Constructing a Criterion Referenced Test for the Eleventh Grade Student in Physics

Dr. Moh. S. ETOOM
Dr. Waleed SHDOOH
JERASH UNIVERSITY

Article Info

Received: 22.04.2016
Accepted: 14.05.2016
Published online: 01.06.2016

ISSN: 2231-8968

Abstract

The study aimed to construct a Criterion Referenced test to measure the outcome of the function for the eleventh grade student in physics , and identified the cut score .

The final form of the test which consist of (45) items was applied on a sample of (120) students male and female of Jerash district of education.

The test was handed to (10) arbitrators to determine the degree of cut score for passing by (Angoff's method) which was (46%).

The reliability coefficient was measured by using KR-20 and (Livingston) which was (0.85 & 0.87) in sequence .

The result showed a low performance achieved on the test ,so the researchers advice for more research on other subjects and more in the sample of the students.

Introduction

The tests are one of the most important means to measure and evaluate students abilities with different levels, from which we can reveal the success elements, to enforce it and to avoid the weakness elements in addition to developing their abilities in order to raise their levels. Also tests have an important role in the formation of the human behavior and the evaluation of the effectiveness of the educational environment components to achieve the goals . despite its drawback it is the most frequently used to assess student's scientific capabilities.

Tests are divided into two types: Standardized reference and Criterion reference , where the origin of the terms goes back to (Glaser , 1963), when he tried to distinguish between the two types of standards in the term of interpretation of the results for each standard . One of the Criterion reference is empowerment tests (master) and adequacy test (competency) . The second type is the standard reference tests. Both of the two types are put to measure a specific teaching goal but the Criterion reference need a detailed targets to be developed , in addition to that the result of the Criterion reference is more than of the standard reference results.

(Oudeh 2010) refers that the primary difference between the two types in term of the interpretation of results. The standard reference tests compares the student's performance with the achieved targets, it is important to reflect the students answer for each paragraph his ability to answer.

The problem of the study and its importance.

The education advantage from the measurement and evaluation is growing up in the field of measuring student's learning to make wise decisions concerned with the field related to tests, recruitment or measurement of achievements. Variety in decision shows two kind of tests, it is standard reference and Criterion reference. At the beginning the standard reference test has greater attention than the Criterion reference tests, the education institutions focused on comparing the individual performance with the group performance average which the individual belongs to . Because of the failures of these standard reference tests, the need arises for a new kind of tests to help teachers , and school managers to take the appropriate measures to avoid the previous failures. The importance of this study comes from the researcher attempt to focus on the Criterion reference tests in terms of preparing and analysis , also this study participates in viewing and applying the steps necessary to prepare tests and analyzing its paragraphs.

The goals of the Study.

1. Showing the best method to prepare tests.
2. Using different methods to determine the degree of taking a decision in the Criterion reference tests.
3. Using this test to measure the achievements of the students in physics course for first secondary (11) grade.
4. And the fact that it seeks to evaluate and assess learners to determine point of strength and point of weakness to be avoided.

The Study determinants:

This study is limited to the first secondary grade students in Jerash governorate schools for the academic year 2014\2015.

The theoretical framework and previous studies.

Oral tests were still used for a long period in Italy and France , while written tests goes back to (Cambridge University)1800Ac then spread to Oxford University , in 1845 Poster exam was used in America.

Tests are known as a measuring tool, it is prepared according to an organized method consists of several steps which are subject to specific conditions and rules to determine individuals capabilities (Oudeh 2010) .

The origin of the two terms (standard reference and Criterion reference) goes back to (Glaser , 1963) , when he tried to distinguish between two types of standards in terms of interpretation of each standard results. For the standard reference it concentrate on the tests results interpretation to determine the individual level according to his classmate . For the Criterion reference , it focus on the tests results interpretation to compare the individual level with the absolute performance level .

(Haertel,1985) added that most of Criterion reference tests includes cutting score , so that the individual will pass if he achieve the cutting score and he will fail if he doesn't achieve it . These tests are called (Mastery test) , this is in line with (Cronbach , 1970) who confirms that the standard is the domain which the test is designed to measure .

(Wang, 2003) says that the Criterion reference tests is an achievement test used to measure the presence or absence of a particular behavior, this behavior is a topic in a specific educational goal. (Berk , 1980) confirms that the difference between the standard reference tests and Criterion reference tests lies in the difficulty of paragraphs .In the standard reference tests , paragraphs are designed for broader degrees of variation then we approach from the normal distribution form but this is not normal in Criterion reference tests , it uses only paragraphs which are characterized by the ability to distinguish the student who achieved the goals that we need to measure and the student who failed.

In (Ojer, 1982) study , he compared four methods to estimate the stability of the Criterion reference tests, one of them is Livingston in which the test was applied in Kansas State in reading and in math . The result was that all of stability indicators have a high and stable values in all samples .

In a research "Criterion-Referenced Measurement for educational Evaluation and Selection" (Christina Wikström, 2005)

The research investigates the consequences of using criterion-referenced measurement for both educational evaluation and selection purposes. The research comprises an introduction and four papers that empirically investigate school grades and grading practices in Swedish upper secondary schools. The results show that schools that are exposed to competition tend to grade their students higher than other schools. The results show that grades have increased every year since the new grading system was introduced, which cannot be explained by improved performances, selection effects or strategic course choices. The conclusion is that the increasing pressure for high grading has led to grade inflation over time. The results show small but significant size effects, suggesting that the smallest schools (<300 students) are higher grading than other schools, and that the largest schools (>1000 students) are lower grading than other schools. The results show that students in vocationally oriented programs are higher graded than other students, and also favored by their programs' course compositions, which have a positive effect on their competitive strength in the selection to higher education. In the introductory part of the thesis, these results are discussed from the perspective of a theoretical framework, with special attention to validity issues in a broad perspective. The conclusion is that the criterion-referenced

grades, both in terms of being used for educational evaluation, and as an instrument for selection to higher education, are wanting both in reliability and in validity. This is related to the conflicting purposes of the instruments, in combination with few control mechanisms, which affects how grades are interpreted and used, hence leading to consequences for students, schools and society in general.

An other research by T. Kumazawa (2007) titled "Criterion-referenced test administration designs and analyses" .

The paper mentions the differences between norm-referenced and criterion-referenced tests and introduces one possible criterion-referenced administration design. Two forms of a (25-item) multiple-choice criterion-referenced vocabulary test were developed and administered to two groups of Japanese university EFL students (n=87) for diagnostic and achievement purposes in a counterbalanced pretest/posttest design. The dependability indexes for these tests were low or moderate and an item analysis of the criterion-reference tests suggests there was a slight increase in score gain after a period of thirteen weeks of instruction. This suggests that most of the students mastered a modest amount of the target vocabulary.

A study by T.J.Frain (2009) "A Comparative Study of Korean University Students before and after a Criterion Referenced Test."

The study aims to determine the perceptions of first year university students to criterion referenced testing. The students have been tested using norm referenced testing for most of their English language education and this has culminated in The College Scholastic Aptitude Test(CSAT).

The poor communication skills of the students has prompted the researcher to question why CLT methodology is not complemented by a communicative 4test that reflects real life situations practiced in the classroom. The attitudes and perceptions of the students may support a different method of testing that complements a communicative approach to learning. It seems that backwash from the CSAT, which emphasizes only reading and listening, is negatively affecting communicative competence (Flattery, 2007). The experimental approach will be action research as this is a single case study of students and their attitudes and perceptions about language testing. It seeks to understand the effect that backwash from a test has on them. The students were tested using a paired criterion referenced test during their first semester at university. They are surveyed twice, before and after the criterion referenced test, to determine their opinions about this new testing method and norm referenced testing. The survey items reflected the qualities of a good language test, namely, inter activeness, practicality, reliability, validity, practicality, and impact. The results seem to indicate that the Students question the reliability of norm referenced testing while criterion referenced testing created positive backwash.

The students perceive the use of real world tasks as being more relevant in assessing their abilities in English compared to de contextualized multiple choice exams. They also perceive that they are no longer being compared with each other but in their ability to perform a task, which seemed to create a positive attitude toward language learning

Criterion reference tests preparing steps:

This type of tests are used to determine the level of students performance according to behavioral system of knowledge and basic skills, without comparing this performance with other students performance , so we need to prepare a detailed plan before start writing paragraphs.

The first stage is determining the content that we need to measure to match the vocabulary with the content, taking into account the student nature. The second stage is determining the general target.

In the third stage we analyze the general goals to a procedural goals, by describing a sample of the behavioral goals which can be aprove on achieving every goal of the general goals.

The fourth stage is determining the behavioral scale which the test can measure. The fifth stage is forming the vocabulary of the test, this stage is divided into two steps, first is the selection of the appropriate vocabulary, second is the determination of the appropriate number of vocabulary, by taking in account the importance of the behavioral scale, so whenever the test specifications are clear and specific, the paragraph will be more honest to measure the content. (Allam, 1986).

The psychometric properties of the Criterion reference test.

1-The Reliability of the Criterion reference tests.

The degrees are interoperated by referring to a certain level to mastering skills and concepts. (oadeh, 2014) .

Reliability is defined by Phelan&Wern (2006) as the degree to which an assessment tool produces stable and consistent results. They (Phelan&Wern) consider four types of reliabilities:

1. **Test-retest reliability** is a measure of reliability obtained by administering the same test twice over a period of time to a group of individuals. The scores from Time 1 and Time 2 can then be correlated in order to evaluate the test for stability over time.
2. **Parallel forms reliability** is a measure of reliability obtained by administering different versions of an assessment tool (both versions must contain items that probe the same construct, skill, knowledge base, etc.) to the same group of individuals. The scores from the two versions can then be correlated in order to evaluate the consistency of results across alternate versions.
3. **Inter-rater reliability** is a measure of reliability used to assess the degree to which different judges or raters agree in their assessment decisions. Inter-rater reliability is useful because human observers will not necessarily interpret answers the same way; raters may disagree as to how well certain responses or material demonstrate knowledge of the construct or skill being assessed.
4. **Internal consistency reliability** is a measure of reliability used to evaluate the degree to which different test items that probe the same construct produce similar results.
    A. **Average inter-item correlation** is a subtype of internal consistency reliability. It is obtained by taking all of the items on a test that probe the same construct (e.g., reading comprehension), determining the correlation coefficient for each *pair* of items, and finally taking the average of all of these correlation coefficients. This final step yields the average inter-item correlation.
    B. **Split-half reliability** is another subtype of internal consistency reliability. The process of obtaining split-half reliability is begun by "splitting in half" all items of a test that are intended to probe the same area of knowledge (e.g., World War II) in order to form two "sets" of items. The *entire* test is administered to a group of individuals, the total score for each "set" is computed, and finally the split-half reliability is obtained by determining the correlation between the two total "set" scores.

There are several methods can be used to estimate the Reliability of these tests , as the method which is required to present the test one time like (Livingston Indent) which depends on the cutting

score and the student score in the behavioral scale score and the deviation of the student scores from the cutting score .
The coefficient can be calculated by the following equation:

$$K^2(X, T) = \frac{\delta^2{}_X (KR - 20) + (\mu_X - n_i c)^2}{\delta^2{}_X + (\mu_X - n_i c)^2}$$

Where ,$K^2(X, T)$:Livingiston coefficient

KR – 20 : Kuder &Richardson coefficient

$\delta^2{}_X$: variance scores of the test

μ_X : average score

C : cut score

n_i :number of items

Some methods need to apply the test two times like Reliability as (Carver) coefficient. the Reliability is estimated by comparing the skilled student , whenever the consistency increases between the two applications then the tests are more stable.
2. The effectiveness of paragraphs.
There are indicators that show the contribution of each paragraph in the test in general .This effectiveness includes the sensitivity of the paragraph. Vargas Cox suggested a method to find the sensitivity of paragraphs. It is by comparing the results before the teaching process by making a test before teaching and the results after the teaching process by a new test after teaching.
3. Validity.
If the test measure what it has to measure then it is valid, its statistical meaning is the real contrast rate that associated or attributed to the total contrast. Validity has several meaning one of them is descriptive validity that used to describe the students' performance according to the behavioral scale .The second is career validity , means test accuracy which is necessary to achieve the purpose of this test so, we have several methods to reveal it , the most common is the correlation coefficient (oadeh , 2010) .
Validity, is defined by Phelan & Wern (2006) as the test measure what it has to measure.
Why is it necessary?
While reliability is necessary, it alone is not sufficient. For a test to be reliable, it also needs to be valid. For example, if your scale is off by 5 lbs, it reads your weight every day with an excess of 5lbs. The scale is reliable because it consistently reports the same weight every day, but it is not valid because it adds 5lbs to your true weight. It is not a valid measure of your weight.
Types of Validity

1. Face Validity ascertains that the measure appears to be assessing the intended construct under study. The stakeholders can easily assess face validity. Although this is not a very "scientific" type of validity, it may be an essential component in enlisting motivation of stakeholders. If the stakeholders do not believe the measure is an accurate assessment of the ability, they may become disengaged with the task.
2. Construct Validity is used to ensure that the measure is actually measure what it is intended to measure (i.e. the construct), and not other variables. Using a panel of "experts" familiar with the construct is a way in which this type of validity can be assessed. The experts can examine the items and decide what that specific item is intended to measure. Students can be involved in this process to obtain their feedback.
3. Criterion-Related Validity is used to predict future or current performance - it correlates test results with another criterion of interest.

4. Formative Validity when applied to outcomes assessment it is used to assess how well a measure is able to provide information to help improve the program under study.

5.Sampling Validity (similar to content validity) ensures that the measure covers the broad range of areas within the concept under study. Not everything can be covered, so items need to be sampled from all of the domains. This may need to be completed using a panel of "experts" to ensure that the content area is adequately sampled. Additionally, a panel can help limit "expert" bias (i.e. a test reflecting what an individual personally feels are the most important or relevant areas).

The study methodology:

1- The study population .

It consists of the first secondary grade \ scientific branch , in Jerash schools for 2014\1015, they are (779) students as shown in table (1).

Table (1) Population of the study schools of 11th grade students in Jerash district

No.	Female schools	No.	Male schools	No.	sum
1	Mersea	8	Bormeh	20	28
2	Sakeb	34	Al-mestabeh	4	34
3	Beleala	9	Mersea	16	25
4	Jerash Camp	25	Jebeh	10	35
5	Al- hadadeh	6	Jerash	160	166
6	Nehleh	10	Souf	20	30
7	Al-keteh	30	Qafqafa	8	38
8	Al-mestabeh	34	Ased	9	43
9	Bormeh	11	Sakeb	15	26

10	Koferkhel	23	Al-kabesi	17	40
11	Qafqafa	26	Beleala	8	34
12	Souf Camp	46	Al-mesharefeh	4	50
13	Deher Al-seroo	14	Souf Camp	20	34
14	Souf	31	Jerar alnoman	12	43
15	Jerash	69	-	-	69
16	Al-khensa	80	-	-	80
sum	-	456	-	323	779

Table (2) Sample of the study schools of 11th grade students in Jrash district

No.	Male schools	No.	Female schools	No.	sum
1	Souf	10	Jerash Camp	10	20
2	Al-keteh	10	Koferkhel	10	20
3	Jerash	20	Sakeb	10	30
4	Al-kabesi	10	Qafqafa	10	20
5	Bormeh	10	Al-khensa	20	30
sum	-	60	-	60	120

3. The study tool.

The researcher uses the Criterion reference test in physics course to show the students clarity of concepts and mastering the basic skills.

The test procedure:

A. Determine the test objectives: it's the formation of Criterion reference tests to measure the student's skills in physics for the first secondary grade \ scientific branch.

B. Analysis the physics course content\ first semester. through studying teacher guide book , students book and the specific goals that students must achieved .

C. Preparing a specification table, appendix (1) shows that,

D. Paragraphs drafting : the researcher formed the paragraphs depending on his experience which is more than 20 years , in addition to analyzing the content of the physics and studying some of previous tests .

E. paragraph arbitration: Paragraphs were displayed to a group of specialist and teachers from the directorates of education in Jordan to determine the relation between the paragraph and the targets. Also paragraphs were viewed by specialists in evaluation &measurement and language to determine the strength of alternatives, the researcher takes in judges suggestions and modifications.

F. The first application: This test was applied on a sample consists of (40) students to know the degree of clarity and to determine the time that students need to answer .

G. cutting score determination : There are many methods used to determine the cutting score. (Angoff) method is the most common , because it suite most of test, and used to applied judge opinions.

The cutting score which is necessary for success is the score at which the students can answer without guess .

H. Test application in all of the samples : The test (which consist of 45 items) is applied on all sample (which consists of 120 students) , after the modifications and determining the cutting score . Then the results are entered to the computer by using SPSS program , in which we can analyze the result through (Angoff)method to determine the cutting score , calculate the paragraph sensitivity coefficient and the stability coefficient by using KR-20.

Discussing the results :

First question results : what is the cutting score which is necessary for success in Criterion reference test in physics course for the first secondary (11th) grade . Depending on the opinion of specialists and experts , (3) male teachers , (3) female teachers , (2) supervisors , (2) specialists in measurement and evaluation . Table (3) shows the results . The skilled students will answer the paragraph with the correct answer while the student who is not skilled does not have a correct answer .

(3)Table opinion of specialists and expert

Item	1	2	3	4	5	6	7	8	9	10	sum
1	1	1	1	1	1	0	1	1	0	1	8
2	1	1	1	0	1	0	0	0	1	1	6
3	1	1	1	1	0	0	1	0	0	0	5
4	1	1	1	0	1	1	1	0	1	0	7
5	0	0	1	0	1	0	1	0	1	0	4
6	1	0	1	1	0	0	1	0	0	0	4
7	0	0	0	0	0	1	0	1	1	0	3
8	1	1	1	1	1	0	1	0	0	0	6
9	0	0	0	0	0	1	0	1	1	0	3
10	1	1	0	0	0	0	1	1	0	1	5
11	0	0	1	1	1	1	0	0	0	1	5
12	1	1	1	1	0	0	1	0	1	0	6
13	1	0	1	1	1	1	0	0	0	0	5
14	1	1	1	1	0	0	0	1	0	0	5
15	1	0	0	0	0	0	0	1	1	1	4

16	0	1	0	0	1	0	1	0	0	1	4
17	0	1	0	0	1	0	1	0	0	1	4
18	1	0	0	0	0	1	1	1	1	1	6
19	0	0	0	0	0	0	0	1	1	0	2
20	0	1	1	1	1	0	0	0	0	1	5
21	0	0	0	0	0	1	1	1	0	1	4
22	1	1	1	1	0	0	0	0	1	0	5
23	1	0	0	0	1	1	0	1	0	1	5
24	1	1	1	1	1	0	1	0	1	0	7
25	0	0	0	1	0	0	0	1	0	1	3
26	0	0	0	1	0	0	0	0	0	1	2
27	0	1	1	1	1	1	0	0	0	0	5
28	0	0	0	1	0	0	1	1	1	0	4
29	0	0	0	0	1	1	0	0	0	1	3
30	1	1	1	0	0	1	1	0	0	0	5
31	0	0	0	0	0	0	0	1	1	0	2
32	0	0	1	0	0	1	1	0	0	1	4
33	0	1	1	1	1	0	0	1	1	0	6
34	1	0	0	0	0	1	1	0	0	1	4
35	0	1	1	0	0	1	0	1	1	0	5
36	0	0	0	0	1	0	1	0	0	1	3
37	1	1	1	1	1	0	0	1	0	0	6
38	0	0	0	1	0	1	1	0	0	1	4
39	0	0	0	0	1	1	0	1	1	0	4
40	1	1	1	1	1	1	1	0	0	0	7
41	0	0	0	0	0	0	0	1	1	1	3
42	0	1	1	0	0	1	1	0	0	0	4
43	0	0	1	1	0	0	0	0	1	0	3
44	0	0	0	1	0	1	1	0	0	0	3
45	1	1	1	0	1	0	1	0	1	0	6
sum	19	19	18	23	19	20	21	24	21	20	204

From the result shown in table (3) cut score is (20.4) which means that ; the student who answers (20.4) or more out of (45) of the items he is mastery other wise not.

Item sensitivity: means the Item ability to distinguish between the sample members in the Criterion reference test .

Through the table (4) we note that the sensitivity coefficient value is between (0.32) and (0.53) which is acceptable for studying purposes .

Table (4)Coefficient of sensitivity for each item

item	sensitivity Coefficient	item	sensitivity Coefficient	item	sensitivity Coefficient
1	0.28	16	00.45	31	0.24
2	0.48	17	0.45	32	0.45

3	0.51	18	0.48	33	0.48
4	0.38	19	0.24	34	0.45
5	0.45	20	0.51	35	0.51
6	0.45	21	0.45	36	0.24
7	0.24	22	0.51	37	0.48
8	0.48	23	0.51	38	0.45
9	0.24	24	0.38	39	0.45
10	0.51	25	0.24	40	0.38
11	0.51	26	0.24	41	0.24
12	0.48	27	0.51	42	0.45
13	0.51	28	0.45	43	0.24
14	0.51	29	0.24	44	0.24
15	0.45	30	0.51	45	0.48

Third question results : what is the psychometric properties (Validity & Reliability) of the test ? The test Items were viewed to the group of specialist and supervisors and teachers for comments . to make sure of the reliability of the Items . the students scores were compared to the average of the scores for the other courses ,the correlation coefficient was (0.86), for the Reliability indicators , the researchers used the Reliability coefficient KR\20 by using the equation KR\20 ={ n/(n-1)}(1- pq/S^2) whereas.

n= number of the Items

p= the rate of the right answers

q= the rate of the wrong answers

S^2= the total variation

KR -20 =45\44(1- 0.15) = 0.87

Also the used of Livingston method .

It was (0.89), it is acceptable value for the purpose of this test.

The result of the fourth question :

What is the level of students performance for the first secondary grade in this test ? to answer this question the researchers calculate the rate of the students who exceed the cutting score and those who didn't exceed it .The table (5) explains that.

Table (5) Students performance level in the test

Level	Students number	Rate
Skilled	50	41%
Not skilled	70	59%

We note in the previous table (5) that the rate of not skilled students is more than the rate of skilled students , the reason to the high brainpower which is required in physics course .

Recommendations :

1 Preparing a Criterion reference test in other courses .

2. Preparing a special banks for questions with a good specifications to be used .

3. Determining the cutting score by using other method .

References:

Allam,salah adeen (2000) **diagnostic test , in the psychological educational ,and training fields** , Cairo, Dar al fiker al arabe ,

Allam,salah adeen (2001) **educational and psychological measurement and evaluation** , Cairo, Dar al fiker al arabe.

Al–khatatneh , ruqaya (2014) **preparing a Criterion reference test in trigonometric ratios for the ninth grade** , , un published master thesis . mu`atah university , al-karak , Jordan.

Frain , T. Joseph (2009) , **A Comparative Study of Korean University Students before and after a Criterion Referenced Test** , University of Southern Queensland, Australia.

Kumazawa,Takaaki (2007) **Criterion-referenced test administration designs and analyses** , Kanto Gakuin University.

Oudeh; Ahmmad (2010) **measurement and evaluation in the teaching process** .Dar al amal , Irbid.

Shqerah, hana (2011), **preparing a Criterion reference test in maths for the first grade and determining the cutting score** , un published master thesis . mu`atah university , al-karak , Jordan.