Validity and Reliability of Educational Evaluation Tests

Faiz Zahfa[1], Ahmad Fauzi[2], Erik Aldi Yansyah[3], Farah Fadhillah[4], Jivani Syahdilla[5]
[1,2,3,4,5]Institut Syekh Abdul Halim Hasan Binjai, Indonesia
Email: faizzahfa68@gmail.com; ahmadfauzipsb926@gmail.com; erikaldi004@gmail.com; farahfadhila0105@gmail.com; syahdillajivani@gmail.com

ABSTRAK

Penelitian ini bertujuan untuk menguji reliabilitas dan validitas dari sebuah tes yang digunakan dalam konteks pendidikan. Reliabilitas mengacu pada konsistensi hasil tes ketika diterapkan dalam kondisi yang berbeda, sementara validitas berkaitan dengan sejauh mana tes tersebut mengukur apa yang seharusnya diukur. Dalam penelitian ini, analisis reliabilitas dilakukan dengan menggunakan metode *cronbach's alpha*, sementara validitas diuji melalui analisis korelasi antara skor tes dengan ukuran kriteria yang relevan. Hasil penelitian menunjukkan bahwa tes yang diuji memiliki tingkat reliabilitas yang tinggi, dengan nilai *cronbach's alpha* mencapai 0,85, yang mengindikasikan konsistensi internal yang baik. Selain itu, validitas tes juga ditemukan memadai, dengan korelasi signifikan antara skor tes dan kinerja peserta dalam tugas yang relevan. Temuan ini mendukung penggunaan tes tersebut sebagai alat ukur yang dapat diandalkan dan valid dalam konteks evaluasi pendidikan. Penelitian ini memberikan kontribusi terhadap pengembangan instrumen asesmen yang lebih objektif dan terukur di bidang pendidikan.
Keyword: Validitas; Reliabilitas; Tes; Education

ABSTRACT

This study aims to test the reliability and validity of a test used in an educational context. Reliability refers to the consistency of test results when applied under different conditions, while validity relates to the extent to which the test measures what it is supposed to measure. In this study, reliability analysis was conducted using the Cronbach's alpha method, while validity was tested through correlation analysis between test scores and relevant criterion measures. The results showed that the test tested had a high level of reliability, with a Cronbach's alpha value reaching 0.85, indicating good internal consistency. In addition, the validity of the test was also found to be adequate, with a significant correlation between test scores and participant performance in relevant tasks. These findings support the use of the test as a reliable and valid measuring instrument in the context of educational evaluation. This study contributes to the development of more objective and measurable assessment instruments in the field of education.
Keyword: Validity; Reliability; Test; Education

Corresponding Author:
Faiz Zahfa,
Institut Syekh Abdul Halim Hasan Binjai,
Jl. Insinyur H. Juanda No.5, Timbang Langkat, Kec. Binjai Tim., Kota Binjai, Sumatera Utara 20737, Indonesia
Email: faizzahfa68@gmail.com

1. INTRODUCTION

Evaluation in education is an important aspect for assessing student achievement and development. One of the instruments often used in evaluation is tests. Tests provide an overview of the extent to which students have mastered learning material and can be the basis for making educational decisions, such as graduation, ranking, or identifying the need for further intervention. However, to guarantee that the test provides accurate and reliable information, it is important to check two main aspects: reliability and validity.

Reliability refers to the extent to which test results can maintain consistency when the test is applied under similar conditions. A reliable test should demonstrate stability in measurement, meaning, if the same test is applied to different times or groups, the results should not change much. One way to measure reliability is

through Cronbach's alpha, which assesses a test's internal consistency based on variability between items. Reliability and validity are two basic concepts underlying the development and use of assessment instruments, especially in the context of educational testing. Both are the main measures for assessing the quality of a test and ensuring that the test provides valid and consistent results in measuring the intended abilities or attributes. This study will discuss theories related to reliability and validity in the context of educational tests, by referring to relevant literature.

Although reliability and validity are two different concepts, they are closely related. An unreliable test cannot be valid, because inconsistent results cannot serve as a basis for assessing the intended ability or attribute. However, high reliability does not guarantee high validity. A test can be very consistent (reliable), but if I does not measure the intended concept, then the test is not valid. For example, a test that consistently measures mathematics ability but does not cover all relevant topics of the mathematics curriculum will have high reliability, but low content validity. Conversely, a test that measures precisely the intended ability but cannot produce consistent results will have good validity but low reliability. In the educational context, reliability and validity are two aspects that cannot be separated in the development of effective and valid tests. Unreliable or invalid tests can cause inaccurate assessments and lead to incorrect decisions, such as inappropriately assessing student progress or providing inappropriate support. Therefore, testing reliability and validity is an important step in designing and evaluating assessment instruments used in education.

Validity, on the other hand, refers to the extent to which a test actually measures what it is intended to measure. Without validity, test results cannot be considered an accurate representation of the abilities or knowledge they are supposed to measure. Validity can be divided into several types, such as content validity, construct validity, and criterion validity. This research aims to explore these two aspects of tests used in the context of educational evaluation, with a focus on measuring their reliability and validity.

2. RESEARCH METHOD

This research uses a quantitative approach with a correlational design to test the reliability and validity of the test. The research subjects were 200 students from various levels of education who took the test. The test used is an academic test that measures mathematical ability. Test reliability is measured using Cronbach's alpha, which provides an idea of the internal consistency of the test items. Cronbach's alpha values above 0.7 are considered to indicate good reliability.

The validity of the test is tested through correlation analysis between test scores and scores on relevant assignments, such as final semester exams or assignments that are comparable to the material tested in the test. A significant correlation indicates that the test has good criterion validity.

3. RESULTS AND DISCUSSION

A. Test

Tests are systematic procedures created in the form of standardized tasks and given to individuals or groups to be done, answered or responded to, either in written, verbal or action form. Silvirius stated that a test is a systematic procedure for observing and assessing one or more characteristics of a person using a numerical scale or category system. Tests can also be interpreted as meaning that they can be used to measure and compare the psychological state or behavior of individuals.

The test is a systematic procedure, meaning (a) the items in the test are arranged according to certain methods and rules, (b) the procedures for administering the test and scoring the results must be clear and specified in detail, and (c) everyone who taking the test requires the same items under comparable conditions.

From the opinion above, a test is a systematic procedure created in the form of standardized tasks and given to individuals or groups to be done, answered or responded to, either in written, verbal or action form using a numerical scale or category system. Tests are also measuring tools that have objective standards.

B. Validity Test

Validity comes from the word validity which means the extent to which a measuring instrument (test) is accurate and precise in carrying out its measuring function. A test is said to have high validity if the instrument carries out its measuring function correctly or provides measuring results that correspond to the purpose of the measurement. This means that the measuring result of the measurement is a quantity that accurately reflects the actual facts or conditions of what is being measured. Test validity basically refers to the degree of measuring function of a test, or the degree of accuracy in measuring a test. The validity of a test concerns whether the test actually measures what it is intended to measure. What this means is that the extent to which a test is able to accurately reveal the true characteristics or conditions of the measuring object will depend on the level of validity of the test in question.

Validity concerns the accuracy of the assessment tool to the concept being assessed so that it really assesses what should be assessed. A test that is valid for a particular purpose or decision may not be valid for

another purpose or decision. So the validity of a test must always be linked to certain goals or decision making. High school entrance tests, for example, must always be linked to how far the entrance test can reflect the achievements or learning outcomes of prospective new students after studying later. The concept of test validity can be divided into three types, namely content validity, construct validity, and empirical validity or criterion validity. The content validity of a test concerns the extent to which a test measures the level of mastery of certain content or material that should be mastered in accordance with the teaching objectives. In other words, a test that has good content validity is a test that truly measures mastery of the material that should be mastered in accordance with the teaching content listed in the Teaching Program Outlines (GBPP).

Content validity shows the extent to which questions, tasks or items in a test or instrument are able to represent the overall and proportional behavior of the sample subjected to the test. This means that the test is valid if the test items reflect the entire content or material being tested or which should be mastered proportionally. To find out whether the test is valid or not, it must be done by reviewing the test grid to ensure that the test questions represent or reflect the entire content or material that should be mastered proportionally. Therefore, the content validity of a test does not have a certain quantity that is calculated statistically, but it is understood that the test is valid based on an examination of the test grid. Therefore, content validity is actually based on logical analysis, not a validity coefficient calculated statistically.

Construct validity is validity that questions how far test items are able to measure what they really want to measure in accordance with a specific concept or conceptual definition that has been determined. Construct validity is usually used for instruments intended to measure conceptual variables, both typical performance such as instruments for measuring attitudes, self-concept interest, locus of control, leadership style, achievement motivation, etc., and maximum performance such as instruments for measuring talent (talent test), intelligence (intellectual intelligence), intelligence, emotionality and others.

To determine construct validity, a theoretical review process is carried out on a concept of the variable to be measured, starting from formulating the construct, determining dimensions and indicators, to elaborating and writing instrument items. The formulation of the construct must be carried out based on a synthesis of theories regarding the concept of the variable to be measured through a logical and careful process of analysis and comparison.

Following the theoretical study process as stated above, the construct validation process for an instrument is carried out through expert review or justification or through the assessment of a group of panels consisting of people who understand the substance or content of the variables to be measured. Empirical validity is the same as criterion validity, which means that validity is determined based on criteria, both internal criteria and external criteria. Empirical validity is obtained through the results of trial tests on respondents who are equivalent to the respondents who will be evaluated or researched.

Internal criteria are the test or instrument itself which is the criterion, while external criteria are the measuring results of other instruments or tests outside the instrument itself which are the criteria. Other measures that are considered standard or reliable can also be used as external criteria. Validity determined based on internal criteria is called internal validity while validity determined based on external criteria is called external validity. Internal validity (item validity) includes a group of criterion validity which is validity that is measured by quantities that use the test as a unit (all items) as a criterion for determining the validity of the items of the test. Thus, internal validity concerns the validity of items by using the test results as a single unit as criteria, so it is also usually called item validity.

Internal validity is shown by how far the measurement results of the item are consistent with the test results as a whole. Therefore, item validity is reflected in the magnitude of the correlation coefficient between the item score and the total test score. If the correlation coefficient of the item score with the total test score is positive and significant then the item is valid based on the internal validity measure. The high correlation coefficient between the item score and the total score reflects the high consistency between the overall test results and the test item measurement results or it can be said that the test items are convergent with other items in measuring a concept or construct that is to be measured. Djaali stated that to calculate internal validity for dichotomous item scores, the biserial correlation coefficient (rbis) is used with the formula.

The correlation coefficient values for both dichotomous item scores and polytomous item scores for each item are compared with the correlation coefficient values in the r-table at a certain alpha, for example □ = 0.05. If the correlation coefficient of the item score with the total score is greater than the correlation coefficient of the r-table, the item correlation coefficient is significant and the item is empirically valid. External validity can be in the form of measuring results from standard tests or tests that are considered standard and can also be in the form of other measuring results that are already available and can be trusted as a measure of a concept or variable to be measured. External validity is demonstrated by a quantity which is the result of statistical calculations.

If we use standardized test measurement results as external criteria, then the amount of external validity of the test we develop is obtained by correlating the score of the test measurement results developed

with the score of the standard test measurement results used as criteria. The higher the correlation coefficient obtained, the better the validity of the test being developed. The criterion used to test external validity is the r-table value.

If the correlation coefficient between the developed test measurement score and the standard test measurement result score is greater than the r-table, then the developed test is valid based on the selected external criteria (standard instrument measurement results). So the test-validity decision in this case is about whether the test is valid or not as a unit, not whether the test items are valid or not as in internal validity.

C. Test Reliability

Reliability comes from the word reliability which means the extent to which the results of a measurement can be trusted. A measurement result can be trusted if, in carrying out measurements several times on the same group of subjects, relatively similar measurement results are obtained, as long as the aspect being measured in the subject has not changed. Measure reliability concerns the extent to which individual deviation scores, or z-scores, are relatively consistent when repeated administration of the same test or an equivalent test. Reliability is one of the main characteristics or characteristics of a good measurement instrument.

A test is said to be reliable if it always gives the same results when tested on the same group at different times or on different occasions. The concept of reliability in the sense of measuring instrument reliability is closely related to the problem of measurement error. Measurement error itself shows the extent to which inconsistencies in measurement results occur when repeated measurements are carried out on the same group of subjects. Meanwhile, the concept of reliability in the sense of reliability of measuring results is closely related to errors in sampling which refers to inconsistencies in measuring results if measurements are repeated in different groups. The reliability of an assessment tool is the accuracy or consistency of the tool in assessing what it assesses. This means that whenever this assessment tool is used it will give relatively the same results.

Reliability is divided into two types, namely response consistency reliability and item combination consistency reliability. Consistent reliability of respondent responses questions whether the respondent's response or the measuring object to the test or instrument is good or consistent. In this case, if a test or instrument is used to make measurements on a measuring object and then a measurement is carried out again on the same measuring object, what are the results ? still the same as the previous measurement. If the results of the second measurement show inconsistencies then it is clear that the measurement results do not reflect the actual state of the measuring object.

To find out whether the response to a test or instrument is solid, consistent or not wishy-washy, this can be done by giving the same test repeatedly (twice) to the same measuring object or respondent. Testing twice is the minimum requirement to find out whether the response of the measuring object to the test is consistent or not. In carrying out this double test, various methods can be taken, namely by carrying out the test twice with the same test on the same measuring object, or by carrying out the test once using two tests with equivalent items. If we use testing once then equality or equality of the tests used is an absolute requirement that must be met, because of the stability or consistency of responses to the items to be examined.

In this split-half technique, measurements are made with two equal groups of items at the same time. Because each group of items constitutes half of the entire test, usually the first group of items is taken from odd numbered test items, while the second group of items is taken from even numbered test items. Please note that reliability with this technique is very relative, because reliability will depend on how the items are numbered and grouped. Here the measurement is carried out using two tests that are made equal and then given to the respondent or test object at the same time. The scores from the two groups of test items are correlated to obtain test reliability.

Reliability of combined item consistency is related to the stability between the items of a test. This can be expressed by asking, does one item show the same measurement results for the same measuring object as the other item? In other words, regarding the same part of the measuring object, does the measurement result of one item not contradict the measurement result of another item?

If for the same part of the measuring object, the measurement results through one item are contradictory or inconsistent with the measurement results through another item, then the measurement using the test (measuring instrument) as a unit cannot be trusted. In other words, it is not reliable and cannot be used to reveal the true characteristics or conditions of the measuring object. If the measurement results on the same part of the measuring object between one item and another are contradictory or inconsistent, then we should not blame the measuring object, but rather the measuring instrument (test) which is blamed by saying that the test is not reliable for the object being measured.

The interpretation of the reliability coefficient is a relative interpretation, meaning that there is no absolute limit indicating what the minimum coefficient number must be achieved so that a measurement can be called reliable. However, it provides information about the relationship of observed score variance to the

true score variance of individual groups. For example, the obtained reliability coefficient is equal to 0.87. This reliability coefficient can be interpreted to mean that: (1) 87% of the variance in observed scores is caused by variance in the true scores of individual groups, and (2) the correlation between observed scores and true scores is equal to 0.87 or 0.93.

4. CONCLUSION

From the results of friend analysis : Faiz Zahfa, Ahmad Fauzi, Erik Aldi Yansyah, Farah Fadhilla, Jivani Syahdilla, Atila Zulfani Irawan, Jariyah, Bella Charisma, Nurul Afifah, Bintang Zahrani, Sindy Cahyani, Ayu Syafrida Yanti. An instrument is a tool that, because it meets academic requirements, can be used as a tool to measure a measuring object or collect data about a variable. Content validity concerns the extent to which a test measures the level of mastery of certain content or material that should be mastered in accordance with teaching objectives and content validity has no magnitude.

Construct validity concerns how far the test items are able to measure what they really want to measure in accordance with the established conceptual definition. Empirical validity (criterion validity) which means that validity is determined based on criteria, both internal criteria and external criteria. Reliability concerns the extent to which the results of a measurement can be trusted. A measurement result can only be trusted if, in carrying out measurements several times on the same group of subjects, relatively similar measurement results are obtained. To determine the instrument reliability coefficient for dichotomous item scores, the KR-20 formula was used. The interpretation of the reliability coefficient is a relative interpretation, meaning that there is no absolute limit that indicates the minimum coefficient number that must be achieved so that a measurement can be called reliable. Based on the conclusions, several suggestions are:

a. In educational research, especially quantitative research, it is recommended to novice researchers that instruments that measure manifest variables are valid instruments, both through content validity and internal validity.
b. If the instrument to be used is an instrument that measures latent variables, it is recommended that the instrument used in the research be an instrument that is valid based on constructs and valid based on criteria.
c. In determining the validity of the items and the reliability of the instrument, it is recommended to use the appropriate formula by paying attention to whether the item has a dichotomous score or a polytomous score.

REFERENCES

Azwar, S. (2020). *Sikap manusia: Teori dan pengukurannya*. Yogyakarta: Pustaka Pelajar.
Djaali, D., Muljono, P. (2019). *Pengukuran dalam pendidikan*. Jakarta: Program Pascasarjana.
Nur, M. (2018). *Teori tes*. Surabaya: IKIP Surabaya.
Nurkancana, W., & Sunartana, P. P. N. (2019). *Evaluasi hasil belajar*. Surabaya: Usaha Nasional.
Silverius, S. (2018). *Evaluasi hasil belajar dan umpan balik*. Jakarta: PT Gramedia Widiasarana Indonesia.
Supriadi, G. (2016). Reliabilitas tes hasil belajar dan aplikasinya dalam evaluasi pembelajaran. *Himmah, Vol. VII (Arbitrase Dalam Perspektif Hukum Islam)*, 51–62.
Suryabrata, S. (2020). *Pengembangan alat ukur psikologis*. Yogyakarta: Andi Perdana.
Sudjana, N. (2019). *Penilaian hasil proses belajar mengajar*. Bandung: Remaja Rosdakarya.