Constructing a Criterion Reference Test to Measure the Research and Statistical Competencies of Graduate Students at the Jordanian Governmental Universities

Maher Hussein Al-Habashneh Dr. Nabil Juma Najjar
Dept. of Psychology / Mutah University. PO Box 61710 MUTAH – JORDAN.

Abstract
This study aimed at constructing a criterion-reference test to measure the research and statistical competencies of graduate students at the Jordanian governmental universities, the test has to be in its first form of (50) multiple choice items, then the test was introduced to (5) arbitrators with competence in measurement and evaluation to determine the cut-off score using Angoff method amounted to (0.69), the test was applied to an experimental sample of (80) students (males and females from the International Islamic University to check out the difficulty and discrimination coefficients of the items, based on these coefficients, (5) items have been deleted, then the test which consists of (45) items in its final form, was applied to the total sample consisting of (275) students (males and females) from the Jordanian governmental universities (University of Jordan, Al-Yarmouk, Mu'tah), the stability coefficient was estimated using kuder-Ritchardson-20 coefficient (0.83) to be used in the verification of Livingston stability at the cut-off score, its value reached (0.87). The results showed that there is statistically significant differences in the degree of possession of competencies attributed to gender, also showed a statistically significant differences in the possessing of competencies attributed to scientific degree in favor of the doctorate, and there was no effect of the interaction between the two variables (sex and scientific degree) to acquire skills, and the results showed that there is a clear decline in the mastery of graduate students at Jordanian universities for research and statistical competencies, the proportion of the expert students who are over the cut-off score was (0.57).

Keywords: Criterion Reference Test, Research and Statistical Competencies.

1. Background of the study

The scientific research is considered the main portal for evolution and progress in all societies, as also it is a productive tool for planning and enhancing the performance in the many different aspects of life, and the scientific research is an activity done by a researcher to solve an existing problem or to add new knowledge to the human information pool, or for a productive criticism with the goal to expose and clear the truth for others. And so we see that the fields of the scientific research are variable and vast for they include multiple aspects of life (Atwan and Al-Falit, 2011).

The theses is considered as a requirement for the collage's graduation in the higher education system in order to get the master degree or to get the doctorate degree, and it is known to be the main resource for the scientific research for they contain additions to the specialised additional knowledge (Al-nerb, 2010).

And the higher studies are supposed to be a mode of scientific research that contribute to making changes in the real world through implying their results in the aspects that they are concerned with, but they face many obstacles which stands in its way and obstruct its effectiveness and hopeful goal, by lacking objectivity and scientific method which is necessary to lead a safe and sure results that can be decided upon (Al-Lahlah & Abu Baker, 2002).

And the science of statistics in the current age looks into the methods to collect analyse show and read the data and information in order to reach decisions based on this information, it is known that collecting information from a certain society as a whole and including all its members is a very hard to achieve goal in most of the times, as for taking a random sample that represent the society as a whole I easier an le time consuming, and can be in the shape of generalization of results or predictions of rejection and acceptance of hypothetical results (Al-Zoubi & Al-Talafha, 2000).

And from this idea, this study came to be about the degree of satisfaction in the master's degree and the doctorate degree in the public Jordanian universities, and to tell the weakness and strength in the higher education students in the Jordanian universities specifically in the educational sciences collage through a criterion referenced test to measure the degree of knowledge in the research an statistical fields, and that is by comparing to a predetermined level set beforehand which is called the cut off degree.

1.1 Problem of the study and questions

The importance of the scientific research in the current age calls for a good researcher that has enough amount of the skills required for the research and statistical processes and to be capable of using them in the field of scientific research, due to the notice I made through applying to the higher educational program in Mu'tah university that the higher education students has a weakness in preparing their theses and in the skills of the scientific research, and with that the researcher believes that the defects in these theses resulted in weakness in facing the community problems, and with that the problem of the study is set in building a test with a good psychometric properties that it can be used to judge how much does the students have of research and statistical qualities. And this study is aimed to build a criterion referenced test to measure how much does the higher educational students in the governmental Jordanian universities have of research and statistical qualities and that is by answering the following questions:

1. What is the cut off degree to pass a criterion referenced test to measure how much does the higher educational students in the governmental Jordanian universities have of research and statistical qualities?
2. What are the psychometric properties (validity and reliability) to test how much does the higher educational students in the governmental Jordanian universities have of research and statistical qualities?
3. How much do the higher educational students in the governmental Jordanian universities have of research and statistical qualities?
4. Does the degree of research and statistical qualities differ in the higher educational students in the governmental Jordanian universities according to the educational degree and social status?

1.2 Importance and goals of the study:

The importance of this study can be shown by the following:

1. The importance of this study comes from the importance of the scientific research itself, because the success and progression of nations depends on how much it cares, nurture and invest in the area of scientific research.
2. In Building a criterion referenced test to measure how much does the higher educational students have of research and statistical qualities gives a feedback to the universities officials to enhance their programs in the future to focus on those qualities and working to nurture them in their students in a real and professional ways in the programs to follow?
3. Determine the cut off degree to pass a criterion referenced test to measure how much does the higher educational students in the governmental Jordanian universities have of research and statistical qualities as well as showing the psychometric qualities (validity and reliability) for the test, the study also aims to determine How much do the higher educational students in the governmental Jordanian universities have of research and statistical qualities?

1.3 Procedural and conceptual definitions

- Criterion referenced test (CRT): which is defined as it is that test which is used to evaluate the performance of an individual in comparing to a slandered determined level of performance without the need to compare to the performance of others (Allam, 2006)

Procedurally the criterion test is defined as the test which is built based on the scientific steps which are known and agreed to by the statistics and measurement experts, which is used to measure how much the higher education students in the governmental Jordanian universities have of research and statistical qualities, which in its final form consists of 45 items.

- Mastery: defined as the ability to get a grade that is equal or surpasses the cut off degree which is considered competent, and if the grade is lower it is considered incompetent.
- Competency: it is the sum of behaviours which includes the knowledge and skills acquired in a specific learning program which reflects on performance and skills, and it can be measured by a Criterion referenced test (Jamel, 1998)
- Cut off score: Hambleton (1978) defined the cut off score with that it is a point on the curve of the score points that is used to classify the students into two categories that reflects the various performance skills in accordance to the goals measured in the test.

1.4 The study limits

- Space limits: this study is limited to the higher education students registered in the governmental Jordanian universities specifically in the educational sciences collage.
- Time limits: the test will be implied in the second semester in the university year of 2014/2015.
- Objective limit: the results of the study are set on how reliable and valid the response of the study society to the Criterion referenced test.

2. Theoretical framework

The scientific research is considered the main pathway to reach knowledge and scientific facts, in the frame and rules of scientific theories (Afanah, 2011).

The scientific research emerged as a result to the multiple attempts made by scientists across the ages to face and overcome problems, and due to the effect of the accelerating knowledge gain that we are living through, it controlled all the aspects and sides of our lives, it is not excluded to the study of social, psychological, and human studies but include life as a whole (Addas, Obidat, AbdAlhaq, 2005).

- Concept of scientific research:

The definitions of the scientific research are various and multiple in concepts and the scientist did not agree on a single and full definition, but despite the difference in details and some procedural aspects they agree on the core concept, and an example for these definitions is:

- Leedy (1980) defined the scientific research as it is the method of which we solve problems through.
- Tuckman (1978) defined it as it is an organised attempt to reach answers or solutions to the questions and problems which faces the individuals in their line of work and positions and the various aspects of their lives.
- Kerlinger (1973) defined the scientific research as it is an organized, well-adjusted, experimental search for information, which criticizes the assumptions about the nature of relationships between variables concerning a certain phenomenon.
- The importance of scientific research: according to Al-Ma'aitah (2011).

The scientific research has a major importance that reflects of the society and the individuals, to all aspects of life and is represented by the following:

1. The scientific research is one of the most important standards to measure the advancement and progress of the society.
2. The scientific research helps in correcting the plans and humane programs, which leads to enforcing the positives and avoiding the negatives in the future.
3. The scientific research helps all the aspects of society in developing and evolving their management methods, and the ways of production and working.

- The steps of scientific research:

The scientific method in researching depends on a number of organized steps of which the researcher follows in dealing with the problem which is being studied, despite the difference between researchers in the number and methods of these steps and its order, never the less there is a general agreement amongst the researchers about the primary steps for scientific research includes the following:

1. Identifying the problems
2. Reviewing the previous researches concerning the study's problem
3. Forming hypothesis which form a possible solutions to the problems
4. Setting the appropriate program plan for the research, and the suitable data resource and its collecting method, choosing the study sample and the study society.
5. Testing the hypothesis through collecting information objectively to make sure of how correct those hypotheses are and to reject or accept them.
6. Analysing the data using various statistical methods.
7. Explaining the results and putting together valid reasonable hypothesis depending on the result of the research. (AL-Lahlah and AbuBaker, 2002).

- Criterion Referenced Tests (CRT) & Norm-Referenced Tests (NRT):

Regarding the fact that psych-metering is a relative measurement, where the mark which the student gets in the test is meaningless unless compared to an appropriate standard to be explained through it, two main curves for testing each differ from the other in the assumption which the reference frame stands on to explain the student's results and they are:

1. The Norm-Reference Test (NRT):

And they are the tests which the student's result is explained in comparison to the mean performance of his normal standard sample group in the test (Odeh, 2005).

As mentioned by (Ababneh, 2009) and (Majeed, 2007) that the norm-reference tests

Had some sort of criticism, because it depends on comparing the student performance with his class group, and therefore it is possible for the student's results according to the difference in his reference group, and due to that it doesn't give the teacher accurate information to help make educational decisions about the satisfaction level of the students.

2. Criterion Reference Tests (CRT):

And they are tests in which the performance of the student is compared to a predetermined levels of preferences or qualities, and the judging of performance depends on reaching this level, or a certain number of goals which is expected to be achieved by the learner without regarding the performance of the group he is in, in this case the teacher can judge the real progress in the learner (Al-Zghoul, 2002).

And the focus these tests is on the student achieving a level of performance in the areas which this test is covering (Al-Najjar, 2010).

- Steps to building a criterion reference test: (Murad and Suliman, 2002), (Ababneh, 2009).

The steps to build a criterion reference tests include all of the following steps:

1. Setting the educational goals and classify them in a procedural behavioural ways.
2. Analysing the content of the subject involved in testing to its basic elements, and writing the contents in a form of behavioural goals that can be measured.
3. Reforming the paragraphs of the test for each goal wanted to be measured, then writing a number of questions about each goal to be measured, and that s before deciding the final form of the test paragraphs.
4. Verifying the validity of the contents to judge the questions of the test that was reviewed by a group of specialized judges in the area of goals and related contents.
5. Setting the cut off score, to classify the students into the category of competent and non-competent which is the lowest level required as a condition for competency.
6. Putting the final form of the tests where the length of the test and number of questions for each behavioural goal, and clarifying the process of marking.

- Statistical analysis to the paragraphs of the criterion reference test:

The analysis of this kind of tests depends on the statistical indicators to show the effectiveness of the paragraph as mentioned in the following:

1. Item difficulty: which indicated the percentage of the tested individuals who answered the paragraph correctly and its symbol is (P), and the value of this cofactor ranges between (0-1) and the closer it is to (1) the easier the paragraph, and the closer the value to (0) the harder the paragraph (Allam, 2007).
2. Item sensitivity: it is known as the ability of the paragraph to tell the difference between the individual who received education and those who didn't receive education.

And there are multiple ways that can be used to test the sensitivity of the paragraph:

- A method that depends upon having a pre-test and an after-test on the same group, and in this case we use this mathematical formula:

$$DI = P_{post} - post_{pre}$$

Where:

- DI is the sensitivity factor
- P_{post}: the percentage of the tested who answered the paragraph correctly in the after test
- P_{pre}: the percentage of the tested who answered the paragraph correctly in the pre-test

And the value of DI varies between (-1 to 1) (Al-Nabhan, 2004).

- A method which implies using the test on two test groups, one who received the educational material and the other group didn't receive the education, in this case we can use this mathematical formula:

$$D = P_{G1} - P_{G2}$$

- Where PG1 is the percentage of the tested who answered the paragraph correctly in the First group which received the educational materials.
- Where PG2 is the percentage of the tested who answered the paragraph correctly in the second group which didn't received the educational materials.
- Using the Brenan index (B-index), and that is through using the cut off score in the group testing, doing so by setting two groups of the tested which are the competent individual group, and the in competent individual group, and this index is given through the following mathematical formula;

$$B = (U \backslash n_1) - (L \backslash n_2)$$

- Where N1 is the number of tested in the group which got a higher score than the cut off score or so called the competent.
- N2 is the number of tested in the group which got a lower score than the cut off score or so called the incompetent.
- U: represent the number of the tested individual who answered the paragraph correctly from the sum of the tested individual of both groups who got higher score than the cut off score (the competent)
- L: represent the number of the tested individual who answered the paragraph correctly from the sum of the tested individual of both groups who got lower score than the cut off score (the incompetent) (Ababneh, 2009).

1-The indices of agreement : it is used when the question about the performance of a group of tested individual on two different paragraphs, the indices of agreement provides an experimental support to the judging which was done by the specialists about the paragraphs in the qualities table, and an example of the statistics used in this study :

○ Kai square to test for independence: if the interest was focused on developing the test by picking paragraphs from a sum of paragraphs randomly, and the developer would like to test wither all the tests are the same, or if a paragraph can be substituted with another paragraph.

$$X^2 = n(ad - bc)/(a + b)(c + d)(b + d)(a + c)$$

	+	-
+	A	B
-	C	D

○ Kai square for McNamara's test : which is used if the developer of the test would like to know if the difficulty level of two paragraphs is equal or that the differences in difficulty is due to certain mistakes (Ababneh, 2009).

$$X^2{}_{mc} = (\,|b-c|-1)^2/b+c$$

The psychometric properties for the criterion reference test:

There must be a set of properties in the criterion reference tests and those properties are represented as following:

1. The validity of the criterion reference test: the concept of validity indicates how much the test is serving the function that it was formed to do (Al-Najjar, 2010).

And as Popham (1978) known as which was mentioned in (Allam, 1995) reported that what is as functional validity is the accuracy of the criterion reference test in performing the function which it was created for.

2. The reliability of the criterion reference test: (Hambleton & Novick, 1971) sees that the reliability of the criterion reference test is basically the harmony in making decisions through the use of parallel tests.

And Hambleton (1978) classified the methods to estimate the criterion reference test according to its use into two groups which are :

First: assessing the consistency of the student's grades in a certain behavioural frame, and the ways of this group cares about decreasing the variable error which is the result of the difference in the scores of the students in the behavioural frame and an example of these methods are as follows:

Index-Livingston

This index is used to find the deviation of the individual's scores from the cut off score, also in takes care in the concept of the mean sum of the squares of deviations in the observed and expected scores and that is by calculating the value of the deviation of scores of each student from the cut off score in the test, and that is calculated according to the following mathematical formula:

$$k^2(X,T) = \ O^{2}{}_{x}(KR-20) + \frac{(\text{ɥ}_x - n_i c)^2}{O^{2}{}_{x}} + (\text{ɥ}_x - n_i c)^2$$

Where:

- $k^2(X,T)$: represents the index- Livingston
- X: represents the mean sum of the square of deviations in the observed scores from the cut off score (c).
- T: represents the mean sum of the square of deviations in the real scores from the cut off score (c).
- ɥ: The mean of the student's scores in the behavioural frame which the test measures.
- n_i: represents the number of questions
- $(KR-20)$: represents the consistence factor of Koder-Richardson
- c: represents the cut off score

Brennan & Ken Dependability coefficient: it was based on the works of both Brennan & Ken in their derivation of the Dependability coefficient for the criterion reference tests on the fundamentals and concepts of the generalization theory of Cronbach; such a coefficient is called the generalization possibility coefficient.

And as Cronbach explains in his theory that there are two types of the error deviations one of them is about the deviation of the full score, which means that it takes interest in estimating the difference between the individual score in the comprehensive behavioural frame which the test measures and the ultimate performance level which is the cut off (Allam, 1995).

Second: estimating the consistency in grouping the student into groups according to the scale of their qualification in the behavioural frame:

This method takes interest in determining the mistakes resulted from not harmonizing in the classification when applying two parallel tests or in the case of applying the test on the study sample individual, and examples for those methods are as follows:

1. Methods that depends on applying the test just one time and from those we mention the followings:

➢ Harris method

Harris (1974) mentioned that this method doesn't depend on the variable of the length of the test, but it relies on the link between a variable which represents the sum of the students score in the test, and another binary variable which represents the classification of students into competent and incompetent in accordance to the cut off score, this method is referred to as the Harris coefficient and known with the symbol (Mc) and it is calculated in the following way

$$Mc = SSB/(SSw + SSb)$$

Where: Mc: the Harris coefficient

SSw, SSb: it represents the sum of the squares (inside, between) the groups and the coefficient's value varies between (0—1).

➢ Huynh Kappa coefficient :

The coefficient of Huynh Kappa can be estimated in applying one test or two parallel tests, and the estimation of this coefficient relies on multiple assumptions which are:

1. For the distribution of the tested individuals scores to be represented in a beta form

2. For the scores of the test questions to be in a form of 0 or 1
3. For the answers to the questions to be statistically independent in a form so no paragraph affects the answering of any of the other paragraphs.
4. For the difficulty coefficient for all the paragraphs to be almost equal (Majeed, 2007).

2. Methods that depends on applying the test two time:

➢ Carver method

It is one of the suggested methods to estimate the consistency of the criterion reference test which is related to the harmony in the grouping decision, and this method relies on performing two parallel tests on a single group of individuals, next comes the comparing of the percentage of the students grouped as competent in both tests, if the value gotten is similar or close to each other, then the tests are considered to be constant and stable.

And we can calculate Carver coefficient by using the following mathematical formula:

$$\frac{A+D}{N}$$

Where:

- A: the number of students applying to the two tests
- D: the number of students who was classified as non-competent
- N=A+B+C+D

But the problem with the method is that it only reflects if the percentage of the individuals would not change in the two times the test was applied (Allam, 2007).

➢ Kappa coefficient

Allam (1995) mentions that Kappa coefficient require applying the test on one single group of individuals twice, once before the teaching and introducing the learning material to the student, and once after, or applying two parallel tests on two groups one of them is competent and the other is incompetent, and the stability coefficient is estimated by using the kappa coefficient which ranges in value between (-1,+1), and the negative value indicating the disharmony or instability, and so its value is affected by the change in the cut off score, as well as it is affected by the number of paragraphs in the test and how difficult they are, and he study sample size and number of individual participating, and that is needed to be able to interrupt the value of the coefficient better we put the cut off score beside the kappa coefficient to be a guide in explaining those values, and it is calculated using this mathematical formula:

$$K = (P - Pc)/(1 - Pc)$$

Where:

- K: kappa coefficient
- P: the percentage of harmony noticed in the grouping
- Pc: the percentage of harmony expected in the grouping

Cut off score:

Hambleton (1978) mentioned the cut off score as the point on the curve of the test scores which is used to classify the students into two groups that reflects the different levels of performance in according to a specific goal (or goals) that was meant to be measured by the test.

As defined by (Halpin, G, Sigmon, 1983) “it is the scale of how much the student’s performance is in accordance with a specific goal”

The names and standards of proper success are variable, sometimes it is referred to as the mastery level, or the passing score, or the minimum competency level, or the criterion level, or the cut off score (Ababneh, 2009).

And considering the importance of the performance level and the success standards in taking educational decisions regarding individuals, the educational science specialists took great care in devising methods to rely on, and following is some examples of these methods:

A. Judgmental methods

These groups depends on a number of methods to set the performance level based on the specialist judgment wither it was single or group judgments, these judgments are related to the contents of the test and the word choice, so they are called the rational methods, the most common of these methods are the Nedlesky method, Angof method, Ebel method and Jaeger method (Allam, 2005).

And as follows, the Angof method will be explained because it is the method used to tell the cut of score in this study, and it is one of the methods used in the multiple selection tests, in this method a set of tests is used as a guide, by showing the paragraphs of the test to a group of specialists and each judge is asked to visualize a group of students with the minimum requirements of competency to live up to what the test is measuring , and then the estimate the percentage of students who are possible to answer each paragraph of the test correctly, and the mean of these numbers taken from the judges is used as a cut off score (Shepared, 1984) .

And the steps of this method can be summed up to the following steps:

1. Assigning a group of specialists and experts in the subject.
2. Each judge is asked to check every paragraph of the test, and to set a percentage of the examined from a

group of minimum requirement to pass the paragraph correctly.

3. All the estimated values are summed up and the mean is calculated for all the paragraphs regarding the minimum requirement to pass the paragraph.
4. The mean is found for all the judges and that mean is set up to be the minimum requirement for passing or what is known as the cut off score.

B. Empirical judgmental method:

This method depends on the actual performance of the student in the test, and also it depends on the statistical analysis where the judge's role is limited to picking up the students and grouping them into competent and incompetent, an example of these methods includes the following:

The border groups method, the paradoxical group method, the criterion group method (Berk, 1982).

C. Judgmental- Empirical method:

This method depends on the specialist and experts with the presenting of data and information regarding the actual performance of the students for those specialists to judge based on this information, for the data drawn from applying the test on a proper sample of individuals to make the judges to decide in a more realistic way, and the most important examples of this method are as follows:

The information supported judging method, the modified Angof method, a method which consists of both absolute and relative ways (Allam, 2007).

2.2 Previous studies

Which represents the previously done and performed studies that are related to the current study's topic?

Al-Kasasbeh (2013) performed a study that aimed to build a criterion referenced test to measure how much the higher education students in Mu'tah university can accomplish in the field of scientific research, the test included in its final form (40) paragraphs of multiple selection type, the test was applied on an experimental sample consisting of (70) students both males and females to check the difficulty factor for the paragraphs, and based on these results all of the test's paragraphs were kept, then the test was reviewed by a set of (6) specialists of experience to set the cut off score, and that is by using Angof method and it was set to be (29.2) which is equivalent to (73%), then the test was applied to the actual complete sample which consists of (300) male and female students, the stability factor was calculated using the Koder-Richardson factor -20 and it was found to be (0.80), to be used to check the Livingston index at the cut off score which turned out to be (0.82), and the results shown no statistically significant differences in the degree of owning quality according to the educational degree between doctorate then master's degree then for the bachelor, and there were no effect what so ever between the factors in owning the qualities, the study also reviled a noticeable decrease in the higher education students in Mu'tah university for the qualities of scientific research, where the percentage of the competent students who passed the cut off score reached a percentage of only (31%).

As for Atwan and Al-Falit (2011) they performed a study that aimed to identify the qualities of the scientific research in the higher education students in the educational collages in the Palestinian universities in Gaza, where the researcher used the method of descriptive research, and the required qualities where set, and classified into personal qualities, scientific qualities, artistic qualities, procedural qualities and linguistic qualities, and they were inserted in a survey that was applied to a sample of 98 individual divided between 34 university professor from the three universities and 64 male and female students of the higher education students in these universities, the study resulted in the facts that the degree of these qualities was on average and reached (64.16%), and most of the research qualities that was tested and surveyed got nearly similar percentages varying between (62.21 to reach up to 65.71), and stating by order as, linguistic qualities, personal qualities, artistic qualities, procedural qualities, and last was the scientific qualities, and the study shown a statistically significant differences in the degree of availability of qualities between the students and the professors in favour of the students, and there were no statistically significant differences in the degree of availability of qualities due to the difference of collage speciality.

And in the study of Al-Banna (2011) which aimed to build a criterion referenced test to measure the statistical qualities in the higher education students in the educational collages in the Yamane universities, and to check the validity of the test scores (descriptive validity and behavioural frame selection validity) and also to check the stability of the test scores according to the Livingston index, as well as setting the appropriate cut off score for each statistical quality on which the students are divided into the groups of competent and incompetent in this quality, and to see how much the higher education student in the Yamane universities in the field of statistical qualities which was included in the test which was constructed for this purpose, and the use of the adjustment scaling method was used for it is the appropriate method to achieve the survey goals, the study sample consisted of (157) male and female students in the master's education stage, and the test was constructed by going through 4 stages (the analysis stage, the construction, the trial period, the finalizing and output) where the test consisted of (77) paragraphs that tests (7) statistical qualities which are: (concepts identifications, basic statistical concepts, using and deducing the descriptive statistics, identifying and explaining the correlation coefficient, using the parametric statistical methods, using the non-parametric statistical methods, reading and interpreting the results extracted from the statistical program of (SPSS), and choosing the suitable statistical method).as this study resulted that this test has a good descriptive validity and good behavioural frame selection validity, and shows that this test characterized by a high stability according to the Livingston index which was (0.99) as for the cut off score it was provided by using the Nedlesky method to be (0.63), it was shown that the students has an apparent shortage in the statistical qualities in the higher education students in the educational collages in the Yamane universities.

The study of Mac and Thames (1996) shows the importance of guidance to the higher education students in the University of Florida in the United States of America, to give them background knowledge regarding the designing of researches in the master's degree and train them to apply it using a computer program in statistics and educational research, for this goal the researcher have prepared a guidance program and giving basic instructions for the student to extract them from the computer and deal with them step by step, in doing so the researcher noticed a substantial progress in the designing of researches in the higher education students which was shown in the definition of the concepts in the research, the presence of morals in the research , identifying the problem, reforming the assumptions, identifying the research limitations, collecting information, displaying results, writing a research summary and abstract, as well as adding a statistical analysis for the information, and to have the ability to design the research and choosing the sample, developing the scales and tests, turning the information into the value form, using the computer as a tool in the statistical analysis while taking in concern the conditions of each type of statistics.

In a study done by Guziano (1995) which dealt with the mistakes common in designing the scientific research, where the researcher analysed (34) scientific research out of (58) studies that was published in (1984) and that is to examine the main element that should be present in the scientific research and he found out that there are holes in the structure of (12) studies, and most of the mistakes were made in the field of forming the assumptions, and in (22) studies of those studies were un realistic, but the researcher found that there is a clear understanding for the educational phenomenon in the researchers, the least percentage of mistakes where in the performance of those researches and the ability to design (applicable research).

In a study by Jonthan (1991) which aimed to find out the common mistakes in designing the master's degree and doctorate degree theses, the researcher has done an analytical study on a sample of (24) theses that was discussed in one of the united states of American universities, and resulted in the presence of many mistakes in all of the elements of the research, the most common and most percentage of these mistakes was in choosing the research method from the various types (descriptive, experimental and comparative), also the study shown that there were difficulties in dealing with the basics of problem solving and the ways to attend them.

Commenting on the previous studies

Despite the difference in goals between those studies from the construction of criterion referenced tests or developing them, or regarding the psychometric properties of these criterion referenced tests, all of those studies show the importance of the criterion referenced tests and that it has a good psychometric properties, and their importance in diagnosing the strengths points as well as the weak points in the research skills and qualities, and some of the studies that was concerned with the statistical and research qualities pointed to the fact that the level of these qualities are pretty low in the higher education students.

An addition that this study offers is a criterion reference test, to measure how much the higher education students in the governmental Jordanian universities have of the qualities both statistical and research in the collage of the educational sciences, and to expose the strength points and weak points and to diagnose them, with the goal to put a repairing and healing plans that helps in curing the weak points.

3. Design of the study

This chapter includes a description to the study population and its sample, the methods of choosing the sample, and the way to building a study tool, as well as the performance explaining standards (the cut off score) on the test's paragraphs, and a description of the statistical engine which was used in the study.

3.1. Population of the study

The targeted population of the study was formed from the higher education students in the governmental Jordanian universities (the Jordanian University, Al-Yarmook University, and Mu'tah University) in the study year of 2014/2015, which counted to (2716), divided among the governmental Jordanian universities in the educational sciences collage in all three universities. And table (1) explains the distribution of the study population according to the educational degree and the university:

Table1. The distribution of the population according to university and scientific degree in the Faculty of Educational Sciences

University / Degree	Master	PhD	Total
Jordan	587	495	1082
Yarmouk	705	407	1112
Mutah	471	51	522
Total	1763	953	2716

3.2. Sample of the study

1. Student sample: the sample was selected through the method of multi-level observation (aciniform method), and its percentage is 10% of the study population, where the population was divided into three sub-populations, according to the university, and the student in each university the students is divided into two groups according to the educational level into master's degree and doctorate degree, and the study tool was distributed randomly amongst the master's and doctorate students in each of the universities and the table (2) mentions the distribution of the individuals from the study sample.

Table2. The distribution of the study sample according to university and scientific degree in the Faculty of Educational Sciences

University /	Master		PhD		Total
Degree	Male	Female	Male	Female	
Jordan	32	27	22	27	108
Yarmouk	28	42	17	24	111
Mutah	25	21	4	6	56
Total	85	90	43	57	275

2. Judges sample: (10) randomly selected judges were selected from university professors who specialize and have expertise in related subjects of scientific research to provide their opinion about the paragraphs of the test and how clear the multiple selections and the alternatives are, and how linked each paragraph to the its correlated goal, also (5) highly specialized personals in the field of measuring and adjusting were selected to determine the minimum required score (cut off score) which the higher education students in the Jordanian governmental universities should get to have the minimum required statistical and research qualities, where their opinions were used to set the cut off score for the criterion referenced test which was prepared for the purpose of this study.

3.3 Instrument of the study

The building of the criterion referenced test to measure the statistical and research qualities for the higher education students in the Jordanian governmental universities went through the following procedures:

1. Identifying the goal of this test: identifying the goal of this test which is being constructed for this study to measure the statistical and research qualities for the higher education students in the Jordanian governmental universities to determine how much those students have of those qualities, for this test works as a diagnostic tool, through it we can determine the weak points in those qualities, and also the strength points in the qualities which the students perfected.
2. Setting the frame work for the qualities: the fields regarding the statistical and research qualities for the higher education students, which the students can perfect within the student's studying period in the university in the higher education period (master's degree and doctorate degree) in the following specialities: and table (3) shows the analysis of the qualities into stage goals.

First: the field concerning with measuring the basic qualities in statistics
Table3. Statistical and research skills to benchmarks.

First, measure the area of basic skills in statistics.
Basic concepts in statistics.
Educational data tab to psychological and represented graphically.
Determine measures of central tendency values.
Identify measures of dispersion grades of crude values in the educational and psychological data distribution
Interpretation of psychological and educational data using bivariate correlation coefficients and linear regression.
Validated statistical assumptions on the average standard deviation using the percentage distribution of critical community (Z).
Validated statistical assumptions on the average standard deviation community is unknown and the differences between the two means two independent samples or interlinked using dividend rate T T-test)).
Validated statistical hypotheses concerning the inference about the differences between the averages of combined or more by using one-way analysis of variance method.
Validated statistical hypotheses concerning the inference about the differences between the averages. Combined or more by using two-way analysis of variance method.
Validated statistical hypotheses concerning the inference about the differences between the averages of combined or more in the case that there is more than one dependent variable using multivariate analysis of variance.
Second, measuring the basic skills in research methodology.
Selection and definition of the problem.
A review of previous studies.
Sampling.
Choice of study tools.
The study methodology.
Writing a research report (documentation).

3. Building the paragraphs of the test:
   The building of the paragraphs of the test was completed through multiple steps from those we can mention the following: depending on the content analysis, looking for the already prepared tests in this field which was mentioned in the books specialized in the statistical ways and research methods as well as reviewing the Arabic and western studies which are related to the topic of this study, and the conditions required to building the objective tests of the multiple selection type, with taking in consideration to form the good paragraphs and fitting them with suiting multiple choices and how appropriate each paragraph is to its goal, the test at the beginning may contain up to 50 paragraphs of multiple choice type with 4 choices each, the supplement (A) shows the paragraphs in a primary test before finalizing its form.
4. Content validity:
   The progress of forming the paragraphs of the test and printing it, and to check the validity of the content the test was reviewed by a group of (10) specialized judges who have experiences and well-known qualities in the field of scientific research and that is to judge how linked the paragraph is to the goal set for the appropriate alternative and forming of the paragraph, and based on the judge's opinion the alteration in the paragraphs because some multiple choices alternative appeared repeated but with other words, also some paragraphs were altered because it would need a huge amount of time to solve and based on the judge's opinion the test was good in built and representing to the goals to be tested and the supplement (B) shows a list with the names of all the judges and the cut off score.
5. Primary trials of the test:
   The test was applied to a recon sample which consisted of (80) male and female students from the international university of Islamic sciences to know the primary indicators for the test paragraphs from difficulty and discrimination and how clear the paragraphs are and how much time is needed to apply the test and according to the procedures done paragraphs were deleted and removed from the test based on the differential values, because of that the differential values of the paragraphs that was removed from the test were (0.035, 0.075, -0.395, -0.064, -0.073) and so the total number of paragraphs was (45) paragraphs and in table (4) it is explained that the statistical indicators for all the paragraphs, and the paragraphs which was deleted was numbered (6, 7, 12, 26, 32).

Table4. Difficulty transactions and discrimination paragraphs testing experimental sample

Paragraph	difficulty	discrimination	paragraph	difficulty	Discrimination
1	0.90	0.32	26	0.90	0.064-
2	0.72	0.43	27	0.87	0.50
3	0.78	0.40	28	0.84	0.24
4	0.65	0.28	29	0.86	0.30
5	0.76	0.31	30	0.75	0.21
6	0.73	0.035	31	0.69	0.47
7	0.57	0.075	32	0.51	0.073-
8	0.75	0.34	33	0.71	0.31
9	0.81	0.44	34	0.74	0.23
10	0.68	0.41	35	0.82	0.48
11	0.82	0.48	36	0.88	0.25
12	0.19	0.345-	37	0.62	0.29
13	0.69	0.44	38	0.56	0.59
14	0.81	0.36	39	0.81	0.51
15	0.82	0.32	40	0.74	0.53
16	0.83	0.21	41	0.65	0.29
17	0.62	0.27	42	0.72	0.38
18	0.69	0.54	43	0.76	0.30
19	0.73	0.23	44	0.69	0.49
20	0.85	0.34	45	0.93	0.27
21	0.83	0.32	46	0.65	0.25
22	0.85	0.23	47	0.21	0.43
23	0.90	0.55	48	0.76	0.30
24	0.73	0.23	49	0.72	0.43
25	0.62	0.20	50	0.81	0.44

We notice from table (4) that the differential factors values ranged between (0.20) and (0.59) and we notice that almost all the paragraphs within the accepted range except the paragraphs that was deleted, where the accepted differential factors values must not be less that (0.20).

6. The cut off score
   This study followed the method of Angof to set the cut off score which is one of the methods used to determine the cut off score in the objective tests of the multiple choices type.
7. Applying the test of the study sample
   After finishing the steps of building the test, and assigning the cut off score, and applying some modifications to the paragraphs, and deleting five paragraphs based on the differential factors values in the experimental sample, the final form of the test which consists of (45) paragraphs was applied on the study sample in the second semester of 2014/2015 and the answers of the students were collected then corrected using a premade model answers form where the maximum score to be gotten in the test is (45) marks and the minimum score to be gotten in this test is (0) marks, after that a statistical analysis for the answers will be applied for the results to be analysed based on the study questions, and the supplement (C) shows the test in its final form which was applied in the governmental Jordanian universities (the Jordanian University, Al-Yarmook University, and Mu'tah University), and the supplement (D) shows the answer key.

3.4 statistical analyses:

To answer the questions of the study the following statistical processes were performed:

1. Calculating the difficulty factor and the differentiation factor for each paragraph of the test.
2. Determining the cut off score by using the Angof method.
3. Calculating the stability factor by using the Koder-Richardson coefficient (KR-20), and the Livingston index.
4. Calculating the correlation coefficient between the test's paragraphs and the total score to find out the vitality of the internal structure of the test.
5. Using the two way ANOVA to expose the differences in two variables, the social type and the scientific grade, and the interaction between them.

4. Displaying the results and analysing and recommendations

This chapter includes the displaying of the results that came from the study in the lights of its asked questions, the goal of this study was to construct a criterion referenced test to measure how qualified the higher educational students in the Jordanian governmental universities for the statistical and research qualities. And the results were shown according to the questions of the study as shown following:

4.1 Results display

The results related to the first question: "What is the necessary cut off score required to pass a criterion referenced test to measure the statistical and research qualified the higher educational students in the Jordanian governmental universities?"

To answer this question the test was reviewed by a group of experts and specialists in the field of measurement and adjustment and also on specialists in the field of scientific research to decide the score of which the higher education students are capable of the statistical and research qualifications which helps him prepare researches and studies, and the method of Angof was used to determine the cut off score, the method and the procedures which the researcher used is explained in the third chapter, the cut off score was set at (69%).

The results related to the second question: "What are the psychometric properties (validity and reliability) to test how much does the higher educational students in the governmental Jordanian universities have of research and statistical qualities?"

Validity: the paragraphs of this test was reviewed by group of experts and specialists in the field of measurement and adjustment and also on specialists in the field of scientific research to check the harmony of those paragraphs with the qualities they measure and how clear the paragraphs are, and how appropriate are the alternatives for the multiple choice questions, some paragraphs were removed and some alternative were adjusted and reformed based on the judges' opinions.

Reliability: the stability coefficient (Livingston index) was used to estimate the stability of the test where each sub-quality's stability had been calculated separately, then the whole stability was calculated for the test and table (5) shows those results

Table5. Reliability Livingstone to test the research and statistical skills transactions

Competencies	number of items	Livingstone coefficient
Statistical	26	830.
Research	19	820.
whole test	45	870.

We notice from table (5) that the sub-quality's stability for the statistical qualities reached (0.83) and for the research qualities (0.82) and for the test as whole reached (0.87), and that represents that the test has a high stability level.

The results related to the third question: "How much do the higher educational students in the governmental Jordanian universities have of research and statistical qualities?"

To answer this question the percentage and the repeat for the number of students both competent and incompetent according to the cut off score set in this study which is (69%) which is equivalent to (31) knowing that the full score of this test is (45) and table (6) shows those results.

Table6. Percentages and frequencies for the number of students who were classified into expert and non expert According cut-off-score

percentage of expert	no of non Expert	number of Expert	Degree
%53	82	93	MA
%63	37	63	PhD
%57	119	156	Total

We notice from table (6) that the number of competent students amongst the master's degree tested reached to (93) students with the percentage of (53%) while the number of competent amongst the doctorate degree tested reached (63) students with a percentage of (63%) and the total number of competent students from both categories reached (156) student with a percentage of (57%).

The results related to the forth question:" Does the degree of research and statistical qualities differ in the higher educational students in the governmental Jordanian universities according to the educational degree and social status?"

To answer this question the two-way ANOVA analysis was used to tell the difference in a variable from the social type and a variable from the scientific status and the interaction between them, also the mathematical means and standard deviations for both the variable from the social type and a variable from the scientific status, and table (7) shows those values.

Table7. Average, standard deviations for the variables of gender and grade

Gender	average	standard deviation
Males	26.29	6.93
Females	28.59	5.95
MA	26.74	6.73
PhD	29	5.81

We can notice from the table (7) that the average of male students performance in the criterion referenced test reached (26.29) while the average of the female students reached (28.59), also the average of the master's degree student's performance reached (26.74) while the average of the doctorate degree reached to (29).

Table8. Analysis of variance duo of the impact of differences in sex and scientific degree.

source deviation	degrees of freedom	sum of squares	average squares	) F(	level of significance
Gender	1	261.59	261.59	6.51	0.011
Degree	1	300.05	300.05	7.47	0.007
Interact	1	11.95	11.95	0.3	0.59
Error	271	10892.15	40.20		
Total	274	220429			

We notice from table (8) that there is difference in the performance in the criterion referenced test which can be related to the social type where the value of (F) reached (6.51) which is a statistical value at the level of (α ≤ 0.05).

By going back to table (7) we can see that the differences were in favour of the females where the mathematical average for the females was (28.59) while the mathematical average for the males was (26.29), we also notice from the table that the value of (F) for the educational degree was (7.47) which is a statistical value at the level of (α ≤ 0.05), also by returning to the table (8) we notice that the difference was in favour of the doctorate degree where the mathematical average reached (29) in compare to the master's degree which was (26.74). As for the interaction between the gender and the scientific degree the value of (F) reached (0.3) which is statistically irrelevant at the level of (α ≤ 0.05).

4.2 Discussions of the results

Discussion of The results of the first question which was texted in the form: "what is the necessary cut off score required to pass a criterion referenced test to measure the statistical and research qualified the higher educational students in the Jordanian governmental universities?"

The results of the first question showed that the cut off score of the criterion referenced test according to the opinions of the judges reached as following in order (0.6188, 0.6833, 0.7733, 0.7388, 0.6288) and through calculating the mathematical average of the five judges the cut off score almost reached (69%) which is equivalent to (31) having known that the full score of the test is (45) and this value is fairly close to the set cut off point to the passing score of the curriculum of the higher education students which is (0.70).

Discussion of The results of the second question which was texted in the form: "What are the psychometric properties (validity and reliability) to test how much does the higher educational students in the governmental Jordanian universities have of research and statistical qualities?"

The studies related to the validity of the test indicated that the test measures what it was meant for and that is through guidance of the judges opinions, where some paragraphs were removed and some paragraphs were altered along with some alternatives until the final form of the test was ready to be applied on the study sample or any sample similar to the properties of the current sample, and the results regarding the stability indicated that the test has a relatively high stability result for the total stability factor measures to be (0.87) and for the research qualities to be (0.82) and for the statistical qualities to reach up to (0.83).

Discussion of The results of the third question which was texted in the form: "How much do the higher educational students in the governmental Jordanian universities have of research and statistical qualities?"

The results of the third question indicated that the percentage of the competent master's degree students reached (0.53) which is a low percentage up to some point; the condition was the same for the students of the doctorate degree were the percentage of competent reached (0.63) which is also low, this percentage is different from the study of (Al-Kasasbeh,2013) were the results of that study shown that the competent percentage amongst the carriers of the master's degree reached to only (0.24) and the percentage of competent amongst the doctorate degree carriers reached up to (0.81) and that may be due to the difference in the study population and also due to that in this study the statistical and research qualities were focused on, were in the study mentioned before the focus of the researcher was on the research qualities only, and to correct for this obvious shortage in the statistical and research qualities a review in all of the plans and programs which are provided to the higher education students, where these programs focus on the theoretical aspects more than the applicable research aspect, the research skills must be nurtured through assigning the students to perform renovated researches and to publish them in the available scientific journals.

Discussion of The results of the third question which was texted in the form: "Does the degree of research and statistical qualities differ in the higher educational students in the governmental Jordanian universities according to the educational degree and social status?"

The results related to the forth questions indicated that there is difference in the performance of the criterion referenced test depending on gender where there is a statistical relevant value at the level of (α ≤ 0.05), the results were favouring the females, the researcher sees that this favouring is due to the general female over male supremacy, and that is due to the females taken more care than males in general.

Also the results shown that there is a difference in the performance which can be related to the scientific degree, where the doctorate degree students performed better than the master's degree students which is only natural because the doctorate degree has more experience in preparing the research theses due to the collection of experience in the field, which is correlating with the study of (Al-Kasasbeh, 2013).

5. Recommendations:

Based on the results of this study the researcher recommends the following:

1. Working or reviewing the programs and plans for the study materials in order to increase the skills of scientific research and that is by preparing researches and studies as a requirement for specific curriculums.
2. Rebuilding similar tests through the modern theories in measurement.
3. Using such tests as this one in the universities qualification tests for the master's and doctorate degree for what it has of good psychometric properties.

References

Atwan, Asad & Al Fallit, Jamal (2011). Efficiencies of scientific research with graduate students in colleges of education Palestinian universities. Proceedings of the Conference on Scientific Research, May 11, the Islamic University of Gaza.

Al Imam, Wifqi (2008). Scientific research (research project and writing the final report) setting. Egypt, Mansoura: modern library for publication.

Al Ashari, Ahmed (2007). Summary of the methods of scientific research. Saudi Arabia, Jeddah: King Fahd National Library for publication.

Al Banna, Mamoun (2011). Building test whom one speaks to measure statistical skills among graduate faculties of education in universities Yemeni students. Unpublished MA Thesis, King Saud University, Saudi Arabia, Riyadh.

Berk, Roland (1982). Criterion-Referenced Measurement The state of The Art. Baltimore and London: The Johns Hopkins University Press. http://dx.doi.org/ 10.1604/9780801822643

Crocker.I.& Algina,j. (1986). Introduction to Classical & Modern Test Theory. New Yorkholt. Ringeholt & Winston.

Guziano, Cecilie (1995). A Twenty five years review of Knowledge gap research, paper presented at association for public opinion research, Minnesota.

Halpin, G, sigmon,44. G,(1983). Minimum competency standard set by three divergent groups of raters using three judgmental procedures: implication for validity. Educational & Psychological Measurement. 43:185-196.

Hambleton, R.K (1978). One the use of cut off scores with criterion referenced tests in instructional setting Journal of Educational Measurement. 15:277-290. DOI: http://dx.doi.org/10.4135/9781446263549

Hmbleton,R.K & Novick. M. R. (1971). To words Theory of Criterion-Referenced test. American College Testing Technical Report, Iowa City.

Gamel, Abdul Rahman (1998). Educational skills in measurement and evaluation and acquisition of self-education. Jordan, Amman: Dar almareek for publication and distribution.

Jonthan, p. Lewies (1991). General overview of bais and validity issues in cross-cultural research, Coloroda state university, FT. Collins, office for applied research.

Kerlinger, F.N. (1973). Foundations of behavioural research seconded. New York: Holt, Rinehart and Winston.

Al Zubi, Bilal & Tlafhh, Abbas (2000). SPSS statistical system understanding and analysis of statistical data. Jordan, Amman: Dar Wael for printing and publishing.

Al Zghoul, Imad (2002). Principles of Educational Psychology. United Arab Emirates: University Book House.

Majeed, Sawsan (2007). Foundations of psychological and educational testing and standards. Jordan, Amman: De Bono for publication and distribution.

Al Maaytah, Abdul Aziz (2011). New trends in scientific research. Jordan, Amman: Dar haneen for publication and distribution.

Mac, Farland, Thernes, W (1996). Computr-hased research and statistics study guide, Nove University, fort Landerdale, ft,center for computer and information.

Murad, Salah, & Soliman, Amin (2002). Tests and measurements in the psychological and educational science preparation steps, characteristics. Egypt, Cairo: Dar al-Hadith book.

Al Nabhan, Mossa (2004). Measurement Fundamentals in the Behavioral Sciences. Jordan, Amman: Dar Al Shorouk publishing and distribution.

Al Najjar, Nabil (2010). Measurement practical perspective and calendar with Spss software applications. Jordan, Amman: Dar Al-Hamed for publication and distribution.

Al Nairab, Fareed (2010). Imagine a proposal to develop the academic productivity of Higher Studies in Palestinian universities in Gaza in light of the development plans. Unpublished doctoral dissertation, Arab Research and Studies Institute, Egypt, Cairo.

Al Kasasbeh, Hanan (2013). Building test whom one speaks reference to measure the proficiency of graduate students at the University of Mutah competences of scientific research, unpublished Master Thesis, Mutah University, Karak, Jordan.

Allahlah, Ahmad. & Abu Bakr, Mustafa (2002). Scientific research definition, his steps, its methods, statistical concepts. Egypt, Cairo: University House.

Ababneh, Imad (2009). Criterion reference tests philosophy and the foundations of development. Jordan, Amman: Dar almaseerah for publication and distribution.

Adas, Abdul Rahman, Obedat, Thuqan, Abdul Haq, Kayed (2005). Scientific concept, tools and methods of research. Saudi Arabia, Riyadh: Dar Osama for Publishing and Distribution.

Odeh, Ahmed & Malkawi, Fathi (1987). Basics of scientific research elements and methods and statistical analysis of the data. Jordan, zarqa: Manar library for publication and distribution.

Odeh, Ahmed (2005). Measurement and Evaluation in the teaching process(5th ed). Jordan, Irbid: House of Hope for publication.

Popham, W.J.(1978), Modern Educational Measurment practical Guidelines four educational Leader (3ed). Bosten: Allyn and Bacon.

Shepard, Lorrie, (1984). Setting performance standards, In Berk (Ed.), Aguide to criterion- Referenced tests construction. Hoblcins university press.

Allam, Salah al-Din (1986). Contemporary developments in psychological and educational measurement. Kuwait: Printing business Qabas.

Allam, Salah al-Din (1995). Diagnostic tests reference stake in the educational, psychological and training fields. Egypt, Cairo: Dar Arab Thought.

Allam, Salah al-Din (1999). Search psychological and educational science curriculum. Egypt: Publishing House of the universities.

Allam, Salah al-Din (2005). Diagnostic tests reference stake in the educational, psychological and training fields. Egypt, Cairo: Dar Arab Thought.

Allam, Salah al-Din (2006). Educational and psychological tests and scales. Jordan, Amman: dar thought for publication and distribution...

Allam, Salah al-Din (2007). Diagnostic tests reference stake in the educational, psychological and training fields. Egypt, Cairo: Dar Arab Thought.