Global Education Journal
Vol. 4, 2 (May-August, 2026), pp. 231-235
EISSN: 2986-6529, DOI: https://doi.org/10.59525/gej.1669

Theoretical Review: Understanding Test Reliability in Learning

Yunisa Habibah[1], Ilda Pulungan[2], Rizki Hannum[3] Nabila Zahra NST[4], Nabilla Chairunnisa[5], Ahmad Adiwan Bincar[6]

[1,2,3,4,5,6] Sekolah Tinggi Agama Islam Negeri Mandailing Natal, Indonesia; yunisahabibah@gmail.com

ARTICLE INFO

Learning Evaluation, Psychometrics, Reliability.

Article history:

Received 2026-04-14
Revised 2026-05-12
Accepted 2026-06-13

ABSTRACT

Evaluation is a crucial instrument in the Independent Curriculum instructional cycle for measuring the achievement of learning objectives and providing pedagogical feedback. However, in practice, the provision of valid and consistent evaluation tools still faces challenges, characterized by fluctuations in test scores due to weak control over measurement error. This article aims to examine in depth the basic concept of reliability, visualize its operation, and reconstruct techniques for estimating test instrument reliability coefficients. The research method used is library research, analyzing classic and contemporary literature in the fields of educational evaluation and psychometrics using a descriptive-qualitative approach. The results of the discussion indicate that reliability reflects the level of consistency, dependability, and stability of measurement results, which can be estimated through three main approaches: stability (test-retest), equivalence (parallel forms), and internal consistency (such as Split-Half, KR-20/21, and Cronbach's Alpha). Through simulation of the Spearman-Brown (Split-Half) formula calculation on a multiple-choice test, a reliability coefficient of 0.79 was obtained. Based on Guilford's criteria, this value is included in the high category and exceeds the minimum limit of feasibility for learning evaluation instruments (0.70), so the instrument is highly reliable.

Corresponding Author:
Yunisa Habibah
Sekolah Tinggi Agama Islam Negeri Mandailing Natal, Indonesia; yunisahabibah@gmail.com

1. INTRODUCTION

The modern educational paradigm no longer views the learning process as a one-way transfer of information from teacher to student, but as a complex and interconnected structured system. In the implementation of the student-centered learning Independent Curriculum, the essence of learning is shifted to strategic facilitation that encourages students to construct their own knowledge independently, actively, and relevant to the reality of life (Ministry of Education and Culture, 2022). The interaction that is systematically designed between students, teachers, and various learning resources boils down to mastering skills, instilling scientific attitudes, and comprehensive internalization of competencies. In the instructional cycle, evaluation plays a crucial role as an instrument to measure the level of success of learning objectives, as well as a trigger for feedback for the continuous improvement of teachers' pedagogical quality (Sudijono, 2018).

However, the reality in the world of education shows that there is a large gap in the provision of valid and consistent evaluation tools. The majority of exam instruments compiled by practitioners and novice researchers are often used immediately without passing a comprehensive psychometric test. Empirically, there are often sharp fluctuations in students' scores when taking tests at different times; This phenomenon is generally not triggered by changes in student competence, but because of the poor quality of the question items presented (Khumaedi, 2012). Matondang (2007) also reminded that weak control against measurement error both due to multi-interpreted sentences, assessment bias, and environmental factors can make learning outcome data biased and invalid. Without consistency testing, teachers risk giving birth to erroneous conclusions about the actual map of students' abilities.

Theoretically, the effectiveness of the educational evaluation system is highly dependent on the fulfillment of two fundamental psychometric aspects: validity and reliability (Arikunto, 2013). If validity serves to ensure that the instrument measures the right aspects, then reliability demands the consistency, reliability, and stability of test results when retested on the same subjects at different times (Azwar, 2019). Referring to the Classical Test Theory, an instrument is considered ideal if it is able to reduce the error variance to close to zero, so that the observed score in students really represents their true ability (true score) (Suryabrata, 2014). Therefore, understanding the reliability coefficient is not just a mathematical calculation routine, but a moral responsibility to present a fair evaluation for students. This article will dissect in depth the basic theory of reliability, its operational visualization, and reconstruction of various reliability coefficient estimation techniques so that they can be applied practically by educators and researchers.

2. METHODS

This article was prepared using the library research method by conducting an in-depth study of classical and contemporary literature in the field of educational evaluation and psychometrics. Primary data sources were obtained from scientific journals related to reliability and research methodology. Data analysis is carried out in a descriptive-qualitative manner through data collection techniques, comparison of experts' theories, conceptual synthesis, and conclusion drawing to construct a systematic understanding of the estimation of the reliability of test instruments.

3. FINDINGS AND DISCUSSION

Meaningfully, the concept of reliability comes from the word reliability which means the extent to which a measuring instrument can be trusted, reliable, and consistent in photographing the phenomenon being measured (Ramadhan et al., 2024). Gronlund emphasized this concept on score consistency, which is how the instrument's score maintains its position from one test to the next. Meanwhile, Wiersma emphasizes reliability as the instrument's ability to minimize internal changes when measuring what should be measured. Empirically, high and low reliability is indicated by a number called the reliability coefficient, the magnitude of the reliability coefficient ranges from 0 to 1, where the higher the reliability number means the more consistent the measurement results, if the coefficient is close to 0, then the instrument is considered unreliable because it is dominated by random errors, but empirically the reliability coefficient that reaches the number 1 is rarely found (Khumaedi, 2012:26). In the understanding of psychometrics, the profound meaning of reliability cannot be separated from the Classical Test Theory (CTT) (Allen & Yen, 2002).

This theory postulates that every visible score or observation score obtained by a student (X) consists of two main components that are mutually additive, namely a pure score that reflects true ability without error intervention (T) and random measurement error/error (E). This theoretical relationship is formulated as follows:

$$X = T + E$$

Figure 1. The basic equation of Classical Pure Score Theory (X = T + E).

To facilitate a structural understanding of how this concept of reliability works and is tested in educational research, the following concept map summarizes its path:

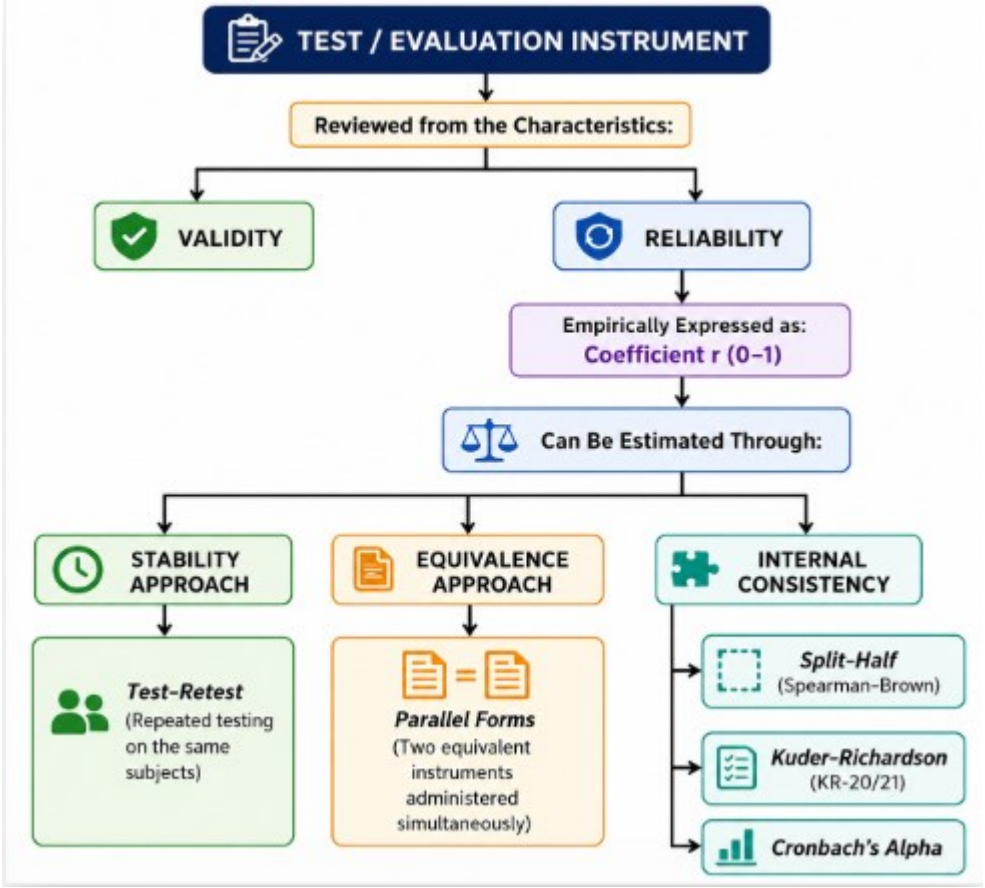

Figure 2. Test/Evaluation Instruments)

To determine the high or low reliability of the instrument, researchers or educators can choose the test technique that best suits the characteristics of the data form and format of the instrument used. (Algina, J & Crocker, L. 2008). The following table theoretically reconstructs the four popular techniques along with their specific designations and characteristics:

Table 1. Theoretical Reconstruction of Reliability Estimation Techniques

Name of Technique / Method	Main Characteristics & Formula	Characteristics of Data Scores	Practical Advantages & Challenges
Test-Retest	One instrument was tested twice on the same group at a certain time interval. Using Product Moment (rxy) correlations)	Continuous Score / Polotomy / Dichotomy	Pros: Very practical, no need for a companion instrument. Challenges: Prone to carry-over effect (participants remember questions) & mood swings
Internal consistency: Spearman -Brown (Split-Half)	One test is split into two equal (odd-even) parts. Total formula: $rnn^1 = 2\ r\ xy\ /\ 1+rxy$	Dichotomy and Polotomy Scores (must be able to be halved in a balanced manner)	Advantages: Sufficient one-time test administration, time efficient. Challenge: Cleavage must be completely homogeneous and equal

Kuder- Richardson (KR-20&KR21)	Analyzes the consistency between items simultaneously without division. Used for objective questions	Dichotomy Score (Score 1 for True answer, and 0 for False)	Pros: Highly accurate for multiple-choice objective tests. Challenge: The calculation is more complicated because it requires a proportion of graduates per item
Cronbach's Alpha	Measure the average of the inter-correlation of all non-dichotomy questions	Polotomy / Gradation Score (Example: Likert Scale 1-5, or description/essay questions)	Pros: Very popular for attitude questionnaires and description format tests. Challenge: Demanding consistency of the unidimensionality of the construct
Inter-Rater Reliability	Measure the average of the inter-correlation of all non-dichotomy questions	Categorical or Continuous Score of the rater assessment results	Advantages: The objective is to suppress subjective assessment bias (such as practice tests). Challenge: Demanding equalization of the perception of rubric references between assessors. (Mardapi, D. 2017).

Case Examples of Scale Analogy and Calculation Applications

Someone weighs a sack of rice on Monday, the needle of the scales points to the figure 5 kg. When the same sack of rice was weighed again on Thursday with the scale, the needle remained consistent at the figure of 5 kg. This condition proves that the scale has a high level of reliability (stable and consistent) (Saputri et al., 2023). In the world of education, "scales" are test question sheets, and "rice" is the cognitive competence of students. If the math test is given today and gives a high score, then when it is retested next week (without any additional learning process) the results plummet, then the question sheet is declared unreliable.

As an applied example, an educator tested the reliability of a multiple-choice objective test consisting of 10 questions. The initial step taken is to divide the total score of students into two groups: Odd Hemisphere (X) and Even Hemisphere (Y). (Dwipayani, 2022) For example, after calculating using the Product Moment (rxy) correlation formula, it was found that the correlation value between the two hemispheres was r xy = 0.65. This figure of 0.65 then reflects the relationship of half the test. To estimate the overall reliability coefficient of the test (r nn[1]), educators enter the value into the Spearman-Brown formula.

$$r_{nn'} = \frac{2 \times 0,65}{1 + 0,65}$$

$$r_{nn'} = \frac{1,30}{1,65} \approx \mathbf{0,79}$$

$$r_{nn'} = \frac{2 \times r_{xy}}{1 + r_{xy}}$$

Figure 3. Spearman-Brown Formula for Estimation of Reliability Coefficient of Split-Half Method

Based on the classic reference classification criteria from Guilford, the reliability coefficient of 0.79 is in the range of 0.60 - 0.79, which means that the learning outcome test instrument has a high level of reliability. (Scott, 2002) In general, the minimum limit of the instrument coefficient to be used in learning evaluation properly is 0.70 so this test is very reliable. (Purwanto, 2016).

4. CONCLUSION

A high-quality educational evaluation system must be supported by instruments that meet the parameters of validity and reliability. Based on Classical Test Theory, the ideal instrument must be able to reduce *the error variance to close to* zero so that the visible score obtained by students reflects the pure score of their competence. Reliability is an indicator of the extent to which a measuring tool is trustworthy, reliable, and consistent in capturing the phenomenon of students' abilities. Understanding the reliability coefficient is not just a mathematical calculation, but a methodological responsibility to ensure the fairness of the evaluation. Diversity of Estimation Techniques: Reliability coefficient estimation can be performed through a variety of approaches tailored to the data characteristics and format of the instrument, such as the Test-Retest method for stability testing, Parallel Forms for equivalence testing, as well as Split-Half, Kuder-Richardson (KR-20/KR-21), and Cronbach's Alpha for internal consistency testing.

The minimum limit of reliability coefficient for an instrument to be used properly in learning evaluation is 0.70. Application trials using the Split-Half technique with the Spearman-Brown formula resulted in a coefficient value of 0.79, which proves a high level of reliability and is very reliable in field assessment practices.

REFERENCES

Khumaedi, M. (2012). Reliabilitas Instrumen Penelitian Pendidikan. Jurnal Pendidikan Teknik Mesin, 12(1), 25-30.

Mardapi, D. (2017). Pengukuran, Penilaian, dan Evaluasi Pendidikan. Yogyakarta: Parama Publishing.

Matondang, Z. (2007). Validitas dan Reliabilitas Suatu Instrumen Penelitian. Jurnal Tabularasa, 4(1), 87-97.

Purwanto. (2016). Evaluasi Hasil Belajar. Yogyakarta: Pustaka Pelajar.

Ramadhan, M. F., Siroj, R. A., & Afgani, M. W. (2024). Validitas and Reliabilitas. Journal on Education, 6(2), 10967-10975.

Retnawati, H. (2016). Analisis Prasyarat Evaluasi: Validitas, Reliabilitas, dan Karakteristik Butir. Yogyakarta: Parama Publishing.

Saputri dkk. (2023). Analisis Instrumen Assesmen : Validitas, Reliabilitas, Tingkat Kesukaran Dan Daya Beda Butir Soal. Jurnal Ilmiah PGSD FKIP, 09, 2986–2995.

Siregar, S. (2013). Statistik Parametrik untuk Penelitian Kuantitatif. Jakarta: Bumi Aksara.

Sudijono, A. (2018). Pengantar Evaluasi Pendidikan. Jakarta: Rajawali Pers.

Sugiyono. (2019). Metode Penelitian Kuantitatif, Kualitatif, dan R&D. Bandung: Alfabeta.

Sumardi. (2020). Teknik Pengukuran dan Penilaian Hasil Belajar. Yogyakarta: CV Budi Utama.

Suryabrata, S. (2014). Metodologi Penelitian. Jakarta: Rajawali Pers.

Yati afiyanti. (2002). Validitas dan reliabilitas dalam penelitian kualitatif. Keperawatan Indonesia, 12, 137–141.

Yen, W. M. & Yen, W. M. (2002). Introduction to Measurement Theory. Long Grove: Waveland Press.

Yusup, F. (2018). Uji Validitas dan Reliabilitas Instrumen Penelitian Kuantitatif. Jurnal Tarbiyah: Jurnal Ilmiah Kependidikan, 7(1), 17-23.