US-China Education Review A 3 (2011) 307-322
Earlier title: US-China Education Review, ISSN 1548-6613

Test Item Analysis: An Educator Professionalism Approach

Mohd Sahandri Gani Hamzah
University Putra Malaysia,
Serdang Selangor, Malaysia

Saifuddin Kumar Abdullah
Ministry of Higher Education,
Selangor, Malaysia

The evaluation of learning is a systematic process involving testing, measuring and evaluation. In the testing step, a teacher needs to choose the best instrument that can test the minds of students. Testing will produce scores or marks with many variations either in homogeneous or heterogeneous forms that will be used to categorize the scores based on the level of achievement or what the teachers desire to measure the achievements of students. The achievement measurement is measured using four types of scales normally used, i.e., the Ordinal Scale, Nominal Scale, Intermittent Scale and the Ratio Scale. The evaluation will be done based on the categories of the measurement used in which the item is used to test the effectiveness of teaching and learning in achieving its objective. The principals of authenticity, reliability, linearity, practicality, objectivity and easy-evaluation are to facilitate recording results so that the final results will not cause any questionings of reliability. There is also an example of difficulty percentage calculation and discrimination index manually done for the English Form 1 real test paper that was used in Bachok National Secondary School, Kelantan. In this example of evaluation, the analysis was made based on 40 test items on 40 students.

Keywords: test, item, analysis, professionalism, approach

Introduction

The process of evaluation consists of testing and measuring activities, beginning with determining the desired learning objectives and ending with evaluating how far the results obtained can be improved. It is a continuous process and not just by giving attention to certain components in teaching and learning, which means that the evaluation has to happen and be thought of during the beginning of teaching but not only during the end of the process.

Testing, measurement and evaluation are the few important concepts in the world of education. These terms are usually used indiscriminately replacing the true meaning of each term. For example, when some teachers give out test papers, they may say that they are measuring the performance of students or evaluating the performance of students in a classroom without thinking about what the real meaning or the specific meaning of the terms are. Specifically, the three terms have their own meanings and have differences from one another.

The Concept of Testing, Measuring and Evaluating

Evaluation, a concept which is used in many fields, is done because of different specific aims and focuses. By having an evaluation done, it can enable an expectation and also prove whether the determined target is

Mohd Sahandri Gani Hamzah, Ph. D., associate professor, University Putra Malaysia.
Saifuddin Kumar Abdullah, Department of Polytechnic Education, Ministry of Higher Education.

achieved or otherwise. An evaluation can also be seen as a way to retrieve correct proofs about an implemented process. The information and proofs are very important in solving any problems happening later.

Testing

Testing is the official procedure to measure and evaluate the students' performance (Mohamad Sahari Nordin, 2002). Testing can also be defined as the process of submitting a standard set of questions which are required to be answered or a set of instrument or systematic procedure to measure the sample of change or a student's or individual's behaviors. According to Murphy (2003), test is a method to obtain the sample of behaviors that can be seen on a student in a controlled or determined situation. The information retrieved from the test can become a basis in making evaluation or referee.

Testing is a systematic procedure to observe the behavior of an individual and elaborate that behavior with the help of a numbered scale, or a categorized scale (Cronbach, 1982). Examples to explain the numbered scale are the number 20/100 for sight test 100 for IQ (Intelligent Quotient) test and 60/100 for an achievement in a subject like life skills. Besides these, examples of a categorized system are like being "extrovert" or "introvert" for the personality test or the color blind test.

The process of testing usually always begins with the preparation stage, followed by the implementation (test administration) and ends with the answer script inspection. Through this testing process, a teacher can understand whether or not his/her students have mastered the skills learnt.

Based on the response by the students, a measurement known as mark will be given to the individual. Through this testing, it will be easier to explain how good the students are in the achievements.

Testing criteria. ***Reliability.*** The general definition of the principle of reliability states that reliability means how far the measurement apparatus can produce consistent readings by Linn and Gronlund (2000). A test will be reliable, if it can measure something with consistency. For example, a student will obtain the same score in a test, if the student's ability is unchanging or the measured trait will not change although it has been measured time and time again with the same instrument.

Back to the implementation of tests in the classroom, the level of reliability of an achievement test is high, if the test results in consistent scores. For example, a student who obtained the score of 75% in the first test and obtained the same score on the same test one week later, thus, this test has a high level of reliability.

The reliability of the test will have some measurement erratum. The measurement erratum is defined as all forms of errors, weaknesses and mistakes that exist in an apparatus or measurement process. Generally, measurements in the physical science field possess few significant measurement errata and produce consistent results. For example, by using the ruler, we can obtain a consistent measurement on the length of a pencil. For the concepts of psychology and other human sciences, they are more abstract.

The process and items used to measure concepts are likely exposed to many measurement errata. The concepts in classrooms like achievement, motivation, interest and students' perceptions are hard to be consistently measured except by using the Likert scale as a measure. Teachers are also not provided the measurement tools like a ruler to help in measuring concepts without mistakes. Teachers are forced to prepare the measurement tools themselves in the form of tests, tasks and observation. Thus, these tests are also seen as "rubber rulers" as stated by Nunnaly (1978). The more "elastic" the ruler is (to illustrate the high measurement erratum), the harder it is to measure the real score of students' achievements.

The measurement erratum can be decreased if the tests are being planned, administered, checked and construed in order. It means that the reliability of a test in a classroom is high, if the test is planned, prepared and implemented carefully to the point that the probability of a measurement erratum of happening is very minimal.

Validity. Nimmer (1984) stated that the level of validity of a test depends on the compatibility of the process. The validity of test in the context of testing the performance of students refers to the compatibility of the aim of test with the contents of test. The degree or level of validity will be improved if the content questions (questions used) can measure and test the achievement of students.

However, if a teacher uses the measurement tools (test) in an incorrect situation, thus, the mistake is not on the instruments used but on the methodology of use of the tool. Thus, the score of a test is considered valid when the test functions towards its aims of being built. Mohamad Sahari Nordin (2002) stated that a valid test can cause information or a retrieved score to help teachers and students to make assessments, inference and figurative about the quality of achievements.

The degree or level of validity of the test is low, if the questions that are asked only touch on a small part of a title. The validity of the test will be even worse if there is an item that questions a concept or a skill that is not discussed in the stated topic. It is clear that the validity of an implemented test refers to how far the contents of test encompass holistically on the knowledge and skills that have been taught, relatively following the importance of each topic.

As a conclusion, there is a significant relationship between reliability and validity. The validity will ensure the created test fulfill its aims or objectives while the reliability ensures the consistency in measurements of students’ achievements in a test. Sometimes, a test will show the consistency but not its earlier objectives. In other words, the test can be reliable but not valid. Thus, a test may not be valid even though it can be reliable. A test must be reliable before it is valid. An accepted test must be valid and reliable.

Measurement

Measurement is defined as the process of gaining explanation numerically on how many certain criteria an individual have. A measurement can also be termed as a process to obtain a numerical data on the level of achievement of students in a skill.

According to Mohamad Sahari Nordin (2002), in the process of education, a measurement focuses more on the quantitative measurement procedures on the levels of achievements of positions of students. Normally, a teacher will use tests to measure the achievements of the students. With the quantitative value, this measure measurement is frequently compared with the maximum value that was allocated for the tool and it is the comparison that illustrates the achievements among students.

The collection of empirical data is very much dependant on the measurement. Measurement is a process of giving numbers to the existence of an object, concept and idea for identification. A valid and reliable measurement is based on an operational definition of an object, concept and idea.

The more detailed the measurement is, the more authoritative and reliable the empirical proof collected becomes. As a result, the expansion in the fields of measurement will be concentrated towards the most detailed measurements.

According to Lee Shok Mee (1993), the measurement is a tool to determine the level of achievement and the positions of students in learning. It is usually done through tests or exams to collect proofs or data in numbers.

It is true that measurement is the determinant of the level of achievement and position of a student in an aspect that is assessed. The achievement is usually compared with the maximum mark allocated and the performance among students. This comparison will illustrate the level of a student in terms of normal performance or achievement in the criterion determined.

The measurement of a criterion can be implemented without the use of a test. The measurement of the process of observation or the use of the leveling scale is examples, in which information can be retrieved in the quantitative form (numeric) without testing. A measurement can also refer to the amount of scores or marks obtained in the process used.

The level of measurement. The expansion in the fields of measurement is concentrated on the making of the most detailed measurement. It results in four levels of measurement, frequently used to measure object, concepts and ideas of research that is from the lowest nominal measurement followed by the ordinal measure, will be continued in gaps to the highest measurement that is the ratio measurement.

Nominal measurement level. The nominal scale is the most basic measurement for the existence of an object, concept and idea of the attitude, and variables of measurement involves classification of the variables into different categories.

A nominal measurement possesses similarities and differences. The male group has same traits as the female group, but it is different from the female group. Thus, the grouping of the variables, such as gender, place of living and race of students, is done by categorizing and analyzing with frequency.

The ordinal measurement level. The ordinal measurement is a measurement that involves arranging into categories. One category is not just different from the others like the nominal measurement; it can also be arranged in other categories according to a required arrangement.

For example, the agreement towards the statement “I really like to learn the science subject” can be categorized in the arrangements, “do not agree”, “partially agree”, “agree” and “very much agree”.

Thus, the collection and analysis of data has to take into account the arrangement traits into categories. The ordinal variables, such as the level of agreement and education need to be collected in an orderly manner using leveling and analyzing with the technique that shows traits of arrangement, such as median or mean.

Interval measurement levels. The interval measurements are measurements, in which the categories are not only arranged, but the distance between the categories can be measured, unlike the ordinal measurement where the researcher does not know the distance between “agree” and “very much agree”.

Since the distance variable has already reached the quantitative level, the data for the variable are usually collected in the form of measurement, and a researcher needs to measure the students’ masteries in performance based on the test marks.

Ratio measurement level. Ratio measurement is the highest level of measurement, in which every category is absolute and meaningful. For instance, a student who is seven years old is actually younger than a student who is 14 years old.

Thus, distance variable data collection needs to be done with measurement. The arithmetic analysis like the mean calculation can be used for the ratio measurement, because all the numbers have absolute and meaningful value.

Evaluation

Evaluation means the activity of determining the value on the measured matter. This activity involves creating the meaning of quality that is the desired traits and on a certain achievement aspect. Anastasi and Urbina (1997) defined evaluation as a process that is considered systematic during the collection and analysis of data to find out whether or not a determined objective has been achieved. It allows the teachers to make correct comparisons or decisions that are related to teaching and learning. It is different from the measurement that involves giving a certain numeric that is quantitative base.

Evaluation is a step in interpreting whether or not the achievement of a student is a “pass” or “fail”, “excellent” or “poor”, “good” or “bad” and “satisfactory” or “unsatisfactory”, and the teachers generally determine the achievement criteria. For that, the achievement criteria used by teachers are the standards used to interpret the levels of achievement, such as “pass” or “fail” and “excellent” or “poor”. Besides, based on experience, experts’ opinions and test contents, a teacher can interpret an achievement that is more than 80% correct answers in the test as an excellent achievement. On the other hand, an achievement that is less than 50% correct answer shows that the students failed in mastering the content of the subject tested.

According to Popham (2002), he explained that a systematic evaluation requires a formal interpretation on the education phenomenon value. Overall, the education evaluation encompasses the aspects, such as learning outcomes, teaching programs and aims towards teaching efforts determined. For example, in a classroom, when a teacher gives a test, is the teacher measuring or evaluating the achievements of the students? Because of this, we can say that the teacher is actually measuring the achievement of the student through the tests. Next, the follow-up action was done by the teachers using the test results with other information related to evaluating the performance of the students.

Interpretation also involves the process of making a decision based on a rule or a standard. It is a part of the learning process. Interpretation is also considered as a learning process that includes elaborating, collecting, recording, scoring and translating information on the learning process of a student for a determined aim in the context of education (Strein, 1997). The interpretation is then stated in the form of scores. A sore must be valid and reliable to ensure fairness to the students and maintain the authority of interpretation institutions. According to Popham (2002), he believed that the term of measurement, testing and interpretation are often used alternately and considered synonymous where these interpreting terms have begun their usage in making evaluations today.

Basically, to achieve an effective teaching and learning is to go through effective testing, measuring and evaluating steps practiced by teachers in classrooms. The efficiency of teachers in developing the instruments of testing and evaluation, and also the teachers’ mastery towards the curriculum being taught are very important and need to be emphasized to ensure the effectiveness of teaching and learning in the classroom.

The purpose of evaluation in the classroom. Evaluation and testing in the teaching and learning sessions are equally important for the process of teaching. The main aim of evaluation is to improve learning. If we see it from a wider aspect, an evaluation can provide huge advantages for the teachers and students and also for the school and the ministry itself when correct and effective information is given.

As stated before this, the outcome of evaluation and testing can be used as an indicator for teachers to make decisions relating to the teaching and learning activity in the classroom. As a teacher teaches, he/she will use the information taken to make necessary decisions to improve the effectiveness of teaching and learning.

To simplify, the purpose or importance of evaluation in education specifically in the classroom is explained for a few purposes like for the purpose of diagnoses, observing achievement, making choices, and placing and giving guidance and counseling to students.

Diagnosis

Testing and evaluation done by teachers in classrooms can provide feedback to the teachers related to the mastery level of the students on a skill that has been touched in the classroom, and also observe the problems that arise in the teaching sessions. With that, a teacher can find out the level of improvement of a student in a classroom on whether the student is in the “very satisfactory”, “moderate”, “poor” or “no improvement whatsoever” category. From the evaluation done, the teachers can also determine active students who need enrichment and also the weaker students who need enrichment activity. The teacher will also make a decision on whether to change the strategy of teaching so that it is more suitable with the students’ needs or repeat same strategies or not.

Evaluation and testing implemented along the teaching and learning processes will show the improvements of students in a certain period of time. The continuous detection of the students’ levels of improvements will allow the teachers to know the problems that students face. Are the students having problems in learning and mastering a certain skill taught in the classroom? If they do face a problem, a teacher will research on the aspects or skills that have caused problems for the students and indentify the problems that happen, this information are crucial for the teachers to provide special training or more attentions towards the students so that the weaknesses or problems faced can be solved. Without evaluation and testing, students who face problems are hard to assist and are being left out from the premier mainstream of the schools in the end.

Achievement

The process of evaluating and testing implemented in the final stage of the teaching and learning can provide information on how much the mastery of a skill have been taught. Besides evaluation done in the end of a teaching and learning session, the evaluation process implemented at the end of a determined studying process, such as at the end of the school session year or for the PSA (Primary School Assessment), Lower LCE (Secondary School Examination) and MCE (Malaysian Certificate Education) students, is aimed to determine the grade that will enable the viewing of the overall level of achievement that the students have mastered. The evaluation at this level is more formal by focusing on giving certificates and awards or any sort of acknowledgement to the students for the achievements that they have shown in examinations.

As a result of this, people can see the achievements of students based on the determined grades whether it is excellent, moderate or fail. A teacher will be able to differentiate students who are excellent, average or poor based on the overall achievements shown in the exams taken.

Selection and Placement

Evaluation and testing is important beginning even before starting teaching and learning. Before a teacher begins his/her lesson in the classroom, it is crucial to have an evaluation and test the students to see how far the students have mastered a skill and also obtain the aspects that will be learnt along the way. For example, when a Malay language teacher teaches a non-native speaker, the teacher needs to know which aspects of skills that have been taught to them in terms of grammar, vocabulary and so forth. This information is important for the teachers to determine the level that is suitable to begin teaching.

Besides, a teacher will also determine whether or not all students have been chosen and grouped into one big group based on different levels of mastery or grouped into similar groups. Due to the difference in mastery level, it is appropriate for the teachers to place the students and focus the teaching based on the level of mastery of each student. This situation shows the purpose of evaluation and testing, that is to select and place students based on the mastery level of a skill.

Besides beginning a lesson, the purpose of evaluation is also to choose and place students based on certain streams. For example, a student who is good at science subjects makes him/her probable to be placed into science streams or based on the students' own tendency.

Guidance and Counseling

We can also use the results from evaluation and testing carried out for guidance and counseling purposes for affected students. Ever students have their motivations to study and it can be intrinsic. So, this motivation comes from the student himself/herself. For example, those who study because of the desire to know about something more deeply will obtain good results, further their studies to higher levels and so forth. Besides intrinsic motivation, there is also extrinsic motivation coming from external influence, such as parental support and encouragement, friends' supports, advice from teachers and so on.

This step of enabling evaluation is to improve motivation for the students to study. The students who have excellent results will continue to maintain their achievements with high spirits. On the other hand, students who are weak can identify their aspects of weaknesses and improve their weaknesses for the upcoming evaluations so that they can achieve better results. Besides, teachers can also guide and teach students who are having problems on the ways to handle their problems.

Other Purposes

Besides the purposes above, evaluation and testing is also important in determining the readiness of students in learning a skill. It is done by teachers before the beginning of the teaching and learning sessions. This type of evaluation is done by teachers through the observation of the physical readiness, such as hands, eyes, fingers and the students' concentrations towards teachers. Through the reaction shown, a teacher can foresee how ready the students are in accepting lessons and vice versa. Beginning the lesson, based on how ready the students are, is very important to ensure students' confidence in following teaching and learning sessions, and thus, allow them to master learnt skills.

Through evaluation and teaching done by teachers, we can also give teachers input on the effectiveness of his/her lessons. Through the response shown by the students, the teachers can evaluate whether or not the lessons are effective. If most of the students cannot master something which is taught, thus, it is clear that the teaching strategy is not effective and needs prompt alterations to better the situation. For the students, this information of evaluation shows the level of mastery towards a lesson. If they are weak in one aspect, they will know about it and give more focus towards this less mastered aspect.

For parents, evaluation and testing done will provide the response towards the development and achievement of their children in the classroom. Parents can take necessary actions, such as sending their children to tuition classes or paying more attention to their children to improve performance and assisting in increasing the mastery level of their children as much as possible.

Evaluation done can also provide clear information for the policy-makers especially the curriculum makers in the name of the MOE (Ministry of Education) of Malaysia on the curriculum being used in schools and the whole nation. The ministry will obtain accurate responses on the achievements of objectives of the syllabus of lessons created whether they have been achieved successfully or otherwise. If responses show that there are still significant weaknesses in achieving the objectives, thus, adjustments need to be made. From this, the teaching and learning process of a subject will be planned in a better, more complete and systematic way.

Generally, the evaluation done in classroom can be seen in three forms, i.e., (1) observation; (2) oral; and (3) writing. These three forms of evaluation are aimed at identifying students, thus, instantaneously identifying the improvements and achievements of the students. In a simpler manner, the table below shows the forms of evaluation.

Level of Cognitive Taxonomy for Test Questions, Discrimination and Difficulty for 40 Questions That Are Given to 40 Students

How do we determine the quality of a set of questions in the form of multiple-choice questions? This question can only be answered if the question makers analyze the items created from the discrimination index value aspect, the percentage of difficulty of the distracters and so on.

The analysis of test and item is explained as a systematic method that uses the statistical technique to evaluate the quality of the test items. The analysis is done to improve the quality of the test technically to determine whether the items are defunct, can be repaired or cannot be used.

According to Secolsky (1983), for the questions which are synthesized by schoolteachers, the analysis of test and item has a few uses that can be simplified as such:

(1) Determining whether the item is functioning as desired by the teachers;

(2) The information from them as a response to the students and the basis to discussion in the classroom;

(3) Response towards teachers on the difficulty faced by students can be known;

(4) The information is complex and discreet;

(5) Identifying the fields which are hard for students so that the improvement of curriculum can be improved further;

(6) The transfer of items can be done;

(7) Improving the skills of writing items.

Mohamad Sahandri Gani Hamzah (1995) suggested the item analysis to be done in three levels, which are:

(1) Determining the degree of difficulty;

(2) Analyzing the item discrimination;

(3) Analyzing the distracters.

Pilot Study

A pilot study was planned and executed in a scientific and systematic manner for the purpose of determining the level of difficulty and the discrimination index where analyzed based on three groups of students and their test performance. Calculation is done based on the basics of principal and procedure, in which only two groups were chosen and analyzed. According to Mohamad Sahari Nordin (2002), the item discrimination index refers to the ability of a question to discriminate two groups of students (i.e., the excellent students and the weak students).

The Background of Sample

The test sample was the English Language Test Set—First Identification of Form One 2007. It was prepared by the MSSPC (Malaysian Secondary Schools Principals Conference) Kelantan Branch. The information on the analysis of the sample is as follow:

(1) Total items: 40 multiple-choice questions (objective);

(2) Twenty students from Form One Cekal of Kota National Secondary School, Jalan Salor, 15100 Kota Bharu, Kelantan;

(3) The date of test: April 1, 2007.

Analysis Procedure

The following procedures were done manually (by hand) because the sample was small (40 students):

(1) The answer sheets were arranged from the highest marks to the lowest;

(2) The answer sheets were divided into three groups (one third for one group), that was the excellent group (higher), average group and the weaker group (lower);

(3) The sample used for this analysis was the 13 answers of the excellent group (higher) and 13 answers from the weaker group (lower);

(4) The students' answer sheets from the higher and lower group were marked as either correct or wrong for each of the test items;

(5) The analysis results on the items were recorded (the table is enclosed);

(6) The difficulty percentage, the item discrimination index and the level of questions were identified based on the raw data.

The Central Tendency Analysis

The central tendency analysis is an analysis that involves the measurement of mean, mode, median, range and so on towards the sample in analyzing the achievements of students. The following is the results of test on the students of Form One Cekal (see Table 1):

After marking all the answer sheets of the students, it was found that 18 out of 40 students have passed the test. In that percentage, only 45% of the whole class succeeded the test. It shows that the students in the classroom have low mastery level in English.

The highest mark in the test is 27 (68%) while the lowest is 20%. The score average mark of the students is 40.62%. The range of score of students of Form One Cekal is 45% and the frequency of the score of 16 marks is the highest mode and showed that the mastery level of the students in the English language is uniform.

Discriminative Index Analysis

Discriminative index analysis of items refers to the potential of an item to discriminate two groups of students (i.e., the excellent students and the weaker students). In this analysis, one third of the total students that obtained the highest marks were classified as the excellent group, whereas one third of the total students that obtained the lower scores were classified as the weaker group.

A question will be considered as having a high discrimination number, if the more intelligent students succeed in answering the questions correctly and the less intelligent students failed to answer the question.

However, if these two groups are able to answer the same questions correctly, the value of discrimination is zero. Also, if the less intelligent students answer a same question more than the more intelligent students, the value of discrimination is negative. These two situations are the ones we wish to avoid for a text item.

Kolstad et al. (1984) drew the guidelines for evaluating an item from the aspect of discrimination index as follows:

Table 1

Students Achievement

	Number	Name of students	Total correct answers out of 40 questions	Marks score percentage (%)
(H) Higher group	1	Aziean Syazana Binti Zakaria	27	68
	2	Siti Najibah Binti Shukri	26	65
	3	Nurul Azunie Binti Zulkifly	25	63
	4	Nur Shahira Binti Mohd Nor	25	63
	5	Nor Shahida Binti Sharuddin	22	55
	6	Nur Atiqah Binti Mohamad	20	50
	7	Mohd Faiz Bin Othman	48	48
	8	Nor Shahiera Binti Razali	48	48
	9	Nur Nadiah Binti Mohd Kamil	18	45
	10	Tuan Mohammad Bin Tuan Hassan	18	45
	11	Wan Mohd Zakir Bin Wan Ismail	17	43
	12	Nur Aishah Binti Mohd Fauzi	17	43
	13	Nor Fitri Binti Haris	17	43
(A) Average group	14	Nurul Dyna Binti Aziz	16	40
	15	Sofea Izati Binti Mohd Noor	16	40
	16	Wan Nurfadhilah Binti Wan Daud	16	40
	17	Mohd Firdaus Bin Sabri	16	40
	18	Wan Zahidah Binti Wan Juhan	16	40
	19	Ahmad Firdaus Bin Maat	16	40
	20	Tuan Nor Aizan Binti Tuan Ismail	15	38
	21	Mohd Nazrul Bin Adanan	15	38
	22	Tuan Muhamad Bin Tuan Mansur	15	38
	23	Mohd Syaza Bin Abdul Aziz	15	38
	24	Mohd Afiq Bin Mohd Yusri	14	35
	25	Ahamad Asef Bin Hamdi	14	35
	26	Mohd Razin Bin Ramli	14	35
	27	Mohd Fathihi Bin Jamali	14	35
(L) Lower group	28	Fahizad Rawi Bin Farok	13	33
	29	Syed Amin Bin Syed Omar	13	33
	30	Nor Syazani Binti Mohamad	13	33
	31	Haris Syazwan Bin Nudman	13	33
	32	Nor Syazidah Binti Mahadi	12	30
	33	Abdul Rauf Binti Hussin	12	30
	34	Amira Izlima Binti Shukor	12	30
	35	Abdul Hamiki Bin Arif	11	28
	36	Mohd Khairul Bin Otman	11	28
	37	Zetti Atirah Bin Omari	11	28
	38	Siti Amira Binti Zainuddi	10	25
	39	Siti Noor Aishah Binti Nordin	9	23
	40	Nurul Fazrin Binti Che Mat Yusof	8	20

Note: For calculation purposes T (total number of student H + L or 13 + 13 = 26).

Table 2

Interpretation of Discrimination Index of the Test Item

DI (discimination index) value	Item evaluation
0.40 and above	Excellent item
0.30-0.40	Good item
0.20-0.29	Average item
0.10-0.19	Unsatisfactory item, needs to be improved
Below 0.10	Poor item, needs to be omitted or changed

The discrimination index and percentage of difficulty for each item is as shown in Table 3.

Implication of Analysis

Based on Table 3, the highest discrimination index value falls in item nine of the analysis sample. The Discrimination Index of the item, which is 0.77, is very good because it can really discriminate the items amongst the two groups.

Eight items have been identified as having discrimination index values between 0.40 and 1.0, which shows that the items are very good. The forms and patterns of items brought upon by the item makers are clear and not too complex for comprehension.

Another eight items have been identified having discrimination index values of between 0.30 and 0.40, which shows that those items are good. These items should be maintained: 3, 5, 10, 13, 15, 19, 32 and 34.

However, 11 out of the 40 items prepared by the item makers need to be omitted or changed, because eight of them have item discrimination index of 0.08 which is under 0.10 and three items have discrimination index of infinity.

Items which need to be changed and omitted are items 1, 6, 9, 11, 12, 14, 116, 18, 26, 29, 33 and 34, because the items cannot discriminate between two groups of higher scorers and lower scorers. After analyzing, it has been found that there are ten items out of 11 that uses high level of language and is hardly understood by Form One students. Item 1 is seen otherwise to be too simple and does not require the need of high level reading and comprehension to answer, thus, also failed to discriminate the two groups.

Difficulty Percentage Analysis

The difficulty of an item shows the percentage of students who answered the items correctly (calculation is enclosed in the results), according to Mohamad Sahari Nordin (2002).

A good item, which is a question that contributes to the improvement of reliability of the test, is the item that has a moderate degree of difficulty. Teachers are advised to keep and use the items which consist of difficulty levels between 50% and 80%. According to Ee Ah Meng (1993), it is the difficulty percentage and the classification of questions made.

The results of analysis towards 40 items that is prepared by the National Conference of Secondary School Principals of Malaysia (PKPSM) Kelantan Branch showed that 21 items consist of item difficulties between 50% and 74% which means in terms of difficulty the items are good.

However, nine items namely item 1, 2, 3, 4, 6, 21, 30, 31 and 38 are items that are identified as easy items. It is possible that items 1 to 4 have been designed to be easy so that they can motivate the students to answer the other questions that follow.

Table 3

Discrimination Index and Percentage of Dififculty (H = 13 & L = 13; Σ(H+L) = 26)

Item/ Question number	Answer	Total correct of the higher group (Htc)	Total correct of the lower group (Ltc)	Total correct (Htc + Ltc)	Difference (Htc-Ltc)	Discrimination index (DI) (Htc-Ltc)/H or L	Taxonomy level	Difficulty percentage $\frac{\sum(H+L)-\sum(Htc+Ltc)}{\sum(H+L)}\times 100\%$
2	B	9	6	15	3	0.23	A	42.31
3	A	10	5	15	5	0.38	A	42.31
4	D	12	4	16	8	0.62	B	38.46
5	B	6	1	7	5	0.38	B	73.08
6	A	12	11	23	1	0.08	A	11.54
7	A	10	7	17	3	0.23	C	34.62
8	C	8	1	9	7	0.54	B	65.38
9	A	11	1	12	10	0.77	A	53.85
10	B	0	4	4	4	0.31	D	84.62
11	C	3	2	5	1	0.08	D	80.77
12	A	1	0	1	1	0.08	E	96.15
13	B	6	2	8	4	0.31	B	69.23
14	C	2	1	3	1	0.08	B	88.46
15	A	7	5	12	2	0.15	C	53.85
16	C	5	5	10	0	0.00	A	61.54
17	C	8	5	13	3	0.23	A	50.00
18	A	3	3	6	0	0.00	B	76.92
19	B	1	5	6	4	0.31	B	76.92
20	C	7	5	12	2	0.15	C	53.85
21	D	12	6	18	6	0.46	C	30.77
22	A	5	1	6	4	0.31	B	76.92
23	D	4	2	6	2	0.15	A	76.92
24	B	6	3	9	3	0.23	A	65.38
25	D	6	1	7	5	0.38	C	73.08
26	A	3	3	6	0	0.00	B	76.92
27	C	6	3	9	3	0.23	C	65.38
28	B	10	2	12	8	0.62	C	53.85
29	D	6	5	11	1	0.08	A	57.69
30	C	11	5	16	6	0.46	B	38.46
31	B	10	7	17	3	0.23	A	34.62
32	D	6	1	7	5	0.38	C	73.08
33	B	7	6	13	1	0.08	A	50.00
34	B	8	3	11	5	0.38	A	57.69
35	B	6	5	11	1	0.08	B	57.69
36	A	5	3	8	2	0.15	B	69.23
37	C	10	2	12	8	0.62	C	53.85
38	D	12	5	17	7	0.54	C	34.62
39	D	9	3	12	6	0.46	D	53.85
40	C	5	3	8	2	0.15	D	69.23

Notes. Level of thought process based on Bloom's Taxonomy Symbol, A: Literal, B: Comprehension, C: Application, D: Analysis E: Synthesis, F: Evaluation.

Table 4

Classification of Difficulty Items

Summary of questions	Difficulty percentage(%)
Too difficult and cannot be used	85% and above
Difficult but acceptable	75% to 84%
Good question	50% to 74%
Easy question	25% to 49%
Too easy	0% to 25%
Question does not test what is intended to be tested	Ltc > Htc

There is only one item that can be classified as an item which is too easy with the difficulty percentage of only 11.54%. This item is item 6, in which 23 students from the higher score group and the lower score group succeeded in answering the question correctly.

From the findings, it can be concluded that the questions that the item designers prepared are categorized in the lowest level in testing cognitive abilities of students that is the literal level of thought processes.

Item 12 is identified as the item which is the most difficult to answer by Form 1 Cekal students with the difficulty percentage of 96.15%. The high difficulty percentage of item 12 puts it in the "too difficult" category based on the difficulty percentage table. Item 12 should be omitted or rejected from the set of test because it will be an item which is only suitable for students of higher cognitive levels like Form 2 and 3 students.

Analysis of Distracters in Multiple Choice Items

Objective questions by multiple choice items are very popular and used frequently. One problem or difficulty faced by the item designers is choosing choices besides the correct answers. These choices are used to distract the attention of test candidates that do not know the correct answers, thus, they are characterized as distracters or disturbers of the items.

A distracter is said to be effective, if the candidate, who does not know the answer, chooses the distracter as the answer.

This is the analysis of distracters for the first ten items (see Table 5).

Ten out of 40 questions have been chosen to be analyzed for the distracter level in the items provided. The items involved in this analysis are the ten earliest items in the test sets.

Based on the analysis, item 5 is an item with the highest level of difficulty which is 1.0. It is because all four of the questions have equal chances to be chosen. The skills of the designers in making difficult questions are significant in that item as the students were confused when answering that question.

After the analysis, item 5 is seen as the item that is categorized in the Bloom's taxonomy as application level and 25% of students failed to answer with the choice A because they have failed to differentiate between what is asked in the question.

Item 6 is the easiest question for students to answer as 85% of students answered the question with the correct answer. The literal taxonomy level on the item is the factor that influenced the tendency of students and choosing the answer.

Only 15% of students chose the incorrect answers for item 6. It shows that the question possesses a very low level of distracters and the item is too easy for the students to answer.

Overall, the items prepared in the set of test possess good levels of distracters because only a few items have difficult levels of distracters. These items show that the item designer is skillful and experienced in designing good items.

Table 5

Analysis Distracters of the Item

Items	Correct answer	Group	Frequency of choices			
			A	B	C	D
Item 1	D	Higher group (13)	2	2	1	8
		Average group (14)	5	2	0	7
		Lower group (13)	6	0	0	7
Correct percentage of Item 1 (%)			32.5	10	2.5	55
Item 2	B	Higher group (13)	1	9	3	0
		Average group (14)	2	8	3	1
		Lower group (13)	2	6	3	2
Correct percentage of item 2 (%)			12.5	57.5	22.5	7.5
Item 3	A	Higher group (13)	10	0	0	3
		Average group (14)	8	2	0	4
		Lower group (13)	5	1	3	4
Correct percentage of item 3 (%)			57.5	7.5	7.5	27.5
Item 4	D	Higher group (13)	0	0	1	12
		Average group (14)	1	3	0	10
		Lower group (13)	1	9	0	4
Correct percentage of item 4 (%)			5	30	2.5	65
Item 5	B	Higher group (13)	4	6	3	0
		Average group (14)	0	3	3	8
		Lower group (13)	6	1	4	2
Correct percentage of item 5 (%)			25.0	25.0	25.0	25.0
Item 6	A	Higher group (13)	12	0	1	0
		Average group (14)	11	0	3	0
		Lower group (13)	11	1	1	0
Correct percentage of item 6 (%)			85.0	2.5	12.5	0
Item 7	A	Higher group (13)	10	2	1	0
		Average group (14)	8	5	1	0
		Lower group (13)	7	5	1	0
Correct percentage of item 7 (%)			62.5	30	7.5	0
Item 8	C	Higher group (13)	1	4	8	0
		Average group (14)	1	4	8	1
		Lower group (13)	2	3	1	7
Correct percentage of item 8 (%)			10.0	27.5	42.5	20
Item 9	A	Higher Group (13)	11	2	0	0
		Average Group (14)	9	5	0	0
		Lower Group (13)	1	11	1	0
Correct percentage of item 9 (%)			52.5	70.0	2.5	0
Item 10	B	Higher group (13)	1	0	11	1
		Average group (14)	1	1	10	2
		Lower group (13)	2	4	6	1
Correct percentage of item 10 (%)			10	11.25	67.5	10

Comment Types That Improve the Quality of Learning Assessments

Making relevant and good quality questions requires technical and creative skills. Many decisions related to students are done based on the results of test, and from that, teachers are responsible in preparing tests discreetly and systematically so that the quality of test can be controlled.

The making of questions is easy, if the learning objectives, learning outcomes that intend to be tested are clear and the test determinant table is prepared. The questions made should obey the rules of the test determinant table that is prepared based on the following aspects:

(1) The topics that will be tested and their weighting;

(2) The skills that will be tested and the weighting;

(3) The level of difficulty of questions based on the taxonomy;

(4) Type of question.

In determining the type of question for a test, a teacher has the choice to make objective or subjective questions. Objective questions are questions that require the students to give correct and brief answers so that they can be checked easily and precisely.

Because of that, objective test are said to have the highest objectivity, because all the testers will give similar marks to all the questions that are proposed. The subjective questions consist of limited response (short answers) type questions and extended response (unlimited answer) type questions.

Overall, a good set of tests will have the following criteria:

(1) Validity: the ability of the test to measure what should be measured;

(2) Reliability: the ability of the test to obtain marks that are free from contradictions caused by the difference in consideration, opinion and emotion of the testers;

(3) Practicality: practical quality that enables and simplifies smooth testing;

(4) Interpretability: the ability of the test to change the test result retrieval into useful and meaningful information.

Conclusions

Overall, to evaluate a learning outcome, a teacher measures using tests as a tool to make good considerations. The relationship of the three elements (i.e., measurements and evaluation) in education is very important to interpret the result of the students' learning. The outcome of interpretation will determine the success of the objectives of learning being achieved, the improvement of the performance of students and the students' understanding towards the absorbing ability in learning. In ensuring the interpretation of a test to be of good quality, we need to ensure that the standard value of discrimination index, difficulty percentage level and distracter level are at the best levels. It is extremely important because the tests are tools to process measurements. If it fails to function, it will affect the interpretation process of learning outcomes.

References

Anastasi, A., & Urbina, S. (1997). *Psychology testing* (7th ed.). N. J.: Upper Saddle River.

Bhasah Abu Bakar. (2003). *The basics of classroom measuring*. Quantum Books Publication.

Cronbach, L. J. (1982). Designing evaluation of educational and social programs. In D. L. Stufflebeam, & A. J. Shinkfield (Eds.), *Systematic evaluation: A self instructional guide to theory and cpractice*. USA: Kluwer Nijohoff Pub.

Ee Ah Meng. (1993). *Pedagogy—An integrated approach* (2nd ed). Budiman Group Sdn. Bhd.

Kolstad, R. et al. (1984). The application of item analysis to classroom achievementtests. *European Journal of Education, 105*(1), 70-72.

Lee Shok Mee. (1993). *Testing and evaluation in education*. Kuala Lumpur. Budiman Group Sdn. Shd.

Linn, R. L., & Gronlund, N. (2000). *Measurement and assessment in teaching* (8th ed.). Upper Saddle River, N. J.: Prentice-Hall.

Mohamad Sahari Nordin. (2002). *Testing and interpreting in classrooms*. International Islamic University of Malaysia Publications.

Mohd. Sahandri Gani Hamzah. (1995). A study of relationship between academic achievements with the mastery of of practical work and academic reflexive concepts within students of integrated life skills (Master's Thesis in Education, UKM).

Mok Soon Sang. (2005). *Education knowledge for KPLI (primary schools: component 1& 2)*. Budiman Group Sdn. Bhd.

Murphy, J. (2003). Task-based learning: The interaction between tasks and learners. *ELT Journal, 57*(4), 352-359.

Ng See Ngean. (1992). *Measuring and evaluating in education*. Penerbit Fajar Bakti.

Nimmer, D. N. (1984). Measurement of validity, reliability, and item analysis for classroom tests. *Clearing House, 58*(3), 138-140.

Nunnally, J. C. (1978). The study of change in evaluatioon research. Principles concerning measurement, experimental designs and analysis. In L. E. Struening, & M. Guttentag (Eds.), *Hand book of evaluation research*. Beverly Hills, California Sage.

Popham, W. J. (2002). *Classroom assessment: What teacher need to know* (3rd ed.). Boston: Allyn & Bacon.

Secolsky, C. (1983). Using examinee judgements for detecting invalid items on teacher-made Criterion-referenced tests. *Journal of Educational Measurement, 20*(1), 51-63.

Strein, W. (1997). Grades and grading practices. In G. Bear, K. Minke, & A. Thomas (Eds.), *Children's needs-II*. Bethesda, M. D.: National Association of School Psychologists.

Yap Yee Khiong et al. (1987). *Measuring and testing in education*. Kuala Lumpur. Heinemann Asia.