Quality standards for achievement tests and steps for building a good achievement test

خالد فوحمة [1*] رابحي اسماعيل [2]

ismail rabhi[1*] khaled fouhma

khaled.fouhma@univ-biskra.dz

smail.rabhi@univ-biskra.dz

مخبر المسألة التربوية في الجزائر في ظل التحديات الراهنة كليةّ العلوم الاجتماعية والإنسانية، جامعة محمد خيضر بسكرة .

The Laboratory of the Educational Issue in Algeria in Light of Current Challenges

Faculty of Social and Human Sciences, University of Mohamed Khider Biskra

تاريخ الاستقبال:2024/05/18؛تاريخ القبول:2024/06/03؛تاريخ النشر:2024/09/10

Abstract:
In this paper, the author attempts to define achievement tests, identify their types, and clarify the process and steps of constructing the ones that meets the required conditions, justifying its importance and purpose. Needless to say that the process of constructing a good achievement test requires passing through successive steps, starting with defining its objectives and purpose, all the way to producing its initial form and testing it to conduct statistical analyses on it, and then modifying it if any defects are found. Paying attention to each step in its preparation contributes greatly to reaching an achievement test built on sound scientific foundations and principles, with the required conditions for its construction, and thus obtaining a valid evaluation tool that enables making sound judgments and decisions.

Keywords: Strategy, Deployment, Human Resource, Challenge, Developmental Action.

* المؤلف المرسل .

.

Introduction

Since the emergence of Bloom's Taxonomy of Educational Objectives in 1956, there has been an increasing interest in achievement tests. Alaouna (2005) indicates that "preparing achievement tests is one of the main tasks of the teacher, which he/she must undertake if he/she intends to fulfill his/her role as an effective organizer of student learning" (Al-Satari, 2010, p. 118). A teacher who seeks to identify areas of weakness in order to address them and improve their teaching methods should prepare achievement tests that enable this process. These tests serve as a tool for measurement and evaluation.

Therefore, in this paper, we attempt to define achievement tests, identify their types, and elucidate the process of constructing them, justifying their importance and purpose.

1-Achievement Tests

Learning is considered a hypothetical process that can be inferred from its impact on learners, specifically through their academic achievement. This achievement is verified and the cognitive and practical abilities of learners are assessed through the processes of educational measurement and evaluation. Achievement tests are among the most commonly used tools in educational institutions for this purpose. In this study, we attempt to define achievement tests as follows:

> - An achievement test is defined as "a tool used to measure knowledge, understanding, and skill in a particular academic or instructional subject, or a group of subjects" (Al-Faqih, 2005, p. 464). According to this definition, an achievement test is a tool adopted to evaluate the extent to which individuals have mastered the knowledge and information presented to them in a specific educational or training program and their understanding and comprehension of it. For instance, whenever a course is offered to a group of individuals, an achievement test is administered to determine the extent of their retention and understanding of the material covered.

Additionally, an achievement test is defined as "an organized procedure to determine the extent of student learning" (Malham, 2001, p. 433). After students have been taught, they undergo an organized procedure represented by the test to assess their achievement of the learned material. This view is also echoed by Airasian (1997), who describes the achievement test as "an organized method for assessing the level of student achievement in knowledge and skills they have learned" (Qatami, 2009, p. 354). This method aims to measure student achievement after learning, where "achievement" refers to "the extent to which a student has achieved the learning objectives as a result of studying a subject matter" (Jalal, 2008, p. 101). Through the process of education, specific objectives are expected to be achieved by the student, and the student's achievement indicator is the extent to which they have been able to accomplish these desired objectives.

Furthermore, Saadah (1948) defines achievement tests as "an organized procedure in which the behavior of students is observed to determine the extent to which they have achieved the set objectives, by presenting a set of items or questions to which students are required to respond, with these responses described using numerical scales" (Abu Jado, 2005, p. 411). According to this definition, Saadah suggests that this procedure involves presenting a set of questions for students to answer, and then assigning numerical estimates to their responses. These estimates reflect the extent to which the desired educational objectives have been achieved. The subsequent section will clarify the meaning of educational objectives. Therefore, based on the definitions of achievement tests, it can be said that they consist of a set of questions posed to students to assess their level of

achievement, aiming to ensure the attainment of educational objectives outlined in the learning process.

1-Steps for Constructing Achievement Tests

The steps for constructing a good achievement test that meets the required conditions must go through the following specific steps:

1. **Identifying the Purpose and Objectives of the Test**: In this step, the aim to be achieved by conducting the test is determined (Ibrahim, 2008, p. 56). For example, if the purpose is diagnostic, aiming to identify areas of weakness, the test should be designed accordingly. Additionally, it is essential in this step to identify the topics and learning outcomes that the test aims to measure among learners, ensuring adequate coverage.

2. **Preparing the Specifications Table**: Anastasi (1988) suggests that the specifications table consists of two dimensions: the first dimension indicates cognitive objective levels such as understanding, application, analysis, etc., while the second dimension represents units or content of the subject matter. Through this table, the teacher can analyze the content of the subject matter into subtopics, determine the level of cognitive objectives for each part of the content, and assign relative importance to them. Consequently, the teacher can select test items representing the content of the subject matter and the objectives from various cognitive levels (Al-Zaghloul, 2009, p. 325). To ensure the sound construction of the specifications table, it is incumbent upon the teacher to follow the following steps, as seen by Ahmed Yaqoub Al-Noor:

- ✓ The teacher should list all the topics covered in the curriculum that need to be assessed.
- ✓ The teacher should calculate how many lessons or class periods are devoted to teaching each topic.
- ✓ The teacher should determine the Relative Weight of the Subject Matter Topics using the following formula:

$$\frac{\text{Number of Lessons Required for a Topic} \times 100}{\text{Total Number of Lessons Required for All Topics}} = \text{Relative Importance of the Topic}$$

-Specify Behavioral Objectives to be Measured: Determine the behavioural objectives to be measured at different levels and assign their relative weights using the following formula:

Relative Weight of Objectives at a Certain Level = Number of Objectives at that Level / Total Number of Objectives for the Subject times 100

- Determine the Total Number of Test Questions: Based on the allowed time for answering and other student characteristics.
- Determine the Number of Questions for Each Topic at Each Level: This is done using the following formula:

Number of Questions for a Topic = Total Number of Questions x Relative Weight of Objectives for the Topic

-Determine the Marks for Each Topic's Questions at Each Level: Using the following formula:

Marks for Questions on a Topic = Final Test Score Relative Weight of Topic's Importance x Relative Weight of Topic's Objectives

- Follow the Specified Steps in Preparing the Table of Specifications: Adhering to these steps ensures the test's validity by distributing questions fairly across topics based on their importance. This prevents overemphasis on minor parts of the studied topics, ensuring that the questions cover all aspects of the curriculum (Noor, 2007, p. 128).

- Specify the Appropriate Question Types for the Topic: Whether essay or objective type.

- Draft the Instructions and Prepare the Test in Its Initial Form: Provide model answers for the test items.

- Pilot the Test: After designing the test, it should be piloted on a random sample similar to the target group. Conduct the necessary statistical analyses to evaluate the test's validity and psychometric properties and make appropriate adjustments based on the results (Zagloul, 2009, p. 325).

From the above, it is evident that constructing a good achievement test involves several sequential steps, from defining its objectives and purpose to piloting it for statistical analysis and making necessary revisions. Paying careful attention to each step significantly contributes to creating an achievement test based on sound scientific principles and meeting the required standards, resulting in a reliable assessment tool that enables accurate judgments and decision-making.

Characteristics of a Good Achievement Test

Achievement tests are the most commonly used assessment tools in our educational system and perhaps the only ones. Therefore, any shortcomings in their preparation will inevitably lead to deficiencies in the educational evaluation process, affecting the teaching and learning process. Hence, it is essential that these tools have specific characteristics that make them suitable for assessment. According to Ahmad Yaqoub Noor, these characteristics include objectivity, reliability, validity, comprehensiveness, discrimination index, ease and difficulty index, and usability. These will be explained in this chapter.

3-1 **Objectivity**

Objectivity refers to the independence of results obtained from assessment tools from subjective judgment (Zeinab, 2009, p. 198). In achievement tests, objectivity requires the test designer and scorer to remain neutral during test design and scoring. Neutrality in test design means not selecting test items from parts of the curriculum the teacher prefers or is biased towards, and neutrality in scoring means not being influenced by extraneous factors such as poor handwriting or the student's personality. Objective scoring means giving the same grade to the same test paper regardless of who scores it, especially in essay tests. Sabbah Abu Libda (1987) noted a study in the

UK where history test papers were given to fifteen scorers, resulting in significant grade variations. When the same answers were re-evaluated by the same scorers after a year and a half, the grades varied drastically. The assessment was reflected for 92 cases, changing from fail to pass. (Noor, 2007, p. 172). This indicates a lack of objectivity in scoring. Using clear and specific criteria can reduce such differences, ensuring consistent scoring across different scorers or the same scorer at different times.

To ensure objectivity in achievement tests, the following procedures related to test design and scoring can be considered:
- The test questions should be a representative sample of different parts of the subject, adhering to the table of specifications.
- The questions should be clear, unambiguous, and at the students ' language level to avoid limiting their ability to answer due to complex language.
- Design objective questions whenever possible.
- For essay questions, provide a model answer to guide scoring and create a scoring rubric.
- Review a sample of student answers and compare them with the scoring rubric, adjusting the rubric, if necessary, before use (Abu Libda, 2008, p. 209).

Although following these procedures cannot completely eliminate scorer subjectivity, especially in essay questions, they can promote objectivity and reduce subjectivity, minimizing bias from personal opinions and preferences.

3-2 **Reliability**

Reliability refers to the ability of a test to yield consistent results when administered to the same sample more than once. It means that the individuals in the sample maintain their rank order if the test is re-administered after a certain period. The reliability coefficient of a test ranges from 0 to 1, with values closer to 1 indicating higher reliability. However, achieving a reliability coefficient of exactly 1 is unlikely due to various uncontrollable factors. This section will discuss methods for calculating the reliability coefficient of achievement tests, the acceptable value of the reliability coefficient, and factors that can affect reliability.

3-2-1 Methods for Calculating Reliability: Reliability can be estimated using the following methods:

3-2-1-1- Test-Retest Method:

This method involves administering the same test to the same group twice, with a certain time interval between the two administrations, ensuring the same conditions for both. The reliability coefficient is calculated by correlating the scores from the two administrations. A reliability coefficient of 0.7 or 0.8 is acceptable, while values below 0.7 indicate low reliability (Noor, 2007, p. 176). However, this method has drawbacks, such as test familiarity and recall of answers from the first administration, especially if the interval is short. Longer intervals can introduce factors like forgetting or learning (Noor, 2007, p. 229). Therefore, careful consideration is needed when estimating the time interval between administrations to avoid significant impacts on results.

- Split-Half Method:

Instead of re-administering the test, reliability is calculated from the results of a single administration by splitting the test into two equal halves. The reliability coefficient is obtained by correlating the scores of the two halves. (Allam, 2007, p. 235). The split can be based on odd and

even questions or other criteria ensuring equal difficulty and variance. The reliability of the entire test is then estimated using formulas such as Spearman-Brown (Abu Libda, 2008, p. 231) or others:

r_test = (n * r) / (1 + (n - 1) * r)

- r: Estimated reliability coefficient for the entire test.
- n: Number of times the test length should be increased.
- r : Split-half reliability coefficient obtained through experimental statistical methods.

- **Rulon's Formula:** According to Abu Libda (2008, p. 232), Rulon's formula for calculating reliability is as follows:

$$r = 1 - (\sigma^2_d / \sigma^2)$$

Where:

r = Reliability coefficient of the entire test
σ^2_d = Variance of the differences between scores of the two halves of the test
σ^2 = Total variance of the test scores

- **Guttman's Lambda Formula:** This formula is applicable when the standard deviations of the two halves of the test are not equal, and it is also suitable when these deviations are equal. According to Noor (2007, p. 181), the formula is as follows:

$$r = (n/(n-1)) * (1 - (\Sigma\sigma^2_1 / \sigma^2_t))$$

\sigma^2_t = Variance of total test scores
k = Number of test items

4-2-1-3 Internal Consistency Method:

According to Bushra Ismail (2004), this method for calculating reliability is based on the consistency in individuals' performance on the test from one item to another. In this method, the test is divided into a large number of parts, with each part consisting of a single test item. The greater the consistency between these items, the higher the reliability of the test as a whole (Al-Nour, 2007, p. 184). Some of the methods used for this include:

Kuder-Richardson Formula:

Kuder and Richardson developed a formula to calculate the reliability coefficient of a test by analyzing the responses to the items and calculating their variances.
This formula is known as KR-20 and is expressed as:
σ^2_1 = Variance of individual item scores
σ^2_2 = Variance of paired item scores

4-2-1-3 Internal Consistency Method

Bushra Ismail (2004) states that this method for calculating reliability depends on the consistency of individuals' performance on the test from one item to another. The test is divided into a large number of parts, with each part consisting of a single item from the test. The greater the consistency between these items, the higher the overall reliability of the test (Noor, 2007, p. 184). Methods used for this include:

- **Kuder-Richardson Formula:** Kuder and Richardson developed a formula to calculate the test reliability coefficient by analyzing the item responses and calculating their variances, named KR-20 (Noor, 2007, p. 185):
r = (2 * n * σ^2_2) / ((n-1) * σ^2_t)

M : Mean of the test scores.
sigma^2_k: Variance of the total test scores
Kuder and Richardson also developed another formula, "KR-21" (Murad and Suleiman, p. 365).
r_21 = (n/(n-1)) * (1 - (nM(n-M)) / (n-1)*σ^2_k))
- Cronbach's Alpha Coefficient:

Cronbach's Alpha α is considered a special case of the Kuder-Richardson formula. It was proposed by Cronbach in 1951 and by Novak and Lewis in 1967 (Noor, 2007, p. 365). This formula relies on the variances of test items and is used when test items do not have binary responses such as yes or no (1, 0), unlike the Kuder-Richardson formula which is used for binary response items. The formula is as follows (Murad and Suleiman, p. 366):

$$r = - [1 - (\Sigma\sigma^2 / \sigma^2_t)]$$

N=Sum of item variances
N1Variance of total scores}

3-2-2 The Acceptable Value of Reliability Coefficient in Achievement Tests:

In this regard, Sab' Muhammad Abu Libda (2008) mentions that the reliability coefficients for standardized achievement tests are typically around 0.50.

Therefore, it is found that teacher-made tests tend to be less reliable compared to standardized achievement tests. This is particularly true for essay tests, where reliability coefficients tend to be low unless the teacher pays careful attention to constructing them according to specific standards and a clear grading rubric.

3-2-3 Factors Affecting the Reliability of Achievement Tests

Several factors can influence the reliability of achievement tests, including the following:
- **Grader:** A student's score depends not only on their knowledge but also on who grades their paper, especially in essay tests. This was confirmed by studies such as the one conducted by researchers Stark and Elliot, who sent a geometry exam answer sheet to 138 geometry teachers for grading. The scores ranged from 28 to 95 (Abu Libda, 2007, p. 236). This indicates that grading does not only reflect the student's answer but is also influenced by the grader's opinion, psychological state, reaction to the student's handwriting, and other factors.
- **Test Construction**: Achievement tests are significantly affected by chance factors. The questions might cover parts of the material that the student is particularly good at, resulting in a high score, or they might cover parts the student is less familiar with, leading to failure. This is due to the questions not accurately representing the curriculum. Additionally, unclear questions can affect test reliability.

- **Number of Items:** Tests with a small number of items tend to have lower reliability. This is evident when applying the Spearman-Brown formula in split-half reliability, where the reliability estimate for the whole test is higher than for half the items. Therefore, increasing the number of

items in a test improves reliability. However, more items require more time, which can cause student fatigue, reducing the benefits of a longer test. Ideally, tests should include a number of questions that can be answered in a short time, depending on the subject, educational objectives, and grade level. Sometimes, tests might include fewer but more detailed questions (Allam, 2007, p. 243).

- **Variance in Student Abilities:** The reliability of a test is affected by the variance in the abilities of the students being tested. If the group of students is homogenous, test reliability tends to be lower. This is because when scores are less spread out, the likelihood of a student changing their rank within the group upon retesting is high, leading to a lower reliability coefficient. Conversely, greater variance in scores increases test reliability.

- **Difficulty Level of Questions:** The difficulty level of the questions affects the reliability coefficient. If most questions are very difficult, chance plays a significant role in student responses, which affects the test's reliability.

- **Health and Psychological State:** The physical and psychological condition of the student affects test reliability. If a student is tired, ill, or stressed, the reliability coefficient decreases (Murad and Suleiman, 2005, p. 369).

- Test Validity: The validity of a test affects its reliability. A valid test is reliable, but a reliable test is not necessarily valid (Al-Mansi et al., p. 150).

Thus, it is clear that many factors influence the reliability coefficient of a test. To maintain test reliability, necessary measures should be taken considering the influencing factors mentioned above.

3-3- Validity

3-3-1 Types of Validity

There are several types of validity, each with its own purpose and method of measurement. We will discuss five main types here:

3-3-1-1 Content Validity

Content validity refers to the extent to which an assessment accurately represents the content of the subject matter being tested. The higher the degree of representation, the higher the content validity. Salah Murad et al. (2005) state that "content validity is evidence of the comprehensiveness of the instrument and the degree to which it represents the content" (*علام، 2007، 351*). To achieve content validity, careful attention must be paid to constructing a content specification table for the material being tested. This helps to ensure a fair distribution of questions across the learning objectives, based on their relative importance. Content validity is also known as construct validity, where the construct refers to the content.

To verify the content validity of a test, it should be reviewed by a panel of at least five experts in the relevant field. If the level of agreement among the reviewers is high (at least 75%), the content validity coefficient is considered acceptable. However, if the agreement rate is below 50%, the items in the assessment tool should be revised. Generally, the higher the agreement among reviewers, the higher the content validity.

3-3-1-2 Criterion-Related Validity

This type of validity relies on the correlation between the scores on the test being assessed and the scores on another test or measure prepared by the teacher, or the student's semester or

annual average. In this case, the teacher's test or student's average is called the criterion. The criterion must be objective, valid, and stable (Abu Lubda, 2007, 218). Criterion-related validity is divided into two types:

3-3-1-3 Predictive Validity

The predictive validity of a measurement tool relies on calculating the predictive value of the tool based on the assumption that behaviour exhibits a high degree of stability in the future (Drouza, 2005, 175). This means that a student who performs well on an achievement test in a particular subject is expected to succeed in their university studies if they major in that subject. The validity coefficient of this test is calculated using scores and a criterion that gathers information later, such as the student's first semester or first-year average in university.

3-3-1-4 Concurrent Validity

This type of validity is calculated by computing the correlation coefficient between the scores on the test being assessed and the scores on another test, administered at approximately the same time. Student grades can also be used as a criterion to calculate the validity of this test (Abu Lubda, 2007, 219). This is useful when a teacher wants to replace an existing test with a new one, perhaps because the old test is too time-consuming to administer or grade. The new test is considered valid if the correlation coefficient between its results and the criterion results approaches one.

3-3-1-5 Factorial Validity

Factorial validity focuses on determining the extent to which the items on a test are saturated with specific factors (components), whether a general factor or specific factors (Murad & Suleiman, 2005, 355). This type of validity relies on factor analysis, which analyzes the correlation coefficients between the test and various criteria to identify the factors that led to these coefficients (Al-Nour, 2007, 201).

3-3-1-6 Construct Validity

This type of validity refers to the extent to which a test measures a hypothetical construct or psychological concept or trait, such as intelligence, mathematics, etc. (Al-Nour, 2007, 201). Thurstone developed a test of primary mental abilities translated by Ahmed Zaki that measures four abilities: word meanings, spatial perception, reasoning, and numerical ability. All these abilities are used to calculate the intelligence quotient. The test developer prepares questions that measure these components and verifies their validity after testing them on a sample. A factor analysis is then conducted to determine the extent to which the test measures the intended components.

3-3-2 Methods for Calculating Validity Coefficients

Al-Nour (2007) mentions the following methods for calculating validity coefficients:

3-3-2-1 Using the Correlation Coefficient to Indicate Test Validity

For example, if the external criterion method is used, the correlation coefficient between the scores on the test being administered and the scores on the previously validated external criterion must be found. This coefficient indicates the validity of the test. The Pearson correlation coefficient can be used to find the degree of validity, which is calculated as follows:

$$r = (\Sigma x1x2) / \sqrt{(\Sigma x1^2 \Sigma x2^2)}$$

Where:
r: Pearson correlation coefficient.
X1: Deviation of scores on the first test from their mean.
X2: Deviation of scores on the second test from their mean.

3-3-2-2 Judge Rating Method:

In this method, the specialized judge estimates the extent to which each item on the test is related to the trait or ability being measured. This is done after clarifying the operational definition of the trait. The following equation can be used to calculate content validity through judgment:

CVI = (S1 - S2) / N

Where:
CVI: Content Validity Index.
S1: Number of items the judges agreed to measure the objective.
S2: Number of items the judges agreed on does not measure the objective.
N: Total number of items on the test.

3-3-2-3 Cross-Tabulation Method:

This method is based on comparing the frequency distribution of individuals' scores on the test with the frequency distribution of their scores on an external criterion. This is done by creating a cross-tabulation table of the test score categories and the external criterion score categories. This table helps estimate the extent of the test's validity at each level of the external criterion.

3-3-2-4 Extreme Groups Comparison Method:

In this method, the aim is to identify the extent to which the test can highlight individual differences within the sample. Raga'a Abu Allam (1987) states that "criterion-referenced tests should increase the differences between the groups that have mastered the course" (Abu Allam, 2004, 426). To determine this, the following steps should be followed as mentioned by Ma'mariyyah (2007):

The test results obtained by the sample individuals are arranged in descending or ascending order.

27% of the distribution extremes are calculated, resulting in a high-scoring group and a low-scoring group on the test.

The performance of the two groups is compared using an appropriate statistical method, which is the t-test to determine the significance of the difference between the means.

Finally, after obtaining the calculated t-value, it is compared with the tabulated t-value to be able to judge the validity of the test. If the calculated t-value is significant, the test can be considered valid.

3-3-3 Interpretation of the Validity Coefficient

The closer the validity coefficient is to one, the higher the test's validity, and the closer it is to zero, the lower the validity. There is no clear-cut threshold where if the estimated validity coefficient exceeds it, the test is considered valid, and if it falls below it, the test is considered invalid. Saba'a Muhammad Abu Libda mentions that the validity coefficient for a test developed by a teacher and the results of an intelligence test ranges between (0.30) to (0.50), and this decrease is due to the fact that the test prepared by the teacher is influenced by other factors such as the student's effort, the subjectivity of the scorer, and other matters that can affect it. However, if the calculated validity coefficient is between the results of an achievement test developed by the teacher and the results of another standardized achievement test, it would be higher than the previous example. Similarly, the validity coefficient between a battery of achievement tests and the student's average or rank in the class is mostly between (0.60) to (0.70). If the validity coefficient is calculated between two standardized achievement tests, it would range between (0.60) to (0.80) (Abu Libda, 2007, 223).

3-3-4 Factors Influencing Test Validity

There are several factors that may reduce the validity of a test and make it unsuitable for the intended use. These factors can be divided as follows:

3-3-4-1 Factors Related to the Student

These include:

The student's anxiety or fear during the test administration may impair their ability to respond and result in a score that does not represent their true abilities.

The students ' resort to guessing or cheating, or their ability to influence the scorer through their style of expression (Malham, 2007, 336).

- **3-3-4-2 Factors Related to the Test:**

These include:

The difficulty of the test item language can cause the student to be unable to respond, as well as the ambiguity of the language, which can lead students to interpret the items differently from the intended meaning, resulting in incorrect answers.

Formulating test items that contain the answer within the item itself, enables students to obtain high scores.

The ease or difficulty of the test items can lead to an incorrect judgment of the students. If the items are too easy for their intellectual level, they can obtain high scores, leading to an assessment of their superiority, and the opposite is also true. In both cases, the student's score will not be valid, as it does not represent their true ability.

The relationship between the test items and what the student has learned affects the test's validity. If the well-designed test items measure knowledge, understanding, analysis, and application, but the teacher has taught the students the answers and the way to solve them, then these items will only measure the recall of facts. Similarly, if the teacher includes items from a topic that the students have not studied, this will lead to a decline in their scores, and this decline does not represent their true abilities (Al-Nur, 2007, 196).

3-3-4-3 Factors Related to Test Administration

These include:

- ✓ Environmental factors that can negatively affect the student's responses, such as high or low temperatures, or surrounding noise.
- ✓ Lack of clarity in the printing of the test items, the presence of typographical errors, or poor arrangement of the test items.
- ✓ Lack of clarity in the test instructions.
- ✓ Using the test for a purpose other than what it was designed for, such as using a grammar test to measure a student's language proficiency.
- ✓ Using the test with a population for which it was not intended, such as administering a test designed for gifted students to a group of underperforming students.
- ✓ Comprehensiveness requires that the test items represent a representative sample of the behaviour being measured (Mansi et al., 135). In achievement tests, this means that the teacher should not include test items from a specific part of the curriculum and neglect other aspects. The tests should cover the content of the curriculum and what has been taught to the learners in a balanced and comprehensive manner, as this significantly contributes to identifying areas of weakness and facilitates the development of a remedial program based on the obtained results.

3-5- Discrimination Index

The discrimination index indicates the extent to which a test item can differentiate between high-achieving and low-achieving students after administering the test. Students with higher achievement levels will answer the item correctly, while those with lower achievement levels will not. In this regard, Fakhri Khudr (2003) states that test discrimination requires varying the levels of difficulty of the test items, increasing the number of test items, and covering the content of the study material. Additionally, cheating and guessing should be minimized to prevent students from obtaining undeserved scores (Al-Nur, 2007, 207). The formula used to calculate the discrimination index varies depending on its application:

Discrimination Index = (X_upper - X_lower) / N

If the item is an essay-type question, the formula that can be used is as follows:

Discrimination Index for an Essay Question = (Qc - Qd) / (S max * n)

Where:
Qc: The total number of points obtained by the students in the upper group.
Qd: The total number of points obtained by the students in the lower group.
S max: The maximum number of points a student can obtain on the question.
n: The number of students in either the upper or lower group (Alam, 2007, 354).

To determine the upper and lower groups, the following is done:
The student papers are arranged in descending order, from the highest score to the lowest. Then, they are divided into two equal groups, with the first group being the papers with the highest scores, and the second group being the papers with the lowest scores. If the number of students is odd, the paper of the student in the middle of the ranking can be excluded. For example, if there are 41 students, the paper of student number 21 in the ranking can be excluded. If the number of students is large, such as 100, only the top and bottom 25% can be used.

Saba'a Muhammad Abu Libda suggests that the teacher can take any proportion to represent the upper and lower groups, as long as the number of students is sufficient or reasonable. If the number of students in the class is between 30 and 40, it is preferable to include them all in the analysis. However, if the exam is for multiple sections with a total of 100 students, it is possible to take the top and bottom 90%, 25%, or 20%, so that the analysis includes between 40-60% of the students. The larger the number of examinees, the lower this percentage can be (Abu Libda, 2007, 309).

So, after obtaining the upper and lower groups, the discrimination index parameters are calculated by applying one of the two formulas mentioned earlier, depending on the type of item. The interpretation of the discrimination index values can be done as follows, based on Ebel's (1965) guidelines:

Discrimination index ≤ 0.40 - The item meets the purpose or objective.
0.30 ≥ Discrimination index ≥ 0.39 - The item requires minor review.
0.20 ≥ Discrimination index ≥ 0.29 - The item is on the borderline and needs review.
Discrimination index ≥ 0.19 - This item should be discarded or thoroughly reviewed (Crocker, 2009, 418).

3-6- Difficulty Index:

Ahmed Ya'qub Al-Nur believes that every test should have questions that suit the different levels of students. He suggests that the test should have 16% easy questions to suit the weak students, 68% average questions to suit the average student, and 16% difficult questions to suit the high-achieving students. Easy questions do not mean questions where no student is expected to fail, and difficult questions do not mean impossible questions that no student can solve, but rather questions that require a reasonable mental effort. It is recommended to start the test with easy questions and progress to the more difficult ones, as this can help motivate the students to attempt the test. The formula to calculate the difficulty index for an objective test item is as follows:

Difficulty Index = (Number of students who answered correctly) / (Total number of students) x 100
As for the difficulty index, it can be calculated as follows:

Difficulty Index = (Number of students who answered correctly) / (Total number of students) x 100
Or according to the alternative formula

Difficulty Index + Difficulty Coefficient = 1
The difficulty index for an essay-type item is:

Difficulty Index = (Sum of students' scores on the item) / (Number of students x Maximum score on the item) x 100

The calculated difficulty indices range between 0 and 1. Awda (1993) indicates that the acceptable (good) difficulty index varies according to the type of question as follows (Kadhim, 2001, 101):

The appropriate difficulty index for True/False questions is 0.75.
The appropriate difficulty index for 4-option multiple-choice questions is 0.63.

The appropriate difficulty index for 5-option multiple-choice questions is 0.60.
The appropriate difficulty index for essay questions is 0.50.
Ahmed Ya'qub Al-Nur adds the following types of questions:

The appropriate difficulty index for 3-option multiple-choice questions is 0.67.
He also set a general criterion for the difficulty index for all items, which should be between 0.40 and 0.60

3-7- Usability
The test becomes usable if it is practical in terms of:

3-7-1 Ease of administration

The easier the test administration, the better. Difficulty in administration can be an obstacle that prevents achieving the validity, reliability, and objectivity of the test. Ease of administration includes the clarity of the questions and the test instructions. The instructions guide and direct the student in taking the test, informing them of exactly what they need to do, clarifying the time allocated for answering the test questions, and explaining the purpose of the test. These instructions should be written in simple and clear language to explain the test objective, the method of

answering, and advise the student to avoid guessing in case of uncertainty about the correct answer. Allen and Yen (1979) suggest the following to achieve clear test instructions:

- ✓ Instruct your students to read the instructions before starting to answer the test items.
- ✓ Inform your students about the components of the test, introduce them to the test items, and indicate the score value for each test item.
- ✓ Inform your students about the number of test papers.
- ✓ Specify the test duration and estimate for your students the time needed to answer each test item.
- ✓ Write down the method of answering the test questions on the instructions page, especially if the questions are of the same type. If the questions include different types, specify each type of test item for your students.
- ✓ Inform your students that they cannot inquire about anything related to the test questions from the supervisor, except for any typographical errors that may appear in the test paper or its instructions.
- ✓ Advise your students to write clearly and neatly (Melhem, 444).
- ✓ Emphasize to the students the importance of writing their names on the answer sheet as soon as they receive the answer papers.
- ✓ Emphasize to your students the need to avoid guessing in their answers to the test items, especially if you intend to correct for guessing.

3-7-2 Ease of scoring

When designing the test, the teacher should consider the method of scoring it. The more complex the scoring process, the more it allows for the subjectivity of the scorer, leading to errors and requiring more time and effort, especially for essay-type tests. Scoring essay questions is highly complex because the answers can vary from one examinee to another, and it is difficult to adhere to a strict scoring guide, which allows for the scorer's subjectivity.

Therefore, the teacher should provide a model answer key accompanied by a clear scoring rubric so that the scoring can be done with the least possible errors (Melhem, 444).

Conclusion

Achievement tests are the most commonly used tool in educational assessment, as they aim to determine the extent to which learners have mastered the knowledge and skills presented to them through the educational process. Therefore, various types of tests serve the intended purposes, and the process of constructing them goes through specific steps that must be adhered to, to ensure the credibility of these tests, as this provides a higher level of accuracy in their results.

The achievement test is a reliable tool in educational measurement and evaluation, and it can lead to valid results and be used to make appropriate decisions if it meets the conditions and specifications of a good achievement test. Among the most important of these conditions are objectivity and minimizing subjectivity as much as possible. It should also comprehensively cover what has been taught and have appropriate validity and reliability. We can trust its results if its items have an appropriate level of difficulty and discriminative power suitable for the item type and the purpose of the test. Additionally, the test should be easy to administer and score, making it usable and thus achieving the desired purpose.

References

Al-Fatlawi, Suhaylah Muhsin Kazim. (2003). Teaching competencies (1st ed.). Dar Al-Shorouk for Publishing and Distribution.

Al-Fiqqi, Isma'il Muhammad. (2005). Psychological and educational evaluation and measurement. Dar Al-Gharib for Printing, Publishing and Distribution.

Al-Nour, Ahmad Ya'qub. (2007). Measurement and evaluation in education and psychology. Al-Janadiriyyah for Publishing and Distribution.

Al-Satriy, Ra'id Muhammad Ibrahim. (2010). Generalization of achievement tests in the college of physical education and sports at King Saud University according to the standards of good tests. Sharjah University Journal for Humanities and Social Sciences, 2.

Al-Zaghul, 'Imad 'Abd Al-Rahim. (2009). Principles of educational psychology (2nd ed.). Dar Al-Masirah for Publishing and Distribution.

Allam, Salah Al-Din Mahmud. (2007). Measurement and educational evaluation in the teaching process. Dar Al-Masirah for Printing, Publishing and Distribution.

Abu Jawd, Salih Muhammad 'Ali. (2005). Educational psychology (4th ed.). Dar Al-Masirah for Publishing, Distribution and Printing.

Abu Labdah, Sab'i Muhammad. (2007). Principles of psychological measurement and educational evaluation (5th ed.). Dar Al-Fikr for Publishing and Distribution.

Crocker, Linda, & Algina, Gina. (2009). Introduction to classical and modern measurement theory (1st ed., Zaynāt Yusuf Da'nā, Trans.). Dar Al-Fikr.

Darwazah, Afnan Nazir. (2005). Educational questions and school evaluation. Dar Al-Shorouk for Publishing and Distribution.

Ibrahim, Magdi 'Aziz. (2008). Contemporary educational and instructional issues (1st ed.). Dar Nahdat Al-Sharq for Printing, Publishing and Distribution.

Jalal, Ahmad As'ad. (2008). Scientific applications and exercises on the SPSS program. Dar Al-Dawliyyah for Cultural Investments.

Ma'mariyyah, Bashir. (2007). Psychological measurement and instrument design. Al-Hibr Publications.

Mansiy, Mahmud 'Abd Al-Halim, et al. (2003). Educational evaluation and principles of statistics. Modern Republic Company for Conversion and Paper Printing.

Melhem, Sami Muhammad. (2001). The psychology of learning and teaching (1st ed.). Dar Al-Masirah for Publishing, Distribution and Printing.

Murad, Salah Ahmad, & Amin, 'Ali Sulayman. (2005). Tests and scales in psychological and educational sciences (2nd ed.). Dar Al-Kitab Al-Hadith.

Qatami, Yusuf Mahmud. (2009). Principles of educational psychology. Dar Al-Fikr Publishers and Distributors.

Zaynab, 'Abd Al-Karim. (2009). Educational psychology. Dar Usamah for Publishing and Distribution.