PSYCHOLOGY AND EDUCATION (2024) 61(10): 555-575
ISSN: 1553-6939

Psychometric Properties (Validity and Reliability)

MEGA FAIZA[1], Oulidi Rabia[2], Khelil Amina[3] , Souad ben bordi[4]

[1-4]University of El Oued , Algeria

laboratory of social development and society service

Received: 22/06/2024, Accepted: 13/09/2024; Published: 01/10/2024

Abstract:

This study aimed to focus on the psychometric properties of research tools, as they are considered one of the first steps for any study to measure the validity and reliability of the scale adopted for the study. Many studies on this topic, including this study, have focused on trying to provide a sufficient and detailed explanation of the psychometric properties of any measurement tool.

Keywords: Validity. Reliability

Introduction

One who reflects on contemporary society will find that one of its most defining characteristics is, first, the precision in measuring things, and second, the expansion of measurement to encompass areas that, until just a few decades ago, were thought to be beyond measurement. Among these areas are mental and psychological characteristics. Third, the progress made in the sciences has been largely due to the precision in measurement that distinguishes them. The rapid advances in psychology, sociology, and educational sciences since the beginning of the 19th century have been made possible by the development of measurement tools. Since the discovery of individual differences in mental abilities, as well as psychological, social, and educational traits, psychologists and educators have devoted their efforts to designing psychological, educational, and social measures and tests that assess the various phenomena related to these sciences. This has led to the emergence of hundreds of tests used for various purposes in fields such as selection, guidance, vocational training, psychological counseling, educational and clinical diagnosis, and more. In this context, psychological, educational, and social tests and measures have become indispensable tools for any researcher in psychology, education, and sociology. Human behavior, in all its forms, is now subject to measurement and evaluation, as most judgments and decisions related to selection, diagnosis, guidance, and the prediction of future human behavior depend on the results of psychological, educational, and social measurements and tests.

In order to obtain accurate information and data about behavioral phenomena, researchers use a number of psychological tools such as tests, measures, observation checklists, self-assessment scales, and rating scales. When a researcher uses a test to obtain information that helps in making a certain

decision, they face a fundamental problem related to selecting the test that can truly assist in making the decision. There may be several alternative tests available that can be used for decision-making purposes, and many factors should be taken into consideration when evaluating the quality of a given test. These factors are divided into three groups:

The first group: General conditions of the psychological test, which include: comprehensiveness, standardization, and objectivity.

The second group: Experimental conditions of the psychological test, which include: reliability, validity, and norms.

The third group: Practical considerations and the test's usability.

Psychometric properties are important characteristics that must be present in any measurement tool, whether they are personality tests, achievement tests, questionnaires, etc., in order to make them valid for application in a research environment, meaning that they can be relied upon for making decisions. Without these properties, it is impossible to trust the tool's ability to measure what it was designed to measure, and the results obtained from using it to assess various traits cannot be considered highly accurate or objective. These properties include:

- **Objectivity:** This means that the results obtained should not be influenced by the subjectivity of either the examiner or the examinee.
- **Comprehensiveness:** This means that the tool should be comprehensive and representative of all components of the ability or trait being measured.
- **Reliability:** Reliability is essential in psychological and educational measurement. It means that the test yields approximately the same results if reapplied to the same group of individuals.
- **Validity:** Validity is another concept that measurement experts have extensively studied due to its importance in decision-making related to individuals and groups. It is not linked to the measurement tool itself but rather to the way the scores derived from the tool are interpreted.

In this research, we will address the study and clarification of the concept of psychometric properties of measurement, focusing on the most important ones: reliability and validity.

First: The Nature of Psychometric Properties of the Measurement Tool

1. The Concept of Psychological (Psychometric) Measurement:

Measurement, as an empirical concept, has several definitions, and perhaps the simplest one is that it describes data related to the characteristics of objects using numbers or quantitative aspects in describing traits or characteristics of individuals.

In this context, Adam Smith (1972) states: "Measurement, in its broad sense, is the systematic collection of information in a specific order, which includes the process of gathering and organizing information and the outcome of this process."

It is also defined as "the assignment of numbers or numerical symbols to objects or events according to rules. Thus, measurement requires specific procedures or operations based on defined rules used to compare objects or events according to a clearly defined **standard** or **scale**."[1]

As for psychological measurement or psychometrics, it is a field concerned with the theories and mechanisms of measuring perceptions, which include measuring knowledge, mental abilities, attitudes, personality traits, and learning assessment. Those working in this field are involved in preparing **Questionnaires** and verifying their validity, as well as conducting personality assessment tests.

The goal is to transform qualitative, intangible, and invisible psychological and social phenomena into quantitative, tangible, and visible phenomena according to specific rules (e.g., assigning a score for intelligence).

Some of the key assumptions of measurement theory in psychology, educational sciences, and sociology are:[2]

1. Performance can be measured and evaluated, meaning that human performance can be transformed from its descriptive form into a quantitative form using numbers, according to specific rules.
2. Human performance is a function of their characteristics, meaning that every performance or behavior stems from one or a set of characteristics that distinguish an individual from others.
3. The characteristic, performance, and the relationship between them vary from one response to another, and this variation forms the basis of the measurement process (individual differences).
4. Considering measurement error, meaning that every score on the measurement scale consists of a true score and the portion attributed to measurement error.

Psychological, educational, and social measurement has several characteristics, including:

- Psychological, educational, and social measurement is relative, not absolute.
- There are situations where an individual's performance can be compared to an agreed-upon absolute standard.
- Psychological, educational, and social measurement lacks specific units of measurement; for example, there is no unit to measure achievement in mathematics or verbal ability.
- There are no fixed rules for judgment in the field of educational, psychological, and social measurement, even though Stevens mentioned rules for judgment, he did not define the nature or limits of these rules.

2. The Concept of Psychometric Properties:

Psychometric properties are a set of indicators that reflect the reliability of the test results, the stability of its outcomes, and their consistency. They also represent the foundations on which the test relies for interpreting its results. Some have defined the psychometric properties of a test as "those essential

characteristics related to the effectiveness of the test items, as well as validity, reliability, and associated factors such as discrimination indices and levels of ease and difficulty in the case of achievement and aptitude tests, along with standards for interpreting results, which are verified after applying the test experimentally to a representative sample of the population known as the standardization sample. The quality and objectivity of the test depend on the extent to which appropriate levels of these properties are present."

From this definition, the psychometric properties of psychological tests can be analyzed into the following: test validity, test reliability, discrimination indices, levels of ease and difficulty, and standards for interpreting results. All these indicators are extracted after applying the test to a sample called the standardization sample, considering that the quality and objectivity of the test depend on the extent to which acceptable levels of these properties are available.[3]

Second: Reliability

Reliability is considered one of the fundamental concepts in psychometric measurement. It is a relative characteristic of measurement tools, meaning that a measurement tool is reliable under certain conditions but can vary with changes in the sample's characteristics and the surrounding factors. Below, we will address the definition of reliability, the methods for calculating it, and the factors that influence it.

1. Definition of Reliability:

Reliability is defined as "the degree of consistency or uniformity between the results of two measurements in assessing a certain trait or behavior. In light of this, it is expected that an individual's scores remain stable if they are similar under slightly different measurement conditions."[4]

For example, if we measure a person's weight twice, we would hope to get nearly the same results if we use a different scale or weigh the individual after just one day.

From a statistical perspective, reliability is "the ratio of the variance of true scores to the variance of observed scores."[5]

Furthermore, since the goal of measurement is to compare individuals to identify individual differences and to determine the extent of variation between them, and because this variation involves random factors present in some individuals and absent in others, reliability becomes a crucial factor in understanding these differences.

Thus:

Total variance = True variance + Error variance

- **True variance**: This refers to the variation or difference in individuals' scores that is attributed to their personal traits and stable characteristics.
- **Error variance**: This refers to the variation or difference in individuals' scores that is attributed to random factors.

The reliability coefficient provides us with a quantitative estimate of the relationship between individual differences in true scores and differences in observed scores. When we obtain an observed score free of errors, the test reliability is perfect, equal to (+1), meaning that the differences between true and observed scores are equal. Therefore, the value of the reliability coefficient ranges between (0 and +1), reflecting the degree of test consistency. It can be interpreted as the proportion of the observed differences between individuals' scores that is due to true differences among them. In other words, the reliability coefficient equals:[6]

$$\text{Reliability (r)} = \frac{\text{True score variance}}{\text{Observed score variance}}$$

Test reliability is viewed from three aspects:

1. It means that when we measure a certain trait two or more times using the same scale or test, or using two equivalent measures, we obtain the same results.

2. It indicates that the results of the measured trait are indeed what the test measures, referring to the accuracy of the scale.

3. It refers to the possibility of errors in the measurement process that affect the stability of the scale. Accordingly, an individual's score on a particular scale actually consists of a true score and an error score, expressed by the following equation:

$$S = TS + ES$$

Where : **S**: The total score obtained by the individual. , **TS**: The true score. , **ES**: The error score , **OS**: The observed score

$$ES = TS + OS$$

Thus, reliability is an attempt to obtain consistent TS (true scores) because we cannot consider reliability as simply producing the same results every time, as random error (ES) does not repeat itself to the same degree when the test is reapplied. Since the obtained score is not the true score, the raw total score (S) cannot be considered a constant standard because it includes the error score, which changes with the reapplication of the test. Therefore, we must seek the true scores of individuals, rather than the total score. Additionally, the reliability coefficient ranges between (0 and +1), meaning that if the test results approach +1, the test has a high degree of reliability, and conversely, if the results approach 0, the test has a low degree of reliability

2. Methods for Calculating the Reliability Coefficient:

There are several methods for calculating the reliability coefficient for personality tests. The researcher chooses one method, although they may also use more than one method in their selection. However, the essential condition is that the chosen method must be appropriate to the nature of the test. These methods are as follows:

Test-Retest Method:

This method is also called the "stability reliability method." In this method, the test is administered to a sample of individuals, and after a certain period, the same test is reapplied to the same sample under similar conditions to those of the first administration. Then, the appropriate correlation coefficient between their performance in both instances is calculated.[7]

For example, in a study on depression in adults, we might administer the Beck Depression Inventory to the study sample, and after a certain period, we re-administer the test to the same sample and calculate the correlation coefficient between the two administrations.

This method is rarely used by teachers in achievement tests, but it is commonly applied in psychological tests, such as personality tests, as well as when assessing attitudes, interests, values, and even abilities and aptitudes.[8]

In this method, special care must be taken when selecting the time interval based on the nature of the trait being measured and the characteristics of the target population. It is common to calculate the reliability of a test for personality tests with a time interval ranging from two weeks to two months, although some studies have used longer intervals of up to a year.

In summary, reliability according to the test-retest method means obtaining the same or similar results when the test is administered and re-administered. The reliability coefficient obtained here is the correlation coefficient between the scores of a group of individuals on the same test in both administrations. This correlation differs depending on the level of measurement: if the data is ordinal, Spearman’s rank correlation coefficient is used, while Pearson’s correlation coefficient is used for ratio or interval data. The correlation will be high if the individuals' responses are consistent in both instances, and therefore, the reliability of the scores will be high.

In addition to the above, this method faces several difficulties, the most important of which are the following:[9]

- It is difficult to standardize the natural or physical conditions in which the test is administered on both occasions. For example, the first administration may occur early in the day when individuals are mentally and physically alert, while the second administration may take place later in the day, when individuals are fatigued and tired.
- The psychological test situation is a learning experience, as individuals become familiar with the test during the second administration. Therefore, during the second administration, they are not as affected by the novelty of the situation or the emotional tension that might have been triggered by the first administration.
- The second administration may involve a training effect resulting from the first administration. However, since the first administration is a learning or training situation, the opportunity to benefit from the training is the same for all individuals in the sample, meaning that it should not affect the calculation of the correlation coefficient between the two administrations.

- The time interval between the first and second administration can pose a problem. If the interval is too short (a few days), the factor of memory or benefit from the first administration may come into play. However, if the interval is too long (several months), there is a higher likelihood of the trait changing, especially in young children.

Thus, the test-retest method may be suitable for some tests and unsuitable for others, which necessitates the existence of other methods to assess reliability.

The Parallel Forms Method:

This is another method for calculating the reliability coefficient of a test, where two parallel forms of the test are prepared and administered to the same group, and the correlation coefficient between them is calculated. Parallel forms mean that both versions of the test should be equivalent in terms of the number of questions, the components of the construct being measured (for example, a personality test consisting of several subscales measuring different traits like introversion, extroversion, dominance, etc.), the number of questions for each component, the difficulty level of the test questions, the phrasing of the test items, the instructions for administering the test, and the method of scoring. Additionally, the mean scores of individuals on both forms should be equal, as should the variance of their scores.

This means that the two test forms should be identical in all aspects so that if they are administered to the same individuals (sample), they should obtain the same score on both.[10]

This method has an advantage over the previous method as it eliminates the effects of training, memory, and forgetting. It also helps avoid the issue of time intervals and prevents the need to administer the same questions twice in a short period, which is referred to as the **equivalence coefficient**.

If both forms are administered over a long period (not less than a week and not more than six months), the resulting correlation coefficient is called the **stability and equivalence coefficient**.

However, this method has its difficulties, mainly in the inability to create equivalent forms for many tests. This is often because preparing equivalent forms doubles the effort, time, and cost involved in constructing the test. Thus, the parallel forms method is costly in terms of resources and physically demanding, especially in psychological tests that require significant effort. Preparing just one psychological test requires a great deal of work, let alone preparing multiple tests to measure the same trait.

The Split-Half Method:

It may be difficult for a researcher to administer two equivalent tests to students, or it may be impractical to test students twice with the same test. If the test is re-administered, some students may become aware of this and may not cooperate, or they may lose interest in the assessment and answer the questions randomly. Thus, the researcher may resort to using this method.

In this method, the test is administered to individuals in a single session. When scoring, the test is divided into two halves, and each individual is given a separate score for each half. Then, the

correlation coefficient between the scores of both halves is calculated. However, the resulting correlation is only the **reliability coefficient for half the test**.

As for dividing the test, it can be done by splitting the test items into odd and even numbers (e.g., 1, 3, 5... or 2, 4, 6...), or by dividing the test into two equal or equivalent halves (the first half and the second half). The odd-even split is considered better than simply splitting the test in half because questions often progress from easier to more difficult. Additionally, when the examinee is tired, they may be more prone to errors and less inclined to focus.

The correlation coefficient calculated using this method is referred to as the **split-half reliability coefficient** or the **internal consistency coefficient**.

To obtain an estimate of the reliability of the entire test, meaning to adjust or increase the correlation coefficient for half the test to the full expected value for a longer test, this can be done using the following method:

- **Spearman-Brown Formula:**

It is a mathematical formula proposed by Spearman and Brown that can be used to estimate the reliability coefficient of the overall test scores based on the correlation coefficient between the scores of the two halves of the test. The formula is as follows:

$$\text{Reliability coefficient of the test scores} = \frac{2 \times \text{Correlation between test halves}}{1 + \text{Correlation between test halves}}$$

The meaning is that we calculate the correlation coefficient between the first half and the second half of the test, then apply the Spearman-Brown formula. For example, if the correlation coefficient between the two halves of the test is approximately **0.75**, we can then use the formula to calculate the overall test reliability.

$$\text{The estimated reliability coefficient of the overall test} = \frac{0.75 \times 2}{0.75 + 1} = \frac{1.50}{1.75} = 0.86$$

This method is not suitable for calculating the reliability coefficient of tests that cannot be divided into equivalent parts, nor is it suitable for timed tests that rely heavily on the speed of responses, as the large number of unanswered questions at the end of each test affects the correlation between the two parts and, thus, alters the reliability coefficient.

It is essential to ensure that the variance of the scores from both halves of the test is equal when using this method. If the variance differs, other methods can be used, such as the following:

- **Guttman's Formula:**

Guttman suggests that the condition of equal variance between the two halves of the test, which the Spearman-Brown formula relies on, can be disregarded when the variance of each half is calculated separately. He proposed his general formula to correct the reliability values when using the split-half method, which is as follows:

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

Where:

- **R**: The reliability coefficient of the test.
- S_1^2: The variance of the scores for the first half.
- S_2^2: The variance of the scores for the second half.
- S_K^2: The variance of the scores for the whole test .

Thus, this formula is used when the variances of the individuals' scores on the two halves are not equal.

- **Rulon's Formula:**

This method aims to simplify the Spearman-Brown formula by calculating the variance of the differences between the scores of the two halves and the variance of the total test scores. Rulon's idea is summarized in the following equation:

$$\boldsymbol{R} = 1 - \left(\frac{\mathrm{S}^2(1-2)}{\mathrm{S}^2}\right)$$

- **R**: The reliability coefficient of the test.
- $\mathrm{S}^2(1-2)$:Refers to the variance of the difference between the scores of the two halves.
- S^2: The total variance of the test scores.

- **Gulliksen's Formula for Timed Tests:**

To calculate the reliability of timed tests, where a certain percentage of questions are left unanswered due to most individuals being unable to respond within the limited time, we use the formula proposed by Gulliksen, which is as follows:

$$\boldsymbol{R'} = R - \frac{AU}{AU^2E}$$

Where:

- **R'**: The reliability coefficient for timed tests.
- **R**: The reliability coefficient between the two parts of the test as calculated using the Spearman-Brown formula.
- $AU = The\ average\ number\ of\ unanswered\ questions = \frac{Number\ of\ unanswered\ questions\ for\ all\ students}{Number\ of\ students}$
- AU^2E : Error variance, which is calculated by recording incorrect answers and omitted questions by the students.

When administering the test, we may encounter a large number of unanswered questions, which affects the correlation between the two halves and, consequently, alters the reliability coefficient.

In summary, the split-half method, according to some, is used in content-homogeneous tools that measure a unidimensional trait. If the tool is not homogeneous, the homogeneous questions can be grouped together, or the tool can be divided into homogeneous sub tools based on their content, then each sub tool is split into two equivalent halves. The questions in both halves should be arranged according to their level of difficulty.

This means that when we divide a test into two halves (odd-even), the first half should be equivalent to the second half. For example, if the first question in the first half is difficult, then the first question in the second half should also be difficult.

There are several observations regarding the use of the split-half method:[11]

The first half may differ from the second half, especially if the items are taken from (1-50) and then from (51-100). This means that individuals' responses in the second half will be more influenced by factors such as fatigue, boredom, and time pressure compared to their responses in the first half, leading to results that may not be highly reliable.

- When dividing the test into two halves by using odd-numbered and even-numbered questions, it is possible that the variance of scores in the first half may differ from the variance of scores in the second half (see Guttman's formula).
- This method has the advantage of allowing the reliability coefficient to be determined from a single administration, avoiding the need for a retest or the creation of equivalent forms, and eliminating concerns related to the time interval that must be considered.

The Internal Consistency Method – Kuder-Richardson (Kuder & Richardson):

This method relies on the same principle as the split-half method, as it is a method for measuring reliability through homogeneity. Many years ago, in 1937, Kuder and Richardson made an important contribution to the field of reliability estimation by developing methods to estimate reliability from a single administration of the test.

Kuder and Richardson proposed two mathematical formulas known as Kuder-Richardson Formula 20 (K-R20) and Kuder-Richardson Formula 21 (K-R21):

- **Formula 20:** This name refers to the fact that the formula was the 20th in a series of equations included in the famous article published by Kuder and Richardson. The formula is as follows:

$$R = \frac{N}{N-1}\left[\frac{S^2 - \sum(Nc - Ne)}{S^2}\right]$$

Where:

- R: The reliability coefficient of the test.
- S^2: The variance of individuals' scores on the test.

- N: The number of test items.
- Nc: The proportion of correct answers.
- Ne: The proportion of incorrect answers.
- **Formula 21:** This formula is a special case of reliability formulas. It does not require calculating the proportions of correct and incorrect answers for each item. It is represented as follows:

$$R = \frac{N}{N-1}\left[1 - \frac{M(N-M)}{S^2}\right]$$

Where:

- **N**: The number of items.
- **M**: The mean score.
- S^2: The total variance.

This method is used for content-homogeneous tools that measure a unidimensional trait.[12]

- It is used to estimate the reliability of power tests only and is not suitable for tests that depend on speed.
- It is used for tools where the scoring system is binary (either one or zero).
- The number of unanswered questions should not be large.

This means that this method involves administering the test only once, for the purpose of ensuring that all components of the test measure the same trait. It assumes the presence of internal consistency among the items, meaning that the difficulty and ease indices must be very similar.

Cronbach's Alpha Coefficient (∝):

Cronbach's Alpha, usually denoted as ∝, is a measure of the internal consistency of a test, as it relates the reliability of the test to the consistency of its items. An increase in the variance of the items relative to the total variance leads to a decrease in the reliability coefficient.

The mathematical formula for Cronbach's Alpha is:

$$R = N - 1\left[1 - \frac{S^2\sum(N)}{S^2}\right]$$

Where:

- **R**: The reliability coefficient of the test.
- **N**: The number of items.
- $\sum S\text{^}2$: The sum of the variances of each item in the test.

- S^2: The total variance of the test scores.

This method is used when there is a similarity in levels of difficulty and ease, there are not many unanswered items, and multiple-choice answers are used, such as: disagree, agree, strongly agree, neutral.[13]

In summary, when comparing reliability coefficients or mentioning one, we must indicate the method used to derive it, the formula applied, and the conditions under which the experiment was conducted to calculate it. Each reliability coefficient has a specific meaning, and it should only be used within that context. The logic of reliability is based on the correlation of the test. If the difficulty or ease indices of the test items are different and varied, we will find that the correlation and reliability coefficients will be low. Therefore, it is essential to choose the appropriate statistical method based on the nature of the test.

The reliability coefficients of personality tests usually range between 0.70 and 0.80, although values lower than 0.70 or higher than 0.80 can also be found. However, these values are relative, and when interpreting any value, the researcher must consider the following:

- The nature of the decisions they intend to make based on the results obtained.
- The nature of the phenomenon: if it is rapidly changing, a lower correlation value can be accepted. However, if it is relatively stable, a higher correlation coefficient is required for decision-making or interpretation.

3. Factors Affecting Test Reliability:

When interpreting the estimated values of the reliability coefficient obtained using any of the aforementioned methods, especially when comparing the reliability coefficients of different tests, the following factors should be considered:

3.1. Test length:

This refers to the number of items in the test. The relationship between the number of items and the reliability coefficient is a direct one—meaning that the more items a test has, the higher the reliability coefficient, provided the test is not excessively long to the point where the examinee feels bored.

3.2. Guessing:

Reliability decreases with increased guessing, because an answer based on guessing in the first administration of the test is unlikely to be based on the same guess in the second administration with the same group. Tests most affected by guessing are those requiring "yes" or "no" answers, as well as multiple-choice tests.

3.3. Variance:

Reliability is related to the variance of the test. The greater the variance, the higher the test's reliability, and the lower the variance, the lower the reliability. Thus, a test with questions that are all difficult or all easy will have weak reliability, and vice versa.

3.4. Test time:

The reliability coefficient continues to increase as the time allocated for the test increases, up to a certain point. If the test duration is excessively long, the reliability starts to decrease. This optimal point varies from test to test.

Reliability is also influenced by the nature of the sample. For instance, a test whose reliability coefficient is calculated for a university student sample should not necessarily be considered reliable for a sample of workers, due to the fundamental differences between the two groups. Therefore, reliability should be recalculated for the worker sample.

In conclusion, after reviewing the different methods that any researcher might need, along with the scope and limitations of each method for calculating test reliability, we have also discussed the factors that influence test reliability to help researchers control them and minimize their impact on reliability.

Third: Validity

Validity is one of the psychometric properties of psychological tests, particularly personality tests. It is an essential criterion for judging the appropriateness of these tests. Below, we will discuss the definition of validity, methods for calculating it, and the factors that influence it.

1. The Concept of Validity:

There are several definitions of validity, including:

- "It is the extent to which the test truly or accurately measures the trait or phenomenon for which it was designed."[14] For example, an agility test is considered valid if it accurately measures agility, and invalid if it measures something else.
- "It means that the test measures the function it is supposed to measure and gives a score that represents the individual's true ability in the trait or function being measured."[15]

In general, there are several key concepts related to test validity, meaning that a test is not considered valid unless the following conditions are met:

1. The test must be capable of measuring what it was designed to measure. This means that the test must be closely related to the ability it aims to assess. For example, a test designed to measure mathematical ability must clearly assess that ability, which can be determined by how well it relates to the components and elements of mathematical ability.
2. The test must measure only what it was designed to measure. In other words, the test should be able to distinguish between the ability it assesses and other abilities that might interfere or overlap with it. For example, a test of mathematical ability, besides measuring this ability, should measure it exclusively, without being influenced by other abilities such as linguistic ability.

A test may measure what it was designed to measure but still not achieve the intended purpose. Even in the case of physical objects, validity can never be 100%, so how could it be for psychological phenomena? It is impossible for a single test to encompass all indicators of a particular phenomenon,

as it is difficult to identify all indicators, and even if we could, it would be challenging to create a test that covers them all. Therefore, the first characteristic of validity is that:

- **A relative attribute**: There is no test that is completely invalid or perfectly valid if it is used for the purpose it was designed for. For example, an intelligence test designed for second-grade students may have high validity when used with second-grade students, but its validity may decrease when used with another sample. Thus, validity is not absolute but varies from one test to another. We cannot say that a test is valid or invalid, but rather that it is valid to a certain degree.
- **A specific attribute**: Validity is related to a specific use (the purpose for which the test was designed). For instance, an ability test may have weak validity if we use it to assess mechanics. This means that validity is connected to the type of phenomenon being measured.[16]
- **An attribute related to test results, not the test itself**: However, we often associate it with the test for the sake of convenience or simplification. The more accurate approach is to talk about the validity of the results or, more precisely, the validity of our interpretations of the results.

3. The test must be able to differentiate between the extremes of the ability it measures, thereby distinguishing between strong, average, and weak performance. If the test scores are all very similar, this indicates weak validity, as the test has not fulfilled its primary function in the measurement process, which is to reveal individual differences among the sample. Similarly, if the test does not clearly distinguish between the extremes of the ability it measures and does not reveal individual differences, it is neither valid nor reliable.

Since validity is the process of gathering evidence and proof to support the inferences and conclusions reached by the researcher through data collection, as an empirical concept, validity refers to the inferences or conclusions about the tool's uses rather than the tool itself. The evidence for validity can be summarized as follows:

- **Evidence related to content validity**: This refers to the nature of the content included in the tool and its relevance to the trait being measured.
- **Evidence related to criterion validity**: This involves comparing the results obtained using the tool with the results from another similar tool in terms of outcomes.
- **Evidence related to construct validity**: This refers to the nature of the psychological construct or trait being measured by the tool, specifically how well this construct explains and interprets individual differences in behavior or performance.

2. Types of Validity:

The American Psychological Association, in its 1966 *Manual of Recommendations*, identified three types of validity: content validity, criterion-related validity, and construct validity. However, before reviewing these types and the methods for verifying them, it is important to mention another type of validity often discussed in books on psychological and educational measurement, which is:

Face Validity:

This refers to what the test appears to measure, meaning that the test includes a set of items that seem to be related to the trait it was designed to assess, and the content of the test aligns with the purpose for which it was created.

It encompasses the overall appearance of the test or its external form, including the type of items, how they are worded, and how clear they are. It also considers the test instructions, their accuracy, and the degree of objectivity they possess.[17]

Often, studying the test in terms of face validity helps reveal weak items that are not closely related to the function being measured.

Thus, it can be said that face validity, also referred to as superficial validity, can be useful at times, but there are other methods that are more reliable for calculating validity, which are as follows:

2.1 Content Validity:

Also referred to by some as logical validity or substantive validity, content validity refers to "the extent to which the items in the test or measurement tool represent the content that was originally selected to be included in the test."

Here, a logical analysis of the test materials and items is conducted to determine the represented aspects and levels, and the proportion of each in relation to the test as a whole. Then, a survey of the behavior domain to be measured is conducted to identify its factors, the importance of each factor, and the influence of this factor on the behavior that represents the overall function. The relative importance of each aspect is assessed, and the test is matched with the function it measures based on these foundations to determine how well the test represents the required function, its factors, components, and proportions.

In other words, a test is considered valid if the content being measured is accurately represented by the test items. This method is often used in the early stages of developing any test.

This method is typically used to measure school achievement tests and training programs for performance and professional competence. However, it is difficult to calculate content validity for psychological tests, such as those measuring personality traits, attitudes, values, and so on.

For example, in personality assessment, there may be an item that measures extroversion/introversion, such as: "Would you prefer going to a party with your friends or staying at home to read a book?" If the individual responds with "read a book," they may be considered introverted. However, they may answer this way simply because they enjoy reading. This shows that the item may actually measure preferences, making it difficult in personality tests to conclude that an item measures the intended trait.

2.2 Criterion-Related Validity:

This type of validity focuses on comparing test results with certain criteria. The criterion could be individuals' performance levels in other activities, such as university achievement or performance on another test. This is usually done by examining the correlation between the test and the criterion.

There are two types of criterion-related validity:

2.2.1. Predictive Validity:

Predictive criterion validity involves finding the relationship between test results and the results of a criterion that will be obtained in the future. The purpose is to determine the extent to which the scores from one test can be used to predict the scores on another measure, called the criterion.

The steps for predictive validity are summarized as follows:

1. Identify the appropriate criterion behavior and how it will be measured.
2. Select a sample of examinees who represent the population for which the test will be used.
3. Administer the test (predictor) to the sample and keep a record of the scores for each individual.
4. Once the criterion data is available, collect data on each individual's performance on the criterion.
5. Estimate the strength of the relationship between the test scores (predictor) and the criterion scores.

Predictive validity in a test is measured by finding the relationship or correlation between the test scores and the criterion scores, which are obtained after administering the test. After collecting data on the criterion and calculating the test scores, we find the relationship between them using one of the following three methods:

- **The Percentile Method**: In this method, individuals are divided based on their ranks on the criterion measure into two opposing groups, such as pass versus fail. Then, the percentages of individuals who obtained high, medium, or low scores on the test in each group are calculated to see if higher test scores are associated with a higher success or failure rate. If the difference is significant, it indicates that the test is valid.
- **The Averages Method**: This method involves calculating the significance of the difference between the mean scores of two extreme groups of individuals on the test. One group receives a high rating on the criterion measure, and the other group receives a low rating. If a statistically significant difference is found between the mean scores of these two groups on the test, it indicates that the test is valid, meaning it measures the trait that predicts future success in work or study.
- **The Correlation Coefficient Method**: The correlation coefficient is calculated between the test scores of a sample of examinees and their scores on the criterion, which is applied after a significant period of time. If a strong correlation is found between the test and the criterion scores, it indicates that the tool has the ability to predict future performance for the statistical

population from which the sample was drawn. The correlation coefficient between the test and criterion scores is called the predictive validity coefficient.

2.2.2. Concurrent Validity:

This type of validity "is calculated by finding the correlation coefficient between the results of the new test and the results of a criterion, which may be another test or the average grades of the same students. The two tests or the new test and the criterion data are collected during the same period or within a short time frame."[18]

This type of validity uses several methods to estimate the value of the validity coefficient, which include the following:

- **Contrasting Groups**: This method is used to calculate the validity of personality tests, particularly by comparing the performance of individuals in leadership positions with those in clerical positions on a specific test of social traits. The assumption is that individuals in certain professions requiring social interaction will perform differently. Specific groups, such as normal individuals versus neurotics, can also be used when calculating the validity of personality tests related to emotional or social adjustment.
- **Ratings**: In this method, the examiner is asked to provide a rating for the examinee on specific characteristics such as dominance, leadership, originality, or creativity. This method is used when it is difficult to obtain objective criteria in the field of personality testing, especially for social traits, which can only be judged through personal relationships. These ratings themselves become the criterion for calculating concurrent validity.
- **Correlation with Other Measures**: The correlation coefficient is calculated between the test and other measures that assess the same trait. It is important that the correlation is positive and high, as this indicates the concurrent validity of the new measure.

Concurrent validity is particularly suitable for personality tests, as it serves as a middle ground when predictive validity requires a long period to establish.

2.3. Construct Validity:

Also known as construct validity or trait validity, this type of validity refers to "the extent to which a test measures a specific hypothetical construct or trait," such as intelligence, mechanical comprehension, verbal fluency, and so on. This type of validity relies on a broader description and requires more information about the trait and behavior being measured.[19]

Construct validity focuses on three main aspects: the score of the measurement tool, the relationship of this score to the hypothetical construct, and the traits being measured. This is linked to the interpretation of the score, and finally, it focuses on what the test measures from the perspective of the test designer, emphasizing (the construct, interpretation, and theory).

A hypothetical construct is a psychological characteristic or trait that we assume exists to explain certain aspects of individuals' behavior. The type of validity suitable in this case is called "construct validity." Estimating this type of validity requires several methods, the most important of which are listed below:

- **Correlations with Other Tests**: In this type, correlations with other tests should be moderately high but not too high. If the new test has a very high correlation with an existing test without adding advantages such as brevity or ease of administration, the new test represents an unnecessary repetition. Correlations with other tests may also be used differently to demonstrate that the new test is relatively free from the influence of certain extraneous factors. For example, a neuroticism test should not have a high correlation with an intelligence test.
- **Factor Analysis**: Factor analysis helps identify the basic components (common factors) that determine an individual's score on the test, in addition to determining the degree of each item's loading on these factors. These loadings represent the correlations between the test items and the factors, known as **factor validity**.

Factor validity of a test is essentially the correlation between the test and the common factor that a group of tests share. Factor analysis provides a quantitative estimate of the test's validity in the form of a statistical coefficient representing the test's loading on the common factor, which is essentially the correlation between the test and the factor.

- **Internal Consistency**: Here, the criterion is nothing more than the total score on the test itself. It is clear that internal consistency coefficients, whether based on items or subscales, are measures of homogeneity. These have some relation to construct validity for these scales. However, the contribution of internal consistency methods to estimating validity is quite limited. In the absence of external data for the test itself, we can know little about what the test actually measures, and this is often used in personality tests.
- **Convergent and Discriminant Validity**:
    - **Convergent Validity** refers to examining the relationship between the new test and another test that has already been validated for measuring the same hypothetical construct or trait that the new test aims to measure. If the correlation between the two tests is high, it serves as evidence that the new test measures the same trait as the validated test.
    - **Discriminant Validity** involves calculating the correlation coefficients between the new test and other tests that measure different, independent hypothetical constructs. Construct validity is determined by examining these correlations. If the correlation values are low, it indicates that the test has high discriminant validity.

In summary, test validity is calculated by finding the validity coefficient, which is a correlation coefficient between the results of the new test and those of another test known for its validity in measuring the same trait. The tests are administered to the same group, and the coefficient should be close to one. Generally, there is no fixed rule stating that validity coefficients above or below a certain threshold are considered high or low.

3. Factors Affecting Test Validity:

There are several factors that influence the validity of a test, including the following:

- **Factors related to the test itself**: For instance, if the test's language is above the level of the students, some of them may not understand the questions, which would prevent them from answering and thus lower their scores. Similarly, unclear test questions may lead students to interpret them differently and provide incorrect answers, which also lowers their actual performance on the test. The ease or difficulty of the questions, as well as how they are phrased, can result in students earning grades they do not deserve, which can mislead our judgment of them as excellent, average, or weak students, without reflecting their true ability.
- **Factors related to test administration and scoring**: Environmental factors, such as heat or cold, can affect students' performance and either reduce or enhance the validity of the test. Additionally, issues such as unclear printed questions, ambiguous instructions, using the test for purposes other than what it was designed for, or applying it to a population it was not intended for can all weaken the validity coefficient and reduce the examinees' ability to achieve higher scores.
- **Factors related to the personality of the examinee that influence their responses**: Guessing, cheating, attempts by the student to influence the examiner, or the student's anxiety during the test can affect their ability to answer questions accurately, resulting in a score that does not reflect their true abilities, which in turn impacts the validity of the test.

Conclusion:

The psychometric properties of psychological, educational, and social measurement tools serve as evidence for the validity of the results obtained. These properties provide the ability to generalize the results under certain agreed-upon conditions. If a researcher uses measurement tools, whether borrowed from previously developed tools that have been verified for their psychometric properties, or tools developed by the researcher without confirming their psychometric properties, the researcher may find themselves in a loop of uncertainty about whether the results are valid. This is why verifying the psychometric properties of tools is one of the most important methodological steps in psychological and educational research.

As mentioned, validity and reliability are key psychometric properties. These two concepts must have a close relationship, as one of the qualities of a good test is that it should be reliable, meaning that if the test is administered and then re-administered, similar results should be obtained. Therefore, the correlation between these two properties should be high. It is also well-known that both validity and reliability are significantly influenced by the test itself, including how it is administered, how its items are phrased, how it is scored, its length, as well as factors related to the examinee and the environment.

Reliability depends on validity: a valid test is reliable, but a reliable test may not necessarily be valid. For example, if a valid math test is administered to a certain class, it is expected that the students will achieve similar scores across different administrations. Thus, if both of these properties are present in any tool, it can be applied in psychological, educational, and social research alike.

References:

1. Zidan, Jamila, & Boujerada, Mohamed: *Al-Siraj Journal on Education and Society Issues (The Psychometric Properties of Measurement Tools)*, Issue 1, March 2017.
2. Fatima Zahra Al-Ashraf, & Mishri, Sulaf: *Journal of the College of Basic Education for Educational and Human Sciences*, Issue 35, University of Babylon, 2017.
3. Musa Al-Nabhan: *Fundamentals of Measurement in Behavioral Sciences*, Dar Al-Shorouk for Publishing and Distribution, Amman, Jordan, 2013.
4. Salah Al-Din Mahmoud Allam: *Educational and Psychological Measurement and Evaluation*, Dar Al-Fikr Al-Arabi, Cairo, Egypt, 2000.
5. Mouqaddam, Abdelhafid: *Statistics and Educational Psychological Measurement*, Diwan of Algerian Publications, Algeria, 2003.
6. Bachir Maamria: *Psychological Measurement and Tool Design*, Al-Heber Publications, Algeria, 2007.
7. Mohamed Shehata Rabie: *Personality Measurement*, Dar Al-Maseera for Publishing and Distribution, Amman, Jordan, 2009.
8. Sab'a Muhammad Abu Lubda: *Principles of Psychological Measurement and Educational Evaluation*, Cooperative Printers Association, Amman, 2003.
9. Saad Abdul Rahman: *Psychological Measurement: Theory and Application*, Hibat Al-Nil Al-Arabiya for Publishing and Distribution, Alexandria, 1998.
10. Marwan Abdel Majeed Ibrahim: *Foundations of Scientific Research for Preparing University Theses*, Al-Warraq Publishing and Distribution, Amman, Jordan, 2000.
11. Ahmed Mohamed Al-Tayeb: *Evaluation and Psychological and Educational Measurement*, Modern University Office, Alexandria, Egypt, 1999.

Endnotes

1. Zidan, Jamila, & Boujerada, Mohamed: Al-Siraj Journal on Education and Society Issues (The Psychometric Properties of Measurement Tools), Issue 1, March 2017, p. 211.
2. Zidan, Jamila, & Boujerada, Mohamed: Ibid., p. 211.
3. Fatima Zahra Al-Ashraf, & Mishri, Sulaf: Journal of the College of Basic Education for Educational and Human Sciences, Issue 35, University of Babylon, 2017, p. 32.
4. Musa Al-Nabhan: Fundamentals of Measurement in Behavioral Sciences, Dar Al-Shorouk for Publishing and Distribution, Amman, Jordan, 2013, p. 276.
5. Salah Al-Din Mahmoud Allam: Educational and Psychological Measurement and Evaluation, Dar Al-Fikr Al-Arabi, Cairo, Egypt, 2000, p. 133.

6. Mouqaddam, Abdelhafid: Statistics and Educational Psychological Measurement, Diwan of Algerian Publications, Algeria, 2003, pp. 152–153.
7. Bachir Maamria: Psychological Measurement and Tool Design, 2nd ed., Al-Heber Publications, Algeria, 2007, p. 90.
8. Salah Al-Din Mahmoud Allam: Ibid., p. 93.
9. Mohamed Shehata Rabie: Personality Measurement, Dar Al-Maseera for Publishing and Distribution, Amman, Jordan, 2009, p. 84.
10. Sab'a Muhammad Abu Lubda: *Principles of Psychological Measurement and Educational Evaluation*, Cooperative Printers Association, Amman, 2003.
11. Saad Abdul Rahman: *Psychological Measurement: Theory and Application*, Hibat Al-Nil Al-Arabiya for Publishing and Distribution, Alexandria, 1998, pp. 169–170.
12. Mohamed Shehata Rabie: *Ibid.*, p. 84.
13. Salah Al-Din Mahmoud Allam: *Ibid.*, p. 100.
14. Marwan Abdel Majeed Ibrahim: Foundations of Scientific Research for Preparing University Theses, Al-Warraq Publishing and Distribution, Amman, Jordan, 2000, p. 43.
15. Mohamed Shehata Rabie: Ibid., p. 143.
16. Sab'a Muhammad Abu Lubda: *Ibid.*, 2003.
17. Bachir Maamria: *Ibid.*, 2007.
18. Sab'a Muhammad Abu Lubda: *Ibid.*, p. 219.
19. Ahmed Mohamed Al-Tayeb: *Evaluation and Psychological and Educational Measurement*, Modern University Office, Alexandria, Egypt, 1999, p. 211.