Using Item Response Models to Develop a Criterion-Referenced Test to Measure the Students' Achievement in Educational Evaluation

Numan S. Al-Musawi*
University of Bahrain, Kingdom of Bahrain

Received: 2/7/2016 Accepted: 3/8/2016

Abstract: The purpose of the study was to develop a criterion-referenced test to measure student's achievement in educational evaluation using item response theory. To achieve this goal, the author constructed a 3-option multiple-choice achievement test of 48 items that was later administered to 348 students enrolled at the University of Bahrain. The findings of study revealed that the students' responses to 31 items fit the Rasch model assumptions while 17 items did not fit the model. All items of the final version of the test, however, were located within the range of the model's infit and outfit indicators. Also, the reliability estimates for persons and items were .87 and .93, respectively, indicating a high reliability of the test, and the maximum information extracted from the three-option test is obtained at the average ability levels. Based on these results, the author recommends using the developed test as a reliable measure of the level of university student's achievement in the subject of educational evaluation.

Keywords: Criterion-referenced test, item response theory, Rasch model of measurement, multiple-choice test, educational evaluation, university students.

توظيف نظرية الاستجابة للمفردة في بناء اختبار محكي المرجع لقياس تحصيل الطلبة في مادة التقويم التربوي

نعمان صالح الموسوي*
جامعة البحرين، مملكة البحرين

مستخلص: تستهدف الدراسة الحالية تطوير اختبار محكي المرجع لقياس تحصيل طلبة الجامعة في مقرر التقويم التربوي، وذلك باستخدام نماذج الاستجابة للمفردة. ولتحقيق هذا الهدف، قام الباحث ببناء اختبار تحصيلي من نوع الاختيار من متعدد يتألف من ٤٨ سؤالا، بواقع ثلاثة خيارات لكل سؤال، وتم تطبيقه على ٣٤٨ طالبا يدرسون في جامعة البحرين. وأشارت النتائج إلى مطابقة استجابات أفراد العينة عن ٣١ فقرة لافتراضات نموذج راش، وعدم مطابقة ١٧ فقرة للنموذج المذكور. غير أن جميع فقرات الاختبار بصورته النهائية جاءت ضمن حدود المطابقة بالنسبة لمؤشري متوسطات المربعات الداخلية والخارجية. كما بلغ معامل الثبات للأفراد ٠,٨٧، ومعامل الثبات للاختبار ٠,٩٣، وهذه القيم تشير إلى مستوى مرتفع من الثبات. كما أوضحت النتائج أن الاختبار يقدم أكبر كمية من المعلومات عند مستويات القدرة المتوسطة. وفي ضوء هذه النتائج، يوصي الباحث باستخدام الاختبار المطور في الدراسة الحالية كأداة موثوقة لقياس تحصيل طلبة الجامعة في التقويم التربوي.

كلمات مفتاحية: الاختبار محكي المرجع، نظرية الاستجابة للمفردة، نموذج راش للقياس، اختبار من نوع الاختيار من متعدد، التقويم التربوي، طلبة الجامعة.

*nalmosawi@hotmail.com

The criterion-referenced test (CRT) is a test in which the examinee's score is compared to a certain point of reference called "the criterion" or "the cut score". The scores from the criterion-references tests are reported by indicating how many items were answered correctly. The interpretation is based on "what percentage of items correct indicates a satisfactory level of performance". How students' performances compare to others is not emphasized. Student tests will report "whether mastery of the desired objectives has been demonstrated". If pass-fail information is provided, the examiner should figure out "what has been done to determine that a certain percentage correct of test items is reasonable". Difficult items, for instance, can make a relatively low percentage correct indicate a high level of performance (McMillan, 2007, p. 101).

Accordingly, the construction of criterion-referenced tests requires a high degree of consistency between the behavioral objective and the item measuring it. In addition, the proportion of item sample size to the possible total number of items that cover the tested material in criterion-referenced tests is larger than in norm-referenced tests, the most common types of standardized tests, where national samples of students are used as the norming group for interpreting relative standing of a student against his or her school peers. These tests tend to provide broad coverage of each content area of the national or local curriculum to maximize potential usefulness in as many schools as possible (Shrock, & Coscarelli, 2008).

As far as the interpretation and use of the criterion-referenced test's scores is concerned, the classical test theory (CTT) statistics provide sufficient summary information about the functioning of a psychological or educational test and have historically been viewed as the "gold standard" for summarizing the technical adequacy of a test's scores (Embretson & Reise, 2000). However, item response theory (IRT) offers several advantages over CTT (De Ayala, 2009). One major advantage of IRT involves the interpretation of scores. From a CTT perspective, when analyzing an instrument measuring ability, the item and person statistics are not sample dependent. CTT analysis typically use normative comparisons tied to the performance of the sample (e.g., mean-deviation scores) whereas the parameter estimates derived from item response models, are less affected by the characteristics of the sample, because these models demonstrate a specific objectivity, which is established when the relative comparisons between persons are consistent regardless of the items used to measure them, and the locations of items are consistent regardless of the sample (Borsboom, 2005; Rasch, 1977). This, in turn, creates an optimal correspondence between the raw score and the trait-level estimate, establishing a more robust empirical justification for comparisons among participants.

Another advantage of using item response theory is the ability to evaluate the functioning of particular items. In all forms of item response modeling, it is possible to calculate how well the model fits the data, as well as weak, biased, and redundant items. In contrast, CTT analysis generally pays less attention to the functioning of specific items. A third advantage of item response theory is its conceptualization and measurement of test reliability. In a CTT analysis, usually a single standard error of measurement is calculated and used to evaluate the reliability of measure's scores. Contrary to this, because of some form of maximum likelihood is used to estimate the IRM's parameters the standard error of measurement can change at different levels of estimated value of the latent trait. The change in standard error reflects the amount of information that the test provides about participants at each level of the latent trait. Moreover, information in IRM "can be measured at the item or test level by the test information function, which is the sum of information available across all the test items" (Sussman, Beaujean, Worrell, & Watson, 2012, p. 138).

Over the years, the psychometric properties of the tests and scales developed based on the item response models were an object of interest for numerous studies that examined the underlying assumptions of IRT. Within this context, Jamhawy (2000) compared between classical test theory (CTT) and item response theory (IRT) in the development of a mathematical ability test that was administered to 1061 Jordanian students in the 9th grade. The final version of the test included 39 items. Results showed that the item statistics and examinees' ability of the two procedures were comparable. Thirty three items were selected by CTT procedures, 20 items by Rasch model, 35

by two-parameter, and 38 by three- parameter models. Also, results revealed that the two-parameter model was the most comparable model with CTT.

Galli, Chiesi, and Primi (2008) conducted a study to develop a scale to measure the mathematical ability that students need to study introductory statistics courses in their degree program. The Rasch model was applied to construct the instrument. The principal component analysis of the residual showed a one-dimensional construct and the fit statistics revealed a good fit of each item to the model. The item difficulty measures were examined and the area of ability accurately assessed by the items was identified. The validity of the scale was assessed: the measures obtained by the scale correlated with attitude toward statistics and statistics anxiety (concurrent validity), and a relationship with statistics achievement was found (predictive validity).

The Hamandneh's (2009) study aimed at using the item response theory to construct a criterion-referenced test in mathematics based on the 3-parameter-logistic-model. To achieve this goal, an achievement test in statistics consisting of twenty eight 4-option multiple choice items was designed by the author and was administered to a sample of 411 students enrolled in the first secondary class in Jordan. Results of the study indicated that the IRT assumptions were met in the test data and the students' responses to the 24 items of the test fit the three-parameter model. Also, the item parameters estimations were acceptable within the range of the adopted criteria.

In Bani Yaseen and Al-Barakat's (2012) study, the psychometric properties of a criterion-referenced test in chemistry were assessed according to the modern theory of measurement. A 52-item test in chemistry was developed by the authors and was subsequently administered to 481 high school students in Jordan. Using the fit statistics for persons and items, 39 students and 9 items outfitting the assumptions of the Rasch model were excluded from the data. The results showed that the designed test represents a unidimensional construct, and that the separation person index was 2.89 and the separation item index was 8.86, indicating high reliability of the test.

Adedoyin and Mokobi (2013) conducted a study aiming at the psychometric analysis of 2010 Botswana mathematics Junior Certificate (paper 1) that consisted of 40 multiple choice test items. Ten thousand students who sat for the Junior Certificate math examination in 2010 were selected randomly by the use of SPSS values and the students' responses were analyzed using IRT (3PL) model. The results showed that 23 items fitted the model: 12 items were classified as poor items, 10 items were classified as fairly good test items and one item was considered to be a good test item.

Al-Shumrani (2014) used the item response theory and classical test theory to estimate the psychometric characteristics of the Thinking and Learning Skills Test, which was administered to 402 students in the preliminary year of study at Taif University in Saudi Arabia. The author prepared a 52-item achievement test, which was reduced to 42 items after the pilot study. The findings indicated that the final version of the test consisted of 30 items whose difficulty and item discrimination indices were at average levels, and that the student ability ranged from 2.13-2.84. Overall, the results were compatible with the three-parameter logistic model.

Finally, Bourion-Bédès, et al. (2015) examined the construct validity and reliability of the Life Enjoyment and Satisfaction Questionnaire-Short Form according to both classical test and item response theories. The psychometric properties of the French version of this instrument were investigated on a sample of 124 outpatients with a substance independence diagnosis. Findings showed that the internal consistency and the test-retest reliabilities ranged from.80-.90, respectively, and all items correlated significantly with the total score. The confirmatory factor analysis with one factor model demonstrated a good fit to data.

In conclusion, the review of the previous studies shows that some of them (Al-Masry, 2015; Bani Yaseen and Al-Barakat, 2012; Hamadneh, 2008) were directly devoted to the examination of the psychometric properties of criterion-referenced tests using the item response theory. Yet, none of these studies has used the three-option multiple-choice test to guide the results, a fact indicating the importance of using the item response models with this kind of objective tests.

The Research Questions

Within the ongoing comprehensive educational reform in the Kingdom of Bahrain that aims to improve the outcomes of the system of tertiary education, the College of Education at the University of Bahrain seeks to enhance teaching skills of prospective teachers enrolled in the Bachelor Education Program, to prepare them professionally to teach various subjects at the primary and secondary levels of education, enhance their leadership and decision-making skills and reform their deeply rooted beliefs and attitudes toward teaching and learning (Al-Musawi, 2003).

Consistent with this goal, the Teacher Education Programs at the College of Education consist of an integrated sequence of core courses that draw on concepts from philosophy of education, teaching methods and strategies, evaluation techniques, educational technology, learning theories, and psychology. In line with this notion, the educational measurement and evaluation course is designed to provide prospective teachers with a concise presentation of assessment principals and techniques that clearly and specifically relate to standard-based instruction, and with some realistic examples of how to integrate assessment into the instructional process, focusing on assessment concepts and skills that are essential for effective teacher decision making.

Since the learning targets should be adequately and precisely measured by the proper and relevant types of assessment questions on different levels of mastery of the taught content, the preparation of a good achievement test for classroom evaluation seems to be a task of paramount importance, given the fact that assessment influences student's learning and provides the teacher with the correct feedback about students. Within this context, it is essential for the teachers involved in student assessment for learning to obtain a set of criterion-referenced tests designed to assess students' knowledge and skills related to the field of educational measurement and evaluation. As the previous studies suggest, the item response theory is a promising mathematical and conceptual technique upon which the future generations of tests and scales should be developed and validated with the help of modern educational technology.

Unlike the similar previous studies (Bani Yaseen, & Al-Barakat, 2012), this study attempts to construct an objective, 3-option multiple-choice test to assess the students' knowledge in evaluation, and to use the one-parameter logistic model to maximize the information extracted from this test for the enhancement of student's competency in educational assessment and improvement of his or her learning at university. Hence, this study attempts to answer the following research questions:

1. Are the basic assumptions of the one-parameter logistic model (Rasch model) adequately met in the students' responses to the test developed in this study?
2. Does the criterion-referenced test developed in this study have adequate psychometric properties consistent with the assumptions of the Rasch model?
3. What amount of information does the criterion-referenced test developed in this study provide on different levels of student's ability?

Significance of the Study

Given the importance of educational evaluation, this study uses the *descriptive research methodology* to provide a practical guide for university teachers that detects the level of student's achievement in evaluation, which in turn helps to understand the subject of in-depth as the teachers are responsible for the quality of student learning. This also leads to more effective programs in teacher preparation because the results of the study may benefit in the development of evaluation curricula and textbooks.

Limitations of the Study

The interpretation and dissemination of the results of this study is limited by a sample chosen from students of the College of Education at the University of Bahrain. Furthermore, the generalization of findings of the study depends on the psychometric properties of the achievement test built and administered by the author in this study.

Procedural Definitions

Achievement in Educational Evaluation: is the ability to understand the concepts, principles, standards of evaluation, differentiate between types of tests by the test format, construct and score an achievement test, write the test items using the criteria of good test items, define

validity and reliability of the test, analyze and interpret the results of the test using criterion-referenced and norm-referenced statistical indicators.

***Item Response Models (IRM)*:** is a group of mathematical models used to measure continuous latent variables from categorical indicators (Wilson, 2005). This study uses the Rasch conceptualization of item response modeling, in which the probability of observing a particular response to an item is calculated as a function of the difference between a person's level on the underlying variable that the instrument measures and the location of that item (Bond & Fox, 2007).

Method

Participants

A total sample of 348 undergraduate students (116 males and 232 females) enrolled in the course of educational evaluation in the College of Education at the University of Bahrain participated in this study. The mean age was 22.3 years (*SD*=1.14), and the sample of students was selected based on the availability criteria.

Materials

The Achievement Test. The author took the following steps to develop an objective criterion-referenced test in educational evaluation that consisted of 48 items:

1. *Defining the objectives of the test*: The objective of the achievement test is to determine the extent to which students have mastered skills or knowledge of the essentials of educational evaluation fundamental for further learning of the subject.
2. *Stating learning outcomes to be measured as specific behavioural objectives*: Learning outcomes are the products to which learning experiences and processes in a teaching and learning situation are directed. To measure the learning outcomes in any subject, the general objective is broken down into more specific and measurable behavioural objectives. As such, at the end of the educational evaluation course, the student shall be able to achieve the following main behavioural objectives:
   - Define the basic concepts, types and standards educational evaluation and testing.
   - Select the suitable type of test based on its advantages for class evaluation.
   - Write objective and constructed-response items using criteria for good test items.
   - Build and score a classroom achievement test in a subject matter of interest
   - Define the validity and reliability estimates of the constructed achievement test.
   - Use the educational statistics to analyze the results of the achievement test.
3. *Developing a table of specifications:* To assure the content validity of the constructed achievement test, the author prepared a table of specifications where the desire learning objectives of the course were related to the content being measured.
4. *Writing test items:* The author wrote down 60 multiple-choice items reflecting the basic behavioural objectives (10 items for each objective) using the criteria for good test items (Osterlind, 1998, p. 40). The 3-option format was used because empirical evidence over 80 years of research demonstrated the superiority of 3-option multiple choice items (Rodriguez, 2005). This action was motivated by the need to adequately cover the content domain to certify achievement proficiency by producing meaningful precise scores, which requires many high-quality items. More 3-option items can be administered than 4- or 5-option items per testing time while improving content coverage, without detrimental effects on psychometric quality of test scores.
5. *Defining test validity and reliability:* To assure the content validity of the test, test items were examined by five specialists in educational measurement and evaluation of students. As a result, eight items that were judged by the referees as poor items were deleted, and the resulting 52-item test was administered to a pilot sample of 36 university students to check for the clarity of items and to make sure that no answer of an item depends on the answer

of another item, in line with assumptions of the item response theory. The value of Alpha-Cronbach was estimated at .89, thus indicating a high value of internal consistency reliability of the constructed achievement test.

The item difficulty and item discrimination indices were calculated for the 52-item test and the results are displayed in Table 1. It can be seen from Table 1 that item difficulty values ranged from .17-.95 (M=.56) and the item discrimination values ranged from .11-.84 (M=.48). After deleting four items with high difficulty and low discrimination values, the final test contained 48 items. The value of Alpha-Cronbach was estimated at .92, showing a high level of reliability of the test.

The resulting 48-item test was administered as a final exam to the total sample of 348 students in classroom settings of 30 to 35 students. Students were told to find the answer to the question solely based on their knowledge of the content domain related to that question. They were strongly advised *not to guess* as guessing might lead to a less score than initially anticipated. They were also informed that *the passing score equals 36 out of 48 (75% of the maximum score on the test).*

Data Analysis

In the line with the objectives of study, the SPSS program was used to calculate classical item difficulty and item discrimination and internal consistency reliability indices of the test that was administered to the pilot sample of the study.

The BIGSTEPS program (Linacre & Wright, 1993) was used to analyze the test data with IRT models. Responses to the test items from all students were included in the data analysis because this program automatically deletes the students who obtained a *total score* of (48) or *none* (0), and also the items that *all* students answered correctly or incorrectly. Neither of these two cases, however, applies to this study. Hence, this data analysis covered the fort eight items of the developed test and the total sample of 348 undergraduate students who took that achievement test.

Results

Basic assumptions of the one-parameter logistic model: to answer the *first research question*, i.e., to check whether the assumptions of the Rasch model are met, the achievement test items were scored by giving score 1 for the right answer (correct choice), and 0 for the wrong answer (incorrect choice). No guessing of the correct answer was allowed, so the total score was just equal to the sum of correct answers, thus meeting the requirement of the use of Rasch model (guessing parameter = 0).

Prior to data analysis, persons not fitting the Rasch model must be deleted. To identify infit indicators for persons, the student's ability and the standard measurement error of that ability were estimated. Furthermore, infit indicators and OUTFIT indicators, represented by Z-scores for the standardized fit (outfit) statistics (ZSTD) and the mean square fit (outfit) statistics (MNSQ) were also estimated. Infit indicator is more sensitive to the pattern of responses to items targeted on the student's ability, while outfit is more sensitive to responses to items with difficulty far from student's ability.

Table 1
Item Difficulty and Item Discrimination Indices for the Initial 52-Item Test

No.	D	P	No.	D	P	No.	D	P	No.	D	P
1	0.90	0.21	14	0.73	0.18	27	0.89	0.21	40	0.72	0.46
2	0.32	0.33	15	0.81	0.38	28	0.80	0.38	41	0.31	0.24
3	0.35	0.17	16	0.90	0.36	29	0.79	0.37	42	0.87	0.51
4	0.42	0.35	17	0.62	0.57	30	0.51	0.36	43	0.68	0.44
5	0.36	0.08	18	0.23	0.11	31	0.28	0.12	44	0.80	0.46
6	0.28	0.33	19	0.90	0.36	32	0.52	0.44	45	0.42	0.35
7	0.27	0.12	20	0.86	0.20	33	0.64	0.26	46	0.25	0.31
8	0.39	0.30	21	0.50	0.84	34	0.81	0.46	47	0.50	0.34
9	0.79	0.45	22	0.78	0.37	35	0.61	0.37	48	0.79	0.37
10	0.42	0.35	23	0.42	0.33	36	0.74	0.14	49*	0.95	0.12
11	0.43	0.37	24	0.84	0.23	37	0.30	0.31	50*	0.18	0.06
12	0.21	0.25	25	0.32	0.33	38	0.25	0.16	51*	0.92	0.46
13	0.82	0.40	26	0.42	0.30	39	0.61	0.40	52*	0.17	0.28

Note. D=Item Difficulty; P=Item Discrimination; *=Items excluded from subsequent analysis

Table 2 illustrates the values of infit and outfit indicators for the study sample.

Table 2
Infit and Outfit Indicators for the Responses of the Total Sample

			Infit Statistics		Outfit Statistics	
	Ability	SEM	MNSQ	ZSTD	MNSQ	ZSTD
Mean	-.19	.36	.98	-.10	1.01	-.20
SD	.89	.05	.18	1.20	.28	1.30

Note. SD= Standard Deviation; SEM = Standard Error of Measurement; N=348

It is obvious from Table 2 that the value of the fit MNSQ is close to 1, which is the optimal case expected by the Rasch model. As for the value of the infit ZSTD, it is equal to -.10, and the standard deviation (SD=1.20). Similarly, the value of the outfit MNSQ is close to 1, the value of the outfit ZSTD is equal to -0.20, and the standard deviation equals 1.30, which means that the infit and outfit values are close to (0, 1), respectively, the optimal values that are usually expected by the Rasch model.

Inspection of the responses of all students to the test items, however, revealed that for 17 students, the values of the outfit ZSTD are more than +2 and the values of the outfit MNSQ are more than 1, indicating that the observed responses of these students are far away from the expected responses from them based on their abilities (for example, a student would incorrectly answer an item although its difficulty is below his or her ability level or a student would correctly answer an item though its difficulty is above his or her ability level). This means that these students did not fit the item response model, and hence their responses were excluded from the sample.

Having excluded the students who did not fit the IRT model, responses of 331 students were examined to check whether the items are compatible with the model. To achieve this goal, infit and outfit ZSTD and MNSQ were again calculated for all the test items and the results are displayed in Table 3. It can be seen from Table 3, however, that infit and outfit ZSTD values and also infit and outfit MNSQ values slightly deviate from the values expected by the model. Consequently, it was found that 13 items did not fit the IRT model and thus were excluded. Following this step, the item difficulty and item discrimination indices, infit and outfit ZSTD and MNSQ statistics were calculated for the remaining 37 items. The results showed that some very difficult items (for example, item No. 12 with item difficulty = 2.26 logit) were correctly answered by low ability students. Likewise, some very easy items (for example, item No. 16 with item difficulty = -2.19 logit) were incorrectly answered by some high ability students.

Table 3
Infit and Outfit Indicators after Excluding the Misfit Persons from the Sample

			Infit Statistics		Outfit Statistics	
Item Difficulty		SEM	MNSQ	ZSTD	MNSQ	ZSTD
Mean	0	.15	1.00	-.20	1.01	-.20
SD	.87	.04	.11	1.71	.19	1.40

Note. SD= Standard Deviation; SEM = Standard Error of Measurement; N=331

After deleting the misfit persons and items from the test data, the responses of the remaining 331 students to 35 items that fit the Rasch model were analyzed once more to obtain final person - and item -free estimates. Table 4 illustrates the results of person-free measurement values that ranged from the minimum score (11) and maximum score (31). The mean ability distribution was .10 logit units (SD=.92), ranging from -2.17 for low ability students to +2.54 for high ability students. The low value of the standard error means of student's ability estimates, which is .29, suggests the high precision of the students' locations on the trait continuum.

Likewise, Table 5 displays the results of item-free measurement values that range from -2.64 to +1.24 logit units, with a mean of 0 logit units (SD=.85). Similar to the person-free estimates, the low value of the standard error means of item-free estimates, which is .14, suggests the high precision of the item difficulty estimates on the trait continuum. The consistency of item difficulty calibration indicates that the achievement test measures a wide range of student ability and taps a

Table 4
Person-Free Measurement Estimates, Person Separation and Reliability Indices of the Test

			RMSE		ASD		Person Separation		Reliability Index	
	A	SE	Real	Expected	Real	Expected	Real	Expected	Real	Expected
M	.10	.29	.29	.28	.69	.70	3.54	3.62	.87	.87
SD	.92	.05								
Max	37									
Min	11									

Note. M=Mean; SD= Standard Deviation; A=Ability; SE= Standard Error; RMSE= Root Mean Square Error; ASD= Adjusted Standard Deviation; Max=Maximum Raw Score; Min= Minimum Raw Score, N=331 persons

unidimensional variable, which clearly demonstrates the construct validity of the test.

Psychometric properties of the criterion-referenced test: to answer the *second research question*, the **reliability** of the achievement test is related, in the item response theory, with the estimation of person parameters and item parameters for each ability level. In the context of Rasch model, the concept of reliability refers to the level of precision in estimating the positions of persons and items on the trait continuum being measured. These kinds of reliability are labeled "Person Separation Index" and "Item Separation Index", respectively. Person separation is used to classify people, and if its value is less than 2 (see Table 4), it means that the instrument may not be sensitive enough to distinguish between high and low performers (more items may be needed) and the person sample is not large enough to confirm the item difficulty hierarchy of the instrument. Similarly, an item separation index less than 2 (see Table 5) implies that the test items are not enough to capture the measured trait.

The person separation and item separation indices for the constructed test were 3.54 and 2.97, respectively, indicating that the sizes of both the student sample and the item sample are enough to measure the student achievement in the subject of educational evaluation. Accordingly, the values of test reliabilities related to persons and items were found to be equal to .87 and .93, respectively. These values suggest that the total number of the student sample is enough to distinguish between the items and define the trait continuum measured by them, and that the number of the items is enough to distinguish between different levels of achievement of the student sample.

As for the **validity** of the developed test in this study, the **content validity** was achieved by the previous procedures used to build the achievement test and to ensure the congruence between the behavioral objectives and the items measuring them. The **criterion-related validity** was achieved by calculating the correlation coefficient between the students' scores on the achievement test (the predictor) and their scores in the final exam of the subject of educational evaluation (the criterion). The obtained value of the correlation coefficient between the two variables was .82, an indication of a high level of criterion-related validity. The **construct validity** of the test, as stated above, is characterized by the consistency of item difficulty calibration, which suggests that the test taps a wide range of abilities of students involved in the study.

Table 5
Item-Free Measurement Estimates, Item Separation and Reliability Indices of the Test

			RMSE		ASD		Item Separation		Reliability Index	
	A	SE	Real	Expected	Real	Expected	Real	Expected	Real	Expected
M	0	.14	.17	.18	.77	.77	2.97	3.04	.93	.93
SD	.85	.03								

Note. M=Mean; SD= Standard Deviation; A=Ability; SE= Standard Error; RMSE= Root Mean Square Error; ASD= Adjusted Standard Deviation; n=31 items

Table 6
Distribution of Test Information Function Values by Different Ability Levels

Ability	TIF	Ability	TIF	Ability	TIF	Ability	TIF	Ability	TIF
-2.00	.00	-1.00	1.92	-.13	12.17	.85	14.02	1.72	4.54
-1.85	.25	-.92	2.27	.00	12.94	.91	11.26	1.85	3.18
-1.73	.37	-.84	4.01	.15	13.37	1.00	10.04	1.91	2.65
-1.64	.49	-.76	5.61	.25	13.65	1.12	9.57	2.00	1.04
-1.50	.62	-.66	6.82	.34	13.88	1.27	9.15	2.12	.52
-1.47	.86	-.50	7.04	.43	14.06	1.33	8.77	2.27	.38
-1.32	.98	-.44	7.98	.50	15.00	1.50	7.64	2.38	.29
-1.22	1.33	-.37	8.86	.63	14.81	1.58	6.41	2.44	.11
-1.18	1.67	-.24	10.59	.76	14.55	1.66	5.79	2.00	.00

Note. TIF=Test Information Function

Table 7
Distribution of Items of the Achievement Test by the Content Domain

Domain	Serial Number of the Item							
Main Concepts	1	10	12	17	18	43	46	48
Types of Tests	11	24	25	28	30	31	33	39
Test Construction	2	5	7	8	9	15	19	35
Writing Test Items	21	22	23	26	32	34	37	41
Validity and Reliability	3	13	16	38	40	42	45	47
Statistical Analysis	4	6	14	20	27	29	36	44

The information function of the achievement test: to answer the *third research question*, the test information function was calculated for different levels of student ability and the results are displayed in Table 6, from which it is seen that the maximum information extracted from the three-option test is obtained at the ability level of .50 logit units that corresponds to mean item difficulty = 0 (see the curve in Figure 1). The lower line in Figure (1) shows the standard error of estimating the achievement, which suggests a high precision in the measurement of the targeted trait.

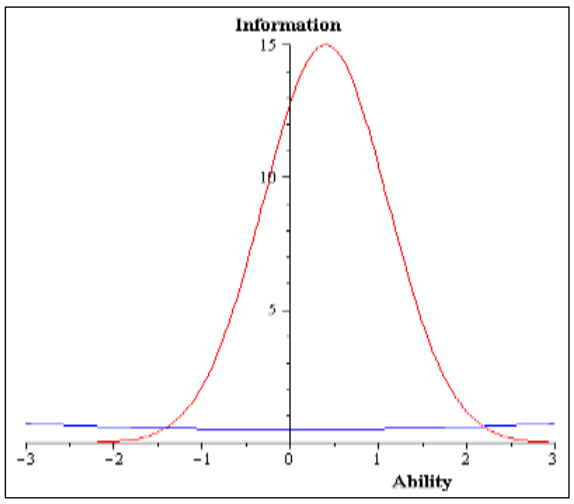

Figure 1: Test information function and standard error of trait estimation

Discussion

Results show that the constructed criterion-referenced test in this study using item response theory models is valid and reliable measure of student achievement in educational evaluation. In general, these results corroborate the findings of similar previous studies with IRT models (Al-Masry, 2015; Al-Shumarani, 2014, Hamadneh, 2009). In terms of dimensionality, the test scores in this study refer to five domains (main concepts of evaluation, types of tests, writing test items, constructing test items, test validity and reliability, and statistical analysis of test results, see Table 7), and each of them contains 8 items. Also, item parameters (item difficulty and item discrimination) remained at acceptable values in the final version of the test.

Furthermore, the developed test in the study has good psychometric properties, establishing a high level of precision in measuring the targeted trait and hence a high level of confidence in the students' scores derived from the test. In item response theory, the concept of reliability is related to the item information function and test information function. Based on this notion, the constructed test provides more effective information about average ability students and less effective information about high and low ability students, in accordance with Rasch model assumptions.

As the chance guessing is one of the problems of multiple choice items (in a 3-option test , it is 33.3%), one would argue that the three-parameter logistic model that take guessing into account would have been more appropriate for this study. In general, this seems to be true. A study on the interpretability of the parameters for the 3PL model (Maris & Bechger, 2009), however, showed that for this model " the parameters are not always identifiable from the distribution of the responses, and two researchers analyzing the same data with either the Rasch model or the 3PL model may end up with equivalent models" (p. 76). This means that there are different ways in which the Rasch model can be represented as a special case of the 3PL model and that different statistical models, with possibly different substantive inferences, may lead to one and the same probability distribution, and, hence, to equivalent results.

Based on the results of this study, the author recommends using the constructed test as a reliable measure of the level of achievement of university students in educational evaluation. Future studies should examine other IRT models with multiple-choice tests containing different number of options in various subjects.

References

Adedoyin, O., & Mokobi, T. (2013). Using IRT psychometric analysis in examining the quality certificate mathematics multiple choice examination test items. *International Journal of Asian Social Science, 3*, 992-1011.

Al-Masry, M. A. (2015). The psychometric properties of a criterion-referenced test in educational research developed according item response theory. *Journal of Basic College of Education (University of Al-Mustansiriya, Iraq), 89*, 701-730.

Al-Musawi, N. (2003). *The effect of student teaching programs in the College of Education at the University of Bahrain on students' beliefs about*

teaching and learning processes. Kingdom of Bahrain: University of Bahrain Press.

Al-Shumarani, M. M. (2014). Use of the item response theory and classical test theory to estimate the statistical psychometric characteristics of the Thinking and Learning Skills Test on students in the preliminary year of study at university. *Journal of College of Education (Al-Azhar University, Egypt), 157*, 717-802.

Bani Yaseen, O. S., & Al-Barakat, S. S. (2012). Psychometric properties of a criterion-referenced test and its items that are used to estimate the domain score according to the modern theory of measurement. *Arab Journal of Education (Tunisia), 32,* 144-167.

Bond, T. G., & Fox, C. M. (2007). *Applying the Rasch model: Fundamental measurement in human sciences*. Mahwah, NJ: Lawrence Erlbaum.

Borsboom, D. (2005). *Measuring the mind: Conceptual issues in contemporary psychometrics.* Cambridge, England: Cambridge University Press.

Bourion-Bédès, S., Schwan, R., Epstein, J., Laprevote, V., Bédès, A., Bonnet, J., & Baumann, C. (2015). Combination of classical test theory (CTT) and item response theory (IRT) analysis to study the psychometric properties of the French version of the Quality of Life Enjoyment and Satisfaction Questionnaire – Short Form (Q-LES-Q-SF). *Quality of Life Research, 24,* 287-293.

De Ayala, R. J. (2009). *The theory and practice of item response theory*. New York, NY: Guilford Press.

Embretson, S. E., & Reise,S. P. (2000). *Item response theory for psychologists*. Mahwah, NJ: Erlbaum.

Galli, S., Chiesi, F., & Primi, C. (2008). The construction of a scale to measure the mathematical ability in psychology students: An application of the Rasch model. *Testing Psychometrics Methodology in Applied Psychology, 15*, 3-18.

Hamadneh, I. M. (2009). Using item response theory in constructing a criterion-referenced test in math according to the 3-parameter logistic model. *Journal of Educational and Psychological Sciences (Bahrain), 10,* 215-238.

Jamhawy, E. (2000). *Comparing item characteristics between classical test theory and item response theory in mathematical ability test.* Unpublished Master Dissertation, Al-Yarmouk University, Jordan.

Linacre, J. M., & Wright, B. D. (1993). *Auser's guide to BIGSTEPS (Computer Program)*. IL, Chicago: MESA Press.

Maris, G., & Bechger, T. (2009). On interpreting the model parameters for the three-parameter logistic model. *Measurement, 7,* 75-88.

McMillan, J. H. (2007). *Classroom assessment: Principles and practice for effective standard-based instruction*. New York, NY: Pearson.

Osterlind, S. J. (1998). *Constructing test items: Multiple-choice, constructed-response, performance and other formats. Boston*, MA: Kluwer.

Rasch, G. (1977). On specific objectivity: An attempt at formalizing the request for generality and validity of scientific statements. In M. Glegvad (Ed.), *Danish yearbook of philosophy* (pp. 58-94). Copenhagen, Denmark: Munksgaard.

Rodriguez, M. C. (2005). Three options are optimal for multiple-choice items: A meta-analysis of 80 years of research. *Educational Measurement: Issues and Practice, 24,* 3-13.

Shrock, S. A., & Coscarelli, W. C. (2008). *Criterion-referenced test development: Technical and legal guidelines for corporate training*. San Francisco, CA: Pfeiffer.

Sussman, J., Beaujean, A. A., Worrell, F. C., & Watson, S. (2012). An analysis of Cross Racial Identity Scale scores using classical test theory and Rasch item response models. *Measurement and Evaluation in Counseling and Development, 46*, 136-153.

Wilson, M. (2005). *Constructing measures: An item response modeling approach*. Mahwah, NJ: Lawrence Erlbaum.