Spor Eğitimi ve Bilim Dergisi
Journal of Sport Education and Science
Cilt: 1, Sayı: 2, 2025
57-75

Standardization of Psychological and Educational Scales and Test

Afifa Djedidi[1]

[1]Akli Mohand Oulhadj University of Bouira, Algeria.

Derleme Makale	Gönderilme Tarihi	Kabul Tarihi	Yayın Tarihi
	09.12.2025	28.12.2025	30.12.2025

Abstract

The current article aims to clarify the various necessary steps that researchers in the psychological and educational fields must follow when using various tests, scales and questionnaires in their research and studies. These steps are collectively called "standardization"; which includes different stages, the first of which is determining the ease or difficulty of the test items for the standardization sample, then the extent to which each item can discriminate in the performance of the sample members as well as its stability. This is followed by the second stage, which includes verifying the psychometric properties of the measurement tool. In the third stage, the researcher must extract the standards that give meaning to the results of the measurement tool, through which he/she can interpret the measurement results.

Keywords: Standardization, item ease-difficulty, item validity, item stability, psychometric properties, standards

Psikolojik ve Eğitimsel Ölçek ve Testlerin Standardizasyonu

Öz

Bu makale, psikoloji ve eğitim alanlarındaki araştırmacıların araştırma ve çalışmalarında çeşitli test, ölçek ve anketleri kullanırken izlemeleri gereken çeşitli gerekli adımları açıklamayı amaçlamaktadır. Bu adımlar topluca "standardizasyon" olarak adlandırılır ve farklı aşamaları içerir. İlk aşama, standardizasyon örneklemi için test maddelerinin kullanım kolaylığını veya zorluğunu, ardından her bir maddenin örneklem üyelerinin performansında ve kararlılığında ne ölçüde ayırt edici olabildiğini belirlemektir. Bunu, ölçüm aracının psikometrik özelliklerinin doğrulanmasını içeren ikinci aşama takip eder. Üçüncü aşamada, araştırmacı, ölçüm aracının sonuçlarına anlam kazandıran standartları çıkarmalıdır ve bu standartlar aracılığıyla ölçüm sonuçlarını yorumlamalıdır.

Anahtar Kelimeler: Standardizasyon, madde kolaylığı-zorluğu, madde geçerliliği, madde kararlılığı, psikometrik özellikler, standartlar

INTRODUCTION

Measurement and evaluation are essential for identifying the categories of normal and abnormal persons, and determining their levels (Groth-Marnat & Wright, 2016: 3), think that: "the central role of specialists conducting assessments should be to answer specific questions and make clear specific and reasonable recommendations to help improve functioning". And without appropriate measurement and evaluation tools for each category, it's difficult for specialists and educational program designers to place these categories in the appropriate place, or to design appropriate educational programs and subsequently evaluate them to determine their effectiveness. The

primary goal of the educational process is to help students achieve comprehensive growth in all aspects of their personalities and achieve the goals set by society (Djedidi, 2018).

Therefore, the tests and scales used by researchers in their psychological or educational studies are among the most important tools used in the measurement and evaluation processes. These tools aim to control study variables, diagnose the phenomenon under study, or find appropriate solutions in light of the results obtained.

Anastasi (In all the editions of her classic textbook on psychological testing, which ranged from 1954 to 1997) defined a psychological test as "an objective and standardized measure of a sample of behavior" (Urbina, 2016). The term "behavior" here may reflect an individual's verbal or mechanical ability, a personality trait, or a set of motor performances.

Cooper (2023: 3) shows that: "Psychological tests and assessments are widely used by practitioners – professionals such as clinical, educational, occupational, sport and forensic psychologists. They are also widely used by researchers and students in psychology and other disciplines, and by organizations seeking to select the most promising applicants or candidates for promotion".

The importance of these tools in academic studies is evident, and they must meet the conditions for a good test (Abdul Rahman, 2008), which are as follows:

a. Objectivity: This means that the subjective aspect does not interfere with the assessment and interpretation of scores, and therefore that correctors do not differ in their assessment of scores. To achieve objectivity, the following must be met (Ahmed, 1960):

- The conditions for conducting the test must be consistent in terms of clarity of instructions, specification of the answer method, and timescale.
- The scoring method must be consistent; That is, the presence of pre-prepared scoring keys.
- The formulation of test questions is clear and specific, so that all individuals understand them with the same meaning.

b. Globalization: This means that the test measures all aspects of the domain (mental, emotional and psychomotor) in the case of psychological tests. It also measures all aspects of the content at different mental levels and in light of the test specifications table of interest in the case of educational tests (Boucenna, 2012).

c. Standardization: This is the standardization of application procedures for all participating individuals and the standardization of scoring methods. It also includes determining the psychometric characteristics that indicate the quality of the test and standardizing the method of interpreting scores (Bernaud, 2014). Standardization also refers to applying the test to a large sample of person's representative of the population for which the test was designed. It is worth noting that standardization contributes to making the test valid in measuring what it was designed for and consistent upon re-administration (Awad, 1998).

METHOD

Spor Eğitimi ve Bilim Dergisi
Journal of Sport Education and Science
Cilt: 1, Sayı: 2, 2025
57-75

The current article is a descriptive theoretical study, which will rely on literature related to the concept of test standardization. This study aims to clarify the practical steps for standardizing psychological and educational tools of measurement, and the necessity of this process to ensure the accuracy and validity of the results. The stages of the standardization process will be described, along with the statistical equations used to verify these steps.

Standardization Process

Standardization is the process by which inappropriate factors that could affect the measurement process are controlled and measurement errors are minimized by applying the test to a sample representative of the study population. In order to obtain psychometric properties consistent with the characteristics of a good test, and then provide appropriate norms for interpreting raw scores, the standardization process is the final stage of designing a psychological test and preparing it for use. This process includes a series of procedures that must be performed well by a measurement specialist so that the test is confidently valid for use. These procedures include a set of theoretical fields, and statistical studies conducted on the psychological test, such as studying its items, validity and reliability coefficients, determining its norms, specifying instructions for its application for examiners and examinees, establishing a method for scoring and obtaining raw scores, and interpreting its results according to the derived norms (Djedidi, 2020).

The standardization process can be summarized in the following illustrative diagram (Djedidi, 2020: 116):

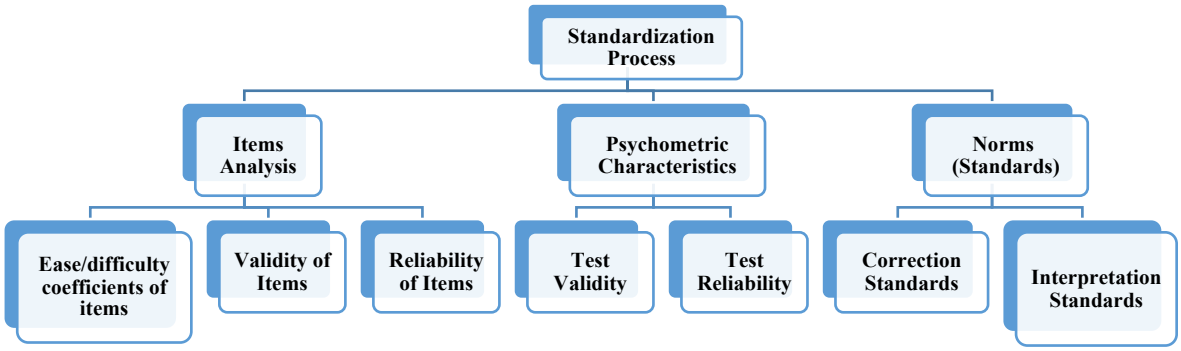

Diagram 1. Components of the standardization process

The next table shows different stages to develop a psychological or educational test (Irwing, Booth & Hughes, 2018: 4):

Table 1. Stages of test development

Stages and Substages
1. Construct definition, specification of test need, test structure
2. Overall planning
3. Item development a. Construct definition b. Item generation: Theory versus sampling c. Item review d. Piloting of items
4. Scale construction-factor analysis and item response theory
5. Reliability
6. Validation

7.	Test scoring and norming
8.	Test specification
9.	Implementation and testing
10.	Technical manual

The previous table identifies the different stages of constructing a test. It is noted that the standardization process takes up the largest part of the construction.

The standardization process goes through three basic stages (as I indicated in the previous diagram). Below is an explanation of each stage through its components:

Items Analysis

Item analysis is a procedure for evaluating the effectiveness of an item in a test. This can involve correlating items with the total score on the test, factor-analyzing items, Rasch scaling the items or subjecting the items to other estimations of their characteristic curves (Kline, 2015).

Once the researcher has completed constructing the test, it must be first test it on a standardization sample. This involves the following practical steps:

Item Difficulty

This refers to studying the ease or difficulty of the item for the standardization sample. This achieves three objectives (Mkaddam, 2003):

- The extent to which the item is appropriate for the sample,
- Placing the item in the appropriate place among the test items,
- Ensuring equal differences in difficulty between the preceding and following items, to ensure the test is graded at approximately equal distances.

The difficulty of the item is studied by applying the items initially designed for the test to the standardization sample. The difficulty coefficient is then calculated using the following equation (Abdul Rahman, 2008):

$$DC = \frac{iA}{N} \times 100$$

Where:
DC: Difficulty coefficient,
iA: Incorrect answers in the item,
N: Number of sample members.

The ease coefficient can also be calculated using the same formula as above. That is (Abdul Rahman, 2008):

$$EC = \frac{cA}{N} \times 100$$

Where:
EC: Ease coefficient,
cA: Correct answers in the item,
N: Number of sample members.

The appropriate difficulty factor for an item is between 40% - 60% (Maamria, 2012).

Correction for guessing

Respondents' answers are usually affected by guessing (chance or luck), making the score insufficient to reflect increased knowledge or reasoning. The guessing factor increases as the number of choices decreases. Items with two answers have a 50% chance of success by guessing, while items with three answers have a 33% chance of success, and so on (Maamria, 2007).

Therefore, measurement experts have proposed the following equation to reduce the effect of guess (Abdul Rahman, 2008):

$$DCc = EC - \left(\frac{EC}{C-1}\right)$$

This equation is used to correct items for guess, where:
DCc: The difficulty coefficient after correction,
EC: Ease coefficient
C: The number of answer choices

Or:

$$CSg = cA - \left(\frac{iA}{C-1}\right)$$

This equation is used to correct an individual's score for guessing, where:

CSg: the corrected score for guessing,
cA: The correct answers,
iA: The incorrect answers,
C: The number of answer choices.

Item Validity (Discrimination Coefficient)

It refers to the item's ability to discriminate between two groups (Anastasi & Urbina, 2006). This is calculated by determining the difference between the correct answers in the upper third and the correct answers in the lower third, using the following equation, known as Johnson's Law (Maamria, 2007):

$$VCi = \frac{CAu - CAl}{n}$$

Where:
VCi: Item validity coefficient,
CAu: Correct answers to the item in the upper third,
CAl: Correct answers to the item in the lower third,
n: 27% of the total sample size.

Item Reliability (Stability)

This means that if the item is re-administered to the same individuals multiple times, it will yield approximately the same results (Hussein & Gasmalla, 2023).

Item reliability is calculated in several ways, most notably by the retest method, which uses the famous Pearson correlation coefficient, or by the quadrilateral correlation coefficient when the item response options are binary (e.g., yes/no). The formula for the equation is as follows (Maamria, 2007):

$$CCq = cos\frac{180}{1+\sqrt{\frac{a \times d}{b \times c}}}$$

CCq= Quadrilateral correlation coefficient
cos= Defines the cosine of an angle (it can be easily found in the calculator)
The following table illustrates how to apply the previous equation:

Table 2. Frequency distribution of scores in the two applications

	Yes=1	No=0
First application	A	B
Second application	C	D

The reliability coefficient of an item with multiple answer choices can also be calculated using the following equation (Maamria, 2007: 122):

$$RCi = \frac{C}{C-1} \times \left(1 - \frac{1}{H}\right)$$

Where:
RCi: Item reliability coefficient,
C: Number of answer choices,
H: Highest relative frequency of these choices.

Psychometric Characteristics

These are indicators of the accuracy of the scale for what it was designed to measure (Thorndike, 1922). Therefore, specialists in psychological measurement attempt to obtain the characteristics of the scale and its items, as these are accurate indicators of its ability to measure what it was designed to measure.

These are the characteristics necessary for a good test, related to validity, reliability, and standards, which are calculated after the test has been administered to a sample representative of the population (Gregory, 2015). The quality of the test depends on the availability of appropriate data for these characteristics.

Validity

This means that the test actually measures what it was designed to measure (Faraj, 2007). Therefore, if we want to measure anxiety, it is not appropriate for the test to include questions about mathematics.

Validity is one of the most important aspects that we must ensure is present in a test before using it. It relates to two important questions (Anastasi & Urbina, 2006):

- What does this test measure?
- What is the degree of accuracy with which the test measures what it claims to measure?

On another side, (Bademci, 2022) explain that: "Validity and validation are two closely related but different concepts used in measurement. Validity is the degree to which evidence and theory support the adequacy and appropriateness of the proposed interpretations and uses of the scores obtained from the test or measurement instrument applied to a particular population or sample. Validation, on the other hand, is the process by which the evidence of the validity of score interpretations is collected. Besides, reliability is the reproducibility or the consistency of the scores obtained from the test or measurement instrument applied to a particular population or sample. It must be borne in mind that score reliability is necessary but not sufficient for score interpretation validity."

The concept of validity is an old one, but it has been subject to frequent modification and change due to the spread and expansion of testing and measurement in the various humanities and social sciences.

The following figure summarizes the main sources of validity evidence (Bademci, 2022: 151):

Spor Eğitimi ve Bilim Dergisi
Journal of Sport Education and Science
Cilt: 1, Sayı: 2, 2025
57-75

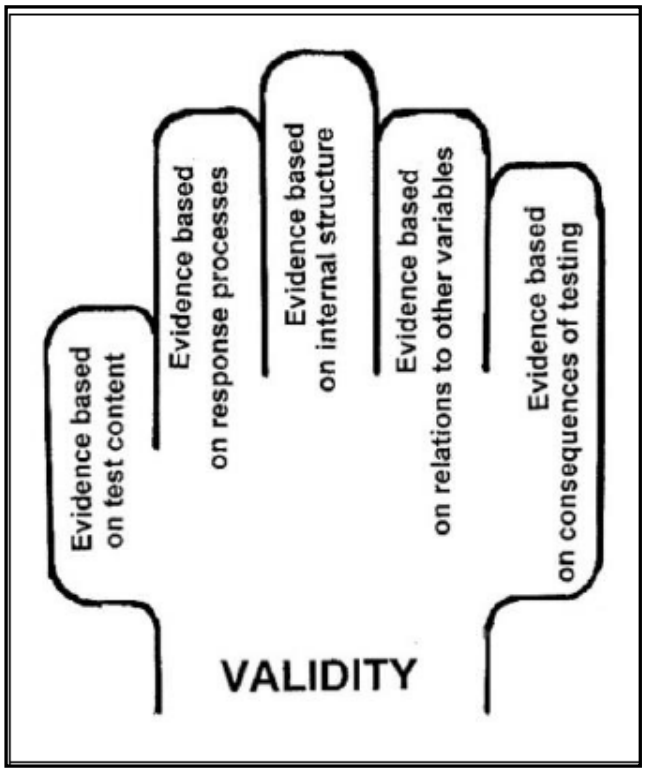

Figure 1. Sources of validity evidence

Face Validity

A test is said to be face valid if it appears to measure what it purports to measure, especially to subjects. Face validity bears no relation to true validity and is important only in so far as adults will generally not co-operate on tests that lack face validity, regarding them as silly and insulting. Children, used to school, are not quite so fussy. Face validity, then, is simply to aid co-operation of subjects (Kline, 2015).

Content Validity

The extent to which a test represents the aspects it is intended to measure. To achieve this, we carefully examine the content of the test to determine the aspects of behavior it measures and the weight of each aspect relative to the aspects of behavior as a whole. For example, when developing a mathematics test, each component of the test must include mathematical operations. We also ensure that all components combined address and cover only mathematical operations and not any other domain. Content validity is assessed by consulting experts specializing in the measured trait. There must be at least five experts. Items with ratings of 80% or more are selected (Ahmed, 1960).

Content validity can also be assessed through internal consistency, by calculating the correlation coefficients between the score on each item and the total test score (Awad, 1998).

(Lawshe, 1975) identified another method for calculating content validity; the CVR method requires experts to determine whether an item is necessary for measuring the content being assessed by the test. A panel of experts is asked to assign a score from 1 to 3, where 1 means the item is unnecessary, 2 means the item is useful but not essential, and 3 means the item is

essential. The CVR value ranges from -1 to 1. The closer the value is to 1, the greater the agreement among experts, and the more items should be included in the test. Conversely, the closer the value is to -1, the less agreement among experts; therefore, items should be removed or revised. The formula for determining the CVR is as follows:

$$CVR = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

Where:
CVR: Is content validity ratio,
n_e: Is the number of panelists indicating "essential",
N: Is the total number of panelists [N_2 is total number of panelists divided by 2].

Criterion-related validity

Criterion validity is the process or procedure taken by the test developer to obtain empirical evidence from the examinee's behavior and actions to support the results obtained using the test. Therefore, criterion validity refers to the judgment reached by the test developer regarding the possibility of inferring about the examinee based on their test results (Mkaddam, 2003).

In statistical terms, criterion validity can be considered the correlation between test scores and an external criterion or standard. Two types of criterion validity fall under criterion validity: concurrent validity and predictive validity (Abu al-Nil, 1987).

There are two types of criterion validity:

Concurrent validity

The correlation between a test and a criterion, provided that data are collected using the test and criterion at the same time or at close intervals. For example, a specialist developing a new scale may achieve its validity by collecting anecdotal data about the subjects in addition to the scores they obtained by administering the test. The criterion may be a clinical interview conducted with the subjects at the same time the test is administered. Therefore, the degree of correlation between the test scores and the interview results is considered an indicator of the test's validity (Faraj, 2007). Based on this concept, it can be said that concurrent validity means: the extent to which an individual's score on a test can be relied upon to judge its validity and put it to the test.

Predictive validity

The ability of a test or scale to predict how an individual will behave in a particular trait or characteristic in the future (Kline, 2015). Therefore, the criterion used to infer the test's validity is not at the same time the test is administered, but rather at a later time, no less than six months and possibly years later. Collecting data on a criterion requires the specialist to wait until the individual's status in the trait or characteristic predicted by the test becomes clear (Anastasi & Urbina, 2006).

One of the most prominent examples of predictive validity of tests is the graduate student test, which is used to predict the success or failure of students admitted to graduate programs at universities. The test is the predictor, while the students' success or failure in the program is the criterion. In this case, data on this test cannot be collected until a long period of time, sometimes years, after the students enroll in the program. They either succeed and complete the program, receiving their degree, or fail and withdraw from the program. Another indicator of predictive validity is when tests are used to predict success or failure at work (Maamria, 2012). A group of employees may be selected based on a personality or mental test. After being selected, they are

given the opportunity to work, and it is then determined who succeeds and who fails according to a criterion set to distinguish between successful and unsuccessful employees. The individuals' scores on the test, based on which their success or failure in the job is predicted, are then compared with the data collected about them in the criterion. Predictive validity is the relationship between test scores and the scores or data in the test. Therefore, the higher the correlation coefficient between them and the test, the more it indicates that the test and test measure the same trait or characteristic. However, if the correlation coefficient between them is low or inverse, this indicates that each measures something different from the other.

Construct Validity

Construct validity was a concept first introduced by Cronbach and Meehl (1955). To demonstrate the construct validity of a test it is necessary to delineate as fully as possible the variable (construct) which the test purports to measure (Kline, 2015)

Psychological characteristics, whether emotional or intellectual, can often be difficult to address directly. Therefore, they must be addressed indirectly through the behavioral effects and outcomes they represent. It has been emphasized that psychological characteristics are the product of cultural, scientific, and genetic development and growth. Therefore, these elements must be taken together when measuring psychological traits and characteristics, regardless of their nature: motivation, intelligence, self-esteem, ambition... These are all psychological characteristics that we cannot address directly, which necessitates addressing them indirectly by identifying a set of behavioral indicators that represent these characteristics. When developing a psychological scale, it may be necessary to verify the concept of this scale.

Construct validity can therefore be defined as the scientific judgment and conviction generated by the scale developer regarding the appropriateness of the conclusions reached through test scores relative to the level and position of the individual within a given characteristic. According to this definition, a psychological concept or construct can be considered a scientific idea employed to describe and explain the behavior we wish to measure (Faraj, 2007). For example, intelligence is a concept used to explain an individual's performance in school or even to explain their performance in life in general. Motivation is a concept used to explain an individual's behavior and performance in the workplace. Adjustment is a concept used to explain and understand an individual's behavior and performance and relationships at work, at home, in society in general, and with themselves in particular.

As a basic premise for achieving construct validity, the test developer must begin with assumptions that constitute a description and explanation of the behavior indicative of the trait or characteristic represented by the measure. This description must include the scale developer's expectations about the behavior of those who score high on the test and those who score low. From these assumptions, the foundations and beginnings of a theory about the nature of the trait or characteristic the scale aims to measure may be formed.

The scale developer must not be satisfied with these efforts; rather, they must retest their hypotheses to confirm the results they arrived at the first time. Many measurement experts view conceptual validity as the cornerstone of any measure developed. This group of specialists even considers other types of validity—content validity and criterion validity—to be types and models of construct validity (Maamria, 2007).

Construct validity can be calculated using several statistical techniques, most notably the t-test and factor analysis, in addition to various correlation coefficients (Ahmed, 1960).

Reliability

This means that a test produces the same results if administered to the same group over a period of time and under similar conditions (Hussein & Gasmalla, 2023).

"The reliability (or precision) of a sample value is the closeness with which it can be expected to approximate the relevant population value. It is necessarily an estimated value in practice, since the population value is generally unknown. Depending upon the statistic in question, and the specific statistical model on which the test is based, reliability may or may not be directly dependent upon the unit of measurement, the population value, and the shape of the population distribution. However, it is always dependent upon the size of the sample" (Cohen, 1988: 6).

Here, it is important to emphasize the need to select a large-sized standardization sample to ensure a high reliability coefficient.

It is noted that a valid test is often stable. Validity is a condition of reliability, as when a test measures the function it was designed to measure, it will produce similar estimates each time it is measured (Irwing, Booth & Hughes, 2018).

Steps have been proposed to ensure test reliability, including (Awad, 1998):

a) Making the test reasonably long to eliminate the factor of chance.
b) Scoring the test objectively.
c) Writing the test items and instructions clearly to prevent errors due to unclear instructions.

Reliability refers to the degree of accuracy, stability, and consistency in the results of an instrument if it is administered twice or more, on the same characteristic on different occasions. Reliability in psychological measurement is ensuring that approximately the same results are obtained if the test is administered again to the same individual or group of individuals (Gregory, 2015).

This means that chance factors are minimally affected by the test results. There are several concepts of test reliability that should be noted (Anastasi & Urbina, 2006; Abdul Rahman, 2008):

- The test must produce approximately the same results if administered repeatedly to the same group of individuals.
- There must be a legal relationship between the test items, indicating consistency in the test's internal structure. This means that the test's reliability coefficient depends on the correlation between the items (inter-correlations), and also depends on the correlation of each item with the total test score.
- Test reliability refers to an individual's actual (true) performance, expressed by the true score they obtain on a given test. This means that there is a possibility of errors in the measurement process that affect the test's stability, as the individual's score is actually composed of a true score and an incorrect (error) score, according to the following equation:

$$iTS = ts + es$$

where:

iTS: Individual's total score,
ts: True score,
es: Error score.

The difference between the total score and the actual score is called the *measurement error* (Abdul Rahman, 2008) (or standard error of the test).

Sources of Test Reliability Error

There are several sources of error that cause variation in the results of a given test. These can be summarized into two main sources:

a. Irregular (random) errors: These are due to chance factors and are represented by (Maamria, 2012):

- **The examinee's state:** When an individual takes a test, their score may be affected by numerous factors beyond the characteristic the test was designed to measure, such as distraction or mood swings. This may result in a score that is lower or higher than intended.
- **The examiner's state:** The way the examiner administers the test may affect the examinee's score. The examiner may be moody or aggressive, making the examinee feel threatened. The examinee may be anxious or not well-trained in test-administration procedures, or they may seek to embody certain attitudes and values, leading the examinee to suggest certain answers.
- **External conditions:** The external conditions surrounding the application process affect the score obtained by the examinee. For example, the location where the application is conducted may be familiar or unfamiliar, making the examinee feel fearful. The prevailing social situation during the application, such as being alone or with a group, may make the examinee feel unsafe or in a competitive situation.
- **Test:** Certain test-related factors affect the score obtained by the examinee, such as the clarity of the instructions, the style of wording of the items, and the time required to complete each item. The longer this time is, the more bored the examinee becomes. The same applies to the degree of similarity of the items in the case of a two-form test.

b. Systematic errors (Maamria, 2007): These are constant, consistent errors related to the validity of the test, as the test may consistently measure a trait other than the trait it was designed to measure. For example, a scale designed to measure extroversion also inadvertently measures anxiety consistently (regularly).

The following table summarizes the sources of error according to types of reliability (Anastasi & Urbina, 2006: 115):

Table 3. Sources of error variance in relation to reliability coefficients

Type of reliability coefficient	Error variance
Test-retest	Time sampling
Alternate-form (immediate)	Content sampling
Alternate-form (delayed)	Time sampling and content sampling
Split-half	Content sampling
Kuder-richardson and coefficient alpha	Content sampling and content heterogeneity
Scorer	Interscorer differences

Types of Reliability

Test-Retest Reliability

This method involves administering the same measurement instrument to the same group twice, then calculating the Pearson correlation coefficient between the two. Cronbach (1970) indicates that if the interval between the first and second administrations is between 1-3 months, a reliability coefficient of 0.70 is sufficient. However, if the interval is one or two weeks, the reliability coefficient should be 0.80 or higher (Maamria, 2007).

Alternate-Form Reliability (or Equivalent Forms)

This involves preparing and administering two equivalent forms of the same test, then calculating the correlation coefficient between them. Equivalence between the two tests means that they measure the same characteristic, have the same number of items in both forms, have the same instructions, and are given the same amount of time to complete. (Mkaddam, 2003)

There are two methods for using equivalent forms (Abdul Rahman, 2008):

- **Immediate method:** The two forms are presented in a single session, with the first form presented and the second form presented after completion.
- **Sequential (delayed) method:** The two forms are presented in two consecutive sessions, separated by a specific period of time (between a week and a month).

Split-Half Reliability

The degree of correlation between test items is estimated by splitting the test into two halves, using the first half versus the second half, or the odd-numbered versus even-numbered items. This creates two sets of scores (Anastasi & Urbina, 2006). The Pearson correlation coefficient is then calculated between these two sets, which is used to determine the reliability coefficient for the two halves of the test.

However, the reliability coefficient for the two halves of the test results in a significant decrease in the correlation coefficient. To avoid this decrease, statistical equations called length correction equations were developed. The most important and common of these equations are (Maamria, 2007):

- **Spearman-brown equation (1910):** This equation is used when the variance of scores across the two halves of the test is equal (in the case of homogeneity). Its formula is as follows:

$$R_{S-B} = \frac{2R}{1+R}$$

R: Is the correlation coefficient between the two halves of the test.

- **Gutman equation (1945):** If the homogeneity condition between test scores is not met, an alternative equation is used that takes into account a significant difference in variance, which is the Gutman equation. Its formula is:

$$R_G = 2\left(1 - \frac{S_1^2 + S_2^2}{S_t^2}\right)$$

Where:
S^2_1: Is variance of first half,

S^2_2: Is variance of second half,
S^2_t: Is test variance.

We also find the Rulon equation (1939), which is similar to the Gutman equation (except that the Gutman equation is easier to apply) (Maamria, 2007).

Internal consistency

The test is administered once, and the consistency of responses (the extent of their agreement and consistency with each other) for each of its items is estimated (Hussein & Gasmalla, 2023). Many equations have been developed, the most prominent of which are:

- **Kuder-Richardson 20 equation:** Kuder-Richardson reliability assumes that the test is uni-dimensional; that is, it measures one aspect of a single characteristic, and that all of its items only measure that characteristic. Kuder and Richardson developed a set of equations, all of which are based on calculating internal consistency between test items, the most prominent of which is the equation known by the abbreviation "K-R20." Its formula is as follows (Hussein & Gasmalla, 2023):

$$KR = \frac{N}{N-1} \times \frac{V - \sum_{i=1}^{n} p_i q_i}{V}$$

Where:
N: Is the number of items in the test,
V: Is the variance of examinees raw scores,
p_i: Is the proportion of correct answers of item *i*,
q_i: Is the proportion of incorrect answers of item *i*.

- **Cronbach's alpha coefficient:** It is considered one of the most important measures of internal consistency, as it links the stability of the test to the stability of its items. An increase in the variances of the items relative to the total variance of the test leads to a decrease in the stability coefficient, while a decrease in them (homogeneity) leads to an increase in the stability coefficient. Its formula is as follows (Maamria, 2007):

$$\propto = \frac{n}{n-1}\left(1 - \frac{\sum S_i^2}{S_t^2}\right)$$

Where:
n: Is number of test items,
S2i: Is total item variance,
S2t: Is test variance.

Norms

These are a type of scale used to interpret the raw scores obtained by individuals after taking a test. Norms are also developed based on the raw scores of a standard sample representing the population for which the measuring instrument was developed.

Raw scores are of two types: item scores and test scores. Although the item score is important as it constitutes a subunit of the test, the final result we often deal with is the total test score. When we obtain this score, we must seek to understand its significance (Fischer & Milfont, 2010).

A raw score does not acquire a specific meaning unless it is attributed to a reference that provides the appropriate interpretation.

There are two references for interpreting raw scores (Osadebe, 2014; Fischer & Milfont, 2010):

Criterion-referenced

This type is usually limited to achievement tests, and its use relies on a prior determination of the acceptable performance level for the examinee; that is, there is a prior plan for what performance should not fall below (the criterion). Therefore, the ability of all or a percentage of examinees to reach or exceed this level, or the failure of any of them to reach it, is not considered a determining factor for the desired performance level as appropriate performance. This means that we are dealing with an absolute concept that does not depend on variable factors related to the level of difficulty, the level of performance, or the nature of the sample.

The general principle in using criterion-referenced is to determine the criterion score (determining the performance level) for passing or failing. This score is governed by criteria that do not relate to the size or level of the examinee's performance or the changing circumstances of performance, but rather is governed by considerations related to the objectives to be achieved.

The criterion-referenced is a priori, absolute, and individual reference; that is, it is determined before the test is administered. It is absolute because all scores are interpreted according to it, and individual because each score is interpreted individually and according to the specified criterion, regardless of the performance of the group to which the individual belongs.

Norm-referenced

A norm is a standardized reference used to compare the performance of the individual to whom the test was administered with the results or performance of the individuals in the standardized sample, which must be representative of the original population. Therefore, the test must be administered to a large sample, increasing the likelihood of its representation.

The norm reference is a posteriori, relative, and group reference; that is, it is identified and determined after the test is administered. Relative refers to the variable that varies from one sample to another, and group reference because each individual's score is interpreted according to their position relative to the group standard of scores of other individuals who took the same test.

Importance of norms

The importance of norms can be summarized as follows (Mkaddam, 2003):

- They are the basis for internally judging a phenomenon.
- They are usually quantitative in nature, indicating the individual's position relative to the group.
- They are determined in light of the actual characteristics of the phenomenon (how far the individual is from the average of the group to which they belong).
- They reflect the individual's current level.
- They are a means of comparison and evaluation.
- They are important in tests that take the form of a battery.
- They can be used to predict and diagnose strengths, weaknesses, and other aspects.

Standard score

Standards are established using two methods: percentage scores and standard scores (Abdul Rahman, 2008). However, the most common method is standard scores, where the standard score is derived by converting the raw score to a standard score, according to the following equation:

$$Z = \frac{x - \bar{x}}{S}$$

Where:
Z: standard score,
x: raw score,
$\bar{x}$: arithmetic mean,
S: standard deviation.

The standard score is defined as a score derived from the raw score and is one of the important tools used by the standard reference to interpret raw scores. Two objectives are achieved through standard score (Anastasi & Urbina, 2006):

- To determine the relative position of the examinee among the individuals in their sample, and to interpret their score based on the performance of examinees from the same sample.
- To enable comparison between individuals, as the raw scores of individuals are ranked against their corresponding standard scores.

When calculating standard scores, we use two fixed derived scores for examinees who took the same test: The arithmetic mean and the standard deviation for the same group of examinees.

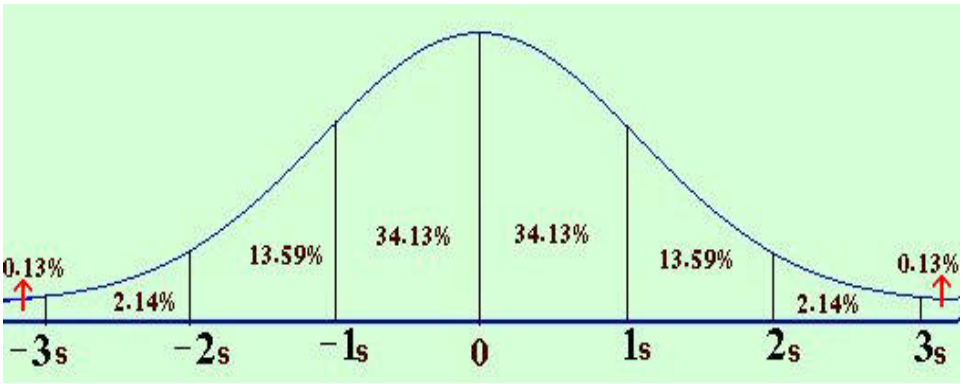

Figure 2. Normal distribution curve (Al-jamasi, 2024)

The previous figure represents the normal curve and the percentages of scores between the arithmetic mean and each standard deviation, and between each standard deviation and the next in both the negative and positive directions from zero, which represents the arithmetic mean.

The normal curve shows that the percentage of individuals falling between -1s and +1s was 68.26%, representing the majority of the sample. The percentage of individuals falling between -2s and +2s was 95.44%, and the percentage of individuals falling between -3s and +3s was 99.72%.

And the next figure shows the normal curve with different scores:

Spor Eğitimi ve Bilim Dergisi
Journal of Sport Education and Science
Cilt: 1, Sayı: 2, 2025
57-75

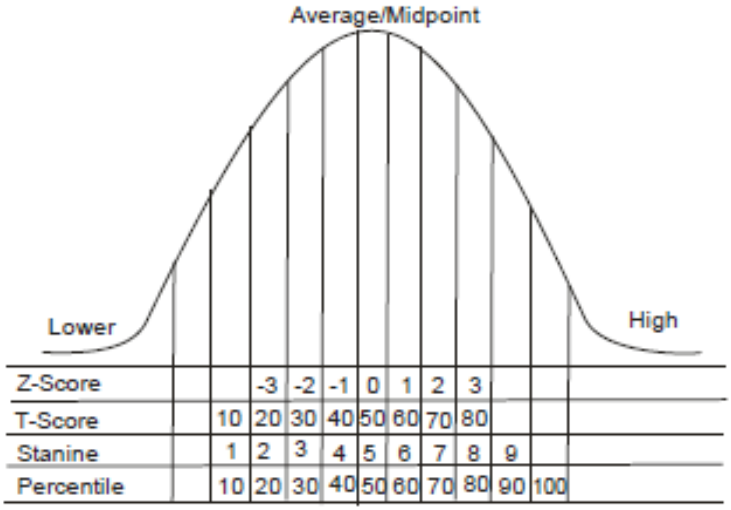

Figure 3. Normal curve with z-score, t-score, stanine and percentile (Osadebe, 2014: 102)

Adjusted Standard Scores

It is noted that the standard scores deviate from the arithmetic mean (represented by the score zero) by only three points, whether in the positive or negative direction. This situation poses a challenge in the scientific interpretation of these scores. Therefore, it was necessary to adjust the standard scores to eliminate small values, negative values and zero, which represents the moderate (average) score.

Numerous models of adjusted standard scores have emerged, all based on the same previous equation: By adding an arithmetic mean greater than zero and a standard deviation greater than one.

One of the most common forms of adjustment is (Maamria, 2012):

The t-score

The arithmetic mean is 50 and the standard deviation is 10; the equation becomes as follows:

$$\mathbf{Z_t = \frac{X - \bar{X}}{s} \times 10 + 50}$$

The modified standard score on the wechsler scale

The arithmetic mean is 100 and the standard deviation is 15; the equation becomes as follows:

$$\mathbf{Z_w = \frac{X - \bar{X}}{s} \times 15 + 100}$$

T**able 4.** Correspondence table between different calibrations

Deviation from the mean (x)	**-3**	**-2**	**-1**	**Mean 0**	**+1**	**+2**	**+3**
		99.72% of the sample					
			95.44% of the sample				
				68.2% of the sample			
%	**0.13%**	**2.14%**	**13.52%**	**34.13%**	**34.13%**	**13.52%**	**2.14%**	**0.13%**
T score	**20**	**30**	**40**	**50**	**60**	**70**	**80**
w.s. score	**1**	**4**	**7**	**10**	**13**	**17**	**21**
W score	**55**	**70**	**85**	**100**	**115**	**130**	**145**

Reference: (Bernaud, 2014: 133)

The table above shows the adjusted standard score according to various common scales in the psychological and educational field, where one of them can be chosen as a standard to convert the raw score into an adjusted standard score; this will allow the possibility of interpreting this score according to the chosen standard, and determining the individual's position within the reference group.

CONCLUSION

Standardizing psychological and educational tests is a crucial step to ensure their validity and accuracy in measuring what they are designed to measure. Standardization guarantees highly reliable and stable measurement results, allows for precise quantitative estimates, and enables objective and impartial interpretation of the results.

Standardization consists of several processes. The first is item analysis, which includes steps such as calculating item difficulty indices and determining the validity and reliability of each item. A very important step is determining the psychometric properties of the test, which involves calculating the test's validity and reliability indices. Another equally important step is establishing test norms. This helps the researcher determine individuals' levels in the trait being measured and their position within the sample. It also helps in providing appropriate interpretations based on the characteristics of the research population.

All of the above steps require considerable time and effort and are also costly. Therefore, some researchers may be content with simply extracting the psychometric properties of a test to save time and money. However, failing to adhere to the standardization process weakens the test's accuracy, which in turn affects the validity of the results and their ability to accurately describe the characteristic the test was designed to measure.

To ensure the standardization process is carried out correctly, by following all the necessary steps, we propose the following:

- Allocating a budget specifically for standardization costs, funded by the university under the auspices of the Ministry of Higher Education and Scientific Research.
- Collaborating between laboratories and research units across different universities in selecting standardization samples. This ensures the test's reliability and its applicability throughout the country.
- Computerizing test standardization projects proposed by researchers. This will prevent different researchers from standardizing the same test simultaneously.

- Providing specialized training in statistics and psychometrics for researchers interested in standardizing tests.

REFERENCES

Abdul Rahman, S. (2008). *Psychometry: Theory and application* (5 ed.). Cairo: Hiba Al-Nil Arab Publishing and Distribution.

Abu al-Nil, M. S. (1987). *Psychological, social, and educational statistics.* Beirut: Dar al-Nahda al-Arabiya.

Ahmed, M. A.-S. (1960). *Psychological and educational measurement* (1 ed.). Cairo: Al-Nahda al-Masriya Librar.

Al-jamasi, M. S. (n.d.). *Standard scores*. Retrieved 01 22, 2024, from Al-Jamasi Family Association. https://jmasi.net/ehsa/normald/z_scores.htm

Anastasi, A., & Urbina, S. (2006). *Psychological testing* (7 ed.). Pearson Prentice Hall.

Awad, A. M. (1998). *Psychological measurement between theory and practice.* Egypt: Dar Al-Ma'rifa Al-Jami'a.

Bademci, V. (2022). Correcting fallacies about validity as the most fundamental concept in educational and psychological measurement. *International e-Journal of Educational Studies, 6*(12), 148-154. https://dergipark.org.tr/en/pub/iejes/issue/73263/1140672

Bernaud, J.-L. (2014). *Méthodes de tests et questionnaires en psychologie.* France: DUNOD.

Boucenna, M. (2012). *Psychometrics: Basic principles* (2 ed.). Algeria: University Publications Office.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2 ed.). USA: Lawrence Erlbaum Associates, Publishers.

Cooper, C. (2023). *An introduction to psychometrics and psychologicql assessment: Using, interpreting and developing tests* (2 ed.). London and New York: Routledge Taylor and Francis Group.

Djedidi, A. (2018). The importance of edumetry (educational measurement) and evaluation and their role in the educational process. *The Didactic Journal, 5*(13), 188-198. https://asjp.cerist.dz/en/article/44186

Djedidi, A. (2020). Necessity of adapting psychological and educational tests and scales for use in academic studies. *Scientific Observer Journal, 7*(2), 108-122. https://asjp.cerist.dz/en/article/128893

Faraj, S. (2007). *Psychometry* (6 ed.). Cairo: Anglo-Egyptian Library.

Fischer, R., & Milfont, T. L. (2010). Standardization in psychological research. *International Journal of Psychological Research, 3*(1), 88-96. https://www.researchgate.net/publication/47386955

Gregory, R. J. (2015). *Psychological testing.* England: Global Edition. Pearson.

Groth-Marnat, G., & Wright, A. J. (2016). *Handbook of psychological assessment* (6 ed.). USA: John Wiley & Sons, Inc.

Hussein, A., & Gasmalla, H. E. (2023). Introduction to the psychometric analysis. In *Written Assessment in Medical Education* (pp. 111-135). https://www.researchgate.net/publication/368340116

Irwing, P., Booth, T., & Hughes, D. J. (2018). *The wiley handbook of psychometric testing: A multidisciplinary reference on survey, scale and test developement.* USA: Wiley Blackwell.

Kline, P. (2015). *A handbook of test construction: Introduction to psychometric esign.* London and New York: Routledge Taylor and Francis Group.

Lawshe, C. (1975). A quantitative aproach to content validity. *Personnel Psychology*(28), 563-575. https://doi.org/10.1111/j.1744-6570.1975.tb01393.x

Maamria, B. (2007). *Psychological measurement and the design of its instruments* (2 ed.). Algeria: Al-Hebr Publications.

Maamria, B. (2012). *Fundamentals of psychological measurement and the design of its instruments.* Algeria: Dar Al-Khaldouniya.

Mkaddam, A. H. (2003). *Statistics and psychological and educational measurement* (2 ed.). Algeria: University Publications Office.

Osadebe, P. U. (2014). Standardization of test for assessment and comparing of students' measurement. *International Education Studies, 7*(5), 94-103. http://dx.doi.org/10.5539/ies.v7n5p94

Thorndike, E. L. (1922). *The psychology of arithmetic.* New York: The Mcmillan Company.

Urbina, S. (2016). Psychological testing. In H. S. Friedman, & C. Reserved. (Ed.), *Encyclopedia of mental health* (2 ed.). Academic Press Publisher. https://www.researchgate.net/publication/286780211_Psychological_testing