Constructing Criterion Referenced Tests in Mathematics for The Basic grades (5th. to 10th.) and investigating their Psychometric Properties

Mohammad W. M. Batsh, Ekhleif Y. Tarawneh, Ferial M. Abu Awwad, Hamzah A. Omari *

ABSTRACT

This study aims at constructing criterion referenced tests in mathematics for the grades (5th. to 10th.) and investigating their psychometric properties, all the mathematical domains and the related intended learning outcomes (ILO's) are identified, the items that best measure these domains and ILO's are constructed, edited, and included in specially prepared templates. The necessary instructions for the administration of the tests are also included in the tests booklets. The population consists of (837471) students from the different grades. The tests are applied on a sample consisting of 3301 students from the six grades.

The cut-off scores used to distinguish mastering students from non-mastering students in math are determined using Angoff method and contrast groups. The validity of the tests is established using content validity, decision validity, and criterion referenced validity/ concurrent. Reliability is established using Cronbach alpha, Livingstone method, Carver coefficient and Kappa coefficient. The results indicate sufficient indicators for psychometric properties of items and tests.

Keywords: Criterion- Referenced Test; Intended Learning Outcomes; Cut- Off Score; Validity; Reliability; Mastering And Non- Mastering.

Introduction

Math is very important for both school and life, because it helps students to deal with problems in a scientific way, and to practice higher order thinking skills, such as critical thinking, creative thinking, reflective thinking, and inductive and deductive thinking. It also helps students to acquire other skills such as objectivity, accuracy, self-planning, organization, and evaluation. Therefore, understanding mathematics and its role in students' learning has a major influence on the development of mathematics curriculum, instruction, and research.

Students' achievement in mathematics can be measured through qualitative methods, such as interviews and classroom observations, or through quantitative methods, such as written tests, rating scales and rubrics or through a combination of both methods (Vos, 2005). However, quantitative methods are not as easy as it sounds, because 'achievement' consists of a large variety of interacting components. Students' scores on an achievement test are in the first place convenient shorthand, especially for reporting on students and comparing between them. It should be noted that these functions are related to norm-referenced tests, and really, this type of test is the common in measuring achievement in math and other subjects.

Popham and Husek (2005) argued that during the past several years, measurement and instruction specialists distinguished between norm-referenced and criterion-referenced approaches to measurement. Traditionally, a norm-reference measure was used to identify an individual's performance in relation to the performance of others on the same measure. A criterion-referenced test, on the other hand, was used to identify an individual's status with respect to an established standard of performance. Criterion-referenced tests also report how well students are doing with reference to a pre-determined performance level on a specified set of educational goals or outcomes included in curriculum. They are

* Department of Measurement and Statistics- The University of Jordan; Department of Educational Administration- The University of Jordan; Department of Educational Psychology- The University of Jordan; and Department of Curriculum & Instruction- The University of Jordan. Received on 16/5/2017 and Accepted for Publication on 13/5/2018.

© 2019 DAR Publishers/University of Jordan. All Rights Reserved.

www.manaraa.com

used when teachers wish to know how well students have learned the knowledge and skills intended by offering a certain program. This information may be used as one piece of information to determine how well the student is learning the desired curriculum and how well the school is teaching that curriculum. For instance, a criterion-referenced test score might describe which arithmetic operations a student can perform, or the level of reading difficulty experienced (Anastasi, 1988). One of the aims of criterion referencing is to focus on individual assessment based on descriptions of performance across a range of levels. Brown cited in (Green, 2002, P. 3) defined criterion referenced assessment as an evaluative description of the qualities which are to be assessed (e.g. an account of what students know and can do) without reference to the performance of others.

Criterion based assessments are meant to determine where students stand in relation to a specific standard, the goal is not to identify winners and losers but, rather, to enable as many students as possible to master the given knowledge and skills. However, while mastery learning uses tests to help students to master discrete bits of content, criterion-based assessments measure student performance in relation to specific learning targets and standards of performance (Conley, 2014)

Early Assessments based on descriptions of levels of performance can be used to provide feedback and give insights into future teaching and learning needs. For such a system to be effective, it is important that teaching programmers be aligned with the expected outcomes which are clearly described in curriculum. In any scale of performance descriptions, it is necessary to define success at a given level. Ridgway -cited in Green (2002, P. 3) commented: "...the definition of mastering is not always clear. Therefore, it would be impractical to expect candidates to attain perfect scores on every aspect of every task on which they were tested...So [we] are faced with the task of making decisions about the level of success which counts for mastering".

One of the principal uses of criterion-referenced measurement is in the assignment of students to mastering states. Typically, this involves the selection of a cut-off score on the criterion-reference test. Students with true scores exceeding this cut- off score are considered mastering; they are deemed to have met the learning objectives and may proceed with the next unit or task. Students below this cut- off score are the non-mastering; usually, they are provided with extra learning time or remedial teaching.

It is important to know that criterion-referenced tests have the ability to determine what students can or cannot do, and not how they compare to others. The performance indicators are inextricably linked with some concrete content area and the respective content standards. For setting the performance standards, therefore, it is not enough specify a grade level and give them the corresponding labels (e.g. A – excellent, B – very good, C – good, D –below good, and F – fail) (Kaftandjieva, 2010). A better solution, therefore, is to introduce a separate cut- off score on the test and to use this for assigning examinees to mastering states (Linden, 1982). The primary purpose of criterion-referenced interpretations is not to determine the rank ordering of examinees, as is the case with norm-referenced interpretations, but rather to determine the placement of examinees in a set of ordered performance standards (García, Abad, Olea and Aguado, 2013).

On the other hand, validity is acknowledged as the touchstone of psychological and educational measurement. Within the context of criterion-referenced tests, validity has not been the focus of attention as reliability. This apparent imbalance may be partially attributed to the fact that criterion-referenced measurement posed some new problems and ways of formulating the issues in those areas that have attracted the most attention. The relative lack of attention to questions of validity may also be attributed to perceptions about what validation of criterion- referenced measures entails and about the inherent strengths of such measures.

One of the important contributions of the criterion-referenced testing movement has been an increased emphasis on content. The absolute interpretations of the measures are dependent upon clear specifications of the content domain and on the degree to which the measure is representative of the domain. These are, of course, the key components of content validity-ones that have often been espoused in other contexts but seldom taken as seriously as they are by proponents of criterion-referenced measurement. Thus, the content validity of a criterion-referenced measure may often seem less debatable than that of a test developed using more traditional methods of content specification and item selection.

المنارة للاستشارات
www.manaraa.com

Furthermore, content validity commonly has been held to be the only, or at least the most important, type of validity that is needed for criterion-referenced measures (Linn, 1981).

Livingston (1980) believes that the greatest challenge criterion- referenced testing has posed to psychometric theory relates to reliability. The general concept of reliability "The extent to which a person's test score is consistent over different occasions of testing forms of a test and so forth" is clearly relevant to criterion-referenced tests, whereas the classical definition of reliability as "a correlation or a proportion of variance" is not consistent. Specialized researchers in measurement and evaluation emphasize the important distinction between the reliability of measurements and the reliability of decisions based on those measurements. They also emphasized the distinction between two kinds of agreement: agreement between two parallel-observed scores and between an observed score and the corresponding true score.

This literature review reveals that math tests, specially using criterion-referenced is an efficient way for measuring achievement. It is clear that these aspects relate to cut-off score, validity, and reliability.

Several empirical studies have shed light on this relationship. For example, Oescher, Kirby, & Paradise (1992) explored methods for validating criterion-referenced test results through correlation with an accessible and easily understood norm-referenced benchmark. Christopherson and Humes (1992) examined some psychometric properties of the Test of Basic Auditory Capabilities (TBAC). Two experiments that evaluate the psychometric functions and the test-retest reliability of the tests comprising the TBAC were described.

Shreim and Sawalmeh (2006) conducted a study which compared Angoff and Nedelsky's methods to determine the cut-off score for a criterion-referenced test in mathematics. The test consisted of 30 multiple-choice items, with four alternatives for each. The sample consisted of 80 male and female raters. The raters were distributed randomly into four equal separate groups. The results indicated that the cut-off score ranged from 0.62 to 0.68 using Angoff's method and from 0.49 to 0.57 using Nedelsky's method. The differences between the reliability coefficients of Angoff's and Nedelsky's methods with or without the raters' knowledge of the values of the item difficulty coefficients were not statistically significant at ($\alpha = 0.05$).

Sawalha (2011) identified in her study the most common mathematical errors and its patterns for students with learning disabilities in mathematics in resources room. The sample consists of 140 male and female students: 69 from 4th grade and 71 from 3rd grade. To achieve this goal the researcher prepared and applied a mathematical diagnostic test to the sample, and there were individual interviews. And to answer the questions of this study means, standard deviations and two ways ANCOVA were used. The results showed there are common errors in fundamental concepts and algorithm and facts of addition, subtraction and multiplication. There is a statistically significant difference in common errors towards 3rd grade, and towards male sex, and there is no statistically significant difference in interaction between grade and sex.

Mahmoud and Sabah (2016) develop a valid geometry test for the fifth grade students. The test was developed and validated using Rasch measurement. The study also measured the fifth graders' understanding of the geometric concepts. a 30-item multiple choice test was developed. After that, it was administered on (216) fifth graders. The results provided evidences that supported the validity of the test and measured students' understanding of geometry. The item reliability index was (0.98) and the person reliability index was (0.77). The item difficulty was estimated; the test items covered a wide range of difficulty (-3.29 - 2.69) logits. The results showed that students lacked basic understanding of a variety of geometric concepts even after instruction: parallelogram, the perimeter, and the relationships between geometric shapes. On the other hand, other concepts were easily grasped by students; some of these concepts were regular polygon, the angle and its measurement.

Abu loum (2016) conducted a study to identify mat misconceptions held by fourth graders, the sample consisted of (300) students, 180 male and 120 female in Jordan, the test was applied to it. The study revealed a set of misconceptions held by primary 4th. Graders, including: number ordering, comparing, addition, subtraction, rounding, approximation, comparing different kinds of fractions.

المنـــارة للاستشارات
www.manaraa.com

Statement of the problem:

This study aimed at constructing criterion-referenced tests to measure students' learning outcomes in mathematics for grades 5-10, since criterion referenced tests are very suitable in diagnosing strengths and weaknesses, they are useful in defining mastery level in outcomes, and those students who are defined as mastered or non-mastered. These tests can be used as indicators for nominating students to attend special programs for talented students, which are offered by the MOE or by other institutions. They can also help teachers and curricula experts to design learning activities and teaching methods that help achieving the intended learning outcomes of school math subject, especially for non-mastering students. In particular, the present study aims to answer the following main question:

What are the psychometric properties of the Criterion- Referenced Tests in mathematics for grades (5-10)?

Questions of the study:

The following sub-questions were addressed:

1. What are the cut-off scores of the criterion-referenced tests in mathematics for grades (5-10), which can be used to classify students into mastering and non- mastering?
2. What are the validity indicators of the criterion-referenced tests in mathematics for grades (5-10)?
3. What are the reliability indicators of the criterion-referenced tests in mathematics for grades (5-10)?

Method:

The main purpose of this study was to construct criterion-referenced tests in math, and to investigate its psychometric properties, population, Sampling procedure, instrumentation, data collection, and data analysis used are detailed below.

The population and the sample:

The population consisted of (837471) students from different grades (5th. To 10th.) at the year 2014. The participants were selected as a stratified sample (based on gender and grade). They were 3301 students in fifty-five Schools in Jordan (25 female school, 22 male schools, and 8 coeducation schools). Table (1) shows the distribution of the sample of the study according to gender and grade level.

Table (1): Distribution of the sample according to gender and grade level

Grades	Gender		Total
	male	female	
5th.	439	366	**805**
6th.	298	208	**506**
7th.	252	238	**490**
8th.	250	250	**500**
9th.	250	250	**500**
10th.	250	250	**500**
Total	**1739**	**1562**	**3301**

Instrument and data collection:

Criterion referenced tests in math for grades (5- 10) were constructed through the following steps:

1- **Identifying the mathematical domains and ILO's to be included in the test:**

The general goals of learning mathematics for grades 5- 10 in Jordan were identified based on the curricula guidelines and students' books and teachers' books (teachers' manuals) which are used by the Ministry of Education (MoE). Those general goals and objectives were developed into specific intended learning outcomes (ILO's) that could be more representative of the mathematical domains for those grades levels. The resulting number of those ILO's was (680) outcomes which were used as a frame of reference to construct the criterion referenced tests of math for the target grade

المنارة للاستشارات
www.manaraa.com

levels. These ILO's were distributed to the sixth grade levels as shown in Table 2 below:

Table 2: The distribution of the ILO's to the six grade levels

Grade	Number of ILO's
Fifth	100
Sixth	111
Seventh	082
Eighth	151
Ninth	111
Tenth	125
Total	680

2- **Identifying the characteristics of the mathematical domains and the related ILO's:**

At this stage, a careful study and analysis of the ILO's and domains specified in the previous step was carried out. Therefore, the characteristics of each ILO are stated in such a way that it can be statistically measured. Accordingly, a template including the test item, the alternatives, the correct answer, and the way the question should be answered was also prepared. To ensure the effectiveness of the template, one test item was provided as an example how other items in the template can be stated, in addition to identifying its psychometric characteristics general layout of the test .

3- **Piloting:**

To facilitate the implementation of the tests, each was divided into two equivalent forms (form A and form B). Each form includes the number of items that test each of the targeted mathematical domains for each grade level as shown in Table 3 below:

Table (3): Distribution of the criterion-referenced tests' items for measuring learning outcomes in math for grades 5-10.

grades	Domains tested	Number of items in Form A	Number of items in Form B	Number of items in the two forms	Number of outcomes tested
5th.	Mathematical concepts	9	10	19	100
	Mathematical generalizations	26	26	52	
	Mathematical applications	8	8	16	
	Measurement, geometry, and data analysis	7	6	13	
6th.	Mathematical concepts	8	8	16	111
	Mathematical generalizations	27	28	55	
	Mathematical applications	8	11	19	
	Measurement, geometry, and data analysis	11	10	21	
7th.	Mathematical concepts	9	9	18	82
	Mathematical generalizations	13	13	26	
	Mathematical applications	6	8	14	
	Measurement, geometry, and data analysis	12	12	24	

المنارة للاستشارات
www.manaraa.com

grades	Domains tested	Number of items in Form A	Number of items in Form B	Number of items in the two forms	Number of outcomes tested
8th.	Numbers concepts and operations	22	23	45	151
	Algebra and analytic geometry	24	23	47	
	Plane and space geometry	25	26	51	
	Statistics and probability	4	4	8	
9th.	Algebra and analytic geometry	35	35	70	111
	Plane geometry	15	15	30	
	Statistics and probability	6	5	11	
10th.	Algebra and analytic geometry	49	49	98	125
	Space geometry	5	5	10	
	Statistics and probability	9	8	17	

The two forms of the test were applied to a pilot group which consisted of 1200 male and female students (two hundred students from each of the six grade levels) who were randomly selected from the population of the study. One hundred students from each grade level took Form A of the criterion referenced test and the other hundred took Form B. The same instructions of administering the test to the respondents were clearly stated on each form. The results obtained of difficulty (the percent of the students who answer each item correctly), and discrimination indices (item- total correlation, and item-subdomain total correlation) are calculated for two categories of students: those who are classified above percentile 90, and those who are placed below percentile 30. Table 4 below shows the results.

Table 4: Range of items difficulty and correlations between the items' scores, sub domains' scores, and total scores of the criterion referenced tests of math for grades 5-10

Grades	Item difficulty for			Correlation between the item score and the domain score	Correlation between the item score and the total test score
	Students above P.90	Students below P.30	All participants		
5th.	0.40-0.90	0.04- 0.38	0.22- 0.64	0.07- 0.88	0.10- 0.73
6th.	0.55- 0.92	0.14- 0.49	0.35- 0.71	0.20- 0.81	0.19- 0.73
7th.	0.46- 0.83	0.10- 0.44	0.29- 0.64	0.23- 0.66	0.14- 0.63
8th.	0.42- 0.85	0.15- 0.72	0.25- 0.74	0.32- 0.77	0.21- 0.67
9th.	0.37- 0.87	0.06- 0.77	0.27- 0.67	0.25- 0.82	0.19- 0.72
10th.	0.39- 0.92	0.08- 0.67	0.38- 0.78	0.36- 0.72	0.28- 0.64

Based on these results, some of the test items were modified in terms of language and clarity of some alternatives, and according to the low values of discrimination coefficients. Specially prepared templates for testing ILO's of all mathematical domains for each grade level (5-10) were finally developed.

4- Test construction

At this stage, five equivalent items were written on each ILO, based on the previous prepared template, then they were edited and distributed into five forms for each class grade, the tests were stated in their final form. Thirty test forms were constructed (five for each class grade). Each one of those forms addressed all the ILO's for each of the six grade levels (grades 5-10). Those forms were designed in such a way that they were consistent with the sample test items included in the exemplary template which was tried out during piloting phase. Table (5) shows the number of the items in each form for each grade, and the total number in the five forms:

المنارة للاستشارات
www.manaraa.com

Table (5): The number of the items in each form for each grade, and the total number in the five forms:

Grades	#items in each form	#items in the 5 forms
S 5th.	93	465
6th.	111	555
7th.	82	410
8th.	75	375
9th.	48	240
10th.	88	440

5- **Layout and editing of the test**

At this stage, all the tests' items on each ILO for each of the six grade levels were reviewed and modified. The resulting forms of the tests for the same grade level almost have equivalent structure, content, item difficulty, and item discrimination coefficients.

Administration of the test:

At this stage, the thirty forms' tests were administered to the sample of the study using the same instructions for administering those tests. The indicators of test reliability, test validity, and the cut- off scores of the criterion- referenced tests were finally derived.

The cut- off scores for each grade were described below in the results.

Results:

Results related to the first question: "*What are the cut-off scores of the criterion referenced tests of mathematics for grades (5-10) which can be used to classify students into mastering and non- mastering?"*

To answer this question, Angoff method and contrast groups method were used. Details of these two methods are shown below:

Angoff method:

This method is used to determine the cut-off scores depending on experts' point of view. It specifies the borderline student and then tries to estimate if a borderline candidate is likely to correctly perform on each of the items. For purposes of setting cut- off scores of the math test for each grade level (5-10), a sample of 7 arbitrators majoring in math education were selected (University professors, educational supervisors, teachers, of mathematics, and graduate students). Then, the following steps were used:

- The arbitrators were asked to provide their initial judgments about the tests items to be included in templates for measuring ILO's for each grade level.
- A brief discussion of the probability of a borderline student who is likely to answer each test item correctly. If the arbitrators' estimates probabilities seemed consistent or close to each other (up to 0.01 higher or lower), the next item was discussed. But, if the estimates were too different, arbitrators were asked to justify their points of view.
- At the end of the discussion, estimates were summed for each test and averaged between the arbitrators to determine the cut- off scores for each grade level. Table (6) shows these results.

المنارة للاستشارات
www.manaraa.com

Table (6): The cut- off scores for each grade level (5-10) based on Angoff method

Grades	Maximum score	Cut-off score	Cut- off scores as percentages
5th.	93	42	45.16
6th.	111	44	39.64
7th.	82	52	63.41
8th.	75	35	46.67
9th.	48	25	52.08
10th.	88	49	55.68

Table 6 shows that the cut- off scores that distinguish between performance levels varied from one grade to another. Those estimates ranged between 39.64% for the sixth grade test and 63.41% for the seventh grade test.

Contrast groups' method:

The cut- off scores were calculated for the two groups of students: high- level performers and low- level performers based on the data available in school records and on teachers' opinions. The sample consisted of 480 students. Half of them were classified as mastering of math concepts and skills, and the other half were regarded as non- mastering. The tests were applied to both groups in each grade level. Two frequency curves were drawn of the scores for each test. The point of intersection between curves was considered an estimate of the level of the required performance (cut-off score). Table (7) shows the cut- off scores obtained for each grade using this method.

Table (7): Cut- off scores obtained for each grade (5- 10) using contrast groups' method

Grades	Maximum score	Cut-off score	Cut- off scores as percentages
S 5th.	93	35	37.63
6th.	111	36	32.43
7th.	82	34	41.46
8th.	75	37	49.33
9th.	48	23	47.92
10th.	88	61	69.32

Table (7) shows that the cut- off scores identified according to the method of contrast groups varied from one grade to another. The range was between (32.43%) for the sixth grade and (69.32%) for tenth grade.

Cut- off scores for identifying outstanding students:

The cut- off scores were estimated for the math-referenced tests to identify outstanding students at math. It was decided that the cut- off score should correspond to the 95th. percentile for each grade level. Table (8) shows the cut- off scores obtained by using this method for each grade level.

Table (8): Cut- off scores used to identify outstanding students at math

Grades	Maximum score	Cut-off score	Percentage form of Cut- off score
5th.	93	68	73.12
6th.	111	93	83.78
7th.	82	55	67.07
8th.	75	49	65.33
9th.	48	35	72.92
10th.	88	64	72.73

المنارة للاستشارات
www.manaraa.com

It can be realized from Table (8) that the percentage of cut- off scores that can be used to identify outstanding students ranged between 65.33 for eighth grade and 83.78 for sixth grade.

The second question: *"What are the validity indicators of the criterion- referenced tests of mathematics for grades (5-10) in Jordan*?**"**

To answer this question, content validity, decision validity, and criterion referenced validity/ Concurrent were used. Results are detailed below:

Validity indicators of criterion referenced tests:

In criterion referenced tests, validity indicates the degree to which the test measures the content domain intended to be measured. Thus, validity depends on the interpretation of the obtained scores. One of these types of validity is content validity which refers to the logical process of constructing items, appropriateness of the items of each domain, and the accuracy of content and form as perceived by referees.

Content validity:

Indicators of content validity were obtained through: identifying the general goals and specific objectives of learning mathematics in grades 5-10 described by the Ministry of Education, preparing a preliminary list of learning outcomes that represent each domain in mathematics, and editing and revising all ILO's for each grade level (grades 5-10). Then, a more detailed list of learning outcomes was developed, covering sub- behaviors to be measured by applying the math tests to the sample of the study. Each test item for each grade level was included in specially prepared templates which were approved by panel of experts in math education. Finally clear instructions followed by an example how to answer were proved for the respondents. All these procedures may express a reasonable degree of tests' content validity

Validity of classification decision:

Indicators of classification decision were identified by applying the tests of mathematics to a random sample of (1440) students. Those students were equally distributed to each of the six grades (240 students in each grade). Nearly half of the students in each grade level were classified as mastering and the other half nearly are non- mastering based on percentile 80 of their school marks. After application of the CRT's, students were classified as mastering and non- mastering based on cut-off scores, then cross tabulation was made to classify them according to the two criteria, and decision validity coefficient was computed as the percent of students classified as (mastering in school, and mastering in CRT's) and (non mastering in school grades, non mastering in CRT's) using this formula:

$$\textbf{Decision Validity Coefficient}= \frac{A+D}{A+B+C+D}$$

A= mastering in school grades, and mastering in CRT's.

D = non- mastering in school grades, and non-mastering in CRT's.

B= mastering in school grades, and non-mastering in CRT's (classification error).

C= non-mastering in school grades, and mastering in CRT's (classification error).

. Table (9) shows the classification of the students as mastering and non- mastering according to the two criteria, and validity of classification decision.

Table (9) shows that the decision validity in students' classification into mastering and non- mastering based on cut-off scores of the criterion- referenced math tests for different grades ranged between 0.75 (in the case of math test of sixth grade) and 0.85 (in the case of math test of the seventh grade).

المنارة للاستشارات
www.manaraa.com

Table (9): The classification of the students as mastering and non- mastering based on test performance, school performance, and decision validity coefficient

Grades	School performance	CRT's performance		Decision validity coefficient
		mastering	Non- mastering	
5th.	Mastering	78	42	0.83
	Non-mastering	0	120	
6th..	Mastering	100	40	0.75
	Non-mastering	20	80	
7th.	Mastering	87	33	0.85
	Non-mastering	3	117	
8th.	Mastering	109	30	0.81
	Non-mastering	14	86	
9th.	Mastering	97	43	0.76
	Non-mastering	15	85	
10th.	Mastering	100	30	0.83
	Non-mastering	11	99	

Criterion- referenced validity/ concurrent:

Indicators of criterion- referenced validity/ concurrent were investigated through finding Pearson correlation coefficient between scores on the criterion- referenced math tests and students' scores in math as shown in school records. Table (10) shows validity coefficients achieved by this method for the criterion referenced math tests.

Table (10): Concurrent Validity coefficients of the criterion referenced math tests for grades 5-10

Grades	Concurrent Validity coefficient
5th.	0.64
6th.	0.53
7th.	0.54
8th.	0.55
9th.	0.48
10th.	0.64

Table (10) shows that the validity coefficients based on criterion- reference/ concurrent validity ranged between 0.48 for grade 9 and 0.64 for each of grade 5 and grade 10. Since the criterion for concurrent validity coefficient and the other methods of validity is nearly 0.70 as it is cited in scientific references; It can be realized that the values are generally moderate and they lie between 0.50 and 0.60.

The third question: *"What are the reliability indicators of the criterion referenced tests of mathematics for grades (5-10) in Jordan***?"**

The reliability of the test is a property of consistency results across time. In order to answer this question, reliability of the math criterion referenced tests was established using four methods: internal consistency or item statistics using kuder Richardson-20 equation, consistency of decision using Carver coefficient and Kappa coefficient, and Livingston method. The results are presented below.

Reliability coefficient using internal consistency/ kuder Richardson- 20:

Reliability coefficient was measured using internal consistency of the scores on the tests by using data derived from the main sample, and computing reliability using Kuder Richardson no.20 using the following formula for a test contains (n) items:

المنارة للاستشارات
www.manaraa.com

$$\text{KR20}= \frac{n}{n-1}\left[\frac{1-\sum_{i=1}^{n} p_i q_i}{\sigma^2_x}\right]$$

Where pi is the proportion of correct responses to test item i, qi is the proportion of incorrect responses to test item i (so that pi +qi = 1), and the variance for the denominator is σ^2_x. Table (11) shows reliability coefficients using this method.

Table (11): Reliability coefficients by using Kuder Richardson- 20 for internal consistency of the scores on the tests

Grades	Reliability coefficients
5th.	0.88
6th.	0.84
7th.	0.77
8th.	0.79
9th.	0.79
10th.	0.87

Table (11) shows that reliability coefficients of the math tests using internal consistency/item statistics with Kuder Richardson no.20 ranged between 0.77 (for grade 7) and 0.88 (for grade 5). It can be noticed that the values of the reliability coefficients obtained by using this method were high, which indicates the low sampling error in the content of these tests.

Reliability coefficient using the method of Livingston:

This method depends on the principles and assumptions of classical theory of measurement where the attention is focused on deviation degree of the respondent score from the sample mean. This is true in standard tests, but in criterion-referenced tests the interest is in score deviation of the cut- off score, using Livingston equation (Crocker & Algina, 1986):

$$\hat{K}^2(x,T)=\frac{\sigma^2 x(KR20)+(Mx-niC)^2}{\sigma^2 x+(M-niC)^2}$$

Where:

$K^2(X,T)$: Livingston coefficient

KR(20): Kuder-Richardson reliability

$\sigma_x{}^2$: Variance of the total scores

μ_x: Means of the total scores

n_i: number of items

C: cut-off score

Table (12) shows the reliability coefficients of the tests derived using Livingston equation for grades (5- 10).

Table (12): Reliability indicators using Livingston equation

Grades	mean score	Cut-off score	Livingston Coefficient
5th.	42	35	0.91
6th.	48	36	0.90
7th.	34	34	0.77
8th.	33	35	0.84
9th.	21	23	0.82
10th.	57	61	0.88

المنارة للاستشارات

www.manaraa.com

Table (12) shows that Livingston coefficients are greater than those calculated using Kuder Richardson no.20 coefficient since the values ranged between 0.77 (for grade 7) and 0.91 (for grade 5).

Reliability indicators of consistency using Carver coefficient:

To identify indicators of reliability for the math tests in terms of decision consistency using Carver coefficient, two equivalent forms were applied to a random sample of (100) students from each of the six grade levels. The time between applying the first and the second forms was two weeks. Data were ordered in a 2×2 table. The percentage of students who were classified as mastering or non- mastering according to the two forms of the test for each grade was calculated, and then Carver coefficient was found using the following formula:

$$\textbf{Carver Coefficient=}\ \frac{A+D}{A+B+C+D}$$

A= mastering in both forms.

D = non- mastering in both forms.

B= mastering in the first form, non-mastering in the second.

C= non-mastering in the first form, mastering in the second.

Table (13) shows reliability coefficient derived using Carver coefficient of the math tests for grades (5- 10).

Table (13): Reliability coefficient derived using Carver coefficient of the math tests for grades (5- 10)

Grades	Form 1	Form 2		Carver coefficient
		Mastering	Non- mastering	
5th.	Mastering	58	4	0.90
	Non-mastering	6	32	
6th..	Mastering	53	9	0.88
	Non-mastering	3	35	
7th.	Mastering	52	0	0.85
	Non-mastering	15	33	
8th.	Mastering	56	5	0.88
	Non-mastering	7	32	
9th.	Mastering	48	5	0.85
	Non-mastering	10	37	
10th.	Mastering	67	2	0.93
	Non-mastering	5	26	

Table (13) shows consistency coefficients in terms of Carver for math tests. They ranged between 0.85 (for grades 7 and 9) and 0.93 (for grade 10). Therefore, this is considered a good indicator of consistency for classifying students into mastering and non-mastering of math.

Reliability indicators of consistency using Kappa coefficient:

The reliability of the math tests was also established using Kappa coefficient for each test based on the same data relating to Carver coefficient for the different grades:

$$K = \frac{P_{observed} - P_{chance}}{1 - P_{chance}}$$

$O_{bserved}$= Decision Validity Coeff.

C_{hance}: Errors in classification.

المنارة للاستشارات
www.manaraa.com

Table (14) shows reliability coefficient derived using Kappa coefficient for grades (5- 10).

Table (14): Reliability coefficient derived using Kappa coefficient for grades (5- 10)

Grades	Form 1	Form 2			Kappa coefficient
		Mastering	Non- mastering	Total	
5th.	Mastering	58	4	62	0.79
	Non-mastering	6	32	38	
	Total	64	36	100	
6th..	Mastering	53	9	62	0.76
	Non-mastering	3	35	38	
	Total	56	44	100	
7th.	Mastering	52	0	52	0.70
	Non-mastering	15	33	48	
	Total	67	33	100	
8th.	Mastering	56	5	61	0.74
	Non-mastering	7	32	39	
	Total	63	37	100	
9th.	Mastering	48	5	53	0.70
	Non-mastering	10	37	47	
	Total	58	42	100	
10th.	Mastering	67	2	69	0.83
	Non-mastering	5	26	31	
	Total	72	28	100	

Table (14) shows the consistency coefficients using kappa ranged between 0.70 (for grades 7 and 9) and 0.83 (for grade 10). It can be noticed that Kappa coefficients are generally lower than those of Carver for the same grades. This result is consistent with the foregoing that the Carver coefficient is less sensitive to the consistency of classification decisions than Kappa. However, the overall reliability using Kappa coefficient is appropriate to make decisions based on the test results.

Discussion

The present study was intended to construct criterion referenced tests to measure the intended learning outcomes in mathematics for grades 5-10 in Jordan. To achieve this purpose, six criterion-referenced tests covering the intended learning outcomes for grades (5-10) were formulated. Each of the six tests appeared in five equivalent forms, which results in 30 forms as a whole. Those forms were applied to a sample of 3000 students representing different areas in Jordan.

The cut- off scores were stated using two methods: Angoff Method and Contrast Group Method.

The validity of the test was established using three different methods: content validity, decision validity, and criterion referenced validity/ concurrent. The results of the study showed that the criterion referenced validity coefficients were less than those obtained through the content validity and decision validity. This result was expected since concurrent validity is generally dependant on different variables such as school policy of permissible pass/ fails percentages, and the nature of teacher made tests.

The results showed that the reliability coefficient using Kappa coefficient was less than the reliability coefficient which Carver method revealed. This result is expected since Kappa is more sensitive to chance variables.

It is worth mentioning that the criterion -reference tests which were constructed by the researchers have addressed all the mathematical domains for all grade levels: mathematical concepts, mathematical processes, mathematical applications, geometry, measurement and statistics. This structure of the test is consistent with the structures of other tests such as that of Georgia State Dept, 1983 which included: (1) concept identification, the basic vocabulary of mathematics and the interrelationships of different kinds of numbers; (2) component operations, focusing on addition, subtraction,

المنارة للاستشارات
www.manaraa.com

multiplication, division, and the use of units of measurement; and (3) problem solving.

In general, the results of the study imply that the criterion- referenced tests of mathematics for grades (5- 10) in Jordan which were constructed by the researchers in the study have sufficient statistical indicators for using them to identify mastering and non mastering students. In fact, the obtained cut- off scores are consistent with those suggested by Shreim and Sawalmeh (2006), which were used for similar purposes. Therefore, the results of this study provide satisfactory indicators of validity and reliability of criterion-referenced test that can be used for testing students' achievement and diagnosing their strengths and weaknesses in math. These criterion-referenced tests can also be used to select students for special programs designed for outstanding students.

Based on these results, the following recommendations can be made:

- Training teachers and stakeholders to apply these tests, analyze students' results, interpret them, and provide students with relevant feedback.
- Applying the criterion referenced tests as tools for classifying students to participate in different activities or programs.
- Conducting other studies to develop other tests in math for the secondary stage of education.
- Conducting other studies to investigate the psychometric properties of the test using Item Response Theory models.
- Conducting other studies to investigate the table of norms of the criterion referenced tests for purposes of using them as norm referenced tests.
- Conducting other studies to develop similar tests in other subjects of education, such as: Science, English language, and Arabic Language.

REFERENCES

Abu Loum, K. (2016). An Analytical Study of Math Misconceptions Held by Primary 4th. Graders and Methods of Treatment. *Dirasat*, 43 (3), 2067-2084.

Anastasi, A. (1988). *Psychological Testing*. New York, New York: Macmillan Publishing Company.

Christopherson, L. and Humes, L. (1992). Some Psychometric Properties of the Test of Basic Auditory Capabilities (TBAC). *Journal of Speech and Hearing Research*, Vol. (35), 929-935.

Conley, D. (2014). A new era for educational assessment. *Deeper learning research series.*

Crocker, L. and Algina, J. (1986). *Introduction to Classical and Modern Test Theory*. Holt, Rinehart and Winston, Inc.

García, P., Abad, F., Olea, J. and Aguado, D. (2013). A new IRT-based standard setting method: Application to eCat-Listening. Psicothema, 25 (2), 238-244.

Green, S. (2002). *Criterion referenced assessment as a guide to learning- The importance of progression and reliability*. UCLES, ASEESA South Africa.

Kaftandjieva, F. (2010). *Methods for Setting Cut Scores in Criterion referenced Achievement Tests- A comparative analysis of six recent methods with an application to tests of reading in EFL*. EALTA | European Association for Language Testing and Assessment.

Linden, W. (1982). Criterion-referenced measurement: its main applications, problems and findings. *Evaluation in Education,* Vol. 5: 97-118.

Linn, R. (1981). Issues of Validity for Criterion-Referenced Measures. *Applied psychological measurement*, 4 (4), pp. 547-561.

Livingston, S. (1980). Comments on Criterion-Referenced Testing. Applied *psychological measurement,* 4 (4): 575-581.

Mahmoud, L. & Sabah, M. (2016). Developing an Achievement Test of Geometry for the Fifth Graders Using Rasch Model. *Dirasat*, 38(Appendix), 1353-1367.

Oescher, Jeffrey; Kirby, Peggy C.; Paradise, Louis V. (1992). Validating state-mandated criterion-referenced achievement tests with norm-referenced test results for elementary and secondary students. *Journal of Experimental Education,* 60(2): 141-150.

Popham, W. & Husek, T. (2005). *Implications of criterion-referenced measurement*. Article first published online: 12 SEP 2005, DOI: 10.1111/j.1745-3984.1969.tb00654.x

Sawalha, A. (2011). Common Errors in Mathematics, Its Patterns for Students with Learning Disabilities in Mathematics. *Dirasat*, 38(Appendix), 2344-2365.

المنارة للاستشارات

www.manaraa.com

Shreim, A. & Sawalmeh, Y. (2006)A Comparative Study of Angoff's and Nedelsky's Models to Determine the Cut-off Score for a Criterion-Referenced Test in Mathematics. *Jordan Journal of Educational Sciences* (JJES*)*, 2 (1), 1- 10.

Vos, P. (2005). Measuring mathematics achievement: need for quantitative methodology literacy. *African regional congress of the international commission on mathematical instruction* (ICMI) University of the Witwatersrand, Education

بناء اختبارات محكية المرجع في الرياضيات للصفوف من الخامس إلى العاشر الأساسي واستقصاء خصائصها السيكومترية

"محمد وليد" موسى البطش، اخليف يوسف الطراونة، فريال محمد أبو عواد، حمزة علي العمري*

ملخص

هدفت هذه الدراسة إلى بناء اختبارات محكية المرجع في الرياضيات والتحقق في خصائصها السيكومترية وفق النظرية الكلاسيكية في القياس، إذ جرى تحليل محتوى منهاج الرياضيات ونتاجات التعلم للصفوف من الخامس الأساسي إلى العاشر الأساسي، وكتابة الفقرات الاختبارية التي تقيس هذه المجالات ونتاجات التعلم وفق قوالب معدة لذلك، وتم تدقيقها وتحريرها، وإعدادالتعليمات اللازمة لإدارة الاختبارات في كتيبات الاختبارات. تكون مجتمع الدراسة من (837471) طالبا وطالبة، وطبقت الأدوات الاختبارية على عينة مكونة من 3301 طالبا وطالبة في الصفوف الخمسة.

تم تحديد درجات القطع المستخدمة للتمييز بين الطلبة المتقنين وغير المتقنين باستخدام طريقتي أنجوف والمجموعات المتضادة. كما تم التحقق من صدق الاختبارات باستخدام صدق المحتوى، وصدق القرار، والصدق بدلالة محك/ التلازمي. تم التحقق من ثبات الاختبارات باستخدام معادلة كرونباخ ألفا للاتساق الداخلي بدلالة إحصائيات الفقرة، طريقة ليفينجستون، ومعامل كارفر، ومعامل كابا. وأظهرت النتائج مؤشرات كافية للخصائص السيكومترية للفقرات والاختبارات.

الكلمات الدالـة: الاختبار محكي المرجع، نتاجات التعلم، درجات القطع، الصدق، الثبات، المتقنون، وغير المتقنين.

* قسم القياس والإحصاء، الجامعة الأردنية، قسم الإدارة التربوية، الجامعة الأردنية؛ وقسم علم النفس التربوي، الجامعة الأردنية؛ قسم المناهج والتدريس، الجامعة الأردنية. تاريخ استلام البحث 2017/5/16، وتاريخ قبوله 2018/5/13.

المنـــارة للاستشارات

www.manaraa.com

Copyright of Dirasat: Educational Sciences is the property of University of Jordan and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

المنـــارة للاستشارات

www.manaraa.com

Copyright of Dirasat: Educational Sciences is the property of University of Jordan and its content may not be copied or emailed to multiple sites or posted to a listserv without the copyright holder's express written permission. However, users may print, download, or email articles for individual use.

المنـــارة للاستشارات

www.manaraa.com