... ZAOULI N°11, Vol. 1, Septembre 2025, pp. 354-371 ISSN : 2788-9343

Soumission : **20/02/2025** **Acceptation : 08/08/2025** **Publication** : **15/09/2025**

Psychometric Properties of Research Instruments in Psychological and Educational Sciences

Nezai ZOHRA
Laboratory for the Analysis of Quantitative and Qualitative Data on Psychology and Social Behavior, Abdelhamid Ibn Badis University, Mostaganem
https://orcid.org/my-orcid?orcid=0000-0001-6796-2456
univ-mosta.dz.@.zohra.nezai

Abstract:
Psychometric properties are among the most important tools in psychological and educational measurement, as they assess validity and reliability and allow for generalization to the population. Therefore, validity and reliability are essential psychometric characteristics that must be present in psychological and educational measurement instruments.
These properties are statistical indicators obtained by subjecting a specific scale to a series of experimental and statistical procedures within a particular context, revealing the strengths and weaknesses of both the scale and the reality it measures. Given their importance as the two essential conditions for a measurement tool in scientific research, especially in the social sciences, this paper discusses validity and reliability in detail.

Keywords: *Educational Sciences, Research Instruments, Psychological, Psychometric properties.*

Propriétés psychométriques des instruments de recherche en sciences psychologiques et éducatives

Résumé :
Les propriétés psychométriques comptent parmi les outils les plus importants dans le domaine de la mesure psychologique et éducative, car elles permettent d'évaluer la validité et la fiabilité et de généraliser les résultats à l'ensemble de la population. La validité et la fiabilité sont donc des caractéristiques psychométriques essentielles qui doivent être présentes dans les instruments de mesure psychologique et éducative.
Ces propriétés sont des indicateurs statistiques obtenus en soumettant une échelle spécifique à une série de procédures expérimentales et statistiques dans un contexte particulier, révélant les forces et les faiblesses à la fois de l'échelle et de la réalité qu'elle mesure. Compte tenu de leur importance en tant que deux conditions essentielles pour un outil de mesure dans la recherche scientifique, en particulier dans les sciences sociales, cet article traite en détail de la validité et de la fiabilité.

Mots-clés *: Sciences de l'éducation, Instruments de recherche, Psychologique, Propriétés psychométriques.*

Soumission : **20/02/2025** **Acceptation : 08/08/2025** **Publication** : **15/09/2025**

Introduction

Psychologists emphasize that reliability and validity are among the most important characteristics and features of a good measurement tool. Without them, one cannot trust the ability of the instrument to measure what it was designed to measure, nor the accuracy of the results obtained when it is used to assess various traits.

There are several conditions that a test must meet to be considered an objective and effective test that fulfills its intended purpose. Such a test is incomplete unless information about its suitability as a measurement tool is available.

Definition of Psychometric Properties

Khitab (2000: 99) defined psychometric properties as the statistical indicators extracted and derived from subjecting a particular scale to a series of experimental and statistical procedures under specific conditions to reveal areas of strength and weakness in both the scale and the reality and the measurement objective.

Validity and reliability embody these properties, and due to their importance as the two required conditions in a measurement tool, they will be addressed below.

Validity

Validity is considered one of the most important psychometric characteristics of educational and psychological tests and measures because it concerns what the test measures and to what extent it succeeds in measuring it (Khitab, 2000: 109).

Validity has a broad concept with several meanings that vary according to the use of the test. The term is sometimes translated as 'soundness' or 'appropriateness.' The variation in translation usually reflects the different issues researchers focus on when discussing validity. A test is valid if it measures what it was intended to measure; in other words, a valid test measures the function it claims to measure.

A test is considered valid to the extent that it measures the trait or property for which it was prepared (Al-Qumash, 2000: 109).

Several definitions of validity exist, among which are the following:

- Robert Ebel defined validity as the accuracy with which a test measures what it should measure.
- Durethreed (or 'Dure Threed'—text variant) defined it as the extent to which the test performs the purpose for which it was intended when applied to the target group.
- Frederick Brown defined it as the extent to which the examination performs the function for which it was used.
- Test validity concerns the purpose for which the test is constructed and the decisions made based on its scores. Test scores are typically used to reach certain inferences, and the question arises as to what can be inferred with high accuracy and confidence (Allam, 2000: 186).

1. Types of Validity

There are multiple types of validity. The American Psychological Association identified three main indicators (aspects) of validity for psychological measures: content

Soumission : **20/02/2025** **Acceptation : 08/08/2025** **Publication** : **15/09/2025**

validity, construct validity, and criterion-related validity. These aspects are not separate but complementary and serve as evidence for the suitability of a measure to aid decision-making.

1.1. Content Validity

Content validity involves an initial examination of the test contents to determine whether the test items are related to the trait intended to be measured, and whether any items can be removed because they do not relate to the target trait.

Content validity refers to the degree to which the items represent the facets or aspects intended. High content validity indicates that the items provide a representative sampling of the behaviour to be measured.

Content validity also refers to the quality of the sample of items included in the measurement tool and their representation of the instructional material or curriculum under consideration. A content-valid test contains items or tasks derived from parts and types of the educational content so that it reflects all instructional units that students had the opportunity to learn.

Judging the content validity of any measurement tool requires a panel of experts and specialists in the field to evaluate the degree to which the test items represent the content from which they were derived. Content procedures begin after the initial drafting of the test items and are supervised by the test developer (Hassani, 2014: 45).

1.2. Criterion-related Validity

This type of validity is also called empirical or practical validity because it indicates the degree to which test results

agree with external facts related to the behavioural function the test measures. It refers to the correlation between individuals' scores on the test and their scores on an independent external criterion whose validity and reliability have been established and which measures the same behavioural property.

Sacks (1984) defined it as the correlation between test scores and an external criterion (Maher, 2006: 163).

Estimating criterion-related validity always involves collecting empirical data about the relationship between test scores and criterion scores. Sometimes this type of validity is referred to as empirical validity, and the main focus is the performance of individuals on the criterion measure. Test scores under examination are not the primary focus but are used to predict behaviour on an important criterion measure. Two types of criterion-related validity can be distinguished: concurrent and predictive validity (Allam, 2006: 108).

1.2.1. Predictive validity:

Predictive validity refers to the test's ability to predict an individual's subsequent performance on an independent external criterion. Abu Hatab and Othman (1985) defined it as the test's effectiveness in predicting an individual's later performance. This type of validity aims to predict, over the long term, an individual's performance in future tasks related to what the test measures and is used for practical purposes.

1.2.2. Concurrent validity:

Concurrent validity refers to the degree to which there is a relationship between individuals' scores on the test and

Soumission : **20/02/2025** **Acceptation : 08/08/2025** **Publication** : **15/09/2025**

their current performance on an independent external criterion that measures the same attribute. Concurrent validity involves collecting data on the criterion at the same time or before administering the test and comparing individuals' test scores with their criterion scores.

1.3. Construct (Hypothetical) Validity

Construct validity refers to the extent to which a test measures the hypothetical construct or trait it is intended to measure; some refer to this as 'construct validity.' This type of validity requires a substantial amount of information shedding light on the nature of the trait being measured (Khitab, 2000: 181).

Construct validity concerns a psychological attribute or property that is assumed to exist to explain certain aspects of individuals' behaviour (Allam, 2006: 110).

This approach to estimating construct validity is similar to the approach used for concurrent validity. Evidence for construct validity typically requires high correlation coefficients between the new test and tests that measure the same construct; high correlations suggest that the new test measures what established tests measure.

Discriminant validity: In estimating construct validity for the test under study, one may compute correlations between it and other tests that measure different constructs. Discriminant validity is supported when these correlation coefficients are low, which indicates the test uniquely measures its intended construct. For example, a test claiming to measure creativity should not correlate highly with established intelligence tests.

1.4. Factor Analysis

Factor analysis is a statistical technique aimed at identifying the minimal set of underlying factors that explain the intercorrelations among a set of tests or among items of the test under study when assessing construct validity.

This method involves selecting a set of external criteria alongside the test under investigation, computing intercorrelation coefficients among this set, and then factor-analysing the correlation matrix to determine the degree to which each test loads on the general factor and other common factors shared among them. The magnitude of a test's loading on the general factor indicates its validity for measuring that factor. Previously, factor analysis was considered a mathematically demanding procedure, but modern computing software has simplified the computational steps; however, interpretation remains a human responsibility (Saad, 1998: 192).

1.4.1. Factors Affecting Test Validity

Factors related to the examinee:

- Disturbance of the examinee during the test administration.
- The seriousness or carelessness of the examinee.
- Poor answering habits.

Factors related to the test:

- Unclear language in the test items, ambiguity in questions, and variations in item difficulty.

Factors related to test administration:

- Environmental factors such as temperature, humidity, and other conditions.

Soumission : **20/02/2025** **Acceptation : 08/08/2025** **Publication** : **15/09/2025**

1.4.2. Reliability

In addition to validity, reliability is one of the most important conditions required of a measure or test. The term 'reliability' indicates that the test is dependable and trustworthy. A reliable test yields consistent estimates: if the measurement procedure is repeated, the results should be consistent for the individual, meaning the individual's score does not change substantially upon repeated administrations or that the individual's position relative to the group does not change substantially between multiple administrations (Abdel Salam).

Reliability refers to the degree of precision, consistency, or stability of the results when a test is applied to a sample of individuals on two different occasions (Muqaddam, 2003: 152).

Test reliability coefficient can be defined as the correlation between individuals' scores on the test across different administrations, between estimates provided by different scorers, or between results obtained when the test is administered to the same group by different examiners.

Robert Emile defined the reliability coefficient as the correlation between one set of scores and another set obtained from equivalent tests given to the same group of students.

Reliability represents the degree of stability of a phenomenon across different occasions. Reliability is a central concept in psychometrics and, together with validity, forms the foundation for test and questionnaire construction for use (Maamria, 2012: 250).

2. Methods for Estimating Reliability

Various methods exist to confirm the reliability of psychological measures and tests. Each method estimates a particular type of error variance based on differing sources of random error. Reliability is calculated as the test's self-correlation. Below are several commonly used methods:

2.1. Test–Retest Method (Stability Coefficient)

In this method, the stability coefficient is computed by administering the same test to a particular sample on two different occasions (e.g., first administration in June and the second in July), and then computing the correlation between the two sets of scores. The correlation coefficient ranges between -1 and +1; the closer it is to +1, the higher the relationship between the two administrations and the greater the test's stability. The choice of correlation coefficient depends on the type of scale.

Several factors affect the effectiveness of this method, including the time interval between the two administrations: the longer the interval, the lower the correlation tends to be, especially for tests measuring cognitive or psychological aspects that change with development and experience. Therefore, special attention must be given to the time interval, which should be appropriate for the rate at which the measured trait changes. Some traits change rapidly (e.g., attitudes) while others change slowly (e.g., personality traits and intelligence), so the interval should be chosen accordingly (Muqaddam, 2003: 154-155).

Test–retest stability coefficients are among the most important reliability estimation methods. In this approach, a test is administered to a sample, then re-administered to the

Soumission : **20/02/2025** **Acceptation : 08/08/2025** **Publication** : **15/09/2025**

same sample under similar conditions, and the correlation between the two administrations is calculated.

2.2. Equivalent Forms Method

Also called the parallel forms method. The researcher constructs two equivalent forms of the measurement instrument that match in specifications such as the number of items, wording, content, difficulty level, objectives, instructions, and time limit. Form A is administered and after a time interval Form B is administered, and the correlation between scores on the two forms is computed; this correlation is the reliability coefficient of the instrument (Bin Safi, 2013: 26).

Sometimes it is difficult to administer the test twice to the same group or when the trait under measurement is unstable; in such cases, two equivalent forms can be administered sequentially to the same group. This method assumes that the two forms are truly equivalent in content, number and difficulty of items, and other characteristics; measurement errors here stem from differences in item samples between the two forms rather than changes in examinees. The equivalence coefficient is used when the purpose of the measure is inferential or diagnostic, such as in psychotherapy and assessment (Fouad, 1997: 122).

3. Split-Half Method (Internal Consistency Coefficient)

If it is difficult to administer two equivalent forms or to test respondents twice, the researcher may use the split-half method. In this method, the whole test is given and then divided into two equal halves such that the means, standard deviations, and difficulty levels of both halves are approximately equal. Often the first half contains odd-

numbered items and the second half even-numbered items. Scores for each half are computed separately, producing for each examinee two scores (one for each half), and the correlation between these two scores is calculated. This correlation is the split-half reliability or internal consistency coefficient (Kawafha, 2005: 87).

Because different scoring or partitioning methods can reduce the observed correlation, length-correction formulas should be used. Common correction formulas include the Spearman–Brown prophecy formula.

Spearman–Brown formula assumes that increasing test length tends to increase reliability and that half-test variances are equal, enabling prediction of the reliability of the full test from the half-test reliability. However, this assumption may not hold perfectly in practice.

Figure 1 (illustrative): Test administration → Re-administration → Compute Pearson correlation between scores.

Figure 2 (illustrative): Test administration → Alternate form → Compute Pearson correlation between scores of two forms.

Figure 3 (illustrative): Test administration → Split items into two comparable halves → Compute Pearson correlation → Apply Spearman–Brown formula to estimate full-test reliability.

Sometimes applying the test twice to the same group is impractical or the measured trait is unstable; in such cases, two equivalent forms are used. Errors in this method arise from differences between the item samples rather than from changes in the examinees. As differences between forms increase, reliability decreases.

Soumission : **20/02/2025** **Acceptation : 08/08/2025** **Publication** : **15/09/2025**

When using split-half reliability, correction formulas such as Spearman–Brown are commonly applied to estimate full-test reliability. Other formulas and corrections include several variants discussed below.

Correction Formulas and Reliability Coefficients

- Spearman–Brown formula: This formula is based on the assumption that increasing test length increases reliability and predicts full-test reliability from half-test reliability. It assumes equal variances across halves and may be less suitable when halves differ markedly (Faraj, 2007: 223).
- Rolon (1939) formula: Rolon argued that the variance in test scores arises from differences in examinees' abilities and from ordinary measurement errors. This formula is a simplified split-half correction that focuses on variance of score differences and test variance (Faraj, 2007: 323).
- Guttman–Flanagan (generalized) formula: The Spearman–Brown assumption may not hold when halves have unequal standard deviations. Guttman proposed a more general formula that can estimate reliability when half variances are unequal; it can also produce estimates without requiring a separate length-correction formula (Faraj, 2007: 323).
- Horst formula: When practical constraints prevent balanced partitioning of a test, Horst's formula can be used as a length-correction method similar to Spearman–Brown while accounting for unequal part lengths (Shakir, 2005: 127-128).

Analysis of variance (ANOVA) is another method for estimating the reliability coefficient of personality and

attitude measures. In this method, the reliability coefficient represents the homogeneity among the test's core items and the degree of association between items and the total test score (Saad, 1998: 170).

Kuder–Richardson method (homogeneity coefficient): This method applies when test items are dichotomous (right/wrong). It examines the homogeneity of all items and is affected by the content sampling and the behavioural domain of the items. KR-20 is a classic formula for dichotomous items that estimates internal consistency under the assumption of tau-equivalence.

Cronbach's alpha: Given the limitation of homogeneity methods to dichotomous items, Cronbach (1951) proposed a generalization now known as Cronbach's alpha, which is suitable for items with multiple score levels (e.g., Likert-type scales). Cronbach's alpha provides a lower-bound estimate of the test reliability and is the most suitable measure for survey research, questionnaires, attitude scales, personality measures, and many achievement tests (Al-Nabhan, 2004: 244).

Cronbach's alpha links test reliability to the reliability of its items: as item variances relative to total variance increase, alpha decreases, and vice versa. A high alpha suggests good internal consistency, while a low alpha may indicate that alternative methods could produce higher reliability estimates (Saad, 1998: 172).

Factors Influencing Reliability

- Test length: Reliability tends to increase with the number of items because more items provide a larger sample of the behaviour, increasing stable

Soumission : **20/02/2025** **Acceptation : 08/08/2025** **Publication** : **15/09/2025**

representation and individual differences, hence higher reliability.

- Test difficulty level: The difficulty index (mean test score relative to maximum score) affects reliability. Extreme ease or difficulty reduces variability among examinees and can lower reliability (Khitab, 2000: 227).
- Independence of items: Including interdependent items (where answering one item makes others easier) reduces the effective number of independent items and lowers reliability (Khitab, 2000: 229).
- Standard error of measurement: Test designers often study the standard error of measurement, which reflects the extent to which an individual's observed score is expected to fluctuate above or below their true score (Awad, 1998: 58).
- Time allowed for the test: Reliability increases with adequate time; insufficient time, especially for difficult tests, leads to guessing and reduced reliability (Al-Tell, 2007: 84).
- Guessing: Guessing negatively affects reliability in objective tests with multiple-choice items.
- Scoring (rater) reliability: Rater reliability refers to the degree of agreement among multiple scorers. Rater reliability increases when scorers are in higher agreement (Khitab, 2000: 233).

Conclusion

There is an expected relationship between a measure's validity and its reliability, especially since both concepts examine the adequacy of the test and its suitability for the fundamental assumptions of measurement theory (Saad, 1998).

A close relationship between reliability and validity is necessary because a good test should be reliable, that is, it should produce consistent results. Therefore, the correlation between them is expected to be high. However, reliability is a necessary but not sufficient condition for validity: one can obtain a reliable measure that is not valid if the test consistently measures something other than the intended construct. Thus, a high degree of consistency does not guarantee that the instrument measures the correct attribute; the test may include items that do not measure what it purports to measure, thereby weakening its validity even if performances on those items are consistent (Khitab, 2000: 243).

References

Al-Qumash, Mustafa; Al-Bawailiz, Muhammad; Al-Ma'aytah, Khalil (2000). Measurement and Evaluation in Special Education. Dar Al-Fikr for Printing, Publishing and Distribution, Jordan.

Al-Jalabi, Sawsan Shaker (2005). Fundamentals of Test Construction and Psychological and Educational Measures. Alaa Al-Din Library, Damascus, Syria.

Soumission : **20/02/2025** **Acceptation : 08/08/2025** **Publication** : **15/09/2025**

Al-Khatib, Ahmed, et al. (1985). Research and Educational Evaluation. Dar Al-Mustaqbal, Amman.

Taysir Mufleh Kawafha (2005). Measurement and Evaluation, Methods of Measurement and Diagnosis in Special Education, 2nd ed., Dar Al-Maseera, Amman.

Hamza Muhammad Doudin (2010). Advanced Statistical Analysis of Data Using SPSS, 1st ed., Dar Al-Maseera, Amman.

Sawsan Shaker Majid (2007). Fundamentals of Test Construction and Psychological and Educational Measures, 1st ed., Alaa Al-Din Publishing, Syria.

Saeed Hassan Al-Ghamdi (2003). The Extent of Variation in Psychometric Properties of Measurement Tools in Light of Variation in Number of Response Alternatives and Educational Stage, Master's Thesis, Umm Al-Qura University.

Allam Salah Al-Din Mahmoud (2000). Educational and Psychological Measurement and Evaluation, 1st ed., Dar Al-Fikr Al-Arabi, Cairo.

Ali Maher Khitab (2001). Measurement and Evaluation in Psychological, Educational and Social Sciences, 2nd ed., Cairo.

Awad Abbas Mahmoud (1998). Psychological Measurement Between Theory and Practice, Dar Al-Maarifa Al-Jami'iyya, Cairo.

Faraj Safwat (2007). Psychometrics, 6th ed., Anglo-Egyptian Library, Cairo.

Muhammad Abdel Salam Ahmed (n.d.). Psychological and Educational Measurement: Introduction to Measurement and Its Concepts, Construction of

Measures and Their Characteristics. Al-Nahda Library, Cairo.
Musa Al-Nabhan (2004). Fundamentals of Measurement in Behavioural Sciences, 1st ed., Dar Al-Shorouk for Publishing and Distribution, Jordan.
Muqaddam Abdul Hafeez (2003). Statistics and Psychological and Educational Measurement, University Printing House, Central Al-Saha, Ben Aknoun, Algeria.
Bashir Maamria (2007). Psychometrics and the Design of Its Instruments for Students and Researchers in Psychology and Education, Hibr Publications Series, Algeria, 2nd ed.
Wael Abdulrahman Al-Tell; Issa Muhammad Qahal (2007). Scientific Research in the Human and Social Sciences, Dar Al-Hamed, Amman.
Anastasi, A. & Urbina, S. (1997). Psychological Testing, 7th ed., Prentice Hall, New York.
American Psychological Association (1985). Standards for Education and Psychological Tests, Washington.