Bulletin of Community Engagement
Vol. 1, No. 1, January 2021
https://www.attractivejournal.com/index.php/bce//index

Test Validity and Reliability in Learning Evaluation

Erlinawati[1*], Muslimah[2]
Institut Agama Islam Negeri Palangka Raya, Indonesia

E-mail : muslimah.abdulazis@iain-palangkaraya.ac,id*

ARTICLE INFO
Article history:
Received
December 05, 2020
Revised
January 05, 2021
Accepted
January 07, 2021

Abstract

The evaluation aimed to get achievement in assessing an educational target by students. In terms of assessment, a quality assessment tool is needed to fulfil two things: validity and reliability. Information that lacks validity and reliability will lead to biased conclusions was not in sync with what it should be and may differ from habit. This article aimed to determine the basic concepts of the validity and reliability of tests in evaluation. This article was also written using literary study methods and qualitative approaches. The discussion of this article, namely: validity is the degree of ability of a test that measures what is to be measured in learning. There are two kinds of validity, namely logical validity and empirical validity. Four types of validity are often used: content validity, construct validity, concurrent validity, and predictive validity. At the same time, reliability is another character of the evaluation results in learning. Reliability can also be interpreted as the same as consistency or consistency. There are two general ways to measure reliability, namely: reliability stability and equivalent reliability.

Keywords: Validity and Relibility, Learning Evaluation

Published by CV. Creative Tugu Pena
Website https://www.attractivejournal.com/index.php/bce/

INTRODUCTION

Today, educational evaluations point to a broader direction concerning the evaluation of educational objectives, program content, and others. Thus, the essence of evaluation is the process of giving or determining the value to particular objects based on specific criteria (Nerita, Maizeli, & Afza, 2017). Therefore, evaluation has an essential meaning in learning activities carried out by an educator. The purpose of the evaluation, among others, is to assess the achievement of educational goals by students, a means of finding out what students already know in learning activities, and motivating students. To evaluate student learning outcomes and learning processes, a teacher uses various evaluation tools or instruments such as written tests, oral tests, observation checklists, questionnaires-interviews, and documentation.

The success in revealing the results and the learning process as it is (the objectivity of the assessment results) is highly dependent on the quality of the assessment tools, and no less critical depends on how it is implemented (Sharder, et al., 2017; Du, Y., Arkesteijn, , den Heijer, & Song,2020). An assessment tool is said to have good quality if the tool has or fulfils two things, namely validity (accuracy) and reliability (constancy or consistency) of the test tool is guaranteed quality. What kind of test tool and what is said to have this validity and reliability, then the writer will describe it in this study by raising the title "Test Validity and Reliability in Learning Evaluation".

METHOD

The research is qualitative study. It includes the process of exploring and understanding the meaning of individual and group behavior, describing social problems or humanitarian problems (Killam & Heerschap, 2013; Conway, 2014; Moser & Korstjens, 2018). This type of research is a literature study or literature study that contains theories that are relevant to research problems. This section reviews the concepts and theories used based on the available literature, especially from articles published in various scientific journals. A literature study serves to build concepts or theories that form the basis of study in research. After all the data has been collected, the next step is to analyze the data so that a conclusion is drawn (Arikunto, 2010). Literature study or literature can be interpreted as a series of activities relating to the method of collecting library data, reading and recording and processing research material (Mulyatingingsih & Nuryanto, 2014). The method of collecting data in this study is documentation, which is collecting data and information from some relevant literature. This means that the researcher examines and/or explores several journals, books, and documents (both printed and electronic) as well as other sources of data and/or information deemed relevant to research or studies. The data analysis technique used is content analysis, which is a technique used to analyze and understand the text.

Content analysis is a research technique for making replicable and valid inferences from texts (or other meaningful matter) to the contexts of their use. As a technique, content analysis involves specialized procedures. It is learnable and divorceable from the personal authority of the researcher. As a research technique, the content analysis provides new insights, increases a researcher's understanding of particular phenomena, or informs practical actions (Krippendorff, 2018). The stages to be carried out in this study are the first is determining the theme. At this stage, researchers conducted more observations of data in the form of documents. Look for topics of interest and in this study topic of interest to researchers are finding a framework for a lifetime education the second is formulating the Problem. This stage is the reason why a topic is decided to be tested. This is done by researchers to formulate the problem with the chosen theme. The third is Collect data and determine research methods. Namely conducting theoretical studies related to the research topic. Literature sources can be obtained from books, journals, magazines, news, research results (thesis, thesis, and dissertation) and other relevant sources. The fourth is Analyze and compile the data findings. The last is draw conclusions. This stage is the answer to the research objectives which are at the conceptual/theoretical level. Researchers regularly arrange the data obtained so that they can conclude from the data that has been collected

RESULT AND DISCUSSION

Validity comes from the word validity, which means the accuracy and accuracy of a measuring instrument in performing its measuring function. A scale or measuring instrument can be said to have high validity if the instrument performs its measuring function, or provides measuring results following the purpose of the measurement (Peijuan, et al., 2017; Zilvinskis, J., Masseria, A. A., & Pike, 2017). Meanwhile, tests that have low validity will produce data that is not relevant to the measurement objectives. Validity is the degree of ability of a test that measures what is being measured. Indirectly it includes tests and scales consisting of some tasks selected to serve as indicators of learning outcomes. Validity relates to the appraisal tool's appropriateness against the concept being assessed so that it assesses what should be assessed—for example, assessing students' ability in mathematics (Günüç, S., Odabasi, H. F., & Kuzu, 2014). For example, questions are given with long and convoluted sentences so that the meaning is difficult to grasp. Finally, students could not answer because they did not understand the question. Validity does not apply because it depends on the situation and the purpose of

the assessment. An assessment tool that has been valid for a particular purpose will not automatically be valid for another purpose.

In using the validity of a test, several things need to be considered, namely: referring to the material to be tested; refers to the results of a test or evaluation instrument imposed on a group of individuals; relating to the degree or with the validation terms high, medium, low; and refers to the use of evaluation results. The validity of an evaluation instrument has several essential meanings including validity relates to the accuracy of the interpretation of test results or evaluation instruments for an individual group and not the instrument itself; validity is defined as the degree to which a category can include categories that can include low, medium and high categories; the principle of a test is valid, not universal (Shen, Chen, & Hu, 2014). The validity of a test that researchers need to pay attention to is that it is only valid for a purpose (Sukardi, 2009). There are two critical elements in the test's validity: the validity of a test must show a certain degree, some are perfect, some are moderate, and some are low; validity is always associated with a decision or a specific goal. As the opinion of R. L Thorndike and H. P Hagen that "validity is always in relation to a specific decision or use (Zainal Arifin, 2011).

There are two kinds of validity, namely logical validity and empirical validity. First, the authors describe what logical validity is. The term "logical validity" contains the word "logical" derived from the word "logic" or logical validity is often referred to as qualitative analysis, which is in the form of reasoning or analysis (Topala, & Tomozii, 2014; Ilker, 2014). With this meaning, the logical validity of an instrument that meets the requirements is valid based on reasoning (Joko Widiyanto, 2013). This reasonable condition is considered fulfilled because the instrument in question has been well designed, following existing theories and conditions. As with the implementation of other tasks, such as making an essay, if the writing follows the writing rules, logically, the essay is good. Based on this explanation, the instruments that have been compiled based on the theory of instrument preparation are logically valid. From this explanation, it can be understood that logical validity can be achieved if the instrument is arranged according to existing provisions.

Thus, it can be concluded that logical validity does not need to be tested for conditions but is obtained immediately after the instrument has been compiled. Second, empirical validity. The term "empirical validity" contains the word "empirical" which means "experience". An instrument can be said to have empirical validity if it has been tested from experience (Mohamad, et al., 2015). Quantitative question analysis emphasizes the analysis of the internal characteristics of the test through empirically obtained data. Quantitative internal characteristics are intended to include parameters about the level of difficulty, distinguishing power and reliability.

Specifically for multiple-choice questions, two additional parameters were seen from the chance to guess or answer the right questions and the answer choices' function, namely the distribution of all alternative answers from the tested subjects. One of the objectives of conducting the analysis is to improve the quality of the questions, namely whether a question is acceptable because it has been supported by adequate statistical data, is corrected because it is proven that there are several weaknesses or is not used at all. After all, it is empirically proven that it does not function at all. In implementing learning, four types of validity are often used, namely:

1. Content validity A test is said to have content validity if it measures individual specific objectives that are parallel to the subject matter or content is given. Because the material taught is stated in the curriculum, its validity is often called curricular validity.
2. Construction validity (construct validity) A test is said to have construction validity of the items that construct the test measure every aspect of thinking as stated in the

specific instructional objectives. In other words, if the items measure the thinking aspect, it is following the thinking aspect which is the instructional goal.

3. The validity "is now" (concurrent validity). This validity is more commonly known as empirical validity. A test is said to have empirical validity if the results are consistent with experience. If there is a term "appropriate", of course, two things are paired. In this case, the test results are paired with the experience results. Experience is always about the past so that the experience data is now there (present, concurrent). In comparing the results of a test, an appeal criterion or tool is needed. Then the test results are something to compare.
4. Predictive validity Predicting means predicting, always predicting things to come, so now it has not happened. A test is said to have predictive validity or predictive validity if predicting what will happen in the future. As a comparison tool, predictive validity is the values obtained after test takers have attended educational institutions' lessons. If it turns out that whoever has a higher test score failed the 1st semester exam than the one whose test score was lower in the past, the admission test or class promotion in question would not have predictive validity. An example that can illustrate the validity of the Islamic Religious Education (PAI) teacher who will assess students' abilities and understanding in the practice of prayer, the teacher should use this type of practical test to obtain test results that are suitable for the purpose. The emphasis here is that a valid test in assessing one group is not necessarily valid if the test used is the same as another group because each member of the group has differences (Sukardi, 2009).

Next, discuss test reliability. Reliability evaluation is used to determine the measuring instrument's stability so that it is reliable and remains stable when re-measuring (Saiful Azwar, 2018). Walizer said that the definition of reliability is the consistency of measurement. According to Masri Singarimbun, reliability is an index that shows the extent to which a measuring instrument is reliable or reliable. If a measuring device is used twice to measure the same symptoms and the measurement results obtained are relatively consistent, then the measuring device is reliable. In other words, reality shows the consistency of a measuring device in the same symptom meter. According to Sumadi Suryabrata, reliability shows to what extent the measurement results with these tools can be trusted. The measurement results must be reliable because they must have a level of consistency and stability. In Aiken's view, a test is said to be reliable if the scores obtained by participants are relatively the same despite repeated measurements.

Reliability is another character of the evaluation results. Reliability can also be interpreted as the same as consistency or consistency. An evaluation instrument is said to have a high-reliability value if the tests made have consistent results in measuring what is being measured. This means that the more reliable a test is, the more confident we can state that test results have the same results and can be used in a school when the test is carried out. Question reliability is a measure that states the level of consistency or consistency of a test question. To measure the level of consistency of this question used Cronbach alpha calculation.

1. Reliability Stability It concerns getting the same or similar value for every person or unit measured every time measuring it. This reliability involves using the same indicators, operational definitions, and procedures for collecting data at any time and measuring them at different times. Obtain stability reliability every time the unit is measured; the score must be the same or almost the same.
2. Equivalent Reliability Concerning the effort to obtain the same relative value with different types of measurements simultaneously. The conceptual definition used is the same, but with different indicators, operational limitations, data collection tools, and observers. Testing reliability by using an equivalent measure at the same time can

take several forms. The most common form is called the middle-slit technique. This method is often used in surveys. When a set of questions measuring one variable is entered in the questionnaire, the questions are divided into two parts in exactly a certain way. (Shuffling or changing is often used for this middle-split technique.)

Each part's results are summarized into the score; then the scores section is compared when the scores for each section are compared. If the two scores are relatively the same, the intermediate reliability is achieved. Equivalent reliability can also be measured using different accounting techniques. Anxiety, for example, has been measured by pulse reports. The relative scores of one indicator of this kind must match the scores of another. So if a subject appears to be anxious at the "restless measure" that person must show the same level of relative accuracy when his blood pressure is measured. After describing the general way to measure reliability in the evaluation of learning, the authors describe the factors that influence the reliability coefficient. At least four factors affect the reliability coefficient of the learning evaluation instrument in the form of a test, namely:

1. Test Length. In general, the longer the test, the higher the reliability. It is because a test with many items will contain quite a lot of measured behaviour.
2. Score Spread. The spread of scores influences the reliability coefficient. The wider the spread of scores, the higher the estimated reliability coefficient. The reliability coefficient will be higher if the individuals tend to remain in their position towards the group.
3. Test difficulty level. Tests that are too difficult or too easy tend to lower the reliability coefficient. These too tricky or too easy tests produce a limited distribution and accumulate at the bottom or top ends.
4. The objectivity of a test shows how far two people who have the same ability get the same score.

A test's objectivity shows how far two people who have the same ability get the same score.

This research was in line with some finding, for instance Miranda, et al., (2016). Who Identification of gifted students by teachers: Reliability and validity of the cognitive abilities and learning scale. Pitkethly, & Lau, (2016). Reliability and validity of the short Hong Kong Chinese Self-Regulation of Learning Self-Report Scale. And Kalk, K., Luik, P., Taimalu, M., & Täht, K. (2014) who discussed about Validity and reliability of two instruments to measure reflection: A confirmatory study.

CONCLUSION

Validity is the degree of ability of a test that measures what is being measured in learning. There are two kinds of validity, namely logical validity and empirical validity. Four types of validity are often used: content validity, construct validity, concurrent validity, and predictive validity. At the same time, reliability is another character of the evaluation results in learning. Reliability can also be interpreted as the same as consistency or consistency. There are two general ways to measure reliability, namely: reliability stability and equivalent reliability. Equivalent reliability can also be measured using different accounting techniques.

REFERENCES

Burhan Nurdiyantoro, *Penilaian Pembelajaran Bahasa*, Yogyakarta: BPFE 2010.

Du, Y., Arkesteijn, M. H., den Heijer, A. C., & Song, K. (2020). Sustainable Assessment Tools for Higher Education Institutions: Guidelines for Developing a Tool for China. *Sustainability*, *12*(16), 6501.

Ilker, E. (2014). A Validity and Reliability Study of the Motivated Strategies for Learning Questionnaire. *Educational Sciences: Theory and Practice*, *14*(3), 829-833.

Joko Widiyanto, *Evaluasi Pembelajaran,* Madiun: Unipma Press, 2013.

Günüç, S., Odabasi, H. F., & Kuzu, A. (2014). Developing an effective lifelong learning scale (ELLS): Study of validity & reliability. *Egitim ve Bilim*, *39*(171).

Kalk, K., Luik, P., Taimalu, M., & Täht, K. (2014). Validity and reliability of two instruments to measure reflection: A confirmatory study. *TRAMES: A Journal of the Humanities & Social Sciences*, *18*(2).

Mohamad, M. M., Sulaiman, N. L., Sern, L. C., & Salleh, K. M. (2015). Measuring the validity and reliability of research instruments. *Procedia-Social and Behavioral Sciences*, *204*, 164-171.

Miranda, L. C., Araújo, A. M., & Almeida, L. S. (2016). Identification of gifted students by teachers: Reliability and validity of the cognitive abilities and learning scale. *RIDPSICLO*, *2*(3), 5.

Nerita, S., Maizeli, A., & Afza, A. (2017). Student analysis of handout development based on guided discovery method in process evaluation and learning outcomes of biology. *IOP Publishing*, *895*(1), 1-4.

Peijuan, Z., Ming, W. C., Zhouhong, Z., & Liqi, W. (2017). A new active learning method based on the learning function U of the AK-MCS reliability analysis method. *Engineering Structures*, *148*, 185-194.

Pitkethly, A. J., & Lau, P. W. (2016). Reliability and validity of the short Hong Kong Chinese Self-Regulation of Learning Self-Report Scale (SRL-SRS-C). *International Journal of Sport and Exercise Psychology*, *14*(3), 210-226.

Saiful Azwar, *Reliabilitas dan Validitas*, Cetakan ke-4, Yogyakarta: Pustaka Pelajar, 2018.

Shrader, S., Farland, M. Z., Danielson, J., Sicat, B., & Umland, E. M. (2017). A systematic review of assessment tools measuring interprofessional education outcomes relevant to pharmacy education. *American Journal of Pharmaceutical Education*, *81*(6).

Shen, W. Q., Chen, H. L., & Hu, Y. (2014). The validity and reliability of the self-directed learning instrument (SDLI) in mainland Chinese nursing students. *BMC medical education*, *14*(1), 108.

Sukardi, *Evaluasi Pendidikan Prinsip dan Operasionalnya*, Jakarta: Bumi Aksara, 2009.

Topala, I., & Tomozii, S. (2014). Learning satisfaction: Validity and reliability testing for students' learning satisfaction questionnaire (SLSQ). *Procedia-Social and Behavioral Sciences*, *128*, 380-386.

Zainal Arifin, *Evaluasi Pembelajaran,* Bandung: Remaja Rosdakarya, 2011.

Zilvinskis, J., Masseria, A. A., & Pike, G. R. (2017). Student engagement and student learning: Examining the convergent and discriminant validity of the revised national survey of student engagement. *Research in Higher Education*, *58*(8), 880-903.