩	٠,٥٧	٤
٢,٥٧	٠,١٩-	٠,٧١	٠,٤٨	٢٠	١,٠١	٠,٥٩-	٠,٤١	٠,٨١	٥
٢,٨٣	١,٣٧-	٠,٣٩	٠,٥٩	٢١	١,٢٩	١,١٧-	٠,٥٥	٠,٤٨	٦
١,٦٩	٠,٠١٧	٠,٤٥	٠,٥٣	٢٢	١,١٩	٠,٩٦-	٠,٤٤	٠,٥٢	٧
١,١٩	٠,٤١-	٠,٣٩	٠,٧٧	٢٣	٠,٩	٠,١٨	٠,٧٠	٠,٤٦	٨
١,٢٤	٠,٩٣-	٠,٤٢	٠,٦٥	٢٤	١,١٢	١,٢٠-	٠,٤٧	٠,٤٥	٩
١,٦٥	٠,٧٧-	٠,٤٣	٠,٥٩	٢٥	١,١٠	٠,٨٣-	٠,١٢	٠,٣٨	١٠
١,٠٩	١,٠١-	٠,٤٠	٠,٧١	٢٦	١,٥٠	٠,٣٣	٠,٥٧	٠,٥٦	١١
١,٢٤	٠,٦١-	٠,٥٩	٠,٥٩	٢٧	١,٢٤	١,٠٦-	٠,٤٩	٠,٧١	١٢
١,٣٣	١,٣٠-	٠,٤٧	٠,٨٣	٢٨	١,٦٢	٠,٥٤-	٠,٤١	٠,٤٢	١٣
١,٤٨	١,١٣-	٠,٥٣	٠,٤٨	٢٩	١,٠٨	٠,٣٣-	٠,٦٦	٠,٤٢	١٤
٠,٧٢	١,٠٣-	٠,٤١	٠,٧٧	٣٠	١,٢٣	٠,٩٨-	٠,٣٩	٠,٥١	١٥

It is noted from the results of Table No. (3) related to the psychometric characteristics of the paragraphs in the light of the classical theory of measurement that the values of the difficulty coefficients ranged from (0.38 - 0.83) with an arithmetic mean of (0.57) and a standard deviation of (0.12). The values of the discrimination coefficients ranged between (0.12 - 0.71), with an arithmetic mean of (0.47) and a standard deviation of (0.11). And after looking at the paragraphs that achieve the statistics used in this research, which are the statistics proposed by (Awdeh, 2010 AD, p.: 281-285), which summarizes that the acceptable range for the difficulty and distinction of the paragraph ranges between (0.20-0.80), and that any paragraph has a coefficient of discrimination higher than (0.39) is considered a good paragraph, and the paragraphs whose discrimination coefficient is less than (0.20) are considered weak. In the light of the previous criteria, the paragraphs whose difficulty coefficient exceeds (0.80) are considered easy in this test, which are paragraphs (5, 28). These two paragraphs have been kept because their discrimination coefficients are good and within the acceptable range, in addition to the need to maintain the

sincerity of the test content. Paragraph No. (10) has been deleted from the test form because its discrimination coefficient is less than (0.20) and equal to (0.12). Thus, (29) items were retained, i.e. (96.67%), while there was one non-conforming item, which is Item No. (10) of the test items, with a rate of (3.33%).

It is noted in the percentage of items matching the traditional theory of measurement in the current research, compared to previous studies, that there is a clear difference, as the number of matching items in Onn's study (Onn, 2021) was (29) items out of (50) items, i.e. (0.58). The results of Yassin's study (2004 AD) showed that (45) items out of (52) items matched, i.e. (86.54%). %). The percentage differs in the current research with the results of the study of Jimelo and Silvestre (2009), where it indicated that (33) items, i.e. (55%) out of (60) items, matched the traditional theory of measurement. The researcher explains the reason for this to the different sizes of analysis samples, and the number of items used in these studies, as well as to the shortcomings of the traditional theory in measurement, where the characteristics of the test items are affected by the sample of individuals, where the values of the difficulty and discrimination coefficients of the test items differ according to the size of the sample.

It is noted from the results of Table No. (3) previously mentioned related to the psychometric characteristics of the test items in the light of the modern theory of measurement using the two-parameter logistic model that the values of the difficulty coefficients ranged from (-1.49-0.54) with an arithmetic mean of (-0.66) and a standard deviation (0.60), and the standard error values for the difficulty parameter estimates ranged between (0.044 - 0.50), with an arithmetic mean of (0.19) and a standard deviation of (0.12). The values of the discrimination coefficients for the test items ranged between (0.30 - 2.83) with an arithmetic mean of (1.32) and a standard deviation of (0.50). . It was observed, using a k-square test for good fit at the level of significance (= 0.01 μ), that the two paragraphs (10, 17) were not identical to the two-parameter logistic model, as the probability value was less than (0.01), and these two paragraphs constituted (6.67%).) from among the test items, and these two items were excluded from the test, while the other items that matched the model used were (28) items, i.e. (93.33%).

Depending on the modern theory of measurement, the items that matched the two-parameter logistic model are statistically independent from each other and from the sample of respondents. It is noted from the results of the analysis that there is a clear difference in the percentage of items matching the two-parameter

logistic model in this research compared to the percentage of items matching in some previous studies that used the two-parameter logistic model, as the number of items matching in Onn's study (Onn, 2021) was (38) items. Among (50) items, i.e. (76%), as the study of (Jimelo and Silvestre, 2009) showed that (33) items, i.e. (55%) out of (60) items, matched the modern theory of measurement. The reason may be due to the difference in the size of the analysis sample and the number of items used in these studies, and the difference in the format of the test items and the content of the test material, in addition to the difference in the analysis programs used in these studies. .

Despite the difference in the quality and objectives of the tests used in previous studies and current research, the test items that are selected according to the two types of models differ if the range of discrimination of the test items is wide. Extreme discrimination (low or high), and this discrepancy in the selection of items may be due to the difference in the goal of analyzing the items in each of the two cases. Estimating the psychometric characteristics of the test as a whole, but in the case of the response models for the item, the goal is directed towards obtaining a grade for the items calibration, i.e. estimating the difficulty of the items items and evaluating the good conformity of their grades to the model used and benefiting from that in evaluating the characteristics of individuals.

The results related to the second question and their discussion: What is the extent of compatibility between the traditional theory of measurement and the two-parameter logistic model in selecting the referenced test items?

The researcher made a comparison between the extracted statistical indicators and F

It is noted from Table No. (4) the following results:

•There is an agreement between the statistical indicators extracted according to the two-parameter logistic model and the traditional theory of measurement in terms of excluding both theories for only one paragraph from the referenced test form.

•There is an agreement between the traditional indicators and the two-parameter logistic model in retaining (28) items from the referenced test items.

•There is no agreement between the statistical indicators of the traditional theory and the two-parameter logistic model for one paragraph, as it was excluded according to the two-parameter logistic model while it was retained according to the traditional theory of measurement, which is Paragraph No. (17) in the referenced test model.

(٢٩) •items conformed to the traditional theory of measurement.

Despite the difference in the model used in this research, as well as

the number of paragraphs used, compared to some previous studies, such as the study of Jimelo and Silvestre (2009), which used the Rasch model in the analysis, but the results of the two studies agree that there are a number of identical paragraphs in both theories. The results of Stage (2003) study showed that the paragraph analysis in the light of the paragraph response theory is better compared to the classical theory in measurement.

The results related to the third question and their discussion: What are the psychometric characteristics of the referenced test according to the traditional theory of measurement and the two-parameter logistic model؟

Conclusions and recommendations

conclusions

.١Is there agreement between the traditional theory of measurement and the two-parameter logistic model in selecting the referenced test items؟

.٢Is there a statistically significant difference at the level of statistical significance (α = 0.05) between the stability coefficients estimated using the traditional theory of measurement and the two-parameter logistic model؟

.٣Is there a statistically significant difference at the level of statistical significance (α = 0.05) between the two validity coefficients estimated using the traditional theory of measurement and the two-parameter logistic model؟

Recommendations:

-١Relying on the two theories of what each one has a use for.

-٢Researchers should be trained on statistical treatments

-٣Interest in modern statistical programs.

the reviewer:

1- Abu Hashem, Al-Sayed Muhammad (2006 AD). A comparative study between the traditional theory and the Rasch model in selecting the items of the search entries list among university students. Journal of the Faculty of Education, Zagazig University, Issue (52), January.

2- Jamhawi, Enas. (2000 AD). Comparing the characteristics of the vertebrae according to the traditional theory and the vertebral response theory in a measure of mental ability. Unpublished master's thesis, Yarmouk University, Jordan.

3-Al-Sharifin, Nidal. (2006). Psychometric characteristics of the reference test in educational measurement and evaluation according to the modern theory of educational and psychological measurement and evaluation. Journal of Educational and Psychological Sciences - University of Bahrain, 7 (4): 8 - 109

4- Allam, Salahuddin Mahmoud. (2001 AD). Diagnostic tests as a reference test in the educational and psychological fields (second edition). Cairo: Dar Al-Fikr Al-Arabi.

5- Allam, Salahuddin Mahmoud. (2000 AD). Educational and psychological measurement and evaluation - its basics, applications and contemporary trends. Cairo: Dar Al-Fikr Al-Arabi.

6- Allam, Salahuddin Mahmoud. (2005 AD). One-dimensional and multi-dimensional test item response models. Cairo: Dar Al-Fikr Al-Arabi.

7- Allam, Salahuddin Mahmoud. (2006 AD). Educational and psychological tests and measures (first edition). Cairo: Dar Al-Fikr Al-Arabi.

8- Eid, Ghada Khaled (2004 AD). True score using latent trait theory and classical theory: a psychometric study. Umm Al-Qura University Journal for Educational, Social and Human Sciences, Issue (2), Volume (16), pp. 230-287.

9- Odeh, Ahmed Suleiman. (2010 AD). Measurement and evaluation in the teaching process. Irbid: House of Hope.

10- Kazem, Amina Muhammad (1988 AD). Using the Rasch model in constructing an achievement test in psychology and achieving objective interpretation of the results. Kuwait: Kuwait Foundation for the Advancement of Sciences.

11- Yassin, Omar Salih Mufdi (2004 AD). The psychometric properties of a reference test in chemistry for first year secondary scientific students estimated according to the classical and modern theories of measurement. Unpublished master's thesis, Amman Arab University for Postgraduate Studies, Amman.