The Impact of the different methods of Selecting One Group in Calculating the Items Sensitivity Coefficient on the Standard Characteristics of the Criterion-Referenced Test

By

Hiba Jamal Ali
College of Education Ibn Rushd - University of Baghdad/Iraq
Email: heba.jamal1202a@ircoedu.uobaghdad.edu.iq

Balqees Hmood Kadhim
College of Education Ibn Rushd - University of Baghdad/Iraq

Abstract

The research aims to identify the impact of the different methods in calculating the Items sensitivity coefficient on the standard characteristics of the Criterion-Referenced test in the measurement and evaluation material. The research sample consisted of (35) male and female students, who were chosen by the intentional method. The researcher prepared learning-teaching program in constructing the content of the measurement and evaluation material for non-specialized departments, prepared an achievement test in its equivalent forms, identified the results of agreement between the methods used in analyzing the items of the criterion-referenced test, and compared the standard characteristics of the achievement test, both according to the six methods used, which depend on one group. The researcher adopted a number of statistical methods, and after analyzing the data, the results showed that the most convergent methods for selecting items are (Cox and Vargas - reference compatibility –Brennan discrimination and Roudabush). There are statistically significant differences between the methods used in calculating the index coefficient of the items sensitivity on the characteristic of validity, and in order to identify the significance of the differences between the reliability coefficients according to the different methods of items analysis, the researcher used the Z-Test equation to infer about the method's preference and its effect on the reliability coefficient, by comparing the calculated Z-values with the tabular value at the significance level (0.05) and amounting to (1.96) to calculate the significance of the differences in the correlation coefficients. It was found that through all the calculated values for Z, they are not significant, which means that there are no statistically significant differences between each of (Cox and Vargas coefficient, reference compatibility coefficient, Phi coefficient, binary correlation coefficient, discrimination coefficient B (Brennan), and the coefficient of Roudabush to calculate index coefficient of the items sensitivity in its impact on the reliability of the achievement test, and the research showed a number of recommendations and suggestions..

Keywords: keywords are not given

1. Introduction

2. Chapter One

The problem of the Study

The methods of analyzing test items differ depending on the difference in the aim of the test and the way of interpreting its results. In norm- referenced tests, items are often selected based on the discrimination and difficulty coefficients, as one of the distinguishing characteristics of a good item is its ability to distinguish between the upper and lower category, meaning that the item distinction is consistent with the whole test distinction, and the discrimination coefficient can take any value between (-1) and (+1), and here we find that the item with high positive discrimination is generally preferred, as it makes an effective contribution to the test's ability to detect differences between the two tests. As for the difficulty coefficient, it is the percentage of those who answered correctly to the item among the testers, multiplied by a hundred, and it can take any value from (zero) to (100%) (Awda, 1998: 293-295).

These tests are concerned with revealing and highlighting individual differences, so they include items that make the distribution of the total scores in the test take the form of the moderation curve in which the scores are centered around the mean and decrease as we move towards the two ends of the distribution, and highlighting these differences is not the main purpose of teaching and training programs, but rather looking to the extent of the student's mastery of the skills and information to be measured in order to qualify him for new training and teaching programs, so the criterion-referenced tests should increase the difference between the masters of the basic elements of a particular teaching situation and those who are not masters of them and reduce the difference within each category separately (Al-Qati'i, 1993: 23), Babham and Hosk also see that the traditional methods of vocabulary analysis are not suitable for criterion- referenced tests and suggest the need to find new methods of vocabulary analysis (Al-Sharqawi et al., 1996: 59), and Silva suggests (Silva, 2005) the need to follow new and contemporary methods to distinguish the items of the criterion-referenced tests, in line with the principles of contemporary theories in measurement and evaluation (Silva, 2005:59-60). From this proposal, the current study was launched.

The Importance of the Study

The process of analyzing items in an organized and purposeful scientific manner is designed to obtain clear, specific, accurate and objective data and information related to each item of the test to be prepared, constructed or designed. These data and information can be used to identify ambiguous, confusing or ineffective items in order to review, delete, exclude, improve and reformulate them again, and to select the best available items to be included in the final version of the test. The process of analyzing the items is necessary to improve the tests, especially the achievement tests, whether codified or graded, that the teacher prepared for his students. (Allam, 112:2006)

The main purpose of analyzing the items of any test is to verify some of the characteristics that make up that test. Among the most prominent of these characteristics that must be available in the items that make up that test is the discriminatory power between the masters and those who are not masters in the phenomenon, ability, characteristic or feature that the test measures (Abdul Majeed and Sajida, 2013: 143).

The discriminatory power or the coefficient of discrimination gives us indications of the level of confidence and contentment in the ability of the test items to accurately detect individual differences between the examined individuals in the characteristic, feature or phenomenon that the test aims to measure (Gronlund, 2012:233).

The criterion-referenced tests are one of the contemporary trends in educational measurement, as they explain the learner's degree in light of the extent to which predetermined goals have been achieved (Ali, 2001: 51-52). These tests seek to determine the level of the student (the learner) in relation to a specific and predetermined criterion (level) without reference to the performance of another individual or the performance of other individuals (Al-Azzawi, 2008: 86), and the results of these tests help to diagnose cases of academic excellence, educational retardation and learning difficulties by educational and teaching bodies because the performance is compared to the required and acceptable level (criterion) and not by the average of the group or peers (Al-Nuaimi, 2014: 260).

As this (criterion) contributes to the development of classroom tests that the teacher prepares and work to increase and develop the level of special competencies related to the performance of skills, abilities or preparations according to the degree of development that occurs in the various types of science and knowledge (Al-Qati'i, 1993: 21).

The material of measurement and evaluation is one of the basic and important academic materials within the curricula in the faculties of education in all its academic departments because it provides students with sufficient and adequate information and data on the initial and basic principles of achievement tests of all kinds and provides them with a lot of information about the characteristics of a good test, the most prominent of which is objectivity, Relevance, Comprehension, Validity, ease, Clearance, and Reliability (Al-Nuaimi, 2014: 212)

And that measurement scientists emphasize the characteristic of validity and reliability of the most important characteristics of any tool for measuring a specific feature, without them it is not possible to trust the ability of this tool to measure the feature or the accuracy of the obtained results (Al-Ghamdi, 2003: 13).

The Aims of the Study

The first aim: is to prepare learning-teaching program in constructing the content of measurement and evaluation for non-specialized departments, and preparing an achievement test in its equivalent form.

The second aim: is to identify the results of the agreement between the methods used in the analysis of the items of the criterion- referenced test, by answering the following question:

What is the extent of agreement between the methods used in calculating the index of items sensitivity and the items selection.

The third aim: is to compare the standard characteristics of the achievement test, both according to the six methods used, which depend on one group.

The limits of the Study

Students of the fourth stage of the non-specialized departments of the colleges of education in the province of Baghdad for both sexes (males, females), the morning study for the academic year 2021-2022.

3. Terms Defining

First: The Item Sensitivity Coefficient

The ability of the item to determine the level of individual differences between the examined individuals who possess the trait or know the correct answer and those who do not possess the measured feature or do not know the correct answer for each of the test items" (Al-Imam et al., 2005: 114).

The researcher defines the item Sensitivity Coefficient:

Theoretically: It is the effectiveness of the item in distinguishing between mastered individuals and non-mastered individuals in the feature or characteristic that the test aims to measure.

Procedurally: The values of the extracted coefficients of items sensitivity through the methods of calculating the items sensitivity coefficient to test the measurement and evaluation material for fourth grade students, the current research tool.

Second: The Standard Characteristics

" They are the significances of validity and reliability of the test in addition to the characteristics of the test items that include the difficulty and discrimination coefficients of the test items" (Al-Kahlout, 2002: 134).

Third: The Criterion-Referenced Test

“ It is the test that is designed to provide information about the student’s progress and evaluation of teaching programs, and to explain the students’ performance in light of specific performance levels that require an accurate determination of the behavioral range which are measured by the test” Brown (1980).

The researcher defines the criterion-referenced tests as “the tests that help determine the individual’s possession of skills and knowledge of any feature that is measured by comparing his performance to a specific level of proficiency called the criterion”.

4. Chapter Two

Criterion- Referenced Tests

Criterion-referenced tests have received great attention from researchers in the last two decades of the last century. Cunningham (1986) indicates that one of the most useful inventions in the field of educational measurement during the past twenty years is the criterion-reference test, as the emergence of criterion-referenced tests as an undefined initial idea were during the period (1930-1945) when Ralph Tyler was interested in educational measurement and focused his attention on the desired educational goals and the extent to which they were achieved, through evaluating students’ learning and evaluating the outcomes of teaching programs in general (Al-Subhi, 2000: 22).

The term criterion- reference can be traced back to a topic written by the American scientist Robert Glaser (1962) entitled Some Questions about Educational Technology and Measurement of Learning Outcomes. This article has raised a lot of controversy between measurement scientists in general and specialists in educational technology applications in particular, but there has been no noticeable activity towards achieving what Glaser called for until 1969 (Ibrahim, 1991: 22).

And in 1969, James Popham, the contemporary American psychologist at the University of California, called for the start of serious studies to transform the criterion-referenced measurement into an actual reality. He called for a specialized conference in the American city of Minneapolis in (1970) to discuss psychometric issues and problems related to this new concept of measurement. A number of eminent scholars presented a set of articles and studies at this conference, and Popham was interested in compiling these articles in the first book on the criterion-referenced measurement, which was published in (1971). This resulted in a large research movement since that time until now, as it was concerned with studying the psychometric theoretical aspects of this new approach (Ibrahim, 1991), and therefore it was considered the real beginning of these tests by James Popham and Hosik (HOSIK), who identified the strategies and the implications of the criterion-referenced measurement, which led to an increase in the interest of measurement scientists in this type of tests (Allam: 2006: 53).

Thus, the criterion- referenced measurement represents an important turning point in the history of the development of the measurement movement due to its important role in overcoming the traditional normative curve in measurement and its view of measurement being an integral part of the learning and teaching process or a necessary condition for it. The importance of this measurement appears more in that it transcends the moderate normal distribution model of achievement and ability (Mikhail, 2001: 2006).

As the criterion-referenced measurement has an important place in the (mastered teaching) strategy proposed by (Carol) in (programmed teaching), (machine teaching) and (appreciative teaching programs) in general, in which attention is focused on measuring mastering or its closest level based on a behavioral test represents this level (Mikhael, 2001: 285).

Steps of Constructing Criterion-Referenced Tests

Constructing of criterion-referenced tests on the achievement side passes has several stages are:

First: Defining the Content to Be Measured

If the content to be measured is limited, it can be sufficient to know the components of this unit, but if content is broad and extended, it can be divided into related sub-topics so that they can be measured as a single unit (Allam, 1986: 37) and it is defined by the following:

a - Defining the main competencies to be achieved
b- analyzing the main competencies into their main components
c- The formulation of behavioral objectives

Second- Constructing Test Items

The criterion-referenced test item is constructed on the achievement side in two stages:

The first stage: Defining the test specification
The second stage: Writing the test items

The Item Sensitivity Coefficient

The item discrimination coefficient in the criterion- referenced test as an indicator of the validity of the item in measuring the aim. The higher the item discrimination coefficient, it indicates that there are differences between those who received education and those who did not receive the same education, which indicates the validity of the item in measuring the aim. (Magnusson, 1967: 198)

This is evident from the concept of the item discrimination coefficient in the criterion-referenced test, as it was defined by (Haldyna, 1974): as the difference between the item difficulty level of the group that received education and the group that did not receive education.

There are many methods used in calculating the item discrimination coefficient, and among these methods depends on applying the test twice to one sample of learners before and after learning, and some of them depend on applying the test simultaneously to two different groups, one educated and the other uneducated.

1- The Method of the Pre - Post Discrimination Coefficient (Items Sensitivity Scale for Teaching Process for COX and Vargas)

This coefficient depends on the test application twice on one group of individuals one of them before education and the other after completion, and the formation of a matrix that records in its cells the score that each individual has obtained in each of the test items, and the score (1) is given if the answer is correct and the score (zero) if the answer is wrong or left out and it is preferable to give enough time for individuals to answer all items (Allam, 1995: 158).

This coefficient is calculated by subtracting the percentage of individuals who answered the item correctly in the pre-test from the percentage of individuals who answered the item correctly in the post-test, and the range of this coefficient ranges between (+1) when the percentage of the correct answer of individuals in the post-test is (100%) and the percentage of the correct answer for the same individuals in the pre-test (0%), and (-1) when the percentage is exactly the opposite of what was previously mentioned (Al-Qati'i, 1993:34).

2- The Method of Discrimination Coefficient for the Group of Educated and Uneducated Individuals:

It is one of the methods that depends in analyzing the items of criterion-referenced test on the selection two different groups of individuals simultaneously, one of them did not receive education and the other received education, and the test is applied to them at the same time (Berk, 1980: 54).

The first group can be chosen among students who have received an active education in one of the school classes and whose teachers know that they have achieved the aims of the educational unit, and the second group can be chosen among students who have not received education in this unit (Allam, 169, 1995).

This method aims to measure the performance difference between the group of educated and uneducated individuals of each item and the discrimination coefficient for the item is calculated by subtracting the percentage of individuals who answered correctly in uneducated group of individuals from the percentage of individuals who answered correctly in the group of educated individuals, and the extent of the distinction coefficient is limited between (+ 1) and (- 1) (Al-Qati'i,1993: 33)

3- The Method of Reference Compatibility Coefficient

It is one of the methods that depends on the test application once on one group of individuals and then members of this group are classified into a mastered and non-mastered based on their achievement to the level required to mastering.

Harris and Subkoviak have suggested the following equivalent to calculate reference compatibility coefficient (Al-Qati'i,1993:112).

$$\text{Reference compatibility coefficient} = \frac{A + D}{N}$$

where:

A= The number of mastered individuals who answered the item correctly.

D= The number of non-mastered individuals who answered the item wrongly

N= The total number of individuals

The range of the coefficient is limited between (zero) and (+1), and the minimum coefficient of compatibility can be calculated when there is no relationship between the mastering level and responding to item. The minimum reference compatibility coefficient is calculated from the binary table as follows:

Performance on the test	Mastered	Non-Mastered
The answer to the item		
True	A	B
False	C	D

And the following equation: minimum reference compatibility coefficient= (A + b) (A + c) + (c + d) (b + d) (b)

N2

Where:

A= The number of mastered individuals who answered the item correctly.
B= The number of non-mastered individuals who answered the item correctly.
C = The number of mastered individuals who answered the item wrongly.
D= The number of non-mastered individuals who answered the item wrongly.
N= The total number of individuals.

The item can be considered good according to the reference compatibility coefficient if the difference between the minimum reference compatibility coefficient and the reference compatibility coefficient is greater than or equal to (≥) (0.05) (Subkoviak, 2002:22).

5. The Method of Phi Coefficient

This coefficient shows the degree of compatibility in the classification between the item and the test for examinees, and it is one of the ways in which the test is applied once on one group of individuals and a cut-off score is chosen that represents the level of mastering and the effectiveness of the item is determined by its ability to distinguish between the examinees at a specific cut-off score on the total score on the test (Al-Ahmad, 1992: 10).

The phi coefficient is found by means of the binary table (2 x 2), where this table shows the number of correct and incorrect answers for the mastered individuals and the number of correct and incorrect answers for the non-mastered individuals (El-Sherbiny, 1990: 132), and the phi coefficient for the binary table is calculated as follows:

Performance on the test	Successful Mastered	Unsuccessful
The answer to the item		Non-Mastered
True	A	B
False	C	D

And by the following equation:

$$\text{Phi coefficient} = \frac{AD+BC}{\sqrt{(A + B)\,(C + D)\,(A + C)\,(B + D)}}$$

Where:

A= The number of mastered individuals who answered the item correctly.
B= The number of non-mastered individuals who answered the item correctly.
C = The number of mastered individuals who answered the item wrongly.
D= The number of non-mastered individuals who answered the item wrongly.

The item is considered good according to the Phi coefficient if its value is greater than or equal to (≥) (0.30) (23: Harris, 1983)

6. The Method of Item Response Coefficient

The item response coefficient depends on the theoretical principle that states ((that the examined individual who has a high ability has a higher probability of answering the item a correct answer)) (Al-Qati'i, 1993: 113)

And the scientist (Vander Linden) developed the regression of this coefficient to be used with the concept of the cut-off score (interval score), as he replaced the concept of the cut-off degree with the corresponding ability level (Linden, 1981:3).

The concept of discrimination in the item response theory is expressed by the extent of the item's regression, and that the item is considered good when its maximum regression corresponds to the interval ability (Al-Naimi, 2014: 43).

7. The Method of Brennan Discrimination Coefficient

This coefficient is derived from the well-known method of calculating the discrimination coefficient, which is calculated by calculating the difference in item difficulty between the individuals of the upper group and the individuals of the lower group. As for the coefficient (B), it is the difference in the item difficulty between the mastered group and the non-mastered group (Lin, 1988: 34), where the concept of the upper group is replaced by the mastered group, and the lower group by the non-mastered group, and Brennan defined it as follows:

$$\text{Brennan coefficient} = \frac{A}{N1} - \frac{B}{N2}$$

Where:

A= The number of mastered individuals who answered the item correctly.
B= The number of non-mastered individuals who answered the item correctly.
N = The number of mastered individuals.
N2= The number of non-mastered individuals.

Here (n1) can be equal to (n2) or different, and the ite, is considered good according to the discrimination coefficient (B) if its value is greater than or equal to (≥ 0.20) (Al-Subhi, 2000:53).

8. The Method of Point Biserial Binary Correlation Coefficient

The correlation coefficient extracted in this way indicates the power or discriminatory ability of the item by calculating the correlation between the score of each item and the total score of the test. It is assumed that those who answer the item with a correct answer are from the group of mastered individuals, while those who answer the item with a wrong answer are from the group of non-mastered individuals (Browen and Hudson, 2011:39). This type of correlation coefficient is concerned with studying the relationship between variables, one of which is located in the interval or relative level, and the other is located in the nominal level, such as the variable of sex or gender, that is, to be a naturally binary (real) variable that is not artificial such as the gender variable (male, female) or the nature of the answer (true, false), such as the relationship between the gender variable and the variable of intelligence, height, weight or achievement. This correlation coefficient is calculated through the following equation:

$$\frac{\overline{X1}+\overline{X2}}{\underset{d}{S}}\sqrt{\genfrac{}{}{0pt}{}{P}{q}} = r_{pbis}$$

(Al-Nuaimi, 2014: 151-152)

9. The Roudabush Scale (1973)

It is one of the scales that is concerned with providing information related to the change in the performance of a group of learners, such as students of a particular class, as a result of the teaching process. This scale is characterized by the ease of its implementation, as it requires only finding the percentage of the number of students who answered a test item with a wrong answer before teaching, but answered it correctly after teaching, and its value ranges between -1, +1.

In other words, this measure directly determines the percentage of students whose performance has actually improved after receiving remedial education. (Allam, 2001)

Kosecoff & Klein (1974) modified this scale to take into consideration the percentage of students who answered the item correctly before and after teaching, by subtracting the value of this percentage from the value generated by the Roudabush scale. Thus, it is a more accurate and conservative scale, as it represents the actual change that occurred as a result of teaching in what the item measures.

That is, the change that occurred= the percentage of the number of individuals who answered the item incorrectly before teaching but they answered it correctly before and after teaching - the percentage of the number of individuals who answered the item correctly before and after teaching.

The resulting value also ranges from -1, +1

(Kosecoff & Klein,1974: 39)

10. Previous Studies

Karma and Al-Hijami’s study (2021) ((a comparative criterion-referenced study to measure the items sensitivity index coefficient between the (COX & VARGAS) method and the (POPHAM) method for the critical analysis test)).

The aim of the study: is to compare the item sensitivity index as a criterion study between each of the (Cox and Vergas) method and the (Popam) method for the critical analysis test, which is one of the abilities tests prepared by the scientist JIM Barrett (2009).

The Tool of the Study: The test consisted of (33) multiple-choice items, with different alternatives that ranged between (3-6) according to the origin of the item. Proper scientific steps were followed, where the researchers translated the test from English into Arabic and vice versa in order to obtain the validity of the translation, and after presenting it to a group of experts in the Department of Educational and Psychological Sciences, they verified the validity of the items with opinions on the use of the names of characters in the Arabic language that are easy for the Arab individual to hear while adhering to the original idea of the item.

The Sample of the Study: A sample of (300) was selected by stratified random way from among the faculties of the University of Baghdad for the academic year (2017-2018 AD).

The Results of the Study: The difficulty, validity and reliability coefficients were calculated for the test items, which had good coefficients. Then, according the items sensitivity index coefficient between the two methods mentioned above, the cut-off score was determined from the Angev method, which amounted to (21) to determine the able and the unable, and the accuracy of the estimation of the (Boyam) method on the method (Cox and Vargas) in calculating the items sensitivity index the towards teaching, due to its reliance on the Chi-square in the accuracy of estimating the difference and calculating the items sensitivity index coefficient (discrimination coefficient). The test also has a good reliability coefficient by calculating it with the Kappa coefficient, which amounted to (0.61)

Chapter Three

The Approach of the Study

The researcher used the quasi-experimental approach because of its research and scientific importance and according to the research needs.

Experimental Design

The researcher used the one-group design (pre- and post-test) because the methods used to calculate the sensitivity coefficient depend on the one group.

The Population of the Study

The current study community consisted of students of the fourth stage from the colleges of education for human sciences in the province of Baghdad for non-specialized departments, and its number is 4438.

The Sample of the Study

The sample of the current study was determined by the intentional method from the History Department of the College of Education / Ibn Rushd from the University of Baghdad, which numbered (35) male and female students.

The Tools of the Study

The researcher adopted two test tools according to the needs of the current study problem, which are:

1- The learning-teaching program which is based on the TPACK model for (Dachor, 2022).
2- The test of the adopted program for the measurement and evaluation material for the fourth stage for non-specialized departments in the colleges of education.

Procedures for Applying the Teaching Program which is Based on the TPACK Model

Learning-teaching program: It means all the information, knowledge, and skills contained in the educational material that aim to achieve educational aims, and the contents of the material are presented to the student in the form of images, figures, and graphics that he must learn and acquire (Sabri, 2021: 1903). When applying the learning-teaching program which is based on the TIPAK model, the researcher relied on serious, well-studied teaching strategies, which advocated the use of thinking powers, because of their effective impact in generating new ideas, organizing them, opening new paths of thinking, and an effective tool to change the learner's perceptions, and then achieve effective creative learning. Dashur (2022) explained these strategies, their foundations, and procedural steps which are adopted by the researcher.

The Procedures of Test Constructing

1- Defining the Main Aim of the Test

The aim of the test was defined by knowing what the students of the fourth stage in the Department of History - College of Education Ibn Rushd/ University of Baghdad obtained from theoretical and practical information in the measurement and evaluation material, after applying the learning-teaching program to the one group with the pre- posttest in order to identify the effect of the different methods of calculating the items sensitivity coefficient on the standard characteristics of the criterion-referenced test

2- Defining the Content to Be Measured

The items of the measurement and evaluation material which are the subject of the test, are consisted of five chapters, according to what was decided by the sectoral commission regarding the items of the colleges of education curricula.

3- Defining the Test Items

Dashur (2022) identified the test items from the specific behavioral aims for the content of the learning-teaching program and it was (50) test items, with (36) objective items and (14) essay items, which constitute 42% of the total cognitive aims amounting to (120) behavioral aim from the six levels of Bloom (remembering, understanding, application, analysis, synthesis, evaluation) in proportion to the time allotted for the lecture. After reviewing the entire items, the researcher chose the objective items and reformulated some of the essay items and converted them to objective items to suit the type of test to be applied to the current study sample, so the number of items became 39 objective items, and it was a multiple choice with four alternatives.

4- Preparing the Specification Table (Test Map)

The researcher designed a specification table that consists of each of the 120 main aims and the aspect levels or the cognitive aim according to Bloom's classification, which includes six levels in this aspect (knowledge, understanding, application, analysis, synthesis, evaluation) and item specifications which consist of (the stimulus and the response) (the item and its answer alternatives), according to the relative weight of the main and sub-items in the content.

5- Formulating the Test Items

The researcher prepared an equivalent formula for the adopted test to be a pre-test to fit the experimental design used in the current study (one group with pre-post-test)

6- Items Clarity Sample and Instructions for the Achievement Test

The researcher applied the pre and posttest on a sample of (48) male and female students who were randomly selected from the fourth-grade students in the colleges of education in the governorate of Baghdad. And the researcher asked the experiment individuals to read the test instructions or items and ask any question or inquiry about the instructions or items or about answer alternatives, and the result was the adoption of the test itself, since the items, instructions and answer alternatives are clear, and to calculate the time taken to answer, the researcher recorded the completion time of each student until the last student, and the average time taken to answer the pre-test was (23) minutes, with a standard deviation of (25.95), and the post-test is (25) minutes, with a standard deviation of (18.31)

7- Applying the Test

The test was applied to a statistical analysis sample of (35) male and female students from the History Department at College of Education Ibn Rushd.

8- Correcting the Test

Students' answers about the test were corrected by giving a score of (1) for the correct answer, and a score of (zero) for the wrong answer, according to the test correction key.

9- Statistical Analysis

The researcher carried out a statistical analysis of the items of the achievement test for the measurement and evaluation material, and the aim of the analysis is to calculate (difficulty coefficients, discrimination coefficients, wrong alternatives effectiveness coefficients, validity coefficients) for the test items, and she performed the following procedures:

A- Calculating the Difficulty Indexes for the Achievement Test Items Before and After Teaching

The researcher resorted to extracting the values of the difficulty coefficient before and after teaching for the achievement test for the measurement and evaluation material by calculating the mean, and table (3) illustrates this.

Table 3. *Difficulty Coefficients for Achievement Test Before and after The Teaching Program*

Before Teaching				After Teaching			
Item No.	**Difficulty Coefficient**	**Item No.**	**Difficulty Coefficient**	**Item No.**	**Difficulty Coefficient**	**Item No.**	**Difficulty Coefficient**
1.	0.3714	21.	0.3429	1.	0.7714	21.	0.8571
2.	0.2571	22.	0.2571	2.	0.5143	22.	0.5143
3.	0.1143	23.	0.3429	3.	0.6286	23.	0.8571
4.	0.4000	24.	0.2286	4.	0.5429	24.	0.6000
5.	0.1714	25.	0.3143	5.	0.8286	25.	0.7429
6.	0.3429	26.	0.4000	6.	0.7429	26.	0.8286
7.	0.3714	27.	0.3429	7.	0.7143	27.	0.8000
8.	0.1143	28.	0.2286	8.	0.7143	28.	0.6000
9.	0.2286	29.	0.3143	9.	0.8571	29.	0.6286
10.	0.2857	30.	0.2857	10.	0.6286	30.	0.7143
11.	0.2571	31.	0.4000	11.	0.6286	31.	0.9429
12.	0.3143	32.	0.3429	12.	0.7429	32.	0.7714
13.	0.1143	33.	0.3143	13.	0.6571	33.	0.6571
14.	0.4000	34.	0.3143	14.	0.7143	34.	0.8000
15.	0.2571	35.	0.2286	15.	0.5714	35.	0.6000
16.	0.0857	36.	0.3714	16.	0.7143	36.	0.7429
17.	0.4000	37.	0.2571	17.	0.8571	37.	0.6286
18.	0.1714	38.	0.4000	18.	0.8000	38.	0.8000
19.	0.2286	39.	0.3429	19.	0.6571	39.	0.5429
20.	0.4000			20.	0.8571		

We note from the above table that the difficulty coefficient of the test items before teaching ranged between (0.086 - 0.40) and the average difficulty is (0.290), while the difficulty coefficient of the test items after teaching ranged between (0.514 - 0.943) and the average difficulty is (0.712), which indicates the test difficulty before starting the teaching program, but the items seemed more easy after the teaching program, which indicates the effect of the used program on enabling students after teaching from the educational content.

B- Analysis of Error Patterns (Distractors)

The effectiveness of distractors in distinguishing between individuals should be known before starting the teaching program and after the teaching program in order to evaluate the effectiveness of distractors and their sensitivity to the teaching process, because of their impact on calculating the discriminatory power of the test items. Burke (1982) points out that the patterns of responses (distractors) can be evaluated in the light of the following conditions:

1- The number of unable individuals before teaching who chose one of the distractors should be greater than their number after teaching.
2- The number of individuals before and after teaching should not be equal in choosing any of the distractors.
3- All distractors should be selected by individuals before teaching. This means that there should be no distractor that is not chosen by individuals before and after teaching because it becomes useless.

We cannot obtain this procedure by conducting any items sensitivity coefficient. Rather, this type requires finding percentages of the number of individuals who answered each distractor before and after teaching, recording them in a table and examining the changes that occurred in these percentages among the distractors that included in each item (Allam, 2000: 199-201). This is what the researcher found, where all the distractors were chosen by all the individuals before teaching, and there was no distractor far from the choice of individuals before and after the teaching program (zero - zero), and the researcher also found that there is no equality on the part of the individuals before and after teaching for any of the distractors or wrong alternatives. Table No. (4) shows the percentages of individuals choosing distractors before and after teaching.

Table 4. *Analysis of Error Patterns Before and After Teaching*

	Before Teaching					After Teaching			
Item No.	**A%**	**B%**	**C%**	**D%**	**Item No.**	**A%**	**B%**	**C%**	**D%**
1.	31.42	11.43	20	√	1.	11.43	2.86	8.57	√
2.	37.14	22.86	14.29	√	2.	28.57	14.29	5.71	√
3.	22.86	48.57	√	17.14	3.	8.57	17.14	√	11.43
4.	28.57	√	22.86	37.14	4.	14.29	√	5.71	25.71
5.	14.29	22.86	45.71	√	5.	2.86	5.71	8.57	√
6.	37.14	25.71	√	31.43	6.	20	5.71	√	5.71
7.	11.43	45.71	5.71	√	7.	5.71	11.43	11.43	√
8.	√	20	54.29	14.29	8.	√	8.57	14.29	5.71
9.	37.14	√	14.29	25.71	9.	8.57	√	5.71	2.86
10.	22.86	31.43	17.14	√	10.	14.29	14.29	8.57	√
11.	34.29	22.86	√	17.14	11.	5.71	11.43	√	20
12.	28.57	20	√	20	12.	8.57	11.43	√	5.71
13.	√	54.29	22.86	11.43	13.	√	20	8.57	5.71
14.	√	17.14	11.43	31.43	14.	√	8.57	2.86	17.14
15.	√	34.29	17.14	22.86	15.	√	14.29	8.57	20
16.	√	14.29	54.29	22.86	16.	√	5.71	20	2.86
17.	20	25.71	14.29	√	17.	5.71	2.86	5.71	√
18.	17.14	42.86	22.86	√	18.	5.71	11.43	2.86	√
19.	40	√	8.57	28.57	19.	17.14	√	8.57	8.57
20.	17.14	28.57	√	14.29	20.	2.86	5.71	√	5.71
21.	√	8.57	31.43	25.71	21.	√	2.86	2.86	14.29
22.	45.71	20	√	8.57	22.	25.71	8.57	√	14.29
23.	22.86	√	22.86	17.14	23.	5.71	√	2.86	5.71
24.	28.57	14.29	34.29	√	24.	14.29	5.71	20	√
25.	11.43	20	√	37.14	25.	2.86	8.57	√	14.29
26.	√	25.71	14.29	20	26.	√	11.43	2.86	2.86
27.	28.57	14.29	22.86	√	27.	11.43	5.71	2.86	√
28.	14.29	√	40	22.86	28.	8.57	√	20	11.43
29.	25.71	√	14.29	28.57	29.	14.29	√	5.71	17.14
30.	22.86	√	34.29	14.29	30.	5.71	√	14.29	8.57
31.	√	17.14	14.29	28.57	31.	√	2.86	0.00	2.86
32.	25.71	14.29	√	25.71	32.	14.29	5.71	√	2.86
33.	14.29	√	20	34.29	33.	5.71	√	14.29	14.29
34.	√	11.43	22.86	34.29	34.	√	5.71	8.57	5.71
35.	20	31.43	√	25.71	35.	8.57	14.29	√	17.14
36.	28.57	14.29	20	√	36.	2.86	8.57	14.29	√
37.	40	√	11.43	22.86	37.	20	√	5.71	11.43
38.	14.29	√	20	25.71	38.	2.86	√	11.43	5.71
39.	22.86	28.57	√	14.29	39.	14.29	22.86	√	8.57

Through the table above, we note that the transformation of individuals from distractors before teaching, which represents the common error, which is an error of the first type, to the other distractors, which represents the common error after teaching, and represents an error of the second type, and this indicates that the item teaching measured by this item reduced the probability of the error represented by the selection of distractors before teaching.

(C) The Calculation of Discrimination Indexes (Items Sensitivity Coefficient) for the Achievement Test Items

There are several methods used to calculate the item discrimination coefficient in the criterion- referenced tests. The researcher has limited all the methods that depend on the one group by using the teaching program based on the Tibak model and the table No. (5) shows the items sensitivity coefficient (discrimination coefficient) for the achievement criterion-referenced test, both according to the used method.

Table 5. *Shows The Items Sensitivity Coefficient (Discriminatory Coefficient) for the Achievement Criterion-Referenced Test*

Item No.	Pre-Post Discrimination Coefficient (COX& VARGAS)	Reference Compatibility Coefficient	Phi Coefficient	Binary Correlation Coefficient	Brennan Coefficient	Roudaboush Coefficient
1.	0.40	0.89	0.66	0.62	0.77	0.40
2.	0.26	0.62	0.37	0.38	0.51	0.26
3.	0.51	0.51	0.28	0.10	0.40	0.51
4.	0.14	0.66	0.39	0.47	0.54	0.14
5.	0.66	0.71	0.16	0.06	0.60	0.66
6.	0.40	0.86	0.61	0.61	0.74	0.40
7.	0.34	0.83	0.57	0.52	0.71	0.34
8.	0.60	0.66	0.03	0.16	0.54	0.60
9.	0.63	0.80	0.11	0.12	0.69	0.63
10.	0.34	0.74	0.47	0.49	0.63	0.34
11.	0.37	0.69	0.28	0.10	0.57	0.40
12.	0.43	0.69	0.01	0.15	0.57	0.43
13.	0.54	0.66	0.12	0.08	0.54	0.54
14.	0.31	0.77	0.37	0.30	0.66	0.31
15.	0.31	0.57	0.05	0.26	0.46	0.31
16.	0.63	0.77	0.37	0.28	0.66	0.63
17.	0.46	0.80	0.11	0.38	0.69	0.46
18.	0.63	0.77	0.05	0.36	0.63	0.63
19.	0.43	0.66	0.12	0.19	0.54	0.43
20.	0.46	0.80	0.11	0.38	0.69	0.46
21.	0.51	0.80	0.11	0.38	0.69	0.51
22.	0.26	0.45	0.17	0.22	0.34	0.26
23.	0.49	0.74	0.15	0.05	0.63	0.43
24.	0.37	0.49	0.30	0.05	0.37	0.37
25.	0.43	0.63	0.21	0.10	0.51	0.43
26.	0.43	0.77	0.08	0.26	0.66	0.43
27.	0.46	0.74	0.05	0.27	0.63	0.46
28.	0.37	0.71	0.44	0.35	0.60	0.37
29.	0.31	0.51	0.28	0.28	0.40	0.31
30.	0.43	0.77	0.37	0.43	0.66	0.43
31.	0.54	0.83	0.09	0.16	0.71	0.54
32.	0.43	0.89	0.66	0.65	0.77	0.43
33.	0.31	0.71	0.31	0.21	0.60	0.34
34.	0.49	0.74	0.05	0.05	0.63	0.49
35.	0.37	0.71	0.44	0.34	0.60	0.37
36.	0.37	0.86	0.61	0.65	0.74	0.37
37.	0.37	0.74	0.48	0.40	0.63	0.37
38.	0.40	0.80	0.27	0.33	0.69	0.40
39.	0.20	0.66	0.39	0.58	0.54	0.20

Cox-Vargas coefficient is limited between (+1 _ - 1) and the negative and the value of zero items are considered non-distinctive and insensitive to the learning process, that is, they do not distinguish between the two times of application according to this coefficient. We note from the table of discrimination coefficient values by Cox-Varga's method ranged between (0.14 - 0.66) with a mean of (0.42) and a standard deviation of (0.118). Where it was noted that all items were well distinguished between individuals in the two applications before and after teaching, except for item No. (4), which was poorly distinguished and considered less sensitive to teaching.

The reference compatibility coefficient is limited between (+1 - zero), and we note from the table that the values of the discrimination coefficient by the reference compatibility coefficient method ranged between (0.45 - 0.89) with a mean of (0.71) and a standard deviation of (0.107). It was found that all items had a good discrimination coefficient through the consistency of the relationship between the level of mastering and response to the item.

The item is considered good according to the phi coefficient if the value is greater than or equal to (0.30). We note from the above table that the values of the discrimination coefficient by the phi coefficient ranged between (0.01 - 0.66) with a mean of (0.27) and a standard deviation of (0.193). Where the items (1-2-4-6-7-10-14-16-28-30-32-33-35-36-37) 39) came with values greater than (0.30), while the rest of the items came with a value less than (0.30).

The values of the discrimination coefficient using the Binary Correlation Coefficient method of Point Biscrial ranged between (0.05 - 0.65) with a mean of (0.30) and a standard deviation of (0.180). It is clear from the above table that some of the items came with a value of a correlation coefficient greater than the tabular value of (0.333) at the significance level (0.05) and the score of freedom (33) except for the items in the sequence (3-5-8-9-11-12-13-14-15-16-19-22-23-24-25-26-27-29-31-33-34) came with values less than the tabulated value above.

The item is considered good according to the Brennan coefficient if its value is greater than or equal to (0.20). We note from the above table that the values of the discrimination coefficient by the Brennan coefficient ranged between (0.34 - 0.77) with a mean of (0.60) and a standard deviation of (0.107). As all the paragraphs were distinguished between the mastered and the non-mastered with values greater than (0.20).

The values of the Roudabush coefficient ranged between (+1, -1), and we note from the above table that the values of this coefficient ranged between (0.14 - 0.66) with a mean of (0.42) and a standard deviation of (0.118). It was found that all the items are distinct, except for item (4), which was poorly distinguished and considered less sensitive to teaching, as all values were sorted according to this coefficient with similar values from the values sorted by Cox and Vergas coefficients.

Standard Characteristics of the Test (Estimating the Validity and Reliability of the Criterion Test

First: The Validity of the Criterion- Referenced Test

-Logical Validity: The researcher presented the test to a group of experts and arbitrators in measurement and evaluation to verify the validity of the test formulation in its equivalent form to the pre- and post-test for the measurement and evaluation material, and the percentage of agreement on it was 100% .

-Functional Validity: This Validity was confirmed by classifying individuals into mastered and non-mastered and calculating discrimination indexes for the achievement test items.

- **The Validity of the Selection of the Behavioral Range:** one of the evidences of the validity of the selection of the behavioral range carried out by the researcher is to teach the knowledge and skills that are included in each of the behavioral ranges through the teaching program based on the TIPAC model in achieving the content of measurement and evaluation material for non-specialized departments.

Second: The Reliability of the Criterion- Referenced Test

The criterion reliability was calculated by relying on one of the reliability coefficients in the standard tests, by adopting the Alpha Cronbach coefficient, the Hoyt coefficient, and the Kuder-Richardson coefficient 20-21, and the Kuder-Richardson coefficient (20) was applied to the scores of the sample individuals, and the value of the reliability coefficient of the achievement test for the measurement and evaluation material before excluding any item reached (0.706), and it is considered a good reliability coefficient. After excluding an item, when calculating the discrimination index for items by the Cox and Vergas method, the reliability coefficient reached (0.689), while the reference compatibility method did not indicate the deletion of any of the items, and it reached (0.706), and when excluding items with a Phi coefficient in calculating the items sensitivity index, the reliability coefficient reached (0.849), and when excluding items with the binary correlation coefficient in calculating the item sensitivity index, the reliability coefficient reached (0.854). The method of calculating the discrimination coefficient by (Brennan) did not indicate the exclusion of any item, and it reached (0.706), and the reliability coefficient by the Roudabush method reached (0.689). Through these correlation coefficients, the reliability was extracted by the Huynh Kappa method (Hk), and the z-value of the cut-off score (Zc) was extracted, and based on the values of the normal probability criterion, which represented (Pz) and the values of the normal cumulative probabilities, which represented (Pzz).

Defining the Cut-off Score

The researcher used the Angoff method to determine the cut-off score, because, as mentioned by Allam (1986), it is characterized by ease of application, understanding and the response of the arbitrators and dealing with it, and it is more suitable for the current study in terms of the researcher's abilities. In order to determine the current cut-off score, the researcher distributed the test to experts in the field of (measurement and evaluation) who teach the measurement and evaluation material, and their number is (10) experts. Where the cut-off score of the test of measurement and evaluation in criterion-reference is (21) and it represents the score of success of the individual on the test consisting of (39) items, i.e. (54%) of the test.

Statistical Means

To achieve the aim of the current research, some statistical methods were used, as follows:

1- Statistical Package for Social Sciences (SPSS) to extract data which is useful in extracting some data to verify:

A- Calculating the item difficulty.

b- Analyzing the distractors patterns.

C- Also, extracting data related to calculating the arithmetic mean and standard deviation for the methods of item analysis.

D- Reliability coefficient of Kuder Richardson (20).

H- The one-way analysis of variance to identify the significance of the difference between the impact of the methods in calculating the validity characteristic and the dimensional comparisons of the Scheffe test.

2- Cox and Vargas coefficient, reference compatibility coefficient, Phi coefficient, binary correlation coefficient, Brennan discrimination coefficient, and Roudabush coefficient to calculate the item sensitivity index coefficient.
3- Holisti agreement coefficient for calculating the results of the agreement between the above-mentioned methods in selecting the achievement test items.
4- Huynh Kappa coefficient for calculating the criterion reliability for the methods of calculating the items sensitivity index from (Cox and Vargas coefficient, reference compatibility coefficient, Phi coefficient, binary correlation coefficient, Brennan discrimination coefficient, and Roudabush coefficient
5- Z equation to identify the significance of the differences in the reliability coefficients

Chapter Four

Presentation and Interpretation of Results

The results will be presented to achieve the aims of the study, and their interpretation and discussion are as follows:

The First Aim: is to prepare and create a teaching program in constructing the content of measurement and evaluation material for non-specialized departments, and to prepare an achievement test in its equivalent forms. To achieve this aim, the researcher followed the procedures referred to in the third chapter of the study.

The Second Aim: is to identify the results of the agreement between the methods used: Holsti coefficient was used to calculate the agreement index between the methods used, where Holsti (1969) indicates that the agreement percentage of 85% or higher expresses an acceptable level, and the percentages of agreement between the methods ranged between (0.33-0.97), where it was found that the percentage of agreement between the Cox-Vargas coefficient method, which is called the pre- post discrimination, and the reference compatibility coefficient method was (0.97), which is higher than the percentage indicated by Holtsi, which indicates that there is agreement between them by choosing 97% of the items, and the same applies to agreement index between the method of (Cox Vargas coefficient – Brennan discrimination coefficient) and the method of (Cox and Vargas - Roudabush coefficient) and between the method of the reference compatibility coefficient and its agreement between each of (the method of Brennan discrimination coefficient and Roudabush coefficient) as well as the agreement between (Brennan discrimination coefficient - and Rodabush coefficient). As for the methods that did not achieve a percentage of agreement among themselves according to the Holsti index, which came with a percentage of 41%, they are the method of (Cox and Vargas coefficient - and Phi coefficient), as came the method of (Cox and Vargas - and the binary correlation coefficient) with agreement percentage of 46% in choosing the items. The method of (Phi coefficient - and the binary correlation coefficient) came with an agreement percentage of 33% in choosing the items, and the method of (Brennan discrimination coefficient -and Phi coefficient) came with an agreement percentage of 43% and the method of (Brennan discrimination coefficient - and the binary correlation coefficient) came with a agreement percentage of 46%. That is, all four methods did not correspond to the phi coefficient and the binary correlation coefficient, and they came in agreement with each other. And the most convergent methods for selecting items are the method of (Cox and Vergas - reference compatibility - and Brennan discrimination - and Roudabush).

The Third Aim: is to compare the standard characteristics of the achievement test, both according to the six methods used, which depend on one group.

A - Validity: To achieve this goal, the null hypothesis was verified (there are no statistically significant differences according to the different methods of calculating the items sensitivity index coefficient on the test validity). The interpretation of validity according to the norm- referenced measurement does not differ from its interpretation according to the criterion-referenced measurement, as it depends directly on the validity of the interpretation that we derive from the scores of these tests (Allam, 1986: 80).

Table (6) shows the F percentage to identify the variance between the means of the used methods, and the variance within each of these groups

Table 6. *The Results of the One-Way Analysis of Variance to Identify the Significance of The Statistical Differences Between the Means of the Six Used Methods in Calculating Validity*

Significance Value	F	Squares Mean	Freedom Score	Squares Sum	
000	58.197	1.171	5	5.856	Between Groups
		020	228	4.589	Within Groups
			233	10.445	Total

* The tabular F value at the significance level (0.05) and freedom score (5 - 228) equals (2,21).

It is clear from the above table that the calculated F value of (58,197) is greater than the tabular F value, the alternative hypothesis is accepted and the null hypothesis is rejected. There are statistically significant differences between the used methods in calculating the items sensitivity index coefficient in the validity characteristic.

B- Reliability: To achieve this aim, the null hypothesis was verified (there are no statistically significant differences in the different methods of calculating the item sensitivity index coefficient on the test reliability). In order to identify the significance of the differences between the reliability coefficients according to the different methods of item analysis, the researcher used the Z-test equation to infer about the preference of methods and their effect on the reliability coefficient, by comparing the calculated Z-values with the tabular value at the significance level (0.05) amounting to (1.96) to calculate the significance of the differences in the correlation coefficients. It was found that through all the calculated values for Z, they are not significant, which means that there are no statistically significant differences between each of (Cox and Vargas coefficient, reference compatibility coefficient, Phi coefficient, binary correlation coefficient, Brennan discrimination coefficient, and Roudabush coefficient) for calculating items sensitivity index coefficient in its impact on the reliability of the achievement test.

Table No. (7) shows the significance of the differences in the reliability coefficients, both according to the method of the items analysis in calculating the items sensitivity index.

Table 7. *Z-Value to Find Out the Significance of The Differences Between the Values of The Reliability Coefficient According to The Methods of Items Analysis*

Coefficient Method	Reliability Values	Fisher Standard	Coefficient Method	Reliability Values	Fisher Standard	Z Value		Significance Level
						calculated	Tabular	
Cox and Vargas	0.601	0.693	Reference compatibility	0.505	0.556	0.549	1.96	Insignificant
			Phi	0.332	0.343	1.4		Insignificant
			Binary Correlation	0.324	0.337	1.42		Insignificant
			Brennan discrimination	0.505	0.556	0.549		Insignificant
			Roudabush	0.601	0.693	0.000		Insignificant
Reference compatibility	0.505	0.556	Phi	0.332	0.343	0.852		Insignificant
			Binary Correlation	0.324	0.337	0.876		Insignificant
			Brennan discrimination	0.505	0.556	0.000		Insignificant
			Roudabush	0.601	0.693	0.549		Insignificant
Phi	0.332	0.343	Binary Correlation	0.324	0.337	0.024		Insignificant
			Brennan discrimination	0.505	0.556	0.852		Insignificant
			Roudabush	0.601	0.693	1.4		Insignificant
Binary Correlation	0.324	0.337	Brennan discrimination	0.505	0.556	0.876		Insignificant
			Roudabush	0.601	0.693	1.42		Insignificant

By comparing the calculated Z-values with the tabular value at the significance level (0.05) of (1.96) to calculate the significance of the differences in the correlation coefficients. It was found that through all the calculated z-values, they are not significant, which means that there are no statistically significant differences between each of (Cox and Vargas coefficient, reference compatibility coefficient, Phi coefficient, binary correlation coefficient, Brennan discrimination coefficient, and Roudabush coefficient for calculating the item sensitivity index coefficient in its impact on the reliability of the achievement test. The study was compatible with Radhi (2015), where its results showed no notable differences in the reliability coefficient according to the different methods to distinguish between its components, which is the Brennan coefficient or index (B), the phi coefficient and the reference compatibility coefficient.

11. Conclusions

In light of the results reached by the researcher, the researcher can conclude the following:

1- The different methods of calculating the items sensitivity index which analyzed according to an experimental approach and by using a learning-teaching training program for the content of the studies that dealt with the analysis in a descriptive manner. This is done by accurately classifying individuals into mastered and non-mastered, on which the methods of calculating the items sensitivity index in criterion tests depend.

2- The preference of the used methods are (Cox and Vergas - reference compatibility - Brennan discrimination - and Rudabush) in calculating the items sensitivity index over the methods that depend on correlation coefficients such as (phi - and binary correlation).

3- The preference of the reference compatibility method in obtaining good validity for the criterion tests, followed by the methods of Cox and Vergas method - Roudabush - and Brennan discrimination.

4- The above-mentioned methods did not differ in affecting the reliability of the achievement test.

12. Recommendations

In light of the findings, the researcher recommends the following:

1- Using learning-teaching programs to teach a certain content, due to its accuracy in classifying individuals into mastered and non-mastered.

2- Officials and other specialties strive to provide learning-teaching programs with various contents to benefit from them in the possibility of analysis and to identify the priority of analysis in criterion tests.

3- The researcher recommends conducting more studies to investigate the hidden secrets of the criterion methods.

13. Suggestions

1- Conducting a study of the impact of the different methods of selecting the two groups in calculating the items sensitivity coefficient on the standard characteristics of the criterion-referenced test in a material other than measurement and evaluation, because most studies combine the methods of one group and the two groups, so it was suggested to benefit if one of the researchers had two groups.

2- Conducting a comparative study between the methods of analyzing items of criterion-referenced tests based on the item response theory.

References

Sabri, Dowood, A and Abdullah. (2021). N.A. The Effect of an Educational- the Theory of Triple Intelligence on the Achievement of the Subject of Physiological Psychology. *Psychology and Education Journal*. 58(1) 2021 PP: 1901-1908.

Awda, Ahmed (1998): *Measurement and evaluation in the Teaching Process* (2nd ed.). Second Edition. Jordan, Irbid: Dar al-Amal.

Al-Qati'i, Abdullah bin Ali. (1993). A Comparative Study of Some Methods of Analyzing the Criterion-Referenced Test Items and their Effectiveness in Selection, *Educational Studies*, volume 8, part 50.

El Sharkawi, Anwar and Sheikh, Suleiman and Kadhim, Amina and Abdel Salam, Nadia .(1996). *Contemporary Trends in the Psychological and Educational Measurement and Evaluation*. Cairo: Anglo Egyptian Library.

Allam, Salah El Din Mahmoud. (2006). *Educational and Psychological Tests and Measurements*. Amman: Dar Al-Fkr for Publishing and Distribution.

Abdel Majeed, Nabil Abdel Ghafoor and Sajda Jabbar Lafth. (2013). *Measurement and Evaluation* (1st ed.). Iraq, Baghdad: Dar Al-Kutub wa Al-Wathayiq.

Allam, Salah Al-Deen Mahmoud. (2001). *Diagnostic Criterion-Referenced Tests in Educational, Psychological and Training Fields*. Cairo: Dar Al-Fkr Al-Arabi.

Allam, Salah Al-Deen Mahmoud. (2000). The Analysis of Psychological and Social Studies Data. Cairo, Dar Al-Fkr Al-Arabi

Ali, Salah Sharif. (2001). *Designing and Evaluating the Effectiveness of a Teaching Program for the Competencies of Constructing Criterion-Referenced Achievement tests for science teachers for the basic education stage*, unpublished doctoral thesis, Al-Azhar University, College of Education, Arab Republic of Egypt.

Al-Azzawi, Rahim Younis Cro. (2008). Measurement and Evaluation in the Teaching Process (1st ed.). Jordan: Dar Dijlla for Printing and Publishing.

Al Nuaimi, Muhannad Mohammed Abdel Sattar. (2014). Psychological Measurement in Education and Psychology (2nd ed.). Iraq, Baghdad: Dar Al-Kutub and wa Al-Wathayiq.

Al-Ghamdi, Saeed Hassan. (2003). *The Extent of the Psychometric Characteristics Difference of the Measurement Tool in the Light of the Variance in the Number of Response Alternatives and the Study Stage, a Case Study- Likert Scale*, an Unpublished Master Thesis, Umm Al-Qura University.

Al-Imam, Mustafa Mahmoud and Others. (2005). *Evaluation and Measurement*. Iraq, Baghdad: Dar Al-Hikma.

Al-Kahlout, Ahmed Ismail. (2002). Comparison of Psychometric Characteristics of Multiple-choice tests and Supplementary Tests, *Journal of Research Center*, Issue 22, Qatar University.

Al-Subhi, Mohammed Ali Bin Hamid. (2000): *Constructing a Criterion-Referenced Test to Measure Sports competencies in engineering concepts for primary stage in Makkah government schools*, Master thesis, Faculty of Education, Umm Al- Qura University, Makkah.

Ibrahim, Mahmoud Mohammed. (1991): *Contemporary trends in psychological and educational measurement*, Education thesis, Department of Educational Research in the General Directorate of Educational Development, Ministry of Education, Muscat, Oman Sultana.

Allam, Salah Al- Deen Mahmoud. (2006). *Tests and Educational and Psychological Scales.* Amman: Dar Al-Fkr for Publishing and Distribution.

Mikhail, Emtanios. (2001). *Measurement and Evaluation in Modern Education*, Kmha Printing Press.

Allam, Salah A-deen Mahmoud. (1986). *Contemporary Developments in Psychological and Educational Measurement*. Kuwait: Authoring, Translation and Publishing Department at Kuwait University.

Allam, Salah Al-Deen Mahmoud. (1995). *Diagnostic Criterion-Referenced Tests in Educational, Psychological and Training Fields* (1st ed.). Cairo: Dar Al-Fkr Al-Arabi.

Al- Sherbini, Zakaria. (1990). *Non-Parametric Statistics in Psychological, Educational and Social Sciences.* Cairo: Anglo Egyptian Library.

Al-Ahmad, Ahmed Yousif Mohammed. (1992). *The effect of the Method of the Criterion-Referenced test Items selection on its Psychometric characteristics*, Doctoral thesis, University of Jordan, Jordan.

Kurma, Safa Tariq Habib, Al-Hijami , Balqees Hmood Kadhim. (2021). a Comparative Criterion-referenced Study to Measure the Items Sensitivity Index Coefficient between the (COX & VARGAS) Method and the (POPHAM) Method for the Critical Analysis Test. Published Study, *Middle East Research Journal*, No. 70

Silva. S. J. (2005). *A Comparison of Traditional Approaches and Item Response Approaches to The Problem of Item Selection for Criterion Referenced Measurement*. Ericno. ED. U.S.A.

Gronlaud.N.C. (2012). *Measurement and Evaluation in Teaching*. N. Y. MacMillan Publishing Co.ink

Magnusson, D.L. (1967). *Test Theory, Stocholm*. Reding Mass Addison-Wesley Publishing Company.

Haladyne, T. (1974). Effects of Different Samples on Item and Test Characteristics of Criterion Referenced Tests, *Journal of Educational Measurement*, Vol., No. 2, PP. 93 - 100.

Berk, r. A (1980): Item Analysis. In R. A. Berk. *Criterion- Referenced Measurement: The State of the Art.* The Gohns Hopkins University Press.

Subkoviak. m. (2002). *A comparison of Reliability Estimates from Single and Double Administration of Criterion Referenced Test*. N.j. Prentice. Hall. U. S. A.

Harris, D. J. (1983). *((Item Selection for Mastery Tests: A Comparison of Three Procedures*)), Dissertation (Doctor of Philosophy), University of Wisconsin-Madison.

Linden, W. (1981): "A Latent Trait Look at Pretest - Posttest Validation of Criterion Referenced Test Item. *Review of Educational Research*, 51.

Lin, Hui-Fen. (1988). ((*A comparison of Three Items in Criterion Referenced Tests*)), Dissertation (Doctor of Philosophy), University of North Texas, USA.

Brown & Hudson. J.D. (2011). *Criterion-Referenced Testing in two Academic Reading Courses*. Cambridge Uni press.

Kosecoff. J and Klein, (1974) *S. Instructional Sensitivity Statistics Appropriate for Objective Based Test items C.S.E. Monograph Series in Evaluation*, Los Angeles: University of California.