College of Basic Education Research Journal

www.berj.mosuljournals.com

Employing Information Function and Standard Error to Evaluate Quality of A Criterion Referenced Tests Based on Lord Model

Waleed Khalid Abdulkareem Baban

Article Information

Article history:

Received: December 29.2024

Reviewer: February 4.2025

Accepted: February 4.2025

Key words :

Correspondence:

Abstract

Achievement tests are designed to provide a clear picture of students' proficiency levels. However, the perspective that focuses on judging degree of their understanding based on comparisons with their peers or classmates became limited, as it does not reflect extent to which these students possess trait measured or their mastery of learning. Therefore, researcher implemented current study within a relatively modern approach of educational measurement compared to traditional approach.

Accordingly, a total of (50) test items were developed to measure abilities of third-year students in Departments of Psychological and Educational Counseling and Special Education at College of Education, Salahaddin University-Erbil, in Kurdistan Region of Iraq. These items assessed their knowledge in Psychological and Educational Measurement and Evaluation. Validity, discrimination, and reliability coefficients were calculated using a representative sample of (865) female students, selected through cluster random sampling method. The study aimed to examine validity of several hypotheses formulated to evaluate effectiveness of Latent Trait Theory, based on Lord's two-parameter model, in scaling items of a criterion-referenced achievement test. The scaling was conducted according to difficulty and discrimination parameters to eliminate any uncertainty in estimating students' abilities in psychological and educational measurement and evaluation, as well as their mastery of subject. Hypotheses were tested using appropriate statistical methods, leveraging the Statistical Package for the Social Sciences (SPSS) software, version 25, and Bilog-MG3 program.

Based on above mentioned , results of study indicated suitability of Lord's model to scale test items. All items of achievement test fell within acceptable limits and conformed to Lord's model. Results also showed that values of difficulty and discrimination parameters for achievement test matched Lord's model within framework of Latent Trait Theory. Furthermore, maximum information values and standard error values for both items and test as a whole supported test's efficiency in measuring ability in psychological and educational measurement and evaluation.

Given these findings, several recommendations and suggestions were proposed by researcher to enhance current state of educational measurement.

ISSN: 1992 – 7452

توظيف دالة المعلومات والخطأ المعياري لتقييم جودة اختبار محكي مبني وفق نموذج لورد

وليد خالد عبدالكريم بابان

قسم الارشاد التربوي والنفسي/كلية التربية/جامعة صلاح الدين-أربيل/العراق

الملخص:

تصمم الاختبارات التحصيلية لتعطي صورة واضحة عن مستويات الطلبة، إلا أن النظرة التي تعنى بالحكم على درجة تعمهم في ضوء مقارنتهم بأقرانهم، أو زملائهم في الصف الدراسي، باتت قاصرة، لكونها لاتنم عن مدى إمتلاك هؤلاء الطلبة للسمة موضع القياس، أو مدى إتقانهم للتعلم، لذلك أتجه الباحث للقيام بالبحث الحالي، وفق إتجاه حديث نسبياً في القياس التربوي مقارنة بالأتجاه التقليدي.

وعليه فقد تمت صياغة (٥٠) فقرة إختبارية لتقيس قدرات طلبة المرحلة الثالثة بقسمي الأرشاد النفسي والتربوي و التربية الخاصة في كلية التربية بجامعة صلاح الدين/أربيل بأقليم كوردستان-العراق، في مادة القياس والتقويم النفسي والتربوي (Psychological and Educational Measurement and Evaluation) حيث تم حساب معاملات الصدق والتمييز والثبات، من خلال عينة ممثلة بلغت (٥٦٨) طالبة، تم إختيارها وفق الطريقة العنقودية العشوائية، تمهيداً للتعرف على صدق عدد من الفرضيات التي تم صياغتها لمعرفة مدى قدرة نظرية السمات الكامنة، ووفق نموذج لورد الثنائي البارامتر، في تدريج فقرات إختبار تحصيلي محكي المرجع، تبعاً لمعالم أو بارامترات الصعوبة، والتمييز، وذلك لعدم ترك فرصة للشك في تقدير قدرات الطلبة في في القياس والتقويم النفسي والتربوي، أو لمدى إتقانهم لتعلمها. حيث تم التحقق من تلك الفرضيات من خلال إستخدام الوسائل الإحصائية اللازمة والمناسبة، ومن خلال إستخدام برنامجي الحقيبة الاحصائية للعلوم الاجتماعية (SPSS) بالأصدار (٢٥)، وبرنامج (Bilog–MG3).

وتأسيساً على ما سبق فأن ما تمخض عن نتائج البحث، أشار الى ملائمة نموذج لورد لتدريج فقرات الإختبار، حيث كانت جميع فقرات الاختبار التحصيلي قد جاءت ضمن الحدود المقبولة والمطابقة لنموذج لورد، كما أشار النتائج أيضاً الى أن قيم معالم الصعوبة والتمييز للأختبار التحصيلي، قد جاءت مطابقة لنموذج لورد لنظرية السمات الكامنة، كما أشرت قيم العظمى للمعلومات، والخطأ المعياري للفقرات والأختبار ككل مؤيدة لكفاءة الأختبار في قياس القدرة في القياس والتقويم النفسي والتربوي. وفي ضوء معطيات البحث تم صياغة عدد من التوصيات والإقتراحات التي وجدها الباحث لازمة للنهوض بواقع القياس التربوي.

الكلمات المفتاحية: دالة المعلومات، الخطأ المعياري، الاختبارات المحكية، نموذج لورد.

1.1.Research Problem:

Achievement tests are among the most commonly used tools and methods for assessing learning outcomes. They are widely employed to determine the extent to which cognitive educational objectives have been achieved. For achievement tests to fulfill their intended functions effectively, they must possess objective characteristics, ease of use, and comprehensiveness in covering the objectives being measured and evaluated (Al-Hayla, 2008: 399–400). Consequently, most researchers tend to assess the effectiveness of education based on various criteria, the most important of which is the achievement outcomes of the educational process (Nashwati, 1998: 159).

This necessitates the establishment of an integrated system of tests across various disciplines and subjects, standardized through empirical evidence to ensure they measure different levels of achievement. The significance of such tests lies in their indispensable role in creating a database of accurate and reliable information. This database can be leveraged to enhance the validity of educational decisions related to curricula, academic programs, and various aspects of the teaching and learning process (Al-Dosari & Al-Mutawa, 1991: 115).

The Latent Trait Theory has become an essential and widely used tool in the construction and development of tests. This has encouraged specialists to advance various logistic models tailored to suit all areas relevant to educational and psychological tests and measurements. The primary aim of these models can be summarized as establishing a connection between item characteristics and one or more logistic parameters. They provide an alternative to Classical Test Theory by estimating individual and item parameters with minimal error. This approach eliminates the need for a random sample of test items from the measured domain or the requirement for a very large representative sample of items.

The primary objective of Latent Trait Models is to determine the relationship between individuals' responses on a specific test and the latent trait underlying those responses. This allows for the quantification of the latent traits influencing individuals' performance on various tests. Such quantification can be utilized to predict their behavior in similar situations and to make informed decisions about them based on this quantitative estimation of traits (Allam, 2005: 53).

The appropriate model is selected based on the purpose and nature of the test, the feasibility of calculating individual and item estimates, and the fit of the data to the model (Suen, 1990, p: 93). Among the most prominent of these models is Lord's Model, which is a Two-Parameter Logistic Model. This model considers two parameters: item difficulty and response discrimination. Additionally, this model allows for the intersection of item characteristic curves (Allam, 2001, p: 206).

Accordingly, the Two-Parameter Model is used to analyze the items of dichotomous tests (binary responses), meaning that each test item is scored as either a "1" for a correct response or a "0" for an incorrect response. When an individual responds to a test item, an interaction occurs between their ability and the difficulty and discrimination parameters of the item. In this context, Hambleton and others (1978) state that the purpose

of applying psychometric tests is to classify examinees into two categories: one possessing the trait and the other lacking it. This objective can be achieved through the use of Lord's Model (Hambleton et al., 1991, p: 126).

The application of measurement tools without verifying and ensuring their psychometric properties threatens the accuracy of results and can lead to incorrect decisions when evaluating examinees. Therefore, it is essential to develop these tools with complete objectivity, utilizing the most precise statistical models and the latest measurement theories. The researcher was thus convinced to develop the test used in this study based on Modern Measurement Theory, specifically the Latent Trait Theory, and by employing Lord's Two-Parameter Logistic Model. This aligns with the observations of Traub and Ackerman, who caution researchers and users of unidimensional models to recognize that unidimensional test items interact with examinees—particularly in cases where multiple skills are required to arrive at the correct response. This necessitates thorough scrutiny (Traub, 1983) (Ackerman, 1994, p: 67).

However, the issue of measurement error and the consistency of individuals' responses should not be overlooked or disregarded without careful investigation and scrutiny. One of the fundamental pillars for determining the presence of any psychological or educational trait is the stability of responses, which must consistently manifest at the same level whenever the opportunity or testing situation arises to reveal its possession. The researcher observed that researchers, in general, tend to focus on validity, often judged based on unidimensional indicators, more than on the reliability of responses or the ability being measured. This observation motivated the researcher to assess these aspects using the information function indicator, both at the level of individual test items and the test as a whole, as well as relying on another crucial indicator: the standard error of measurement (SEM).

Based on the preceding discussion, the researcher will attempt to answer a fundamental question

- To what extent is the criterion-referenced test, constructed according to Lord's model, effective in measuring achievement in the subject of Psychological and Educational Measurement and Evaluation, based on the indicators of the information function and standard error of measurement (SEM)?

1.2. Research Significance

Undoubtedly, the educational process, as an interconnected and multidimensional system, is influenced by numerous variables that impact its success. These variables include factors related to the learner, the parameters, and the subject matter. Naturally, this interplay results in the interference of these variables with the learners' true scores. Thus, there is an urgent need to purify these scores by isolating the extraneous variables, ensuring that the learners' scores are free from influences that do not accurately reflect their true performance. This necessity places experts in educational measurement before the responsibility of advancing their scientific efforts. As a result, educational measurement specialists work diligently to develop methods for designing criterion-referenced

achievement tests, refining their construction techniques, and innovating modern mathematical methods and statistical models. These advancements provide guidelines for analyzing the items in such tests and assessing their quality effectively.

Undoubtedly, the best approach to achieving this level of accuracy in achievement tests is to elevate the use of valid and objective criteria. This means that measurement tools for assessing individuals' abilities and traits should be free from the characteristics of the sample items included in the test applied to them. Thus, the system of objective measurement represents a modern development in psychological and educational measurement, linked to a new approach known as the latent traits approach in measurement. This approach is expressed through the latent trait theory, which includes innovative mathematical models (Allam, 2001: 13-19).

As a result of these important developments in educational and psychological measurement, new trends have emerged in measuring psychological and educational phenomena, including the idometric or criterion-referenced approach. This approach emerged relatively recently compared to psychometric measurement, as it arose from the evolution of the concept of the educational process, aiming to achieve the concept of learning for mastery and comparing the student's performance to a specific performance standard based on the set objectives. (Al-Naimee, 2005: 2).

The importance of using this modern approach is highlighted in the construction of criterion-referenced tests, as they exclude items if their relationship to the content is weak or if there are issues with their wording. In contrast, norm-referenced tests select items of medium difficulty with high discrimination indices. Criterion-referenced tests are built to assess the performance of students or individuals based on specific competencies and to determine the extent to which individuals have acquired skills before and after an educational or training program. On the other hand, norm-referenced tests are designed to determine a student's rank in a certain trait, assign final grades, classify students, and measure individual differences among learners (Allam, 2000: 342).

Based on this, the importance of the Latent Trait Theory, or what some call the Modern Measurement Theory, becomes evident. This theory addresses many measurement issues more effectively than the traditional theory. It assumes the possibility of predicting individuals' performance and interpreting their test results based on one or more characteristics that define this performance. Consequently, this theory aims to estimate individuals' trait scores, which fall within the measurement scope, by relying on individual test items rather than the overall test score (Al-Sharifain, 2006: 84).

Hence, the importance of conducting this research lies in its focus on developing a highly significant tool—achievement tests. These tests must exhibit the highest levels of accuracy and credibility due to their critical role and profound impact on making decisions regarding learners' mastery of current learning and the achievement of objectives, whether educational, behavioral, or preparatory for progression to the next level of the educational program. These reasons have motivated the researcher to adopt this modern approach in constructing achievement tests.

1.3. Research Aimes:

The current research aims to achieve the following Aimes:

1. Identify the Estimates of Difficulty and Discrimination Parameters for the Test, the Information Function of Items, and their Standard Errors According to Lord's Model.
2. Analyze the Information Function and the Standard Error for the Test Constructed Based on Lord's Model.

1.4. Scope of the Study:

This Study is Delimited to a Criterion-Referenced Achievement Test Designed to Measure the Ability of Third-Year Students in Colleges of Education in the Subject of psychological measurement and educational evaluation. The latent trait theory, specifically the two-parameter Lord Model, will be Employed. This Model Incorporates Item Difficulty, Item Discrimination, Item Information Function, Test Information Function, and Standard Errors for All Aforementioned Parameters.

1.5. Definition of Terms:

Test Information Function (TIF):

Conceptual Definition: Represents the amount of information provided by the test as a whole at each ability level. This function is depicted as a curve showing the relationship between two variables: the ability levels (represented on the horizontal axis) and the test's information (represented on the vertical axis) (Allam, 2005: 117).

Operational Definition: The researcher operationally defines it as the probability of different groups of individuals within the same achievement level performing differently, where the average of the maximum value of the information function $(\theta)max$ falls within an acceptable range.

Standard Error (SE): A statistical indicator that reflects the dispersion of parameter estimates around the true value of the item's parameter (Zumbo, 2007: 228).

Criterion-Referenced Tests (CRT): An idiometric standard in which an individual's score is compared to an independent benchmark known as the absolute criterion (Kilani et al., 2009: 37).

Latent Traits Theory: Represents the probability of a correct response to an item with a specific difficulty level, as a function of the latent trait level displayed at each level of the trait or ability (Anastasi & Urbina, 1997: 189-190).

Lord Model: One of the latent trait theory models based on a dichotomous (binary) scale. This model allows test items to differ in both difficulty and discrimination parameters. The two-parameter logistic model is derived when the guessing parameter is set to zero (van der Linden, 1986: 82).

2. Theoretical Framework:

The theoretical framework is a fundamental requirement in the construction of the test and metrics, because it provides the researcher with the theoretical concepts on which most test building procedures are based, especially when he relies on the mental or logical approach that requires the derivation of certain concepts or premises from the theoretical framework, and the researcher has to follow them with whatever these concepts or pronouncements of the procedures impose.The theoretical framework guides researchers towards specific goals, in the sense that it rids researchers of the indiscriminacy and flaw between the elements of the problem they are dealing with, namely the advantage of individuals dealing with psychological problems and uptake within the scientific curriculum.

The theoretical framing of the concept of information function and error will therefore address the standard and model used for current research purposes.

2.1. Information Function

The information function is a crucial statistic in modern measurement theory. It enables the determination of the standard error of estimation by relying on the maximum likelihood estimate of the ability parameter. The variance of the error in ability estimation is equal to the inverse of the information function. Similarly, the variance-covariance matrix of the estimates equals the inverse of the information matrix of the item parameters Therefore, it is a mathematical function that represents the relationship between an individual's ability and the information provided by the test items. It expresses the amount of information represented by the item's discrimination between ability levels of individuals, as determined by the maximum height of the information function curve. This represents the amount of information provided by the item or the entire test when estimating the examinees' information. Through it, the standard error of estimation can be determined (Hambleton & Swaminathan, 1993: 573).

2.2. Item Information Function:

Hambleton and Swaminathan (1985: 163) argue that each item in a test has an item information function. This function is a curve that shows the extent to which an item contributes to the determination of ability. Generally, items with high discrimination contribute more strongly to ensuring measurement accuracy than those with low discrimination. An item makes its best contribution to ensuring measurement accuracy around its difficulty value (b) on the ability continuum. In other words, the information function depends on the slope of the item response curve and the variance at each ability level (q). The steeper the slope, the lower the variance, and the more information the item provides. If the curve is shifted to the right, it means the item is difficult. If its height is high, it means the item has high discrimination, and vice versa. It is calculated according to the following equation:

$$Ii(\theta) = Pi(\theta)\ Qi(\theta)$$

Where:

Ii(θ): Information function of item (i).

Pi(θ): Probability of a correct response to item (i).

Qi(θ): Probability of an incorrect response to item (i).

2.3. Test Information Function:

The test information function is defined as the sum of the information functions of the individual items that make up the test. Therefore, studying the item information function and the variables that affect it provides an opportunity to obtain a test with a desirable function. The test information function is calculated by summing the information of the items according to the following equation:

$$\sum Ii(\theta) = I(\theta)$$

Where:

I(θ): Test information function.

Ii(θ): Information function of item (i).

Hambleton and Swaminathan (1985: 235) emphasized that as the number of items increases, the amount of test information also increases. Additionally, as the item discrimination increases, the information provided by the item increases. This means that items with high discrimination parameters provide more information about the examinees' ability, leading to greater accuracy. Therefore, test items can be selected based on the amount of information each item contributes to the overall test information.

It is easily calculated from the test information curve. The test information function also provides the standard error of measurement for the test at each ability level. Therefore, for a test whose items have been analyzed, the test information function can be calculated by summing the item information functions. The amount of information contributed by a set of items at a specific ability level is inversely related to the standard error of measurement at that level. In other words, when the amount of information is high, the error in estimation is low, indicating an inverse relationship between the two. The standard error of measurement at a given ability level (θ) is calculated using the following equation:

$$SE\ (\theta) = \frac{1}{\sqrt{1(\theta)}}$$

(Warm,1978:73

2.4. Criterion-Referenced Tests:

Hambleton et al. (1977) and Rovinelli & Hambleton (1977) stated that the purpose of criterion-referenced tests is to classify examinees into two categories: those who possess the trait and those who do not. In this regard, Popham (1978) adds that criterion-referenced tests are used to determine an individual's level of performance relative to a well-defined behavioral domain. Therefore, the most important component is to ascertain an individual's standing relative to a specific behavioral domain, rather than determining their relative standing compared to others (Popham, 1978: 94).

When using criterion-referenced tests, we are not concerned with the relative position of an individual among their peers, but rather with comparing their score on the test to a specific performance level (Performance Standard). This performance level serves as a criterion, indicating the acceptable level of the individual's behavior and performance. In other words, if the individual's score meets or exceeds the set performance level, they have achieved the required level. However, if their score is below the set level, their performance is considered inadequate. Therefore, the interpretive framework for criterion-referenced tests does not rely on a reference group of individuals, but instead depends on a level or performance standard within a well-defined behavioral domain measured by the test. In this context, the focus is on what the individual can perform and what they cannot, rather than comparing them to their peers without considering their knowledge or skills, or what they should know or perform. Thus, criterion-referenced measurement requires the pre-determination of performance levels, and performance standards or cut scores can be set for different professions and jobs. Additionally, standards can be established to measure student proficiency at different points in the learning process to gather information about their performance or academic progress (Allam, 2000:262).

To achieve the function of criterion-referenced tests in the accurate interpretation of an individual's performance, the following conditions must be met:

1. The behavioral domain being measured should be clearly defined. This can be achieved by dividing the curriculum into small units, as is the case with programmed instruction, where the focus is on a limited number of learning outcomes.

2. The learning objectives or outcomes must be clearly defined in behavioral terms. Teaching objectives should be framed as statements representing the learning results or outcomes that the individual should demonstrate by the end of the course.

3. Performance standards must be clearly defined. This principle is the cornerstone of criterion-referenced measurement, as the interpretation of an individual's performance is based on these standards.

4. A criterion-referenced test must include a representative sample of the individual's performance. The test should contain a sample that is both representative and sufficient for the learning outcomes.

5. Test items should be selected in a way that reflects the behavior specified in the teaching objectives, so that each item serves as a direct measure of the learner's behavior.

6. The methods for designing and scoring test items should describe the individual's performance level on specific tasks, such as using the percentage of correct answers (Abu-Allam, 2005:147).

2.5. Latent Trait Theory:

It is one of the fundamental features of the modern theory for item analysis. It is a graphical representation of the characteristics of a specific item or can represent the entire test. In latent traits, the total test scores are represented on the horizontal axis, and the proportion of test takers who answer this item correctly within this score range is measured along the vertical axis (Allam, 2000: 694).It can be expressed as a mathematical function that relates the probability of an individual's ability to answer an item correctly to the trait measured by a set of items. This function is a nonlinear regression of the item score on the trait measured by the test. Differences among item response models depend on the mathematical function used to plot the item characteristic curve, which describes the relationship between ability and performance on the item. The shape of the latent traits depends on the item parameters: difficulty (β), discrimination (a), guessing (ci), and the individuals' ability (θ) (van der Linden, 1999: 82).

An individual's correct response to a test item in unidimensional item response models can be mathematically represented by considering the test as a mapping function from the sample of test-takers to the continuum of the latent trait. The probability of a correct response is a strictly increasing function of individuals' positions on the latent trait continuum. Consequently, the likelihood of an individual providing correct answers to test items increases as their level of the trait improves (Allam, 2005: 59).

2.6. Lord's Model:

The Latent Trait Theory gave rise to multiple models that differ in the mathematical functions they rely on, with each model based on a specific response pattern. Some models are based on binary responses (0, 1), while others focus on graded or polytomous scored items (Abu Khalifa, 2004: 23). This model, developed by the American statistician Frederic Lord and his colleagues at Columbia University, belongs to the family of Dichotomously Scored Models (DSM). These models are applied to items with binary scoring, such as multiple-choice questions or true/false items, where examinee responses are constrained to only two possible values: ٠, representing an incorrect response, and 1, representing a correct response. These binary responses are characteristic of test situations where individual examinees' responses are evaluated on a two-dimensional scale (Allam, 2005: 74).

Lord introduced a new parameter in his model called the Item Discrimination Parameter. In the Lord Model, it is assumed that test items vary in their difficulty and in

their ability to discriminate between different levels of ability. Lord added this parameter because it is challenging to find a set of items that uniformly discriminate between levels of ability or the latent trait measured by the test or scale. This contrasts with the assumption underlying the Rasch Model (Allam, 2007: 217–218).

Psychometricians using Latent Trait Theory rely on item statistics, such as the Item Difficulty Index and the Item Discrimination Index: Item Difficulty is simply the proportion of correct responses to an item by examinees. Item Discrimination typically refers to the relationship between the item score and the total test score. It essentially reflects the extent to which individuals with higher or lower total scores are more or less likely to correctly answer a specific item.

Since item difficulty (bi), item discrimination (ai), and examinee ability (θ) are fundamental parameters in this model, the difference between the ability (θ) possessed by an individual (S) in the trait to be estimated (latent ability) and the difficulty level (i) of the item the individual wishes to answer, represented by (b), multiplied by the discrimination power of item (i), represented by (ai), is assumed. This assumes a unidimensional underlying factor for individual differences in students' responses (Al-Taqi, 2008: 22). The mathematical formula for the model is as follows:

$$pi(\theta) = [\ 1+exp\ [\ -ai(\theta -bi)]]$$

Where: *Pi(θ)P_i*(\theta) *Pi(θ):* The probability of a correct response to item (*i*) given ability (*θ*)\theta (*θ*) (Hambleton & Swaminathan, 1985, P: 37)

Wiborg suggests that the discrimination parameter is usually positive, meaning that the probability of a correct response increases as ability increases. However, we may encounter negative discrimination, where the probability of a correct response decreases as ability increases. The function representing Lord's model can be expressed as follows the function representing Lord's model can be visualized as follows (Figure 1):

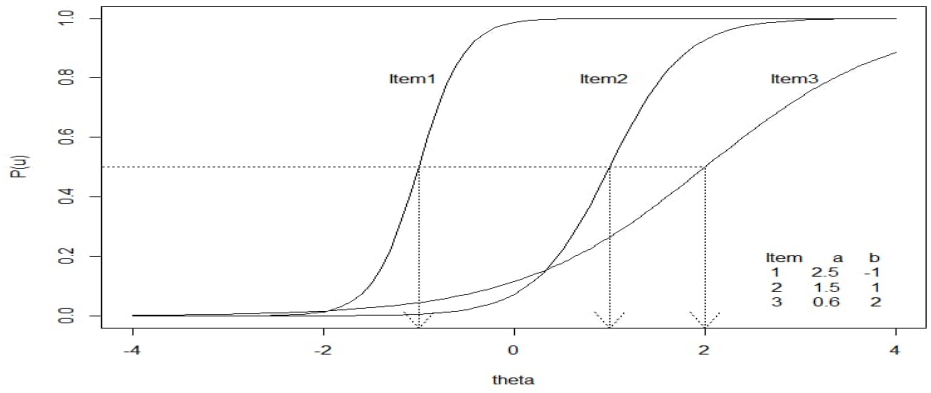


Figure No. (1) shows the Lord model curve.

It is evident from Figure (1) that the two characteristic curves of the items differ in the points where they intersect the horizontal axis. This indicates that the two items differ in their levels of difficulty. Additionally, the curves differ in their slope, with the slope of the characteristic curve for item (2) being steeper than that of item (1). This suggests that item (2) discriminates between respondents to a greater degree than item (1) (Allam, 1986: 109). Consequently, items characterized by two parameters differ not only in their difficulty and discrimination but also in their slope (Allam, 2001: 211). The item response (IRT) is the point where the shift occurs from a positive slope to a negative slope. At this point, the slope of the curve reaches its maximum. Thus, the discrimination index is associated with the maximum slope (Linden & Hambleton, 1997, p: 17–18).

3.1. Research methodology:

Research methodology is a fundamental pillar of scientific research. It is evident that there are multiple scientific methods employed by researchers to conduct their studies. Among these research methods is the descriptive survey method, which focuses on existing conditions and examines what actually exists at the time of the study (Al-Taib, 2018: 227). Therefore, the researcher will conduct their study using the descriptive survey method, as it is considered the most suitable approach for the current research. The researcher will apply the two-parameter Lord model within the framework of this study to investigate the latent trait theory of a criterion-referenced achievement test for third-year college of education students in the field of psychological and educational measurement and evaluation.

3.2. Population of the Research:

The population represents all individuals who possess the observable data relevant to the study or refers to all units or elements of the phenomenon under investigation (Oudah, 1998: 66). The population of the current research comprises third-year students enrolled in the Colleges of Education. Upon reviewing the records of the Directorate of General Registration at the Presidency of Salahaddin University/Erbil, the researcher determined that the total number of third-year students in the Colleges of Education at the mentioned university for the academic year (2022–2023) was 2,310 students, including 1,462 female students and 848 male students, distributed across 18 departments. Table (1) provides a detailed breakdown of the research population.

Table (1): Breakdown of the Research Population

Male Students	Female Students	Total Population
848	1462	2310

3.3. Sample of the Research:

The sample size used in research is one of the most critical factors influencing the accuracy of measurement and the validity of results. Dale (1995) stated that the accuracy of measurement can be determined by the sample size, its characteristics, the degree of homogeneity in the property being measured, and the extent to which the sample represents

the original population (Dale, G.T., 1995:17). In this regard, Nunnally suggests that the sample size for item analysis should range between 5-10 individuals per item of the scale to minimize the effect of randomness (Nunnally, 1978: 262).

For the current study, a representative sample of the research population was required, with a size appropriate for the statistical procedures. The researcher, being an instructor of the subject Psychological and Educational Measurement and Evaluation in the Departments of Psychological Counseling and Educational Guidance, Special Education, and Chemistry, selected a sample of (865) third-year students from the College of Education at Salahaddin University/Erbil using the random cluster sampling method. This approach is considered highly accurate for sample selection, particularly in cases where the population's characteristics are homogeneous. This was the rationale behind the researcher's choice of this method for selecting the sample.

3.4. Research Tool (Achievement Test):

The achievement test is one of the important tools used in assessing learners' achievement due to its ease of preparation, grading, and application (Al-Imam et al., 1990: 59). Below is an explanation of the steps followed in constructing the achievement test:

3.4. Content Specification:

The content was specified based on the topics covered in the subject Psychological and Educational Measurement and Evaluation for the third-year students in the College of Education. These topics are outlined in Table (2).

Table (2): Titles of Chapters Included in the Study Content

Chapter titles	Chapter Number
Principles of Psychological and Educational Measurement	**1**
Educational Objectives	**2**
Achievement Tests	**3**
Psychometric Properties	**4**

1 .Formulating the Achievement Test Items:

To formulate the items of the achievement test, a test blueprint was designed for the required test. As a result, (50) multiple-choice items were developed, each with four answer options. According to Ebel (1972), this number of alternatives aligns with an acceptable difficulty level for tests and is effective in reducing guessing opportunities. Moreover, this number of alternatives is capable of yielding a good reliability coefficient (Ebel, 1972: 268–273).

2 .Preparing the Test Blueprint:

The test blueprint is the optimal tool that enables teachers to establish the foundational elements of the educational material they have taught, within a structured plan from which questions are selected in terms of formulation and type (Al-Shujairi & Al-Zuhairi, 2021: 262). It is used to ensure the content validity of achievement tests, whether criterion-referenced or norm-referenced, and is constructed based on three main steps:

- Determining the relative weight of objectives: These are the learning outcomes that need to be assessed in light of the instructional goals.
- Determining the relative weight of the content: This is achieved by analyzing the different components of the taught material.
- Constructing the test blueprint: This involves creating a two-dimensional table consisting of instructional objectives, arranged horizontally, and content topics, arranged vertically (Ghunaim, 2003: 90).

After determining the content and formulating the behavioral objectives, the researcher selected (50) items for the achievement test. These items were distributed across the topics within the scope of the research material and the behavioral objectives they aim to assess. The weights or focus percentages for both content and behavioral objectives, as well as the number of items at each level, were calculated. Table (3) illustrates this distribution:

Table (3) Displays the Specifications Table for the Test

Chapter Sequence	Content	Number of Hours	Relative Importance (100%)	Level of Behavioral Objectives				Number of Items
				Recall 50%	Comprehension 20%	Application16%	Analysis 14%	
First	Principles of Psychological and Educational Measurement	2	%12	3	1	1	1	6
Second	Educational Objectives	7	%41	10	4	3	3	20
Third	Achievement Tests	2	%12	3	1	1	1	6
Fourth	Psychometric Properties	6	%35	9	4	3	2	18
Total		**%17**	**%100**	**25**	**10**	**8**	**7**	**50**

1. Answer Alternatives:

Each question has four answer alternatives, with one point given for each correct answer. Incorrect answers receive a score of zero. Therefore, the lowest score a respondent can achieve is zero, indicating a low level of the ability being measured, while the highest possible score is 50, indicating a high level of ability.

The researcher performed the test correction using electronic grading technology at the Psychological Research Center affiliated with the Iraqi Ministry of Higher Education and Scientific Research. A separate answer sheet was prepared in advance and used to mark the answers of the students who took the achievement test created for the purposes of the current research.

1. Psychometric Properties of the Achievement Test:

First: Validity of the Test

Validity refers to the ability of a measurement instrument to accurately measure what it is intended to measure (Harrison, 1983, p: 11). Based on this, it is important to verify whether the measurement tool is capable of achieving the goal for which it was designed.
(Awda, 1998, p: 333).

A. Content Validity

A test is considered to have content validity when its items represent the educational objectives and the instructional content. This is achieved by comparing the test items with the educational objectives included in the curriculum or by consulting with experts and specialists (Abu Saleh et al., 1995, p: 213). To ensure the content validity of the achievement test, the researcher consulted with experts and specialists in the fields of educational sciences and psychological and educational measurement. The results of their opinions were as follows:

Table (4) Content Validity of the Achievement Test Items

Item Sequence	Agree		Disagree		Chi-Square Calculated	
	Number	Percentage	Number	Percentage	Calculated	Table
1, 2, 3, 5, 7, 8, 9, 11, 12, 13, 14, 15, 16, 17, 19, 20, 22, 23, 25, 26, 27, 28, 29, 30, 31, 32, 35, 36, 37, 38, 39, 40, 42, 46, 47, 48, 49, 50	10		0	0%	10	3.84
4, 6, 10, 18, 21, 24, 33, 34, 41,	9	90%	1	10%	6.4	

43, 44, 45						
Excluded Items:				**There are no excluded items**		

B. Construct Validity”

In order to verify the validity of the test, the researcher used construct validity as one of the methods to extract the validity coefficient of the test items. The researcher relied on calculating the validity of each item based on the biserial correlation coefficient (rbis) between the score of each item and the total score. It is assumed that this relationship should be significant and positive to serve as an indicator of construct validity. This contributed partially to confirming construct validity as an empirical form of validation. Anastasi (1976) mentioned that the correlation of an item with an internal or external criterion is an indicator of its validity. When no appropriate external criterion is available, the total score of the respondent serves as the best internal criterion for calculating this relationship (Anastasi, 1976: 206).

The researcher analyzed the items using internal consistency by calculating the correlation coefficient between the score of each item and the total score, as the relationship between the item and the total score indicates that the scale measures a single trait (Abdul Rahman, 1998: 215). Table (5) presents these results.

Table (5): Correlation Coefficients Between the Item Scores and the Total Score of the Achievement Test

Item Number	Correlation Coefficient	Item Number	Correlation Coefficient	Item Number	Correlation Coefficient	Item Number	Correlation Coefficient	Item Number	Correlation Coefficient
1	.362**	**11**	.421**	**21**	.605*	**31**	.352**	**41**	.٤٨٥*
2	.421**	**12**	.462*	**22**	.319***	**32**	.430**	**42**	.٥٢٨*
3	.671*	**13**	.358**	**23**	.313***	**33**	.403**	**43**	.٦٠٧*
4	.317***	**14**	.399**	**24**	.272***	**34**	.272***	**44**	.280***
5	.526*	**15**	.514*	**25**	.330***	**35**	.513*	**45**	.378**
6	.443**	**16**	.609*	**26**	.299***	**36**	.335***	**46**	.496*
7	.284***	**17**	.303***	**27**	.376**	**37**	.573*	**47**	.452*
8	.310***	**18**	.528*	**28**	.536*	**38**	.527*	**48**	.594*
9	.513*	**19**	.509*	**29**	.597*	**39**	.341***	**49**	.641*
10	.467*	**20**	.603*	**30**	.335***	**40**	.290***	**50**	.402**

■ The critical value of the correlation coefficient with (863) degrees of freedom at the significance levels:

- (0.432) = 0.05*

- (0.345) = 0.01*

- (0.265) = 0.001*

Based on these results, we can conclude that all the test items have an acceptable degree of validity, as their correlation values were higher than. (0.265).

Secondly: Reliability:

In order to calculate the reliability coefficient of the achievement test, the researcher used both the split-half and variance methods as techniques to extract internal consistency for calculating the reliability of scales and tests. Therefore, the researcher applied the Jackson and Cronbach's Alpha formulas to calculate reliability using the variance method. The results showed the reliability of the scale, and Table (6) illustrates these results.

Table (6) Reliability

Reliability Coefficient Using Cronbach's Alpha Formula	Reliability Coefficient Using Jackson's Formula	Number of Test Items
0.814	0.828	50

From the above values, it is clear that the test exhibits an acceptable level of reliability, as it is considered reliable according to the Foran criterion, which states that an instrument is considered reliable if the reliability coefficient is greater than 0.70 (Foran, 1961: 484).

1. Preliminary Experiment (Pilot Study) **:** To ensure the clarity of the test instructions and the clarity of the scale items in terms of language and content, the researcher selected a sample of 30 male and female students from the research population. These individuals were chosen from the Physics department to avoid repeating the application of the research measurement tools on the same individuals from the research population during the final application. The results of this application indicated no ambiguity or misunderstanding regarding the items and how to respond to them. Thus, the researcher completed all the necessary procedures to apply the test to the research sample.

2. Time Required to Complete the Achievement Test: The researcher utilized the results of the preliminary experiment for the achievement test and calculated the time it took the

students to answer the test. It was found that the time spent ranged between 29 and 58 minutes.

3.5. Verification of the Assumptions of the Latent Trait Theory using the Lord Model:

This theory is based on fundamental assumptions, including: Unidimensionality: The test measures a single underlying trait. Local Independence: The responses to individual items are independent of one another, conditional on the latent trait. Item Characteristic Curve (ICC) Fit: The relationship between the latent trait and item responses follows the expected curve. Freedom from Speededness: The test is not influenced by time constraints or the speed at which individuals answer. (Hambleton, 1991: 151)

To verify these assumptions for the application of Lord's Two-Parameter Logistic Model, the researcher followed the following steps:

3.5.1. Verification of the Unidimensionality Assumption:

Data analysis according to Lord's model requires verifying the unidimensionality assumption, meaning that all items collectively measure a single underlying trait. This ensures the objectivity of the test in measuring the intended attribute or phenomenon (Hulin et al., 1983:79). Based on the researcher's review, factor analysis, specifically the Principal Component Analysis (PCA) method, is among the most common approaches for verifying this assumption. Using the SPSS software (version 25), the researcher entered the data from a sample consisting of (568) individuals, including their responses to the (50) test items. The analysis revealed a single dominant factor for each section included in the test. To enhance interpretability, the factor was rotated using Kaiser's Varimax orthogonal rotation method. This process yielded a general factor. The factor interpretation adhered to Guttmann's minimum thresholds, wherein a factor is considered statistically significant if its eigenvalue is equal to or greater than (1) (Abdel Khalek, 1983:118). Thus, the analysis supported the unidimensionality of the test items.

Moreover, applying the Standard Error of Measurement (SEM) equation by Burt and Banks resulted in an acceptable saturation value for each item with its respective factor for each academic level at (0.163) or higher. Thus, the unidimensionality assumption was verified. According to Reckase (1979: 214), the explained variance for the eigenvalues of the variables should not fall below (20%) to confirm unidimensionality. The results of the factor analysis satisfy this condition, further corroborating the unidimensionality obtained. Table (7) and Figure (2) illustrate these results in detail.

Table (7): Eigenvalues, Explained Variance, and Cumulative Explained Variance

Sequence	Factor	Eigenvalue	Explained Variance (%)	Cumulative Explained Variance (%)
1	Principles of Psychological and Educational Measurement	30.821	30.929	30.929
2	Educational Objectives	27.017	57.240	26.311
3	Achievement Tests	22.120	79.274	22.034
4	Psychometric Properties	20.042	100.000	20.726

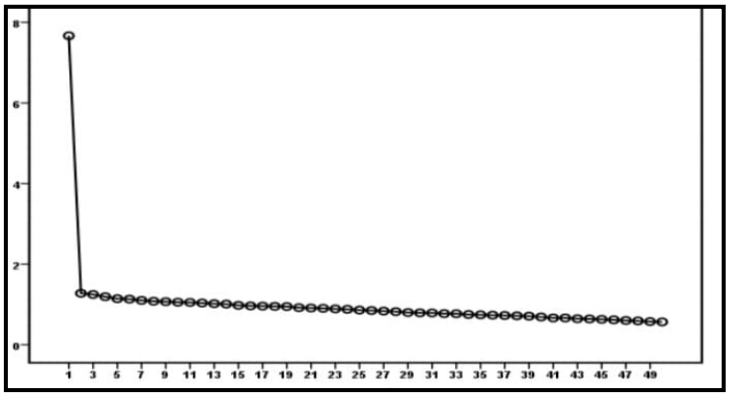


Figure (2) Shows the Graphic Representation of the Achievement Test Factors.

3.5.2. Assumption of Local Independence

This assumption states that responses to an item should be independent of responses to other items, and any observed relationship between items should be explained solely by their shared association with a latent variable (θ). In other words, local independence implies that, at a constant level of the trait being measured, there should be no correlation between responses to different items. A violation of this assumption may lead to parameter estimates that differ from what they would be if the data were locally independent (Reeve, 2003: 12).

In this context, Hambleton and Swaminathan (1985), as well as Warm (1978), suggest that the assumption of unidimensionality is equivalent to the assumption of local independence, though the reverse is not necessarily true. Accordingly, the researcher inferred that the assumption of local independence was satisfied through the validation of the unidimensionality assumption.

3.5.3. Nature of the Item Characteristic Curve (ICC)

This assumption refers to the nature of the Item Characteristic Curve, which describes the relationship between an individual's ability (θ) and their performance on an item. The shape of the ICC depends on the item's parameters, including difficulty (β),

discrimination (a), and the individual's ability (θ). It is expected that the test items will vary in their locations along the ability scale, aligning with the assumptions of the model used in the current study.

3.5.4. Freedom from Speededness

Since the test was allocated sufficient time based on calculations using the time required for responses, any failure by a respondent is attributed to a lack of ability rather than insufficient time or the influence of speed. Thus, the researcher has verified all four assumptions of the latent trait theory, ensuring that the data derived from the achievement test is suitable for analysis under this framework. Subsequently, the data was analyzed using the (Bilog-MG3) software, following the two-parameter logistic model of Lord.

3.6. Verifying the Fit of Individuals' Responses to Test Items with Lord's Model

The researcher utilized the (Bilog-MG3) program to evaluate the fit of individuals' responses to the test items based on Lord's two-parameter logistic model. Results indicated that all individual responses conformed to the model. The chi-square values ranged between (37.88) and (3.86), highlighting that all items were statistically significant at a significance level of (0.05) or higher. An item is considered non-conforming to the model if its probability value is less than or equal to (0.05.) The results are presented in Table (8), which details these values:

Table (8): Chi-Square Values and Significance Levels for Lord's Model

Item	Chi-Square Value	Significance Level (p-value)	Item	Chi-Square Value	Significance Level (p-value)
1	**8.67**	**0.24**	**26**	**37.88**	**0.06**
2	**7.21**	**0.20**	**27**	**10.76**	**0.25**
3	**4.71**	**0.38**	**28**	**27.82**	**0.19**
4	**11.17**	**0.14**	**29**	**35.43**	**0.08**
5	**8.26***	**0.27**	**30**	**28.11**	**0.13**
6	**3.86**	**0.53**	31	**32.61**	**0.10**
7	**17.84**	**0.11**	32	**12.32**	**0.28**
8	**8.10**	**0.36**	33	**32.70**	**0.17**
9	**15.13**	**0.18**	34	**6.87**	**0.36**
10	**4.87**	**0.44**	35	**9.45**	**0.42**
11	**8.55**	**0.21**	36	**17.76**	**0.31**

12	**8.01**	**0.29**	37	**21.98**	**0.22**
13	**8.36**	**0.25**	38	**33.51**	**0.16**
14	**21.72**	**0.15**	39	**38.68**	**0.09**
15	**9.04**	**0.33**	40	**32.57**	**0.10**
16	**17.61**	**0.23**	41	**24.71**	**0.20**
17	**10.08**	**0.17**	42	**7.45**	**0.37**
18	**6.32**	**0.40**	43	**8.50**	**0.32**
19	**6.89**	**0.38**	44	**30.68**	**0.07**
20	**21.60**	**0.13**	45	**5.09**	**0.35**
21	**7.64**	**0.30**	46	**4.21**	**0.43**
22	**6.20**	**0.34**	47	**3.89**	**0.48**
23	**19.56**	**0.12**	48	**4.58**	**0.41**
24	**26.89**	**0.07**	49	**22.84**	**0.27**
25	**31.67**	**0.10**	50	**4.71**	**0.46**

It is evident from Table (8) that all items of the achievement test fall within the acceptable range and conform to the adopted model. This confirms that the researcher has verified the conformity of the test items with Lord's two-parameter logistic model. This conclusion establishes the reliability and validity of the test for further analyses, ensuring that the responses align appropriately with the theoretical framework of the model.

3.7. Statistical Methods:

In order to analyze the research data, the researcher utilized both the Statistical Package for the Social Sciences (SPSS) version 25 and the (Bilog–MG3) program. The following statistical methods were employed by the researcher:

- Chi-square (χ^2) Goodness of Fit Test for calculating:
  A. Content validity.
  B. The alignment of individual responses to test items with the Lord model.
- Gelbson's Equation for calculating reliability.
- Cronbach’s Alpha Equation for calculating reliability.
- Biserial Point-Biserial Correlation Coefficient for calculating the validity of item-total score correlation.
- Exploratory Factor Analysis (EFA) to assess the data fit with the Lord model.
- Equation for calculating item information function.
- Equation for calculating test information function.
- Equation for calculating the standard error.

4. Results Presentation and Interpretation:

4.1. The First Aime: Identify the Estimates of Difficulty and Discrimination Parameters for the Test, the Information Function of Items, and their Standard Errors According to Lord's Model.

The First Aime was to determine the item difficulty and discrimination indices, item information functions, and standard errors for each item based on the Lord model. To achieve this Aime, item analysis was conducted on the achievement test, including the calculation of item difficulty and discrimination indices, as well as the maximum information values (θmax) and their corresponding item information (I(θ)max) values, according to the two-parameter logistic model. The results were as follows:

Table (9) Illustrates the Difficulty and Discrimination Parameters, the Maximum Value of the Item Information Function, and the Standard Errors for the Test Items Using the Logit Model.

Item Sequence	Difficulty	Standard Error (Difficulty)	Discrimination	Standard Error (Discrimination)	Maximum Information Function Value (θ)max	Standard Error (Information)
1	0.77	0.029	1.56	0.041	0.602	-0.344
2	0. 54	0.025	1.14	0.032	1.407	-0.285
3	0.68	0.023	1.74	0.043	1.809	-0.341
4	0.64	0.035	1.39	0.039	1.652	-0.096
5	0.55	0.034	0.99	0.048	1.800	0.10٥
6	0.51	0.025	0.93	0.020	1.203	-0.185
7	0.49	0.021	1.66	0.044	0.809	-0.095
8	0.71	0.036	1.32	0.033	1.406	0.102
9	0.68	0.024	1.82	0.031	1.200	0.265
10	0.36	0.027	0.88	0.037	0.806	0.180
11	0.42	0.025	0.92	0.027	0.539	0.096
12	0.40	0.032	1.01	0.046	0.672	-0.272
13	0.56	0.035	1.37	0.027	0.816	-0.283
14	0.75	0.029	0.98	0.026	1.620	-0.297

15	0.69	0.035	1.17	0.029	1.442	0.039
16	0.70	0.035	1.92	0.042	1.684	0.297
17	0.45	0.030	0.99	0.033	0.922	0.211
18	0.51	0.031	1.42	0.039	0.494	0.086
19	0.58	0.032	1.80	0.042	0.622	0.024
20	0.72	0.034	1.62	0.045	0.834	-0.108
21	0.73	0.033	1.20	0.022	0.951	-0.313
22	0.64	0.033	1.47	0.041	1.228	-0.194
23	0.44	0.024	0.96	0.028	1.726	-0.353
24	0.64	0.031	0.91	0.045	1.427	-0.149
25	0.50	0.033	1.49-	0.036	1.832	-0.349
26	0.69	0.028	1.74	0.030	1.321	0.124
27	0.52	0.032	1.31	0.031	1.820	0.291
28	0.66	0.020	1.22	0.047	0.766	-0.223
29	0.72	0.033	0.93	0.044	0.502	-0.004
30	0.46	0.26	0.89	0.36	1.498	-0.282
31	0.59	0.031	0.96	0.038	1.010	-0.182
32	0.64	0.032	0.97	0.037	1.644	-0.049
33	0.65	0.035	0.82	0.045	1.398	-0.017
34	0.67	0.027	1.81	0.042	1.102	0.017
35	0.47	0.025	1.44	0.043	1.036	-0.279
36	0.69	0.030	1.76	0.032	0.966	-0.219
37	0.56	0.032	1.78	0.040	1.638	-0.126
38	0.34	0.032	1.62	0.048	1.488	0.104
39	0.52	0.034	1.69	0.033	1.269	0.200
40	0.48	0.031	1.03	0.037	0.802	0.072
41	0.64	0.031	1.29	0.028	0.640	0.238
42	0.60	0.022	1.30	0.022	0.860	-0.067
43	0.66	0.031	0.82	0.039	0.580	-0.288

44	0.51	0.027	1.47	0.047	1.140	-0.099
45	0.46	0.033	1.82	0.024	1.660	0.222
46	0.51	0.023	1.61	0.025	1.208	-0.048
47	0.59	0.027	1.87	0.033	0.604	-0.216
48	0.44	0.030	1.94	0.037	1.100	-0.039
49	0.39	0.029	1.97	0.023	1.480	-0.079
50	0.56	0.024	1.93	0.041	1.080	0.364
Mean	**0.573**	**0.034**	**1.37**	**0.042**	**1.162**	**-0.056**
St. Devition	**0.111**	**0.032**	**0.36**	**0.046**	**0.409**	**0.201**

Table (9) shows that the difficulty parameter values ranged from (0.77) to (0.36) Logit, with a mean of (0.57) and a standard deviation of (0.11). The discrimination parameter values ranged from (1.97) to (0.82) Logit, with a mean of (1.37) and a standard deviation of (0.36). The maximum item information values ranged from (0.99) to (1.83) Logit, with a mean of (1.16) and a standard deviation of (0.409). This indicates that the values of the difficulty and discrimination parameters for the achievement test are consistent with the Lord model of latent traits theory, thus achieving the First Aime of the current research.

If there is consistency between the responses observed by the examiners to the Item and their likelihood of success, there is consistency between their responses to the Item and their overall grades on the scale, i.e., their responses to the rest of the Items, indicating agreement between the attribute measured by the Item and that measured by the rest of the Items, across the sample, thus matching the model requirements. (Awadallah, 2000: 158-161).

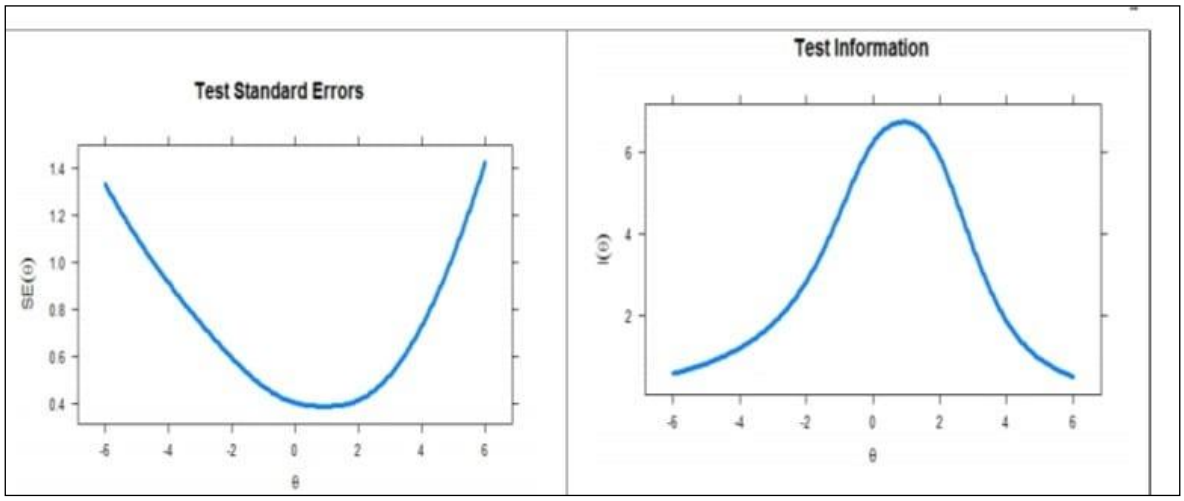
Where the statistician's value is a statistical function, the Item should be deleted, because it Do not express the attribute of the rest of the Items, and these statistics are used Also to exclude examiners with non-model responses. The relative difficulty of Items in these examiners varies in matching results Before deleting the Items, then re-analysing and after deleting the Items that do not correspond to the model, they may be responsible for some Items not conforming to the model. (Wright & Stone ,1979 ; 82)

These indicators also indicate that the characteristics curves of the Items are model-friendly and have a similar tendency or curvature. When the parameters of the Items are independent of the sample, the Items' ability to distinguish is relatively equal. According to the Lord's model.

4.2. The Second Aime: Analyze the Information Function and the Standard Error for the Test Constructed Based on Lord's Model.

Since the information function of the test and the standard error for the test estimates are key characteristics of the latent trait theory, the (Bilog-MG3) program was used to determine these two functions for the achievement test. The maximum value of the information function ranged between (0.5) and (-0.5) Logit on the trait continuum, i.e., at its midpoint, and gradually decreased as it moved away from the midpoint. This aligns with the model's predictions, suggesting that the test provides the maximum amount of information at moderate estimates, where the standard error is minimized and the information function is at its peak. On the other hand, at the extremes, a decline in the test's information is observed. As shown in the figure, the standard error corresponding to the maximum information value was minimized and approached zero, reaching a value of (-0.056), with a standard deviation of (0.201). This indicates high estimation precision, reflecting the test's good reliability. It is well-established that there is a relationship between the test's information function and reliability: as the test's information function increases, the standard error decreases, which in turn leads to higher reliability.

Figures (3) and (4) below respectively illustrate the Information Function Curve and the Standard Error Curve.



Figures (3) and (4) represent the Information Function Curve and the Standard Error Curve for the achievement test, respectively.

The figure (3) shows that the maximum value of the information function was between (0.5) and (-0.5) loggit on the connection of the measured trait, i.e. in the middle and gradually decreasing by moving away from the midpoint. This matches the model's expectations and the test provides the maximum amount of information at the estimate, while at the parties the amount of information for the test decreases.The presumption is that the curves with the distinctive observation of the Item have a common general form or curve, stating (Kadhim et al., 1996: 353-354) that when there is a general form or curve of all curves with the distinctive observation of the Items, that is, that these curves have the same force to distinguish individuals from the attribute.

It is also noted from figure (4) that the test is equivalent to the corresponding statistical estimates. in accordance with the accepted criterion of the standard error of those

estimates with those estimates derived from the analysis of the performance of the total sample personnel, This means that the estimate of an individual with a certain overall score on this scale is not affected by the different performance level of the analysis sample capability ", thus freeing an individual's measured ability from that of other individuals who answer to him, This means that the parity of the corresponding estimates in the overall sample analysis as benchmarking estimates and those derived from sample performance, indicating the liberalization of the parameters of the Items from the sample's ability to which the test was applied.

In light of the findings achieved in the current research, the researcher successfully met all the Aimes outlined, along with the results obtained. Consequently, the researcher managed to construct a criterion-referenced achievement test in psychological and educational measurement and evaluation based on Lord's Model. This test serves as a reliable benchmark aligned with the assumptions of Latent Trait Theory, reflecting acceptable ranges for the Information Function and Standard Error. Therefore, the researcher recommends that specialists use the test developed in this study due to its precise psychometric properties, which provide an accurate representation of students' abilities in psychological and educational measurement and evaluation.

4.3. Recommendations:

In the light of the researcher's findings, he recommends that:

1. Conduct training courses on how to use the Lord's model to develop the methods of preparing and building tests for psychological and educational researchers.

2. Using A Criterion Referenced Tests to Assess Students' Attainment Level.

4.4. Suggestions:

The researcher suggests conducting the following studies:

1. Conducting a Study on Achievement Tests Using the Rasch Model as One of the Models of the Latent Trait Theory.

٢. Conducting a Study on The Tests A Criterion Referenced Using the Birnbaum Model of the Latent Trait Theory.

٣. Conducting a study on Achievement Tests Using the Grading Models of the Latent Trait Theory such as Rating Scales and Partial Credit models, Etc.

References

- Abdulkhaleq, Ahmed Mohamed (1983). Basic Dimensions of Personality. Alexandria: University Knowledge House.

- Abdulrahman, Saad (1998). Psychological Measurement. Kuwait: Al-Falah Library.

- Abu Allam, Raja Mahmoud (2005). Evaluation of Learning. Dar Al-Maisarah, First Edition, Amman.
- Al-Dosari, Rashid Hammad (2004). Modern Educational Measurement and Evaluation: Principles, Applications, and Contemporary Issues. First Edition, Amman: Dar Al-Fikr for Publishing and Distribution.
- Al-Heila, Mohammad Mahmoud (2008). Instructional Design: Theory and Practice. Dar Al-Maseera for Publishing and Distribution, Amman, Jordan.
- Al-Imam, Mustafa Mahmoud (1990). Evaluation and Measurement. Baghdad: Dar Al-Hikma.
- Allam, Salah El-Din Mahmoud (1986). Contemporary Developments in Psychological and Educational Measurement. Kuwait University: Department of Composition, Translation, and Publishing.
- Allam, Salah El-Din Mahmoud (2000). Educational and Psychological Measurement and Evaluation: Basics, Applications, and Contemporary Directions. Dar Al-Fikr Al-Arabi, Cairo, Egypt.
- Allam, Salah El-Din Mahmoud (2001). Diagnostic Criterion-Referenced Tests in Psychological and Educational Fields. Cairo: Dar Al-Fikr Al-Arabi.
- Allam, Salah El-Din Mahmoud (2005). Item Response Theory Models: Unidimensional and Multidimensional Applications in Psychological and Educational Measurement. Dar Al-Fikr Al-Arabi, Cairo, Egypt.
- Allam, Salah El-Din Mahmoud (2007). Alternative Educational Evaluation: Its Theoretical and Methodological Foundations and Field Applications. Dar Al-Fikr Al-Arabi, Cairo, Egypt.
- Al-Kilani, Abdullah Zaid, Ahmad Al-Taqi, and Abdulrahman Adas (2009). Measurement and Evaluation in Learning and Education. United Arab Marketing and Supplies Company.
- Al-Taeb, Masoud Hussein (2018). Scientific Research: Rules, Procedures, and Methods. Arab Bureau of Knowledge, Amman.
- Al-Shujairi, Yasser Khalaf and Haidar Abdul-Karim Al-Zuhairi (2021). Modern Trends in Psychological and Educational Measurement and Evaluation. Arab Society Library for Publishing and Distribution, Amman.
- Al-Shareefeen, Nidal Kamal (2006). Psychometric Properties of Criterion-Referenced Tests in Educational and Psychological Measurement. Journal of Educational and Psychological Sciences, University of Bahrain, Issue (3), Volume.(٧)
- Acherman , T. A. (1994) , using , Multidimcu sional Item Response theory to understand what (Item and test are Measuaring Applied Measurement).

- Anastasi, A & Urbina, S. (1997): (Psychological testing), 7th ed; New York; prentice-Hall.
- Awadallah, Mohammed Abdulrahim. (2000): (Comparison of Rush Model Style and Traditional Method of Building Intelligence Tests Using the Test of Predicting Academic Achievement), (PhD thesis published), Baghdad University, Ibn Rushd School of Education.
- Dale T. (1995),"Classroom Testing for Teacher Who Hat Testing Criterion-Referenced Test, Construction and Evaluation Reports-Research, Technical," US Department of Education Office of Educational Research and Improvement, Washington ,17.
- Ebel , R. L. (1972) :(Essentials of Educational measurement), New Jersey; prentice Hall Inc.
- Foran, T.G. (1961): "A Note on Methods Measuring Reliability," Journal of Educational Psychology Vol.22, No.4.383-387.
- Furqan, Second Edition, Amman, Jordan.
- Ghoneim, Mohamed Abdel-Salam (2004). Principles of Psychological and Educational Measurement and Evaluation. Dar Al-Fikr Al-Arabi, Cairo, Egypt.
- Harrison, A. (1983) :(A language testing handbook). London: Macmillan. ISBN 0-333-27174-2.
- Hambleton, R.K., and Swaminathan, H. Algina, J. (١٩٨٥) (Criterion- Referenced Testing and Measurment :Areview of Technical Issuse and Development). Review of Educational Research. Vol. 48, No. 1.
- Hambleton ,R.K., Swaminathan, H., & Rogers, H.J. (1991). Fundamentals of Item Response Theory, Sage Publications, Newbury Park CA.
- Hulin, C.L.; Drasgow, F.; & Parsons, K. (1983): Item Response Theory Application to Psychological Measurement, Ilinois, USA: Dow Jones-Irwin, Homewood.
- Huynh, H. (1994). A new proof for monotone likelihood ratio for the sum of independent Bernoulli random variables. Psychometrika, 59, 77–79.
- Kadhim, Amina, Anwar al-Sharkawi et al. (1996), (Contemporary trends in measurement, psychological evaluation and pedagogy), Anglo-Egyptian Library, Cairo.
- Linden , W. I. and Hambleton , R. k. (1997) , Item response theory, brief history , common models and extension , New york , sprmger.
- Nashwati, Abdul-Majeed (1998). Educational Psychology. Dar Al.
- Nunnally, J.C. (1978): (Psychological Theory), 2nd Ed., New York: McGraw-Hill.

- Ouda, Ahmed Suleiman (1998). Measurement and Evaluation in the Teaching Process. Dar Al-Amal for Publishing and Distribution, Jordan.

- Popham, W. J. (1978). (criterion referenced measurement). Englewood cliffs، New Jersy. Prentice – Hall.

- Reckase, M. (1997):(The Past and Future of Multidimensional Item Response Theory). Applied Psychological Measurment

- Reeve, Bryce B. (2003). An Introduction to Modern Measurement Theory. Outcomes Research Branch, Applied Research Program, National Cancer Institute.

- Rovinelli, H. & Hambleton, R. (1977): "On the Use of Content Specialists in the Assessment of Criterion-Referenced Test Item Validity", Dutch Journal for Educational Research, 2

- Suen, H. K. (1990) . (Principles of Test Theories) , New Jersey Lawrence Erlbaum Associates , Inc, Hill sadle.

- Van De Vijer, F.J.R.(1986): (The Robustness of Rasch ectimates), Applied Psychological measurement.

- Warm, T.A. (1978). (A Primer of Item Response Theory). Oklahoma: U .S.Coast Guard Institute 73/69.

- Wright, B. D., & Stone, M. H. (1979): (Best test design. Chicago): MESA Press.

- Zumbo, B. (2007):(Three Generations of DIF Analyses: Considering Where It Has Been, Where It Is Now, And Where It Is Going). Language Assessment Quarter**,** 4(2), 223-233.