ISSN: 2253-0347 مجلـة علوم الإنسان و المجتمع EISSN: 2602-781X
Journal of Human and Society Sciences
volume: 15 /N°: 03 /Year:2026 /pp: 930-961

Reliability in Psychological and Educational Tests: Estimation Methods, Interpretation.

Received date:09/01/2026 Accepted date04/09/2026 Published date:11/10/2026

Daili Nadjia[1]

[1]University of Setif 02 , Algeria, ***Email:*** dailinadjia@yahoo.fr

https://orcid.org/0009-0007-1075-4120

Abstract:

This research paper aims to examine reliability as a fundamental statistical indicator in psychological and educational tests by clarifying its concept, importance, estimation methods, interpretation, and influencing factors. The paper adopts the descriptive-analytical method through reviewing and analyzing the relevant literature on the topic. It also addresses the difficulties encountered by some students when applying reliability estimation methods and interpreting their results in academic research, while providing clarifications to promote a better understanding of reliability and its more accurate use. The paper concludes that estimating reliability is not limited to calculating its coefficient; rather, it also requires selecting the appropriate estimation method and interpreting the results in light of the nature of the test, the characteristics of the sample, and the research context.

Keyword: Test, Instrument, Statistical Indicator, Reliability

Corresponding Author: Daili Nadjia, **Email:** dailinadjia@yahoofr

1. Introduction

Scientific research in the fields of psychology and education relies heavily on data collection. Researchers use various tools for this purpose, including interviews, observations, scales, and tests. Among these, tests are particularly important in the data collection phase due to their strong capacity to measure human behavior.

The American Educational Research Association (AERA), the American Psychological Association (APA), and the National Council on Measurement in Education (NCME) define five key elements that determine the validity of a test: test content, examinee response, test structure, item consistency, and test results (Sireci & Faulkner-Bond, 2014, p100).

The validity and reliability of a test are ensured through its psychometric properties, which primarily include validity and reliability. The quality of research instruments depends on the extent to which these two indicators are met, as they are widely recognized both scientifically and internationally.

Reliability refers to the consistency of test scores, the accuracy of results, and their independence from random influences when applied to a specific group of individuals on two different occasions or when the same individuals are assessed using two different sets of test items (Boudlal & Mouras, 2021, p113).

Reliability is determined through various mathematical formulas and statistical models. However, a key challenge lies in selecting the most appropriate formula for a given test. Assessing reliability requires researchers to consider multiple factors, such as the number of test items (whether they are even or odd), the type of response

options (binary or multiple-choice), and the test administration method (single or repeated application). Each of these variations necessitates a specific statistical approach.

This raises an essential question: When should a researcher use one statistical formula over another, and what are the conditions for its application?

To address these questions, this research paper provides an in-depth examination of reliability as a statistical measure in psychological and educational testing.

Based on the issue raised regarding the appropriateness of reliability estimation formulas and equations according to the nature of the tests and the conditions under which they are administered, the main research question of this paper is as follows: When should a researcher use one mathematical formula rather than another to estimate test reliability, and what conditions and considerations govern its selection?

Accordingly, this paper aims to clarify when a researcher should use one mathematical formula rather than another to estimate the reliability of psychological and educational tests, by identifying the conditions for using each method and the considerations that govern its selection according to the nature and characteristics of the test.

The rationale for addressing this topic stems from the observation that, although the relevant literature and previous studies have addressed the concept of reliability and its estimation methods, they do not always provide a clear and comprehensive treatment that helps researchers establish a link between the nature and characteristics of the test, on the one hand, and the appropriate reliability estimation method, on the other. Presenting reliability estimation formulas separately may also be insufficient to clarify the circumstances that warrant the use of one formula rather than another, which may lead to

difficulties in their application and interpretation, particularly among students and novice researchers in psychological and educational research.

From this perspective, the contribution of this paper lies in providing an organized theoretical review that links the various reliability estimation methods with the conditions for their use and the considerations governing their selection, while clarifying how to select the most appropriate method according to the nature of the test, the number of its items, the method of administration, and the nature of the response alternatives. Thus, the paper seeks to move beyond a general presentation of reliability and its estimation methods by providing clarifications that help researchers select the appropriate reliability estimation method and interpret its results more accurately.

2. Reliability:

Test reliability refers to the extent to which an individual's score remains consistent when the test is repeated. Statistically, it is expressed as a correlation coefficient between an individual's scores across different administrations of the test (Al-Dhaher et al, 2002, p 140).

This definition reflects a fundamental aspect of test reliability, namely the stability of results over time. However, the concept of reliability is not limited to this aspect alone; it also encompasses the extent to which the components and items of a test are consistent in measuring the intended construct, while taking into account the amount of error associated with the measurement process. Therefore, the test–retest method represents one approach to estimating reliability, rather than a comprehensive definition of reliability or the only method for evaluating it.

Livingston (2018) defines reliability as follows: is the extent to which test scores are not affected by chance factors—by the

luck of the draw. It is the extent to which the test taker's score does not depend on ...

• the specific day and time of the test (as compared with other possible days and times of testing),

• the specific questions or problems that were on the edition of the test that the test taker took (as compared with those on other editions), and

• the specific raters who rated the test taker's responses (if the scoring process involvedany judgment) (Livingston, 2018, p. 6)

This definition indicates that reliability is not limited to the stability of test scores over time; it is also related to the extent to which scores are free from different sources of error, such as differences in administration time, item content, and raters' judgments. Accordingly, estimating reliability requires selecting a method that is appropriate to the nature of the test and the source of error to be examined.

Reliability can be examined from two main perspectives

- **Stability of Results:** The test should yield the same results regardless of how many times it is administered.
- **Measurement Error:** The observed score of an individual is the sum of the true score and the error score (Mekadam, 1993, p 152).

It can be observed that these two aspects highlight that reliability is not limited to the stability of test results when the test is administered repeatedly; it is also related to the extent to which the scores are free from random measurement errors. More consistent and stable results reflect a higher degree of reliability, whereas a greater influence of measurement error leads to lower accuracy and dependability of the results. Therefore, stability,

consistency, and a low level of measurement error are interrelated aspects in understanding the concept of reliability, while the methods used to estimate it vary according to the aspect they aim to assess and the source of error on which they focus.

Reliability indicators are divided into two categories: stability and consistency.

- **Stability** refers to the extent to which test results remain unchanged over time. A test is considered reliable if it produces stable results when administered multiple times. This stability is typically measured using correlation coefficients.
- **Consistency** relates to two key assumptions: either the test measures a single construct (which requires examining the consistency of its items), or it assesses a composite construct (necessitating an evaluation of how well different dimensions align with each other and with the overall test). Correlation coefficients are also used to assess this aspect (Jaija, 2022, p 311).

It is important to note that internal consistency carries multiple meanings, including *homogeneity, average inter-item correlation, general factor saturation, and internal consistency reliability.* Cronbach (1951) used the terms *internal consistency* and *homogeneity* interchangeably, emphasizing that an internally consistent test should be independent of its length. Revelle (1979) defined internal consistency as the degree to which test items measure the same underlying construct, highlighting that a test with high internal consistency should be unidimensional, meaning it assesses only one trait (Tang et al., 2014, pp 205-206).

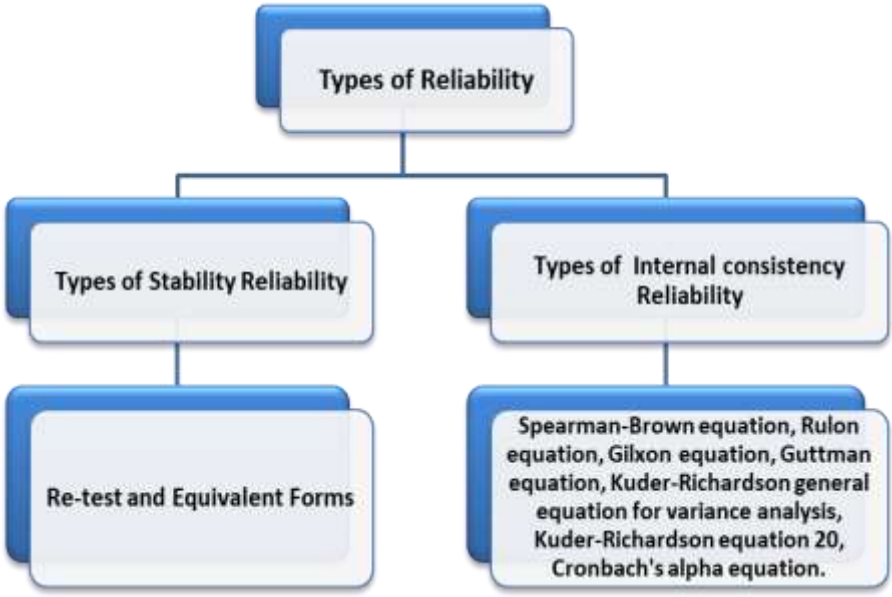
Regarding specific reliability indicators, tests of stability typically involve *test-retest* and *parallel forms* reliability. On the other hand,

measures of internal consistency include various statistical formulas, such as:

- Spearman-Brown formula
- Rulon's formula
- Jackson's formula
- Guttman's formula
- Kuder-Richardson general variance analysis formula
- Kuder-Richardson 20 (KR-20)
- Cronbach's alpha (Jaija, 2022, p. 311).

The following figure illustrates the different types of reliability:

Figure (1): Illustration of the Types of Reliability.



(Source: Prepared by the researcher)

From the figure above (Figure 01), it is evident that the types of internal consistency reliability are more numerous than those of stability reliability. The researcher must select the appropriate indicator based on the specific circumstances of their study. For instance, regarding the duration of implementation, the test-retest method requires approximately six months.

The reliability of a test or questionnaire is established by having it reviewed by professors or experts with extensive experience and competence in psychological or educational research, depending on the nature of the topic. If the questionnaire pertains to school violence, it is preferable for it to be evaluated by specialists in the field of education, such as university professors and teachers across all three educational levels. Teachers, due to their direct interaction with students and their extensive daily classroom experience, can provide valuable insights based on their teaching expertise.

However, if the questionnaire focuses on depression, the experts should be university professors, preferably those specializing in clinical psychology. Additionally, the expertise of clinical psychologists can be leveraged due to their specialized training and practical experience in the field.

2.1 Measures of Reliability:

The reliability of a psychological test is determined using the reliability coefficient, which can be calculated using various statistical methods. Some of the methods to be reviewed are as follows:

(a) Test-retest reliability. (Hamza Hassan, 2008, p. 69)

This method addresses measurement error associated with differences in time and test administration conditions.

EISSN: 2602-781X————————————ISSN: 2253-0347

(b) Equivalent forms reliability. (Hamza Hassan, 2008, p. 69) This method addresses measurement error resulting from differences in item content between test forms.

(c) Split-half reliability. (Hamza Hassan, 2008, p. 69) This method focuses on measurement error associated with differences among the items within the test and the extent of their consistency.

(d) Variance analysis reliability. (Hamza Hassan, 2008, p. 69) This method aims to estimate the sources of variance in test scores and determine the amount of error associated with them.

(e) Judge reliability estimates. (Hamza Hassan, 2008, p. 69)

This type is concerned with error resulting from differences among the ratings provided by judges or raters.

The choice of which reliability formula to apply is influenced by several factors, including the presence of the sample and the nature of the test alternatives (binary or multiple). For example, the test-retest reliability method is not applicable if the sample members cannot be found during the second test administration. For instance, if we want to measure feelings of guilt among drivers responsible for traffic accidents, we may find them in the hospital during the first administration due to injuries. However, during the retest phase, we may not find them, as there is a significant chance they have left the hospital, especially if the time gap between the initial and retest administrations is three weeks or more. Additionally, when it comes to the number of items in the questionnaire or test, if the number of items is odd, the Guttman split-half formula cannot be applied, as it requires an even number of items to allow the researcher to divide them into odd and even items. Below are some of the formulas used to measure reliability:

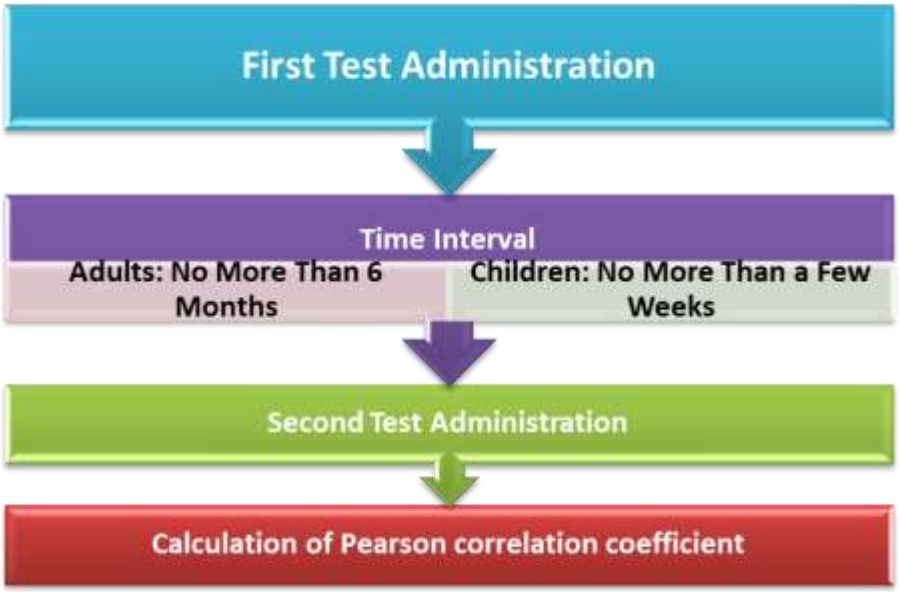
(a) Test-Retest Reliability

Test reliability is assessed by re-administering the test to estimate errors associated with administering it at two different points in time, with a time interval between them (Hamza Hassan, 2008, p. 69).

The core principle of this method involves administering the test to a group of individuals and then re-administering the same test to the same individuals after a certain period. This interval should not exceed a few weeks for younger individuals and should be no longer than six months for adults. The Pearson correlation coefficient is then calculated between the individuals' scores from the first and second administrations. This correlation coefficient, known as the "Reliability coefficient," indicates test reliability, with a value of 0.6 or higher signifying an acceptable level of reliability.

This method is among the most widely used, particularly for measuring **traits** or characteristics that remain stable over time (Zakaria Mohamed Al-Daher et al., 2002, p. 140), as illustrated in the following diagram:

Figure 02: Shows the steps for calculating the reliability coefficient using the test-retest method.



(Prepared by the researcher)

The time required for administering and re-administering a test varies according to the perspectives of scholars and researchers. Some do not strictly adhere to these time intervals and recommend a re-administration period ranging from 15 to 21 days.

It is important to note that the test-retest method is not suitable for all situations, as there must be a minimum time gap of at least four weeks between the first and second administrations. For instance, if the test measures anxiety, an individual's scores may fluctuate significantly between the first and second administrations. A relevant example is stock market traders, whose anxiety levels can fluctuate hourly—let alone over the course of a week or several weeks.

Additionally, test fatigue may occur when respondents are required to take the same test twice, which can affect the reliability of the results. (Crutzen & Ygram Peters, 2015, p.3).

Accordingly, it can be said that determining the appropriate time interval between the two administrations does not follow a fixed time period; rather, it is determined according to the stability of the construct or trait being measured, the purpose of the test, and the characteristics of the sample. The interval should be long enough to minimize the effects of memory, while not being so long that genuine changes in the measured trait may occur.

Advantages of This Method: This method is suitable for both time-limited tests and untimed tests, provided that the time gap between the two application ranges from a few weeks to a few months, depending on the age group.

However, this type of reliability is prone to several errors, though these do not diminish its significance. The administration or application of the test on the subjects during the second round can be influenced by several factors, as follows:

- **Familiarity:** Subjects benefit from repetition, which leads to what is known as familiarity with the testing situations. This reduces emotional disturbance, the shock of novelty, and the process of establishing rapport during the second administration. (Reynolds & Livingston, trans. Mahmoud, 2012, pp. 172-173)

  Additionally, repeating the same test to assess its reliability may not appeal to the subjects, which can result in reduced engagement and interest. This makes the re-test a weaker measure compared to the first administration. (Al-Zaher et al., 2002, p. 142)

- **Training or Practice Effects:** The subject may benefit from learning specific strategies for answering items or questions on

the test or handling the test materials and apparatus. (Reynolds & Livingston, trans. Mahmoud, 2012, pp. 172-173)

- **Experimental Context:** This method is affected by the various factors influencing the experimental context during the second administration, which may differ from the first. This highlights the limitations of this method in achieving the level of scientific accuracy we aim for in different types of research. (Al-Zaher et al., 2002, p. 140)
- **The Length of Interval Between Administrations:** The performance of the subjects during the second application is influenced by the length of time between the first and second tests. If the interval is too long, there is a higher likelihood that subjects' scores in the second Administration may be influenced by developmental changes, especially as their age increases. (Reynolds & Livingston, trans. Mahmoud, 2012, p. 173)

Therefore, it is essential to determine the length of the time gap between **Administrations** based on factors such as the sample nature, individual intelligence, age, and the nature of the test itself, including its length and what it measures. Generally, this period should not be shorter than one week nor longer than six months.

(B) Equivalent Forms:

- In this method, two versions of the same test are created and Applied

either in a single session or across two sessions with a short time gap, but to the same individuals. (Reynolds & Livingston, trans. Mahmoud, 2012, p. 174)

If the equivalent test is Applied after a long period, the reliability coefficient can be regarded as a measure of both equivalence and stability. It accounts for all factors that affect the consistency of results, such as changes in the subject (learning, growth, maturation,

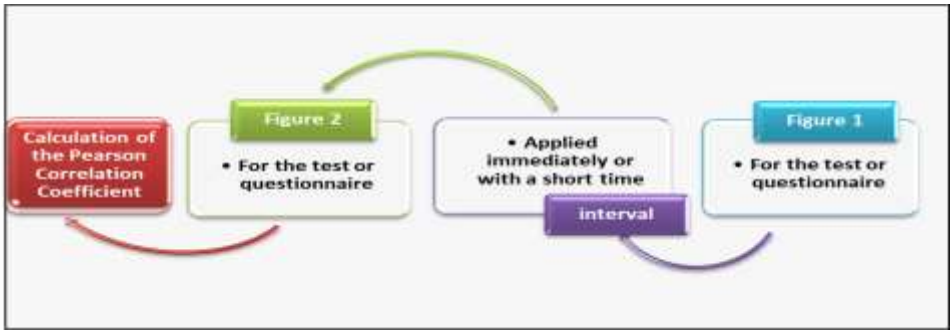
psychological state) and differences in the test application methods between the two sessions. (Al-Zaher et al., 2002, p. 143)

This method has an advantage over the previous one (retesting) in that it eliminates the effects of practice, memory, and forgetting. It also addresses the issue of time intervals and the need to administer the same questions twice. (Reynolds & Livingston, trans. Mahmoud, 2012, p. 174)

It is crucial that the conditions for administering the first test are similar to those for the second test. Additionally, a reasonable time gap should be provided between the administration of the first and second versions of the test to minimize the effects of fatigue on the subjects. (Bertrand & Blais, 2004, p. 51)

The correlation coefficient between the performance of individuals on the first version of the test and their performance on the second version is calculated. This coefficient represents the test's reliability and is referred to as the "equivalence coefficient" in this method, as it reflects the correlation between scores on two equivalent tests. A value of 0.6 or higher indicates the reliability of the test. (Reynolds & Livingston, trans. Mahmoud, 2012, pp. 174-175)

Figure 03: Illustrates the steps for calculating the equivalence coefficient.



(Prepared by the researcher)

In both forms, the following conditions must be met for equivalency to occur:

- The number of questions in both forms must be equal.
- The means and standard deviations must be equal.
- The application time for both forms must be the same.
- The difficulty levels of the questions in both forms must be equal.
- The same method of application and scoring must be followed in both forms.

Advantages of this method:

- It eliminates the effect of familiarity with the test and also removes the impact of training/learning.
- It is suitable for calculating the reliability of memory tests as well as standardized achievement tests.

Difficulties of the Equivalent Forms Method:

- It is not always possible to prepare equivalent forms for many tests, as creating equivalent forms requires ensuring that the means, standard deviations, difficulty gradation, number of items measuring each dimension, and difficulty level of corresponding vocabulary items are all equal. It also requires standardizing the application, scoring methods, and testing time for both forms.
- Another challenge is the increased effort, time, and cost required in constructing the test. (Reynolds & Livingston, translated by Mahmoud, 2012, pp. 174-175)
- Test takers may express dissatisfaction due to the additional tests required to obtain the reliability coefficient. (Al-Dhahir et al., 2002, p. 143)

What we observe in field studies in the area of psychological and educational research, as well as in academic studies such as

Bachelor's, Master's, and Doctoral theses, is the complete absence of this method. The reason is the difficulty of preparing two questionnaires for a single variable, to avoid the effort and time involved.

(C) Split-Half Method: This method involves splitting the test into two equivalent halves, giving a separate score for each half, and then calculating the correlation coefficient between them. (Reynolds & Livingston, translated by Mahmoud, 2012, p. 175)

The researcher uses the fairness method in this approach, where the test is administered in full, and then divided into two equal parts during scoring. This ensures that the means, standard deviations, and difficulty levels of the questions are equal in both halves. (Al-Dhahir et al., 2002, p. 144)

This method relies on administering the test once and then dividing it into two halves, where one half contains the odd-numbered items (1, 3, 5, 7, ...) and the other half contains the even-numbered items (2, 4, 6, 8, ...). This way, the test taker's effort and time are evenly distributed between both halves.

Table (1): Shows the individuals' scores on the odd and even items

Even items	Odd items
2 -	1 - .....
4 -	3 - .....
6 -	5 - .....
8 -	7 - .....
10 -	9 - .....

(Prepared by the researcher)

EISSN: 2602-781X———————ISSN: 2253-0347

The odd-even division method is characterized by ensuring a higher degree of equivalency between the two halves of the test, particularly when the test items are arranged from easy to difficult. (Reynolds & Livingston, translated by Mahmoud, 2012, p. 175, p. 176)

One of the difficulties encountered with the split-half method is the lower correlation coefficient, as it calculates the correlation of half the test length rather than the full test length. The reliability coefficient increases as the test length increases. (Hamza Hassan, 2008, p. 75)

It should also be noted that the correlation obtained through this method represents the reliability coefficient of the test half, not the full test. (Reynolds & Livingston, translated by Mahmoud, 2012, p. 176)

To estimate the reliability of the entire test, the correlation coefficient of the test half must be adjusted or corrected to reflect the expected reliability of a full-length test. This can be done using formulas known as "length correction equations" in the classification or split-half methods. Below, we will use the "Spearman-Brown" formula as an example:

$$rxx = \frac{2r\,½\,½}{1+r\,½\,½}$$

Where:
rxx: Estimated reliability of the full test
r½: Reliability of the half-test

For example, if the correlation of the two halves of a test is "0.60" (r½ = 0.60), the estimated reliability of the full test is calculated as follows:

$$rxx = \frac{2(0.60)}{1+0.60} = \frac{1.20}{1.60} = 0.75$$

(Mehrens, Lehmnn, translated by Al-Zubaidi, 2003, p. 343)

As we can observe from the previous example, the reliability value before correction was 0.60, and after correction, it increased to 0.75, which means it increased by "0.15." For further clarification, refer to the following table:

Table (2): Reliability Values Before and After Correction

Correlation value before correction	Correlation value after correction	Difference
,05	,09	,04
,15	,26	,11
,25	,40	,15
,35	,52	,17
,45	,62	,17
,55	,71	,16
,65	,79	,14
,75	,86	,11
,85	,92	,07
95	97	,02

(Hamza Hassan, 2008, p. 76)

From the table, we observe that the reliability value increases significantly after correction, particularly in the middle section of the table, where the middle values (ranging from 0.35 to 0.75) often represent the reliability of most personality and attitude tests. Meanwhile, well-constructed intelligence tests, such as the "Wechsler and Binet" tests, may exceed 0.80. This method is one of the most commonly used for calculating reliability because it addresses the

criticisms directed at the test-retest method and is preferred over the equivalent forms method due to its cost-effectiveness and speed.

However, a limitation of the split-half method is that it only allows for one test administration. If there are obstructing circumstances, such as health issues, low motivation, or any other external factors that influence performance, the results may be affected. (Reynolds & Livingston, translated by Mahmoud, 2012, p. 176)

It is important to note that the "Spearman-Brown" formula cannot be relied upon in all cases. If the variance of the two halves of the test is not equal, this formula is not recommended. In such cases, the "Cronbach's Alpha" formula can be used to calculate reliability. (Hamza Hassan, 2008, p. 76)

Additionally, this method is not suitable for calculating the reliability of tests that cannot be divided into equivalent parts, where there is a significant numerical difference. It is also not appropriate for time-limited tests that rely on the speed of responses. (Al-Zaher et al., 2002, p. 145)

In general, this method is widely used by researchers, especially when the test or questionnaire alternatives are binary (Yes-No). In such cases, the researcher cannot use Cronbach's Alpha because it requires at least three alternatives (Yes, Sometimes, No).

(d) Analysis of Variance Method

Kuder (G.F. Kuder) and Richardson (M.W. Richardson) relied on variance analysis of test questions in their study of reliability. Their method involves a detailed study of this variance, and the researchers were able to deduce certain equations suitable for measuring reliability. (Fouad Al-Bahy, no date, p. 390). In their method, the test is administered once, and the total test score is estimated. Their approach is based on analyzing the internal structure of the test (internal consistency) to assess the homogeneity of its items by

calculating the Kuder-Richardson 20 or 21 coefficient or Cronbach's Alpha coefficient.

This method is suitable for measuring the reliability of non-timed tests, or what are referred to as power tests, which have binary scores (1, 0), such as multiple-choice or true/false tests. It addresses the issue of split-half reliability. (Reynolds & Livingston, translated by Mahmoud, 2012, p. 177)

- **Kuder-Richardson 20**: Kuder-Richardson developed a set of internal consistency formulas, symbolized as "KR20," attempting to solve the issue of splitting the test into two parts. This method addresses several problems, including variations in the variance between the two halves of the test. (Shavelson, 2014, p. 6). Furthermore, the split-half method requires each half to be independently corrected, adding a burden to those conducting the analysis.

Thus, the Kuder-Richardson method avoids these issues by considering all possible ways to divide the test.

The equation for calculating the reliability of tests that contain binary items, which typically involve answers of "1" or "0" (true or false), is known as Kuder-Richardson 20 "KR20" or KR21. The term is derived from the famous article published by Kuder and Richardson. (Hamza Hassan, 2008, p. 77).

The equations "KR20" and "KR21" are widely used and are considered common formulas. They are as follows:

$$First\ equation\ :\quad K-R20{:}rxx=\frac{n}{n-1}\left[1-\frac{\sum pq}{s^2x}\right]$$

Second equation: $$K-R21 = \frac{n}{n-1}\left[1-\frac{\bar{x}(n-\bar{x})}{ns^2_x}\right]$$

n: The number of items in the test.

p: The proportion of individuals who answered the item correctly (item difficulty). For example, if 6 out of 30 individuals answered item 1 correctly, the p-value for this item would be calculated as follows: 6/30 = 0.20.

q: The proportion of individuals who answered the item incorrectly, calculated as q=1−pq = 1 - pq=1−p. If p=0.20p = 0.20p=0.20, then q=0.80q = 0.80q=0.80.

pq : The variance of the item, corrected for dichotomous scoring (correct or incorrect).
Σ : The summation symbol, indicating that pq is summed over multiple items.

s^2x : The total variance of the test.

x : The overall mean of the test.

The difference between the KR20 and KR21 formulas is that KR21 assumes all test items have the same difficulty level, meaning the p-value is constant for all items. If this assumption is not met, KR21 will yield a slightly lower reliability estimate. However, both formulas are widely used.

Moreover, KR21 requires fewer calculations than KR20. Given the number of test items, only the test's mean and variance need to be

computed. Both formulas provide estimates of test reliability, but KR20 generally produces a higher and more accurate estimate. Therefore, KR20 should be preferred whenever possible, as it involves calculating item difficulty indices, which provide valuable insights.

KR21 also assumes that all items have equal difficulty, with an average item difficulty of 50%. This means that each item is expected to be answered correctly by half of the test-takers, while the other half answers incorrectly. In practical applications, however, tests rarely meet these exact conditions. As a result, KR21 tends to underestimate reliability compared to the split-half method. (Mehrens & Lehmann, translated by Al-Zubaidi, 2003, pp. 343–344; Hamza Hassan, 2008, p. 79)

*** Cronbach's Alpha (α)**

The Alpha coefficient method, developed and refined by Cronbach, is a generalization of the KR20 formula when items are not scored dichotomously. The Alpha formula is identical to KR20, except that the sum of the variances (∑pq) in KR20 is replaced by the sum of the item variances ($\sum s^2 i$), where $\mathbf{s^2 i}$ represents the variance of a single item. (Mehrens & Lehmann, translated by Hatham Kamel Al-Zubaidi, 2003, p. 344)

In Cronbach's Alpha formula, $\mathbf{s^2 i}$ refers to the variance of each item, whether in a dichotomous (right/wrong) format or any other response format. This contrasts with KR20, which specifically differentiates between the variance of correct and incorrect responses.

Cronbach's Alpha is particularly useful for personality tests and attitude scales, where response alternatives are more than two options (i.e., not just correct/incorrect). For example, in an attitude scale measuring opinions on whether "A mother has the right to work to achieve self-fulfillment," response options may include:

- Strongly agree
- Agree

- Neutral
- Disagree
- Strongly disagree

Since these responses are not in a dichotomous (correct/incorrect) format, KR20 would not be suitable, making Cronbach's Alpha the preferred reliability measure. (Hamza Hassan, 2008, p. 79)

Cronbach found that the Alpha coefficient is a strong indicator of test equivalence, providing reliable estimates of both equivalence and internal consistency (homogeneity). (Allam, 2000, p. 166)

- **Cronbach's Alpha Formula**

The **Alpha coefficient** is calculated using the following equation:

Where:

$$\alpha = \frac{n}{n-1}\left(1 - \frac{\sum_{i=1}^{n} s_i^2}{s_x^2}\right)$$

- **n** = Number of items
- **s^2i** = Variance of each items
- **s^2x** = Total test variance (variance of the sum of the items). (Abderrahmane, 1998, p. 172)

A high Cronbach's Alpha coefficient may not necessarily reflect internal consistency of the test items as much as it indicates a high level of similarity between the items in the scale, which may be repeated with only slight variations in wording. We may be cautious about adopting some scholars' suggestions regarding the level at which the reliability coefficient, particularly Cronbach's Alpha, should be set. For instance, Nunnally initially recommended a minimum of 0.70 as an acceptable threshold for reliability in research contexts.

However, in later editions of his famous book on psychometrics, he raised this minimum to 0.80. Nunnally recommended that for clinical purposes, the reliability coefficient should not be lower than 0.90, and ideally, it should hover around 0.95. (Tigheza, 2017, p. 12)

Accordingly, a value of 0.70 should be regarded as a general guideline rather than an absolute criterion, as the acceptable level of reliability varies depending on the purpose of measurement, the type of instrument, the stage of its development, and the context in which it is used.

On the other hand, Streiner argues that a Cronbach's Alpha above 0.90 may indicate unnecessary redundancy in what the items measure, rather than signifying internal consistency. Therefore, he suggests that Cronbach's Alpha should not exceed 0.90 as a maximum. (Tigheza, 2017, p. 12)

Cronbach's Alpha is also influenced by the number of items. For instance, its value may reach 0.70 if the scale consists of 40 items, despite a low average correlation between them. Additionally, the Alpha coefficient is impacted by whether the test is unidimensional (measuring a single trait) or multidimensional. High Alpha values are typically obtained for unidimensional tests. (Laveault, 2012, p. 3)

Dunn, Baguley, and Brunsden (2013) stated that Cronbach's Alpha is the most widely used coefficient in psychometrics, with over 17,600 publications citing its use since the 1950s. (Beland et al., 2017, p. 393)

However, Deng and Chan (2017) note that Alpha is not the only measure of internal consistency. They pointed out that Alpha requires certain conditions to be applied, such as homogeneity among items. As an alternative, they highlighted Omega (Ω), which they considered less commonly used by researchers compared to Alpha. (Gana et al., 2021, p. 19)

In academic studies, Cronbach's alpha coefficient is one of the most widely used reliability coefficients among researchers in educational

and psychological research, as it can be used to estimate the internal consistency of tests and scales. Its calculation is based on estimating the variance of each item in the questionnaire, along with the total variance of the questionnaire or test.

However, the use of Cronbach's alpha coefficient to estimate reliability should be related to the nature of the test and the characteristics of its items, and it should not be considered a comprehensive or general indicator of internal consistency in all cases. It is particularly appropriate for tests and scales that include items with multiple response alternatives, while taking into account the degree of item homogeneity and their relationship to the construct or trait measured by the test. The value of the reliability coefficient may also vary according to the characteristics of the sample and the distribution of scores among its participants, which calls for interpreting Cronbach's alpha in light of the degree of variability in scores rather than generalizing its value independently of sample characteristics. Caution should also be exercised when interpreting its value, as a high Cronbach's alpha coefficient may be influenced by the number of items or by excessive similarity and redundancy among them, and does not necessarily mean that the test is unidimensional or that the measurement instrument has high psychometric quality. Accordingly, selecting and interpreting Cronbach's alpha requires consideration of its conditions of use, the nature and characteristics of the test, the characteristics of the sample, and the objectives of measurement, rather than relying on it alone to judge internal consistency or the quality of the measurement instrument.

(W) Inter-Rater Reliability

In many cases, researchers need an indicator of the reliability of raters' estimates, especially when using some personality measurement methods that rely on observation (such as checklists or rating scales), personal interviews, projective techniques, or creative thinking tests. This is also relevant when assessing academic achievement using

essay-type questions, consecutive structured questions, or open-book questions.

It is important to note that each reliability method examines a specific source of variation (variance). (Hamza Hassan, 2008, p. 69)

For example, if a set of papers is assigned for independent grading by two different raters, Pearson's correlation coefficient can be used to estimate the reliability of each rater's score. If the researcher wants to assess the reliability of the average or total score of the two raters, the Spearman-Brown prediction formula, which was previously mentioned, would be used. In this case, (r½½) represents the correlation between the two sets of scores, and rxx represents the estimated reliability of the total scores. (Mehrens, Lehmnn, translated by Al-Zubaidi, 2003, p. 340)

Accordingly, after reviewing the main methods for estimating the reliability of psychological and educational tests, as well as the characteristics of each method and the conditions for its use, it can be concluded that the selection of an appropriate reliability coefficient depends on the nature of the test, the objectives of measurement, and the conditions under which the test is administered. To clarify the differences among these methods and facilitate comparison, the following table summarizes the main reliability coefficients and the conditions for their use.

Table (3): Comparison of Reliability Coefficients and the Conditions for Their Use

Method of Estimating Reliability	What Does It Estimate?	When Is the Method Used?	Key Conditions or Considerations
Test–Retest Reliability	Stability of Individuals' Scores Over Time	When the objective is to assess the stability of test scores obtained from the same individuals across two different testing occasions.	An appropriate time interval should be selected, taking into account the effects of memory, practice, and genuine changes in the trait being measured.
Equivalent Forms	Equivalence of Scores Across Two Test Forms	When two equivalent forms of a test are available, the aim is to assess the consistency of the scores obtained from the two forms.	Developing two equivalent forms that are comparable in content, difficulty, and psychometric properties.
Split-Half Method	Consistency Between Test Halves	When the reliability of a test is to be estimated from a single administration by dividing its items into two halves and comparing the scores obtained from each half.	Dividing the test items appropriately and adjusting the correlation coefficient to estimate the reliability of the full test using the Spearman–Brown formula, when necessary.
Analysis of Variance Method	Estimating test reliability by analyzing the sources of variance in test scores, such as variance attributable to individuals, items, or testing occasions, depending on the model used.	When the aim is to analyze the sources of variation in test scores and determine the extent to which the variance is attributable to true differences between individuals, as opposed to measurement error.	It requires selecting an appropriate analysis of variance model based on the test design, identifying the relevant sources of variance, and verifying the required statistical assumptions. .
Cronbach's Alpha (α)	Internal Consistency of Test Items	When the aim is to assess the degree of consistency among a set of items measuring a particular dimension or trait, especially in multi-item scales.	The assumptions underlying the coefficient and the nature of the items should be considered, as a high coefficient value alone does not establish unidimensionality
Inter-Rater Reliability	Degree of Inter-Rater Agreement	When scoring involves human judgment, as in the scoring of a researcher-developed questionnaire.	Selecting an appropriate coefficient based on the nature of the data, the number of raters, and the rating design, while distinguishing between agreement and correlation.

(Prepared by the Researcher)

The table shows that methods for estimating the reliability of psychological and educational tests differ in terms of the coefficients used and the conditions under which they are applied, depending on the measurement objectives and the characteristics of the test. Accordingly, selecting an appropriate method should not be based solely on the reliability coefficient value; rather, it requires consideration of the nature of the data, the test design, and potential sources of measurement error to ensure a more accurate interpretation of reliability estimates.

2.2 Interpreting the Reliability Coefficient:

When comparing reliability coefficients or discussing one of them, it is essential to mention the method used to obtain it and the equation applied in its calculation. This is because each reliability coefficient has a specific meaning, and its use is only valid within that context. For instance, when calculating the reliability coefficient using the test-retest method, what matters is the consistency of individuals' scores. In contrast, when calculating the reliability coefficient using the equivalent forms method, the focus is on the equivalence of individuals' scores between the two tests. Similarly, when calculating the reliability coefficient using the split-half method, the emphasis is on the equivalence of scores, as each half of the test should match the other half of the total test. (Abu Labda, 2008, p. 234)

What matters in calculating the reliability of a test or questionnaire is that the quantitative indicators we obtain should be 0.60 or higher, as this indicates an acceptable level of reliability that meets the requirements for applying the measure to the sample individuals.

3. Conclusion:

This paper concludes that the estimation of reliability in psychological and educational tests does not rely on a single mathematical formula. Rather, the choice of the estimation method is determined by the nature of the test, the characteristics and number of its items, the method of administration, and the nature of the response alternatives.

It was also found that each method has specific conditions and considerations that researchers should take into account, and that its selection should be based on the characteristics of the measurement instrument and the objectives of measurement.

As evident from the discussion of Cronbach's alpha coefficient, although it is widely used to estimate internal consistency, its value may be affected by the number of items and the similarity between them. This calls for caution when interpreting high values and for avoiding considering them sufficient evidence of the quality or unidimensionality of the measurement instrument. Accordingly, reliability coefficients should be interpreted in light of the characteristics of the test, the sample, and the context of administration, rather than relying on a single coefficient to judge the quality of a measurement instrument.

Accordingly, the answer to the main research question is that the selection of an appropriate reliability estimation formula depends on the nature of the test, its characteristics, and the conditions under which it is administered. The proper use of these formulas requires an understanding of the conditions and limitations of each method and the interpretation of its results within the context of its application. This enables researchers to select the most appropriate method and enhance the accuracy of findings and the reliability of conclusions in psychological and educational research.

4. Bibliography List :

. 1. Beland, Sébastien et al (2017), Using mcdonald's omega coefficient instead of cronbach's alpha, mcgill journal of education , vol. 52 no 3 fall, p791-p804.

2. Bertrand, & Blais, J. G. (2004), Modèles de mesure: L'apport de la théorie de réponse aux items, Canada : Presses de l'Université du Québec

3. Crutzena, Rik and Ygram Peters, Gjalt-Jorn (2015), Scale quality: alpha is an inadequate estimate and factor-analytic evidence is needed first of all, Downloaded by [University of Maastricht] at 08:36 08 January 2016.

4. Gana, Kamel et al (2021), Current methodo-epistemological frameworks for the construction and transcultural adaptation of psychological tests and measurement scales, published by Elsevier. This manuscript is made available under the Elsevier user license https://www.elsevier.com/open-access/userlicense/1.0/

5. Laveault, Dany (2012), Sixty Years of Good and Bad Uses of Cronbach's Alpha, Measurement and Evaluation in Education, Volume 35, Number 2, ADMEE-Canada - Université Laval , 1-7

6. Livingston, Samuel A. (2018), Test Reliability—Basic Concepts, ETS Research Memorandum, No. RM-18-01, Educational Testing Service, 1-18.

7. Shavelson, Richard. J (2004), My Current Thoughts on Coefficient Alpha and Successor Procedures, Center for the Study of Evaluation (CSE) Graduate School of Education & Information Studies, University of California.

8. Sireci, Stephen and Faulkner-Bond, Molly (2014), Validity evidence based on test content, Center for Educational Assessment, Hills South, University of Massachusetts Amherst, 01003 Amherst (USA), pM100-107.

9. Tang, Wei & all (2014), Internal Consistency: Do We Really Know What It Is and How to Assess It? Journal of Psychology and Behavioral Science, Vol 2, No. 2, Published by American Research Institute for Policy Development , pp. 205-220

10.Abu Labda, Saba Mohammad (2008), Principles of Psychological Measurement and Educational Evaluation, 1st Edition. Amman, Jordan: Dar Al-Fikr.

11. Al-Jarjawi, Ziad ben Ali ben Mahmoud (2010), Methodological Educational Rules for Constructing Questionnaires, 2nd Edition. Gaza, Palestine: Ibn Al-Jarrah Printing Press.

12. Allam, Salah Al din Mahmoud (2000), *Educational and Psychological Measurement and Evaluation*, 1st Edition. Cairo, Egypt: Dar Al-Fikr .

13. Hamza Hassan, Barakat (2008), Principles of Psychological Measurement, 1st Edition, Cairo, Egypt: International House for Cultural Investments.

14. Irvin. J. Lehmann & Mehrens William.A (2003), *Measurement and Evaluation in Education and Psychology* (Trans: Haitham Kamel Al-Zubaidi), 1st Edition. United Arab Emirates: University Book House.

15. Mekadam, Abd el Hafiz (1993), Statistics and Psychological and Educational Measurement with Examples of Scales and Tests, Algeria: National Office of University Publications.

16. daideche , Abd el Salam (2018), Lectures on Psychological and Educational Measurement, University of Mohamed Lamine Debaghine, Setif 02, Algeria.

17. Livingston. Ronald B & Reynolds. Cecil R (2012), Mastering Modern Psychological Measurement (Trans: Salah el din Mahmoud), 1st Edition. Jordan: Alam Dar Al-Fikr.

18. AL Said, Fouad Baha Al din (Under a year), Statistical psychology and measurement of the human mind , Cairo, Egypt: Dar Al-Fikr.
19. Al-Dhaher, Zakariya et al (2002), Principles of measurement and evaluation in education, 1st Edition, Amman , Jordan: Dar Al Ilmia.

20. Boudlal, Fatiha & Mouras, Mahbuba. (January 2021), An Evaluative Study of Measuring Psychometric Properties (Validity and Reliability) in Scientific Research. Dhiaa Journal for Psychological and Educational Research, Vol. 1, No. 2, pp 111-131.

21. Jaija, Omar (2022). Validity and Reliability Indicators Used in Psychometric Studies of Measurement Tools. Rowaiz Journal, Vol. 6, No. 1, pp. 304-321.

22. Tighza, Mohamed (2017). Recent Trends in Estimating the Validity and Reliability of Measurement Instrument Scores: A Theoretical, Evaluative, and Applied Analysis. Journal of Psychological and Educational Sciences, pp. 7-29.

EISSN: 2602-781X———————————ISSN: 2253-0347