Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 August; 6 (9s2): 470-491

Using Dual Criterion Model in the Construction of Teaching Thinking Test among Students of the Departments of Educational and Psychological Sciences

[1]Zahraa Raad Abdul Rasul, [2]Prof. Dr. Safa Tairq Habeeb
[1]zahraaamin77@gmail.com
[2]College of Education Ibn Rushd / University of Baghdad
safa.tairq@ircoedu.uobaghdad.edu.iq

Received: 14- June -2023
Revised: 16- July -2023
Accepted: 21- August -2023

Abstract
The present study aims at using dual criterion model in constructing a diagnostic criterion- Referenced test in the course of teaching thinking for the students at the Department of Educational and Psychological Sciences. To achieve this aim, a test of (80) MCQ items with four alternatives has been constructed. The sample of the present study consists of (604) male and female students distributed on the universities of : (Baghdad, Al-Mustansiriya, Al-Iraqya, Al-Kufa, Wasit, Karbala, Al-Qadisiyah). from fourth-grade students in the colleges of education, and for the morning study, for the academic year (2021/2022), The sample was selected by the stratified random method with a proportional distribution. The items have been statistically analyzed by using the (ICL) program in order to grade the items based two-parameter model. The researcher verified the hypothesis of Unidimensionality by using factor analysis by the principal components Kaizer method, and computing the relationship between the item scone and the test total score. In addition, the assumption of freedom from speed factor as well as the decrease of the Guessing factor have been verified through assuring the individuals performance who were of lower ability on the difficult items. As a result, the items have been matched to the assumptions above. To identify the cut-off score of the test, the opposite groups method has been used, where the score has been found out to be (51) representing (68.5%). The assumptions of this model have been verified by the significance of matching all the test items with the model itself except five items which were not matched. The final version of the test consists of (75) items. Result of the data analysis show that all the items and the whole lest have good psychometric characteristics. Based on the results, the researcher reached a set of conclusions, recommendations and suggestions.

Keywords: Criterion-referenced test , dual criterion model , teaching thinking .

First , research problem:

There are several problems and difficulties related to the cognitive aspect of academic achievement, which vary according to the different philosophy of measurement and evaluation and its objective, whether these problems are related to evaluation tools (achievement tests), problems related to the criteria on which the interpretation of the student's degree is based on a test, or problems related to performance evaluation, and measurement here is considered as the first step in the process of evaluating academic achievement, and for measuring its tools, which are used to estimate the student's performance in a quantitative form (Sergewa, 2016 : 3).

Among these problems in the interpretation of scores derived from measurement tools in the norm-referenced assessment, which depends mainly on the nature of the reference group used to obtain the standard, as the characteristics of this group should be as similar as possible to the characteristics of the group of testers in terms of age, gender, grade and geographical area. Because we cannot evaluate individuals based on the raw scores each received on the test, these scores do not provide us with any information about an individual's standing in the trait or ability that the test measures. The score an individual receives becomes meaningless and difficult to interpret if we do not return it to the norm (Allam, 1986 : 105).

Despite the abundance of studies and research that applied theOne Parameter Model but there is a difficulty in construction of test items that achieve the assumption of equal discrimination for all the items that make up the test, even if we try to do so, but it is within very narrow areas of ability, and on the contrary, we note the lack of

studies that applied the triple- parameters model because it needs very large samples. In addition, there is a difficulty in estimating the parameters of the item , especially the guessing parameter, and from here the researcher used the dual criterion model (Habib and Al-Hajami, 2018 : 103)

The criterion assessment is used as a diagnostic criterion- Referenced test tool to evaluate the educational program in order to increase its effectiveness because there are defects in teaching; because it diagnoses shortcomings or gaps in the performance of the student, which is a possible cause in any part of the educational program, whether in the curriculum, in the teaching method, or even in the test itself or in classroom interaction. The trend to use criterion tests is a new trend in the teaching and learning processes that has many positive effects, including identifying strengths and weaknesses in students' learning, as well as determining the extent of their mastery of the tasks required of them , and determining the mechanisms of treatment if they fail to master them within various methods, including with regard to modifying teaching methods and methods, and in light of benefiting from the positives of criterion tests based on encouraging feedback in using this method in evaluating student learning (Sergewa, 2016 : 20).

Based on what has been mentioned, the problem of the current research is determined by the absence of a reference diagnostic criterion- Referenced test in the teaching of thinking for students of the departments of educational and psychological sciences in the faculties of education, based on the modern measurement theory and the measurement models emanating from it (the dual criterion model) specifically. Because of the accuracy of this test in measuring the educational and psychological characteristics of individuals.

Second: The Importance of Research:

Those interested in education and educators emphasize the need to pay attention to the quality of measurement and evaluation methods that contribute to making objective decisions on scientific grounds, and considering that evaluation is part of the fabric of the educational system, through which the extent of achieving the objectives of the educational system is determined, which provides continuous feedback that helps to modify and develop the system and increase its efficiency and then its quality and the quality of its outputs. The weakness of the educational system (inputs _ processes _ outputs) may be due to a weakness in the tools and means used, and therefore the interest in evaluation and its tools is a good input for reforming education. Raising the quality of its outputs, this is done through the evaluation of all components of the system and the elements associated with it, including the elements of the basic and secondary educational situation. Since the student is one of the basic elements, the evaluation must include all aspects of the learner's personality (cognitive _mental _ physical) (Ghunaim, 2003 : 96).

Educational tests and psychological measures are an important part of the educational process, through which the teacher can judge the extent to which the objectives of the educational and training programs that he teaches have been achieved. They also benefit the student in identifying his level of academic achievement through his performance in the test. There are many uses of tests in educational institutions, some of which are used for measurement, some of which are used for evaluation and some of which are used for diagnosis (Tamimi, 1999 : 147).

The modern theory came with premises that address many of the disadvantages of the classical theory of measurement (Eid , 2004 : 235), and that the importance of modern theory in measurement comes from its focus on the level at which the individual reaches in his performance, and the estimation of his acquisition or achievement, and that the use of modern theory makes the characteristics of the item (difficulty, discrimination, guessing) independent of the sample of individuals used to estimate these characteristics, as well as make the estimation of the capabilities of individuals independent of the items of the tests used to obtain these estimates, as well as improving the accuracy and reliability of the results by identifying and deleting individuals and items that do not correspond to the model used(Hambleton, 1989, :187) , and this is the customary interest of contemporary educational standards and tests.

The importance of this theory is highlighted by identifying the characteristics of the responses of individuals and their characteristics of the test items , and these features or traits can be called latent traits , and called latent, that is, hidden or invisible because it can not be observed directly, but we infer its existence by observing the behavior of the individual by his responses to the test items (Allam, 1986 : 102).

And presenting the item response theory models according to considerations developed by the psychological and educational measurement developers according to empirical evidence to prove the justifications and reasons for selecting each of the item response theory models, where the characteristics of the data contribute to facilitating the selection of the appropriate model, however, after testing the appropriate model, it is somewhat complicated because Appropriate use of forms relies heavily on matching the form to the data. If the wrong model is chosen for the selection data, the results may be incorrect within the individual response theory models (IRT) (Dodouh, 2023:25).

A study (Dabbous ,2016) pointed to the use of theitem's response theory in the construction of the collection of items of the criterion Criterion-referenced test in mathematics according to the dual criterion logistical model, and the results showed that the assumptions of theitem's response theory were achieved. The results of the analysis with respect to the matching of the two-stage item e to the two-marker model also showed that all the item e were identical to the two-marker model (Dabbous, 2016 : 1453).

Because the subject of teaching thinking is one of the modern subjects that began to be taught in universities, and most of the studies that were built according to theitem's response theory - as far as the researcher knows - used the single-parameter model (Rush model), and therefore the researcher decided to use the two-parameter model in construction of a diagnostic criterion- Referenced test in the subject of teaching thinking.

Third: Research Objectives:

The current research aims to construct a reference diagnostic criterion- Referenced test in the subject of teaching thinking according to the dual criterion model among university students.

Fourth: Limits of Research:

The current research is limited to fourth grade students in the Department of Educational and Psychological Sciences in the Faculties of Education for Humanities in Iraqi universities (Baghdad , Mustansiriya, Iraqi, Wasit, Karbala, Kufa , Qadisiyah) for the academic year (2021-2022) for morning study.

Fifth: Terminology:

Test Referenced-Criterion(CRT):

Defined by:

- Papam and Husek (1969 ,Popham and Husek): The test, which does not depend on its interpretation of the student's score on the characteristics of his group to which the test was applied, but depends on balancing the student's performance with a predetermined expected performance levels. These levels are determined in the light of the behavioral objectives to be measured. These tests measure the extent to which the student achieves these behavioral objectives . Thus, the efficiency of each student can be verified from his acquisition of existing skills and knowledge and then the diagnosis of weaknesses and strengths (3 : 1969 ,Popham and Husek).

- Abdul Hafez (1999): It is concerned with detecting the weaknesses of the pupil, the sources of error and the reasons, which may be collective or individual, and each of them has its usefulness and importance. The collective tests stop the teacher on the prevalence of errors among pupils, they are widely used and applied to a large number of students (Abdul Hafez, 1999 : 297).

The researcher adopted the definition of Popham and Husek (1969 Popham and Husek), a theoretical definition of the diagnostic criterion- Referenced test criterion reference because it is commensurate with the purpose of the current research.

The practical definition of the term diagnostic criterion- Referenced test is referenced, it refers to:

A sample of items of the MCQ type, with four alternatives, one of which is correct and the remaining three are wrong, prepared according to the assumptions of theitem's response theory and analysis of the results of students' responses using the dual criterion model in order to achieve the accuracy and objectivity of

measurement and diagnose the degree of mastery of fourth grade students in the departments of educational and psychological sciences by balancing their performance with the degree of pieces specified in the light of the objectives set.

Dual criterion model :

Defined by:

- Hambleton et al. (1991): It is the model in which the item is represented by the logarithmic weighting function, relying on a new parameter, the discrimination parameter, as well as the difficulty parameter used in the single model (Rush model) (Hambleton et al., 1991 : 12).

- Allam (2001): It is a model of item's response theory which assumes that both the difficulty and discrimination factors are variables, and that the guess for all item e is zero (Allam, 2001 : 126).

The researcher adopted the definition of Hambleton et al. (1991 Hambleton et al.,), a theoretical definition of the two-parameter model because it is commensurate with what the research aims to do.
As for the procedural definition of the term bi-parameter model, it refers to:
All statistical analyses that the researcher will conduct for the items of the diagnostic criterion- Referenced test are referential, starting from verifying the basic assumptions of the item's response theory and matching the items with their parameters (difficulty , discrimination) to acceptable ranges, as well as verifying the norm characteristics of the test.

Teaching Thinking:

The subject taught in the faculties of education in the fourth stage of the departments of educational and psychological sciences in the faculties of education in all Iraqi universities, which is one of the basic subjects that provide the student with an integrated set of knowledge and skills about thinking and its components and determinants, types and skills associated with it, and various thinking programs (Sectoral Commission for Educational Sciences, 1990 :59).

- **Criterion-referenced tests:**

Criterion-based Measurement sciences represents an important input in the study and appreciation of various phenomena, which led to increased accuracy in measurement in the educational and psychological aspects, which helped to understand the phenomenon and to the accuracy of prediction and its control (Kadhim, 2020 : 23).
The tracer of the educational and psychological measurement literature notes the multiplicity and diversity of the definitions of the test criterion-referenced tests , and this may be due to the novelty of the concept of criterion-referenced tests (the reference of the test) and the multiplicity of types of these tests, as well as the lack of agreement on the concept of the test to which the grades of individuals are attributed in these tests, as some measurement scientists believe that the test means a comprehensive range of knowledge and skills that are well defined so that the performance of the individual in the test can be balanced with this scope to know what this individual can perform and what can not. This means that the balance here is in the light of the test items themselves. Each single test contains a knowledge or skill that the individual should acquire in order to succeed in performing a work with the successful leader. The successful leader can not be able to do so, so that the signs indicated do not enable him to pass, and not to acquire the skills that are safe (24).
While others believe that the criterion indicates the level of performance or cut-off score , when the teacher would like his students to achieve a certain level of mastery, each of them should answer correctly about a percentage of the test items are determined in advance, and Robert Glaser is the first to develop the term of reference criterion-referenced tests, and many researchers in the last two decades in the seventies and eighties have added new insights about the concept, so we have a large number of data and studies on the uses and technical aspects associated with the concept of reference criterion-referenced testing , and that these studies and multiple data led to the development of the reference criterion measurement to collect double meaning. It is in the sense of returning the individual's performance to a standard, which is a specific level of performance and explains the student's degree of testing in an absolute manner. (Abu Dhabi 1994, 52).

In view of the important role that tests play, especially in determining the fate of students, we find that most educational systems in the world attach great importance to the teacher as the main responsible for the process of evaluating students, especially since preparing such tests, applying them, correcting them, analyzing and interpreting their results, requires the teacher to possess a set of basic competencies , in the field of building tests, especially objective. This trend has become clear in all educational institutions that seek to pay attention to the competencies. (AL-Hijame and Karma 226: 2021).

The achievement of students and their acquisition of certain knowledge and skills are among the priorities of educational institutions, so the measurement of achievement, It is one of the important sources for defining the outcomes of the educational processes in the education sectors with the necessary information needed to make educational decisions such as passing or failing, or classifying students into categories according to their levels of achievement (Al-Jubouri and Al-Kadhimi, 2014: 423).

❖ **Characteristics of Criterion-referenced tests :**

criterion-referenced tests are based on evaluating the learner's performance in the light of a specific test that takes the student's level into account. The use of these tests has become common in the field of education.

These tests are classified by a number of characteristics, the most important of which are the following:

1. Based on a number of behavioral objectives .
2. It is designed to be highly relevant because its contents are related to its objectives.
3. They are representative samples of the actual behavior or performance of individuals, on which performance can be interpreted in light of predetermined interval scores (Mansi, 2001: 224).

❖ **The purpose of Criterion-referenced tests :**

Tests are usually used to make decisions about the capabilities and knowledge of students and decisions taken by teachers or educational administration based on the results of tests Failure or success of students in a difficult subject or subjecting students with low achievement to a specific treatment program or classifying students into certain specialties or changing the teaching strategies adopted by the teacher, and the purposes of Brown (1996) vary between the purpose of the test and the type of test, whether it is criterion or standard, (Ababneh, 2009 : 25). Table (1) shows that pairing:

Table (1) Uses of reference criterion-referenced tests and reference standardized tests

test objective	Test type Criterion or Norm referenced
Admission test Preparing to enroll in a training program	**Norm-referenced** : We want to find the appropriate level for the student in the program by comparing with a number of other students
Appointment	**Norm-referenced**: We want to compare the performance of the individual with the performance of each individual in the group
Diagnostic	**Criterion-referenced :** We want to measure specific aspects of student knowledge that are related to the objectives of the program
Measuring student achievement (**achievement**)	**Criterion-referenced** : Because we want to measure specific points in the knowledge of the student and these points are related to the objectives of the program

Areas of use of Criterion-referenced tests : -

a. **Evaluating mastery learning**: The concept of the test is not limited to describing the behavioral field, but includes determining the level of mastery (performance) in the form of a numerical estimate (the correct response rate in the task should not be less than (80%) until the student is classified as proficient.

B. Tests use reference simulators to make educational decisions such as:

What does the learner know? Where can a student be in the learning continuum? What learning do students need? This is done by making tribal decisions that are made at the beginning of the educational program.

c. **Used in formative assessmen**t: Bloom pointed out that we apply the Criterion-referenced test at the end of each unit, analyzing the results shows us the strengths and weaknesses of learners and suggested remedial methods to overcome deficiencies. A second image of the Criterion-referenced test is then used, after a sufficient period of time.

d. **Used in the post assessment**: where we can measure here the basic skills as they are prerequisites for new learning, and to determine the type and level of the appropriate program with the capabilities of students, in addition to determining the basic level of teaching where to start? After teaching, where did you arrive?
e. In the areas of diagnosis: where the test is based, it deals with models of tasks that have a high probability of error, in addition to the presence of a representative sample of the basic tasks.
F. Evaluation of the achievement in the programs based on the results: Some of the results are simple and others are complex, and the performance on the tests reflects the results that the student has mastered, which is necessarily a specific achievement area, and determine the level of acceptable performance, related to the target decision of test development, and whether educational products have been developed or not? It is a diagnostic training process at the same time aimed at ensuring the achievement of the targeted outputs in the educational program (Ababneh, 2009: 27).

- ❖ Types of Criterion-referenced tests :

1. **Targeted criterion-referenced tests :** They are those tests that are based on a set of behaviorally formulated educational objectives, and there is a mating between the test items and these objectives, but the behavioral scope represented by these objectives is not specific, and therefore the items that the test includes are relatively few in number, because they do not represent the comprehensive range of possible items that measure the set of objectives, and these tests are usually applied upon completion of an educational unit or a specific modular unit, with the aim of classifying the two laboratories in two groups, one of which achieved the objectives and the other did not achieve them in light of a specific percentage of the items that should be answered correctly and identify the objectives that each of them could not achieve (Allam, 1985, 24).
2. **Scope criterion-referenced tests** : These tests are adopted by clearly and accurately defining a comprehensive behavioral range of tasks, skills or requirements, and the items included in the test are selected from this range randomly or by randomly applied methods. The scores of these tests are used to obtain statistical estimates of the probability of an individual or group of individuals answering the items of the comprehensive range represented by the correct answer to the items of the test at a certain time, and this helps to generalize the results of the test to the comprehensive scale (Allam, 2001 : 25).
3. **Proficiency tests**: These tests depend on determining the extent to which a particular individual acquires the behavior that the educational or experimental program aims to develop. The proficiency test helps in making decisions related to the individual's mastery of a specific educational goal, skill, or range of skills. Therefore, this test can be a reference for the behavioral scope of the performance measured by the test (Allam, 2001 : 25).

- ❖ **Steps for construction of criterion tests Reference:**

First: Determining the content to be measured: It depends on the nature and limits of that content, if the content to be measured is specific (such as a specific unit of study, or a clearly defined subject of study), it can be enough to know the components of this unit or this content, but if the content to be studied is wide and large, it is advisable to divide it into sub-topics linked to each other so that it can be measured as a unit. This requires the construction of several tests, each linked to one of these sub-topics. There is no fixed rule for the division of content to be measured, but this division must allow for the construction of a test or tests whose items serve as a sufficiently representative sample of the different topics involved in the content. The nature of the students tested, and the time required to teach this content, must also be taken into account.
Second: Determining the general objectives that the test will measure: Each content has associated objectives aimed at measuring the extent to which they are achieved by students. For example, the general objective applies deductive reasoning in solving mathematical problems. " This general objective expresses expected outcomes from the learning process, but needs to be reformulated to reflect behavioral activities that are evidence that the student has applied the deductive method and so on.
Third: Analyzing general objectives into partial objectives : By this, we mean reformulating general objectives by describing a sample of behavioral objectives that can take evidence of the achievement of each of the general objectives . The test designer usually faces some difficulties when performing this analysis. This depends on the behavioral range to be measured (Chalabi, 2005: 172).

Fourth: Specification of the behavioral range measured by the test: When obtaining procedural objectives from the previous analysis, they may not be clear enough to determine the scope of the test items that measure the desired behavior. Of course, it is difficult to select a representative sample of test items from the range if this range, which measures a particular target, is unclear. We have already mentioned that the performance of the individual in a sample of test items that measure a particular goal is used to estimate the degree of mastery of the overall range of items that measure this goal. Therefore, this comprehensive range of items must be well defined and representative samples can be drawn from these items . If the overall range is difficult to determine in full, the test specification can be prepared using the Papam method that has been shown in determining behavioral ranges. However, if the overall scope of the boundary is well defined, the wording of the items proposed by Hayfley may be used.

Fifth: Composition of the test items : It includes the following steps:

1. Selecting the appropriate types of items to measure objectives : After specifying the specifications of the behavioral range, the best types of items should be chosen that measure the behavior identified in the specifications for each goal. It is known that some types of items are suitable for measuring certain objectives to a better degree than others. For example, MCQ items , mating items , and right and wrong items are suitable for measuring remembering, understanding, and sometimes applying, while the items of the article are suitable for measuring the ability to organize information, deduce, interpret, and reformulate ideas.... and so on.
2. Identify the appropriate number of items : In this step we try to identify the appropriate number of items that measure the objectives represented in the previously defined behavioral range (Majeed, 2014: 190).
3. Writing test items : This step needs great attention, as the items of the Criterion-referenced test are based on the specifications of the behavioral range that was previously prepared. Therefore, you must measure this range with a great degree of accuracy, and that the level of difficulty of each individual is appropriate to the level of difficulty of the goal you are measuring, and its level of knowledge. The sample items should be representative of the behavioral scope of the objectives. The technical principles must also be taken into account in writing the different types of items . The choice of the form of the item to be used in the test depends on the nature of the educational feature or goal that the test is interested in measuring. There are two main types of items : the substantive items and the article items . Both types include different forms of items . The reason for naming the substantive items with this name is the availability of objectivity in correcting this type of items based on a norm correction key. In any case, objectivity is not absolute, as it is related to the correction process. Therefore, both types are available as a certain amount of subjectivity. Technically, the objectivity of the item is related to correcting the item and not construction of that item . We will differentiate between the different types of items by how the item is answered, is the item answered by choosing from several answers or does the answer need to be created by the examiner? (Ababneh, 2009 : 67).

- **Advantages and disadvantages of Criterion-referenced tests :**

Advantage of the Criterion-referenced tests :

1. They are tests originally prepared to be used in making decisions about the levels of student empowerment for a specific subject, so we find them known as tests of mastery or proficiency.
2. Since the tests are absolute reference used to determine mastery or efficiency in relation to a subject, we find that they require the examiner to comply fully with all the details, procedures and instructions contained therein.
3. It requires setting the interval score to determine the level of mastery or efficiency and the interval score is a point or rank on the basis of which the acceptance of the above and the rejection of the results or things such as successful or failing as well as acceptable or rejected.

Defects of criterion tests Ref:

1. The problem of setting the standard.
2. The problem of estimating validity and reliability .
3. You need specialists trained to prepare them.
4. The problem of selecting a behavioral sample representative of the skill to be measured (Ababneh, 2009 , 31).

❖ **Item Response Theory (IRT)**

The efforts of psychometric and educational scientists have led to the development of a contemporary theory of psychological and educational measurement called the item response theory (IRT). Under this theory, a set of models allows objective measurement. This theory has addressed many of the problems facing the classical theory of measurement. Many psychometric scientists such as Hambleton and Lord prefer to call this trend in psychological measurement the item response theory because it links the probability of the correct response of the individual to a particular test item , and the characteristics of this item . For this reason, mathematical models are used to describe this relationship between the observed performance of individuals on the test and those latent traits (Samurai and Khafaji, 2012: 98).

In recent years, the theory of item responsiveness has begun to be relied upon in the analysis of test items. This theory relies on the performance of the individual on each item in the test to give a statistical estimate of the ability of the individual measured by the test (Mclntir & Miller , 2000: 216).

The item's response theory is based on some axioms, and these axioms are:

1. An individual's performance on any test can be predicted by a combination of factors , called latent traits or abilities.
2. The relationship between the performance of an individual on any test item and the set of attributes or potential abilities that are supposed to affect his performance on this item can be described as a direct function and is called the function of the characteristics of the item or the curve of the characteristic of the item (ICC) as this function identifies individuals who have achieved high scores in the attributes that have a high probability of the correct answer to the item from the selectors or examiners who have achieved low scores on the attributes (Hambleton , 1989 : 148), as this function determines the probabilities of the correct answer to the different levels of ability, as individuals with high ability have higher probabilities of the correct answer to the items than the probabilities of individuals with low ability, that is, the greater the ability of the individual increases the probability of his correct answer to the item , Therefore, this theory, which was previously called The Latent Trait Theory Theory, or Theory of Distinctive Curve (Item Characteristics Curve) (Fraley , et al2000: 351).

The use of item -response theory over classical test theory has two distinct advantages:

1. It allows researchers and specialists in the field of educational and psychological measurement to classify test-takers more accurately with regard to their response models and their abilities and attributes.
2. The use of item response theory estimates allows extrapolation of future scores of interested and users while classical test theory scores do not allow extrapolation of future users (Ismail, 2007 : 18).

Some concepts related to item response theory:

The following is a review of the meaning of some of the basic concepts related to item response theory and its models:

1. Item Difficulty

Is a point on the latent attribute continuum represents the probability of reaching the correct answer equal to (0.50) and this point corresponds to the point of inversion of the curve, and Lord (1980) believes that the difficulty coefficient under theitem's response theory refers to the amount of power necessary to become the probability of giving the correct answer to a item (0.50), and the greater the value of the difficulty coefficient of the item , the displacement of the distinctive curve of the item to the right , and the lesser its value, the less the distinctive curve of the item to the left.

2. item discrimination

It is a parameter that is usually expressed by the ability of the item to distinguish between the examinees whose ability to answer the item passes correctly and the examinees whose ability does not enable them to answer the paragraph correctly (Alwan & Jasim, 2022 : 1153).

The ability of the item to distinguish is estimated by the coefficient of discrimination that can be expressed by the relative slope of the distinctive curve of the item on the axis of the feature at the point of inversion. The greater the degree of torsion of the distinctive curve, the greater the value of the coefficient of

discrimination. Thus, the distinction of the item indicates the rate of change in the probability of the correct response of individuals to the item in relation to the level of ability. Therefore, the best distinguishing items are those items that are intermediate in discrimination , and the best slope of the distinctive curve of the item when its slope angle is on the axis of the feature(45 °), then the value of the slope of the curve fluctuates around the optimal value on the attribute continuum, which is one value (Kazem , 1988 : 73).

3. Item guessing

The guessing of the item is measured by the probability of an appropriate answer (a point on the vertical axis) at very low levels of power, and this parameter is used in MCQ questions, and it represents the occurrence of the correct response by chance, and the value of the guessing parameter represents the part that the characteristic curve of the item cuts from the vertical axis that represents the correct answer to the item , as the individual who is located at a low level of ability from the ability communicator, expects to guess the correct answer to the choice item from four alternatives with a probability of (0.25) , and this is the place to start in the event of guessing , and therefore the part that the characteristic curve of the item cuts with the probability axis represents (0.25) (Rust & Golmbok, 1999 , 57).

Item Response Theory Assumptions:

1. Unidimensionality assumption

Dimensions refers to the number of inherent features responsible for the performance of individuals in the test and the attribute is a concept used to describe the behavior of individuals, which is a combination of overlapping and interdependent behavior in an integrated manner, and this means that the attribute is not a single attribute, but an abstract concept that is not tangible, so the identification and definition of the attributes to be measured are among the basic steps in behavioral measurement, and most models of theitem's response theory assume that there is only one attribute or ability sufficient to explain and clarify the differences between the performances of individuals on the test, and these models are called Unidimensional models, while models that assume that there is more than one ability behind the performance of the individual and are called multi-dimensional models (Weiss & Yoes, 1994 , 72).

2. Local independence

It is meant that the individual's responses to the different items are statistically independent, and this means that the individual's performance on the item is not affected by his performance on the other items in the test, and is also not affected by the performance of other individuals who perform this test, that is, the estimation of the parameters of the item does not depend on the estimates of the parameters of the items that make up the test, and does not depend on the estimates of the capabilities of the individuals who answer these items , as well as the estimation of the ability (attribute) of the individual does not depend on the ability of other individuals who perform the test, nor on the estimates of the parameters of the items they perform, that is, we are expected to obtain the same estimate of the ability for any subset of items as long as it is appropriate, as well as for any sample of testers or examiners, it is expected that the estimates of the parameters of the item itself, and therefore, the local independence requires that the probability of the correct or incorrect answer of the individual to a item is independent of any other item at a certain level of ability, and the factors affecting the probability of the individual's response to a particular item are only the underlying characteristic that is measured, and with the item to which the item (Hamble 1989):

3. Item curve assumption

It is one of the basic assumptions underlying theitem's response theory that deals with the Unidimensional feature, which is a mathematical function that links the probability of success of the individual in answering the item with the characteristic or ability measured by a set of items , or measured by a test that contains this item , that is, a non-linear function of the decline of the degree of the item on the underlying characteristic, or the ability measured by the test and the main difference between the models of response to the item depends on the mathematical formula for the curve of the characteristic of the item (Hambleton, 1989 , 151).

There is a consensus that the curve of the characteristic of the item can be determined by four variables , namely the ability of the individual and three variables related to the item , which are called the parameters of the item , and these parameters are the coefficient of difficulty , the parameters of discrimination, the coefficient of guessing, and the difficulty of the item corresponds to a point on a continuum of the underlying characteristic

or ability in which the curve of the characteristic of the item passes at the probability value (0.5), and therefore the difficulty of the item corresponds to a point on a continuum of the characteristic or ability in which the probability of a correct answer is equal to the probability of the wrong answer and equal to (0.5) (Aiken , 1998 : 53).

4. Assumption of Speedness:

Theitem's response theory assumes that the speed factor does not play a role in answering the test items , in the sense that the failure of the test-takers or examiners to answer the test items is due to the decrease in their abilities and not to the impact of the speed factor on their answers. It can be estimated whether the speed factor played a role in the answer by knowing the number of test-takers or examiners who were unable to answer all the test items that were conducted on them. When speed is one of the factors affecting performance on the test , there are two abilities that affect this performance, which are the ability measured by the test , and the ability of the speed of performance , and this is a violation of the Unidimensional assumption.

❖ **Previous Studies**

Table (2) Previous Arabic and Foreign Studies

No.	First: Arabic Studies	
1	Title	Using item's response theory in the construction of criterion Criterion-referenced test items in mathematics with bi-stage and multi-stage items according to the two- parameters logistic model
	Researcher Name and Year	Dabbous (2015) Mohammed Taleb
	Place of carrying out the study:	Istiqlal University/Palestine
	objectives of the study:	Using theitem's response theory in the construction of the collection of items of a Criterion-referenced test according to the dual criterion logistical model
	The research community and its sample	502 male and female students,
	Research tool	Construction of a collection of item e of referenced tests in mathematics of the type of questions MCQ consisting of (50) items for each item (4) alternatives
	Statistical means	Bilog MG3.Multilog.7..
	Remarkable Results	Item Response Theory assumptions are validated and the validity and validity of the test is verified using the two-pronged model
2	Title	Construction of a Criterion-referenced test in mathematics using the theory of response to the singularity according to the dual criterion logistical model for fifth grade students
	Researcher Name and Year	Abdullah bin Mohammed Al-Salami (2018)
	Place of carrying out the study:	University of Tabuk / Saudi Arabia
	objectives of the study:	Using the singular response theory in constructing a criterion-referenced test in mathematics according to the dual logistic model
	The research community and its sample	400 students in the fifth grade of the Tabuk Education Department

	Research tool	Construction of a 40-item criterion -Criterion-referenced test in mathematics
	Statistical means	Bilog MG3 - spss
	Remarkable Results	The results of the analysis with regard to the matching of the two-graded test items of the marked dual criterion model showed that the items were identical to the dual criterion model
Second: Foreign Studies		
3	Title	Education in Estimates Based on Classical Test and Item Response Theories Verification of validity in classical test-based estimates and item - response theory
	Researcher Name and Year	(2010) Adedoyin Adedoyin
	Place of carrying out the study:	Nigeria
	objectives of the study:	Verifying the reliability of estimating the parameters of individuals in the traditional theory and theitem's response theory (two-parameter model)
	The research community and its sample	The research sample consisted of 5000 male and female students
	Research instrument	First sheet of the 40-item mathematician's high school diploma test
	Highlights	The results showed that there are statistically significant differences in the reliability of estimating the parameters of individuals and in favor of the two-parameter model
4	Title	Classical test theory versus item response theory: An evaluation of the comparability of item analysis results Classical Test Theory vs. item's Response Theory: Evaluating Comparability of Element Analysis Results
	Researcher Name and Year	Onn (2013) , Onn
	Place of carrying out the study:	Nigeria
	objectives of the study:	Comparison of classical theory and item -response theory using the dual criterion model
	the Research Community	The research sample consisted of 69 male and female students
	Research tool	Preparation of a test in the subject of physics consisting of 50 items
	Remarkable Results	The results showed that the reliability coefficient of the two-parameter model is higher than the reliability coefficient in classical theory.

Aspects of the benefit of previous studies can be clarified by providing clear indicators on the problem of research and also helped to formulate the objectives accurately. Previous studies have clarified the methods used in the description of the study community and the methods of sample selection. It has also been shown through the presentation of previous studies The steps to be followed to verify the assumptions of the model and calculate the norm characteristics and statistical programs used, as well as discussing the results of the current research in the light of previous studies.

The Research Methodology

The comparative descriptive approach is the appropriate approach for the current research topic in order to describe the characteristics of the test according to the dual criterion model according to the modern measurement theory.

Research Community:

The current research community consists of students of the fourth stage in the Department of Educational and Psychological Sciences in the faculties of education at the universities of (Baghdad , Mustansiriya, Iraqi, Qadisiyah, Wasit, Kufa, Karbala, Maysan, Basra, Dhi Qar) for the academic year (2021-2022), and the total number of students reached (919) students.

Research Sample :

The researcher identified the sample in the random stratified way from the faculties of education, the Department of Educational and Psychological Sciences from each of the University of (Baghdad – Mustansiriya – Iraq - Wasit – Karbala – Kufa – Qadisiyah), which numbered (604) students, with a percentage of (65.7%) of the total research community.

Test construction Procedures:

1. Determining the academic content: The researcher determined the academic content of the subject of teaching thinking by referring to what is prescribed in the body of the sectoral committee for the curricula of the College of Education/ University of Baghdad, for the purpose of determining the chapters and then the topics in the subject of teaching thinking.

2. **Formulation of behavioral objectives :** The researcher formulated behavioral objectives for the specific subject chapters and for the first three cognitive levels in Bloom's classification (knowledge, understanding, and application), as the objectives reached (115) behavioral objectives distributed at the three levels as follows: (65) behavioral objectives for the level of knowledge and its weight reached (56%), (30) objectives for the level of understanding and its weight reached (26%), and(20) objectives for the level of application and its weight reached (17%).

3. **Preparation of the test map:** To build a diagnostic criterion- Referenced test in the subject of teaching thinking among university students, a table of specifications was prepared with the dimensions of content and behavioral objectives , as 80 behavioral objectives were withdrawn in a random manner, representing (70%) of the total behavioral objectives , and the behavioral objectives of (80) behavioral objectives were distributed to the subject chapters for the specific content, and according to the relative importance of each chapter.

4. **Drafting items and alternatives:** The researcher resorted to questions of MCQ with the four alternatives in the construction of all test items , because this formula is the most common and used in diagnostic criterion-Referenced test s criterion reference, which is analyzed according to the dual criterion model and the theory of response to the test item.

5. **Verification of the validity of the test items** : The researcher adopted the value of the Kai box (Ka2) as a test to accept or reject the item , which is equivalent to (80%) of the percentage of expert agreement. Some amendments were made to the test items , behavioral objectives and cognitive levels that measure the test items to be more technically and scientifically appropriate, noting that no test item was deleted from the test

6. **Experiment with the clarity of items** and instructions: The researcher applied the test in its initial form to a sample of (30) students randomly selected from the fourth grade students of the Department of Educational and Psychological Sciences of the University of Maysan, recording the time of completion of each student until the last student, and the average time it took to answer the entire test was (35) minutes

7.**Test Evaluation:** The researcher adopted the method of manual correction using the perforated correction key

8.Statistical analysis experience:

Verification of model assumptions (2PLM,2PLM) :

The main assumptions of the theory of response to the experimental item are : the assumption of Unidimensionality ,as well as local independence and conformity to the distinctive curve of the item , it is very necessary before using the approved model (two-parameter, PLM2) in statistical analysis, to verify the assumptions of the model as follows :

First: Verifying the Unidimensional Hypothesis:

The term Unidimensional has two distinct meanings. First , it can be interpreted in a psychological sense by referring to the underlying ability that influences performance. Second , it can be defined as a psychometric property that refers to one basic measurement dimension, (McNamara 1996: 271). It means that a set of items in the test measures only one ability. In fact , in practice , the Unidimensional requirement is to be a dominant factor, (Hambleton and Cook, 1977: 77) and the Unidimensional assumption is seen as an ideal situation analogous to the heterogeneity assumption in the analysis of variance. A degree of violation of a Unidimensional assumption can be problematic, Ayala, 2015;23).

1. **Factor analysis:** To verify the first step, the researcher subjected the items of the criterion diagnostic criterion- Referenced test in the thinking teaching material to factor analysis, in order to verify the factorial structure of the test. Accordingly, the forms on which the statistical analysis of the items was conducted, amounting to (604), were subjected to factor analysis. After relying on the results of factor analysis using the method of analyzing the basic components, and after conducting perpendicular rotation in the varimax method (maximum variance) for Kayes, one factor was obtained and with a latent root of more than one integer according to the method of minimum confidentiality. The latent root of the test reached (26.228) and contributed to the interpretation of (32.785) of the total variation, as this method considers that the signifying factor or the latent root that can be interpreted is equal to or more than one integer (Abdul Khaliq , 1983: 118) , and all the items of the criterion diagnostic criterion- Referenced test for the material teaching thinking are saturated with the general factor according to the Guilford criterion and at a rate of more than (0.30) (Guilford, 1998: 500).

2. Internal validity : The researcher calculated the correlation coefficient between the score of each item and the total score of the test using the Pont Biceral correlation coefficient, and using the statistical analysis sample of (604) students, and after comparing the values of the correlation coefficients calculated for the criterion diagnostic criterion- Referenced test of the subject of teaching thinking between the score of the item and the total score of the test, it was found that the relationship of all items of all tests to the total score is statistically significant. It is an indication that the test items are consistent among themselves to measure the property , which indicates that the first hypothesis (Unidimensional) is achieved from the assumptions of the theory of response to the test item .

Second: Verifying the assumption of local independence:

This assumption was achieved through the results of the factor analysis by , the underlying root, the interpreted variation and the saturation of the items by one factor, as well as through the correlation coefficients of the item with the total degree, and therefore these results are indicators that achieve the assumption of local independence by verifying Unidimensionality .

Third: The characteristic curve of the item :

Through the use of the computerized ICL program in the analysis of the research sample data, the characteristic curves of all the criterion diagnostic criterion- Referenced test items of the subject of teaching thinking were obtained through the outputs of this program.

Fourth : Freedom from the Speed Factor:

The percentage of students who completed all tests is (87%), and the percentage of items that were answered is (92%), which means that the test is free of the speed factor, and therefore it can be said that the data is suitable for analysis according to the two-tutor model PLM2 .

Fifth : Imposition of low guessing factor :

The researcher has selected a group of individuals (63) students with the least ability in the overall degree, representing (10%) to ensure that the dual criterion model does not include the guessing effect , and then study their performance on the most difficult items using the statistical portfolio (SPSS), and compare the percentages of those less able individuals who answered correctly to those difficult items with the theoretical value of random guessing, which is (0.25) in the case of the four alternatives as in the current test.

Sixth : Verifying the suitability of the data for the PLM2 parameter dual criterion model :

For the purpose of revealing the answers of the research sample to all the items of the criterion diagnostic criterion- Referenced test for the subject of teaching thinking, that is, the individuals whose answers were all correct answers to all items , as well as the individuals whose answers were all wrong answers to all the items

of the test, which are called zero data, and the complete data is called the items to which all the sample members answered correctly , and the researcher has checked the data for the answers of the sample members to the Flip Carter tests for mental fitness, and the researcher did not find such cases mentioned above. The researcher then entered the data of the individuals of the analysis sample amounting to (604) students, on the criterion diagnostic criterion- Referenced test of the thinking teaching material, through which it is possible to obtain accurate estimates of the estimations of the value of the K-square for the good match test, as well as verifying the usefulness of the model to predict the actual scores of the test. If the value of the K-square calculated is greater than the tabular value of the K-square, it is a function and the item is deleted. Table (3) shows the values of the K-square (K2) for the test items to judge their suitability for the dual criterion model PLM2.

TABLE 3 Evaluate the Difficulty coefficient, discrimination factor, and Kai square values of the criterion diagnostic criterion- Referenced test items of the thinking teaching material on the suitability of the dual criterion model PLM2

C-item	Difficulty index	Standard error	Discrimination index	Standard error	K-square value	Degree of freedom
1	0.694	0.180	.904	0.158	19.332	18
2	- 423.	-0.128	1.243	0.232	13.439	18
3	0.905	0.196	1.526	278.	7.427	18
4	-1.992	0.120	1.241	206	20.661	18
5	1.113	0.213	1.416	0.422	19.289	18
6	0.727	0.182	1.207	0.209	18.351	18
7	.760	0.185	1.091	.193	22.917.00	18
8	662	0.178	1.289	0.220	11.878	18
9	.760	0.185	.963	168	17.246	18
10	-1.239	0.115	950	160	10.475	18
11	-1.199	0.115	- 1,420	0.223	22.104	18
12	0.831	0.190	0.892	0.157	15.713	18
13	630.	0.176	1.191	0.176	15.092	18
14	0.905	0.196	1/451	0.319	16.360	18
15	.867	.193	1.897	- 312.	22.512	18
16	0.694	0.180	1.256	0.190	18.133	18
17	.867	.193	1.244	0.169	8.540	18
18	-1.810	117	1.515	0.240	20.418	18
19	.867	.193	1.400	.193	15.987	18
20	.788	0.120	1.198	0.298	17.410	18
21	0.456	165	1.498	260	12.902	18
22	-2.064	0.121	1.203	0.211	25.248	18
23	0.727	0.182	1.370	275	19.587	18
24	.867	.193	1.801	0.366	16.959	18
25	-1.797	117	1.227	0.238	10.795	18
26	0.302	156	.904	0.158	10.072	18
27	0.600	0.174	1.243	0.232	24.674	18

28	271	0.132	1.526	278.	10.912	18
29	.795	0.187	1.241	206	20.234	18
30	-1.213	0.115	1.416	0.422	21.049	18
31	-759	0.120	1.207	0.209	15.119	18
32	390	0.129	1.091	.193	18.509	18
33	612	0.123	1.289	0.220	17.021	18
34	390	0.129	.963	168	11.784	18
35	- 181...	135	950	160	12.762	18
36	439	- 127.	- 1,420	0.223	7.803	18
37	566	124	0.892	0.157	19.070	18
38	-519	0.125	1.191	0.176	10.737	18
39	- 181...	135	1/451	0.319	15.273	18
40	125	0.137	1.897	- 312.	18.915	18
41	439	- 127.	1.256	0.190	7.522	18
42	-519	0.125	1.244	0.169	18.645	18
43	.503	0.126	1.515	0.240	7.637	18
44	-0.253	0.133	1.400	.193	13.020	18
45	125	0.137	1.998	0.298	20.827	18
46	235	0.134	.904	0.158	14.656	18
47	-217	0.134	1.243	0.232	17.542	18
48	390	0.129	1.526	278.	12.595	18
49	-0.253	0.133	1.241	206	18.237	18
50	323	0.131	1.416	0.422	14.768	18
51	181	135	1.207	0.209	14.838	18
52	-0.106	0.138	1.091	.193	15.071	18
53	181	135	1.289	0.220	13.572	18
54	271	0.132	.963	168	11.029	18
55	-008	142.	950	160	10.928	18
56	-0.253	0.133	.904	0.158	13.389	18
57	535	0.125	1.243	0.232	14.917	18
58	144	0.137	1.526	278.	9.880	18
59	.048	0.140	1.241	206	22.496	18
60	181	135	1.416	0.422	22.008	18
61	235	0.134	1.207	0.209	15.306	18
62	271	0.132	1.091	.193	6.311	18
63	-519	0.125	1.289	0.220	17.848	18
64	535	0.125	.963	168	25.307	18
65	271	0.132	950	160	10.985	18
66	-008	142.	- 1,420	0.223	16.876	18
67	.048	0.140	0.892	0.157	15.560	18

68	-288	0.132	1.191	0.176	8,421	18
69	- 423.	-0.128	1/451	0.319	20.425	18
70	-471	- 127.	1.897	- 312.	14.129	18
71	- 423.	-0.128	1.256	0.190	9.840	18
72	0.012	0.143	1.244	0.169	19.648	18
73	390	0.129	1.515	0.240	13.318	18
74	1.374	1.002	1.400	.193	10.554	18
75	1.484	0.250	1.398	0.298	14.182	18
76	0.253	0.154	1.598	260	9.240	18
77	-0.253	0.133	1.203	0.211	15.147	18
78	1.407	0.383	1.570	275	11.449	18
79	0.430	0.164	1.201	0.366	13.336	18
80	1.270	0.359	1.727	0.238	16.693	18

Table (3) shows that all the test items are identical to the two-parameter model PLM2 (difficulty and discrimination), as no statistically significant differences were shown between the distribution of individuals' answers.

Seventh: Measurement Independence :

The Measurement Independence indicates the ability of the sample, which responds to the test , in terms of the fact that the ability of the individual is not based on the ability of other individuals who answer the test , and the estimation of the difficulty of the item is not based on the ability of the individuals who answer the test , and for the purpose of achieving the Measurement Independence , the researcher calculated the estimation of the difficulty of the item independently of the rest of the test items , and the estimation of the ability of individuals independently of the rest of the test items that the sample members answer, and the researcher verified these two aspects as follows :

- **Measurement independence from the capacity of the sample performing the test, and for the purpose of achieving this :**

1. The sample of the statistical analysis of the tests was fragmented and the total sample of (604) individuals was divided into a high-level sample (above the median) , and a low-level sample (below the median) , depending on the grading file

2. The results were analyzed based on the responses of the individuals of each sample to the tests , for the purpose of calculating the difficulty of the items , their norm errors, ability estimates and norm errors.

For the purpose of verifying statistical equivalence, a comparison was made for the teachers of (difficulty , ability) to test the subject of teaching criterion thinking, and the results were derived through the analysis of the total sample and the analysis of the two samples (high level , low level). Difficulty estimates are statistically equivalent; the difference between either estimate did not exceed the sum of the norm error of each. It was found that all the test items were equivalent to the corresponding statistical estimates, and also all the differences were less than the sum of the norm error of the two estimates. This means that the corresponding estimates are equal in the analysis of the total sample by describing them (reference estimates) except for items (22 , 46, 55 , 67 , 73). Their statistical estimates were not equal , and those derived from the performance of the two samples, whether high or low, and this indicates (the difficulty of the items is freed from the ability of the sample to which the tests were applied). Estimates of the corresponding capacity for each potential aggregate score derived from the performance of the total, high and low-level sample, and their norm errors, were extracted and shown in the table.

Journal for Re Attach Therapy and Developmental Diversities
eISSN: 2589-7799
2023 August; 6 (9s2): 470-491

Eighth : Changing zero regression in logit units :

Zero regression changes after deleting non-conforming items before they are deleted . Any displacement of this zero affects the hierarchy of the difficulty of the items and the estimates of the capabilities of the examiners, and of course this does not mean a difference in their quantitative significance, but the displacement of the hierarchy of items and the hierarchy of the capabilities of individuals (Kazem , 1988 : 100-101) The deletion of items that do not match the dual criterion model affects the average difficulty of the test items , and since the average difficulty of items in the program or any analysis of items is zero grading , so the test data was re-analysed again using the computerized program after deleting the items that do not match the model , and a new zero for grading is the average of the difficulties of the remaining 75 items in the test.

The researcher considered relying on the watt unit (Wat), which represents the percentage grading provided by Masters (1984) because it is the most familiar grading in most areas of measurement . Estimates of the difficulty of the item e and the capabilities of the examiners can be converted from the logit unit to the watt unit

Table (4) Estimating the difficulty of the items estimated in logit and watt units for the finalized test

C-item	Item Difficulty coefficient		Standard error		C-item	Item Difficulty coefficient		Standard error	
	Logit Unit Logit	Watt unit (Elec. Eng.) WAT	Logit Unit Logit	Unit WAT		Logit Unit Logit	Unit WAT	Logit Unit Logit	Unit WAT
1	0.264	41	0.126	1	39	0.664	54	0.184	2
2	-383	46	-0.128	1	40	.084	49	0.138	1
3	950	60	0.196	2	41	0.400	46	-0.128	1
4	0.879	59	0.145	2	42	-0.480	45	0.126	1
5	1.159	63	0.213	2	43	- 464.	45	0.126	1
6	-772!	58	0.183	2	44	212	48	0.133	1
7	0.805	59	0.185	2	45	.084	49	0.138	1
8	1.234	64	0.196	2	46	195	48	0.134	1
9	0.805	59	0.185	2	47	177.	48	135	1
10	-1.205	37	0.115	1	48	0.350	46	0.129	1
11	-1.165	37	116	1	49	212	48	0.133	1
12	0.876	59	0.190	2	50	-0.282	47	.131	1
13	0.211	44	0.122	1	51	140	48	0.136	1
14	950	60	0.196	2	52	.065	49	139	1
15	1.267	64	0.134	1	53	140	48	0.136	1
16	-1.178	37	116	1	54	0.941	53	0.149	1
17	722	42	0.121	1	55	754	56	0.134	1
18	0.350	46	0.129	1	56	212	48	0.133	1
19	573	44	.124	1	57	0.341	37	0.125	1
20	0.350	46	0.129	1	58	G-103	49	0.137	1
21	140	48	0.136	1	59	- 254	34	0.264	2
22	0.400	46	-0.128	1	60	140	48	0.136	1
23	231	32	.124	1	61	195	48	0.134	1
24	-0.480	45	0.126	1	62	-0.230	48	0.133	1
25	1.025	52	0.143	1	63	0.687	55	0.157	2
26	.084	49	0.138	1	64	526	52	0.145	2
27	0.400	46	-0.128	1	65	-0.230	48	0.133	1
28	-0.480	45	0.126	1	66	0.034	50	142.	2

29	- 464.	45	0.126	1	67	-006	50	0.141	2
30	212	48	0.133	1	68	248	47	0.132	1
31	.084	49	0.138	1	69	-383	46	-0.128	1
32	195	48	0.134	1	70	432.	45	- 127.	1
33	177.	48	135	1	71	-383	46	-0.128	1
34	0.350	46	0.129	1	72	.054	51	0.143	2
35	212	48	0.133	1	73	0.350	46	0.129	1
36	-0.282	47	0.131	1	74	997	59	0.152	2
37	140	48	0.136	1	75	1.530	67	0.250	2
38	.065	49	139	1					

Standard characteristics of the criterion diagnostic criterion- Referenced test of the subject of teaching thinking according to theitem's response theory :

First : Test Validity:

Stability is one of the basic standard characteristics of psychological scales with reliability which comes first, A reliable scale is considered stable but a stable scale is not necessarily reliable. Therefore, we can say every reliable scale is stable (Abbas. Etal, 2022 : 378).

The researcher has verified the validity according to the theory of response to the item e for the criterion diagnostic criterion- Referenced test of the subject of teaching thinking as follows :

- **Descriptive validity** : This type of validity was achieved by presenting the criterion diagnostic criterion- Referenced test for the teaching of thinking to a group of experts in order to express their opinions in it. (100%) of the experts' opinions were adopted and the amendments requested by the experts were made to the test items .
- **Practical validity** : For the purpose of achieving functional validity , the researcher used the indicators of the suitability of the items for the model used, where the statistics of the K-square values were used, which is the method used in the computer program, which is the program used by the researcher in analyzing his data, and according to the results of this statistical indicators, the appropriateness of the items or inappropriateness was judged to be deleted or retained.

Second: Test reliability:

through which we can identify the range of internal consistency between items of the scale with each other and we can infer the factorial validity of the grand total of the scale if the correlation coefficient among items and the scale are very high (Noori & Jassim, 2022 : 635).

Measurement tools that include good quality items can be more stable, this is confirmed by theitem's response theory (Allam , 2005: 56). There are several indicators to determine the reliability of the test according to theitem's response theory , and therefore the researcher will rely on the variance ratio indicator to estimate the reliability of the test:

Table (5).
Reliability coefficient values to test the thinking teaching material according to the contrast ratio indicator

Test	Source of variance	norm Deviation of Estimation	EstimationVariance ^{2} Tσ	norm Deviation of Error	Error variance Estimation^{2}Eσ	Reliability coefficient (R)
Diagnostic criterion - a criterion-referenced	among the individuals	0.978	0.941	0.034	0.001	0.898

test for the subject of teaching thinking						

- **Determining cut-off score :**

In order to determine cut-off score for the criterion diagnostic criterion- Referenced test of thinking teaching material, the researcher used the method of opposing groups. Thus, cut-off score the criterion diagnostic criterion- Referenced test of thinking teaching material was obtained from the intersection of curves. Cut-off score reached (51), which represents a percentage of (68.5%).

Statistical means:

First : The researcher used the Statistical Package of Social Sciences (SPSS) to extract :

1. To extract factor analysis by the basic components method,(Principles Component) with reanalysis by the Varimax method to verify Unidimensionality .
2. Correlation coefficient (point bicerial) to calculate the correlation of the score of the item with the total score of the tests.
3. Equating the norm error to estimate the norm error of the scale built according to the traditional measurement theory and the latent attribute theory model based on the reliability coefficients calculated according to the two methods.
4. Chi-Square test for good conformance in the calculation of matching item e to the two-parameter model (PLM2) .
5. One Sample T.test to find out the difference between the average scores of the sample members and cut-off score . The diagnostic criterion- Referenced test is a reference simulator for the subject of teaching thinking to fourth grade students.
8. The ATA square law to determine the size of the impact of the mastery of the fourth grade students of the subject of teaching thinking.

Second : Command Language Program for Item Response (ICL), (Item response command language) to analyze the research data according to the item response theory of the two-marker model (PLM2)

Presentation, interpretation and discussion of research results

- To achieve the first goal, which stipulates (construction of a reference diagnostic criterion- Referenced test for the teaching of thinking for fourth grade students of the Department of Educational and Psychological Sciences in the faculties of education in Iraqi universities) and to achieve the first goal, the researcher carried out all the procedures and practical steps necessary to build the tests in accordance with theitem's response theory , and as explained in Chapter Three. Starting from determining the academic content and setting behavioral objectives , and then the procedures of logical and statistical analyzes to verify the assumptions of theitem's response theory , and determine the norm characteristics of the items and the total test, and the command language program was adopted to respond to the item (ICL), (Item response command language) to analyze the research data according to the item response theory of the dual criterion model (PLM2) , and it was found that all items and the total test has good norm characteristics, and all items were retained, and thus this goal was achieved.

Conclusions :

1. The relevance of the PLM2 model in the construction of the criterion diagnostic criterion- Referenced test is a reference to the material of teaching thinking as a tool for current research, by matching the data of the tests to the assumptions of the model.
2. The effectiveness of the statistical program command language program to respond to item (ICL), (Item response command language) in conducting statistical analyses of test data, in order to assess the conformity of these data to the assumptions of the model as stated in the previous point, as well as its reliability in the calibration and grading of items and the capabilities of individuals on one connection to the inherent feature, and determine the norm characteristics of items and the overall scale, through the above textual and graphical outputs of these analyses.

3. The occasion of the diagnostic criterion- Referenced test is a reference to the subject of teaching thinking for fourth grade students in Iraqi public universities, in order to harmonize the capabilities of the sample members and the location of the items in achieving the goal of the test , and through the convergence of the two points of origin of the related trait or the ability of individuals , known as well as the distribution of community capacity estimates (θ). The proximity of the two distributions to the equinox distribution indicates the representation of the research sample to the research community, with the exception of the exclusion of some non-conforming individuals.
4. Excludes students with non-conforming responses to the analysis model, which achieves several purposes, including: achieving a better matching of the test data with the model expectations, and reaching better estimates of the parameters of the items and individuals, and thus achieve the research sample a better representation of the research community.

Recommendations

1. Using the diagnostic criterion- Referenced test as a reference for the subject of teaching thinking to fourth grade students in the Department of Educational and Psychological Sciences as a tool for their academic achievement and diagnosis of their academic problems, and not limited to the current sample.
2. The statistical program uses the command language program to respond to item (ICL), (Item response command language) in conducting statistical analyses of test and metric data.

Suggestions:

1. Adding another variable related to the characteristics of the sample to other studies (such as the characteristics of the sample distribution, or the size of the sample). The design uses repeated mixed measurements. Or any working design, according to the levels of the two variables.
2. Adding another variable related to the characteristics of the items in other studies such as (the order of alternatives, the number of response alternatives, or the type of alternatives) and using any factor design, and according to the levels of the two variables.

References:

1. Abu Nayha, Salah al-Din Muhammad (1994). Educational Measurement. Anglo-Egyptian Office, Cairo, Egypt.
2. Abbas. F. A &. Muhammad. A, & Khalid,J.Jasim. (2022). The Use of Psychometric Scale Theory in Formulating Gilliam Scale GARS-3 for Diagnosing Autism Spectrum Disorder. ALUSTATH JOURNAL FOR HUMAN AND SOCIAL SCIENCES, 61(4), 364-385.
3. Al-Hijami, Belqis Hammoud Kazem and Karma, Tafaa Tariq Habib (2021). A referenced comparative study to measure the vocabulary index coefficient between (COX & VARGAS) method and (POPHAM) method for critical dissection test, Middle East Research Journal, Issue 70, Ain Shams University, Egypt.
4. Al-Tamimi, Khaled bin Hassan Shiban(1999). The effect of both the type of arbitrator and the length of the test on determining cut-off score for the reference simulator test measures the mathematical competencies in the calculations on the numbers in the sixth grade of primary in Jeddah. Master Thesis, Umm Al-Qura University, Mecca, Saudi Arabia.
5. Al-Jubouri, Abdul-Hussein Razzouqi and Al-Kazemi, Ali Muhammad Hussain (2014). Constructing a referenced achievement test for the subject of philosophy and psychology for the fifth grade of literature according to the theory of latent traits, Nasaq magazine, Issue 2, College of Education for Human Sciences (Ibn Rushd), University of Baghdad. https://drive.google.com/file/d/1iv8vizl4AQ53yNdoSCzv3mVBHxe81WIp/view?usp=drivesdk
6. Alwan, A. M., & Jasim, K. J. (2022). The Effect of the Difference in the Distribution of the Level of Ability that is Skeweded Positive for the Parameters of the Items of the Mental Ability Test According to the Item Response Theory. International Journal of Early Childhood Special Education,14.(1)
7. Chalabi, Sawsan (2005). Basics of construction of tests and psychological and educational measures. Aladdin Printing and Publishing, Damascus, Syria.

8. Dabbous, Mohammed Taleb(2016). Using theitem's response theory in the construction of the collection of items of the criterion Criterion-referenced test in mathematics according to the dual criterion logistical model. Al-Najah University Journal of Human Sciences, Volume 30, Issue 7.
9. Ismail et al., Muhammad (1994) . Pre-school child development standards. Psychological Studies, Volume II, National Council for Childhood and Motherhood, Cairo, Egypt.
10. Samurai and Khafaji, Muhammad Anwar Mahmud and Ahmad Muhammad Shakir(2012). Construction of a reference achievement test in the subject of personal psychology for students of the departments of educational and psychological sciences. (n=203). https://www.iasj.net/iasj/download/ed2d7589d58534f4
11. Sergewa, Abdul Salam Awad(2016). Construction of a reference diagnostic criterion- Referenced test in topics from the course of measurement and evaluation at Omar Al-Mukhtar University. Faculty of White Education, Libyan International Journal, Libya.
12. Al-Salami, Abdullah bin Muhammad (2018). Construction of a Criterion-referenced test in mathematics using the theory of response to the singularity according to the dual criterion logistical model for fifth grade students. Psychological and Educational Counseling Center, Faculty of Education, Assiut.
13. Dodouh, Heba Abdel Latif (2023). The effect of the difference in the response model for the item (1PL, 2PL, 3PL) on the differential performance of the item, Al-Ustad Journal for Humanities and Social Sciences, Volume 62, Issue 1, College of Education Ibn Rushd, University of Baghdad. https://alustath.uobaghdad.edu.iq/index.php/UJIRCO/article/view/1964/1470
14. Ababneh, Imad Ghadab(2009). The tests criterion the reference of its philosophy and the foundations of its development. Al-Masirah Publishing House, Amman, Jordan.
15. Abdul Hafez, Shehteh Abdul Mawla (1999) . Evaluate the construction of tests referenced to the touchstone, a norm in singular response theory and conventional theory. PhD thesis, Faculty of Education, Ain Shams University, Egypt.
16. Abdul Khaliq, Ahmed Mohammed (1983). Basic Dimensions of Personality, 6th Edition, Alexandria, University Knowledge House.
17. Allam, Salah al-Din Mahmud(1985). Diagnostic criterion- Referenced test s are the reference of the test in the educational, psychological and training fields. Dar Al Fikr Al Arabi, Cairo, Egypt.
18. _____ (1986). Contemporary developments in psychological and educational measurement. Al Qabas Commercial Press, Kuwait.
19.) 2001. Diagnostic criterion- Referenced test s are the reference of the test in the educational, psychological and training fields. Arab Thought Publishing House, Cairo, Egypt.
20. _____ (2005) . Unidimensional and multi-dimensional experimental singularity response models and their applications in psychological and educational measurement. Arab Thought Publishing House, Cairo, Egypt.
21. Eid, Ghada Khaled(2004). The true degree using latent attribute theory and classical theory – a psychometric study. Umm Al-Qura University, Issue 2, Volume 16.
22. Ghoneim, Mohamed Ahmed Ibrahim(2003). Recent trends in the research of academic achievement evaluation problems. Research presented to the Standing Scientific Committee for Educational Psychology and Mental Health "The level of professors " , Zagazig University, Egypt.
23. Kazim, Amina Mohammed (1988) . A critical theoretical study on the psychological measurement of behavior according to the Rush model, in Anwar Al-Sharqawi and others, contemporary trends in measurement and psychological and educational evaluation. Cairo, Anglo-Egyptian Library.
24. Kadhim, Balqees Hammood(2020). Verifyingthe Criterionreference-based hierarchical Assumptionsfor high-order thinking skills Using ItemResponse Theory, Al-Ustath Journal for Human and Social Sciences Vol.(59) No.(1) (March-2020AD, 1441AH).
25. Majeed, Sawsan Shaker (2014). Foundations of construction of psychological and educational tests and measures. Third Edition, Debono Center, Amman, Jordan.

26. Mansi, Mahmoud Abdel Hakim(2001). Educational Evaluation. Dar Al-Maarefa Al-Jamiya for Printing, Publishing and Distribution.
27. Adedoyin, O. O. (2010). Investigating the invariance of person parameter estimates based on classical test and item response theories. International Journal of Educational Sciences, 2(2), 107-113.
28. Aiken, L. R. (1998). Tests and examinations: measuring abilities and performance. Journal of personality and Social psychology
29. Hambleton & Swaminathan ,H.(1989).. Item Response Theory Principles and Applications.Kluwer-Nijhof publishing , Boston ,U.S.A.
30. _____, Swaminathan, H., & Rogers, H. J. (1991). Fundamentals of item response theory (Vol. 2". Sage.
31. Mclntire , S. A. & Miller ,L. A. 2000)."Foundations Of Psychological Testing". New York ,McGraw-Hill Companies.
32. Noori, Noor Badri & Jassim, Khalid Jamal (2022) Panic Attacks Over COVID 19: A Survey Study on An Iraqi University Sample. Journal of Educational and Psychological Studies. Vol. 19 Iss. 75. 621-647. doi.org/10.52839/0111-000-075-023
33. Onn, D. (2013, May). Classical test theory versus item response theory: An evaluation of the comparability of item analysis results. Joint Admissions and Matriculation Board, 1-23.
34. Popham, W. J., & Husek, T. R. (1969). Applications of criterion-refined measurement 1, 2. Journal of Educational Measurement, 6(1), 1-9.
35. Rust , J & Golombok ,K ,S (1999) .Modern Psychometrics: The Science of Psychological Assessment.British Library Catalogying Inpublication Data.
36. Weiss , D. J. & Yoes , M. E.(1994). . Item Response Theory, IN: H ambleeton, R.k.& Zaal , J.N.(Advances in Educational and Psychological Testing : Theory and Applications. Kluwer Academic Publishers, Boston.