Towards Quality Education: An Empirical Study on the Reliability Index of Final Exam Questions

Khusnul Khotimah*[1], Rusijono Rusijono[2], Hari Sugiharto Setyaedhi[3], Irena Yolanita Maureen[4]

[1,2,3,4]Universitas Negeri Surabaya, Indonesia

khusnulkhotimah@unesa.ac.id; rusijono@unesa.ac.id;
harisetyaedhi@unesa.ac.id; irenamaureen@unesa.ac.id

Abstract. The process of learning should always include some form of evaluation. The outcomes of assessments can be considered by lecturers when making judgments about the learning process, which can ultimately improve both the quality of learning and the learning process students engage in. The estimated reliability index of the final examination taken by undergraduate students will be investigated in this study. The instrument's dependability was evaluated through the single presentation approach in this particular research endeavor. This study is an evaluation study whose purpose is to assess the quality of the items on the final test. The population of this research consists of all of the final exam questions, and the samples taken for this study consisted of seven different courses. The documentation approach was employed to acquire the data for this study. Cronbach's alpha was the method of analysis that was utilized. The study found that the reliability index of the final exam questions was low. The highlights that the poor reliability of final exam questions stems largely from instructors not following established best practices for question development. Although this lapse can be attributed to limitations like time and resources, it's crucial for the integrity of educational assessment that action be taken. To address this, a unified effort from educators and policymakers is recommended. This includes professional development for lecturers in question formulation methodologies and allocating appropriate resources for these initiatives, possibly guided by structured models like CDCAT. This finding also suggests there is a need for collaboration between lecturers and policymakers to allocate adequate time, cost, and effort for ideal question development, even though there may be challenges.

Keywords: learning, assessment, reliability, Cronbach's alpha, final exam questions, education, quality education.

1 Introduction

A crucial component of the learning process is assessment [1]. When there is a learning process, there must be an assessment process so that the assessment process con-

M. Salimi et al. (eds.), *Proceedings of the 7th International Conference on Learning Innovation and Quality Education (ICLIQE 2023)*, Advances in Social Science, Education and Humanities Research 873,
https://doi.org/10.2991/978-2-38476-301-6_40

tinues to be carried out on an ongoing basis [2]. Furthermore, the information obtained from the assessment process can be used to evaluate students and assess the success of teachers and institutions in carrying out learning [2] [3]. Assessment of student learning requires techniques to measure student achievement [4]. However, assessment is not just a collection of techniques but a systematic process that plays an essential role in effective teaching [5]. Assessment can and should provide information that enhances and encourages learning [6] [7]. The assessment results can be used to assist in making decisions about the learning process to improve the quality of education and the learning process of students [6].

Higher education increasingly emphasizes assessment, particularly formative assessment, for learning. Scholars working in education know its critical significance in acquiring skills and abilities [8]. Educators are responsible for instructing students to create learning, verify ongoing education, and improve teaching effectiveness [2]. The assessment begins with identifying learning objectives, monitors the progress achieved by students towards these goals, and ends with an assessment of the extent to which these objectives have been completed [9]. Not only planning to learn, but lecturers are also responsible for planning instruments, which are then used to evaluate learning processes and outcomes. In this case, the competence of educators to determine the proper evaluation is essential [9].

Functionally, assessment involves gathering student information to aid educators in decision-making for learning [10], [11]. Assessment means collecting information about students that can be used to assist in the decision-making process as a follow-up [10]. Therefore, the fundamental problem in assessment is not only collecting data or information but rather a systematic approach so that the evaluation results can play a significant role in an effective learning process [9].

The quality of the assessment is determined by the data or information collected [12]. Good data can be used to make the right decisions; otherwise, insufficient data will result in wrong decisions [13]. Thus, the quality of the data collected is strongly influenced by the quality of the instruments used [12]. Therefore, the assessment is meaningful or can be used to make decisions that can improve the quality of learning and motivate student learning [14]. Before carrying out the assessment, it is necessary to determine in advance the purpose of the evaluation and what aspects will be assessed [2]. Determination of these two things is significant because it will determine the data to be collected, data collection methods, and instruments used in data collection [14]

There are several alternative methods to determine learning progress. Peer assessment and self-assessment are two of them [12], [15]. The system has been implemented and has proven effective in improving skills, promoting active learning in which students act as assessors or assessors, and reducing lecturers' time to provide assessments to each of their students. However, this can also potentially be biased due to the lack of student skills in understanding assessment standards from the expert side and the possibility of bias due to the strength of the relationship[16]. In addition to peer and self-assessment, assessment using exams is one of the most frequently used methods in education.

In this study, according to the rules contained in State University of Surabaya academic guidelines, four different aspects comprise the assessment components utilized to measure the progress made in learning. The student's involvement in the lectures (worth two points), the assignments (worth three points), the midterm exams (worth two points), and the final exams (worth three points) make up these four elements. The lecturer's observations of what the students do during the lecture are used as the basis for evaluating the student's level of engagement. In this context, the term "task" refers to an organized assignment offered by the lecturer to the students after a series of lecture sessions. However, a structured assignment can be given to the students even if there is only one meeting of the lecture. Exams given approximately halfway through the semester (around the 8th meeting) are meant to be understood by the term "mid-exams." Although the "final examinations" are done after the lecture process (15 face-to-face meetings), the content of these examinations covers all that was covered in the lectures.

In determining an excellent measuring instrument, experts have established the main criteria for the instrument used in making measurements. These criteria are the validity and reliability of the instrument. Validity is indicated by the high accuracy and accuracy of the measurement results [17]. A valid instrument is an instrument that produces accurate information or data about the attributes or variables being measured. In comparison, reliability is a translation of the word reliability [17]. Measurement instruments that can produce data that have a high level of reliability are called reliable instruments. Some terms often used for reliability include consistency, reliability, trustworthiness, stability, and constancy. The central concept in all these terms is how much the measurement process results can be trusted [17].

The measurement results can be trusted if several measurements produce relatively the same data. In the concept of measurement, what is meant by being relatively the same means acknowledging the differences between one measurement and another (with the same measurement object). If the difference is vast, then the measurement results cannot be trusted or, in other words, are unreliable. Measurements with unreliable results can be categorized as inaccurate because consistency is a requirement for accuracy [18], [19].

Measurement is the process of determining how much of a quality, attribute, or feature someone or something possesses. Measurement enables more objective trait descriptions and simplifies comparison [5]. Measurement is the systematic ascertaining of a characteristic property or attribute through a numerical device[1]. The device may be an inventory, a checklist, or a test. Measurement is limited to quantitative descriptions of behavior and does not include qualitative reports or judgments of the desired behavior [20], [21]. Therefore, we can define measuring as acquiring data or information and numerically describing it. Measurement in educational environments is concerned with student performance, which is typically expressed in quantitative terms [22]. It happens by applying a measurement tool in a specific learning scenario and results in descriptive data [23].

Understanding the reliability of measuring instruments and the reliability of measuring results are often considered the same. There are differences in the meaning of the two terms that need to be considered. The concept of measuring instrument relia-

bility is related to measurement error (error of measurement) [24]. The idea of measurement error in question refers to the extent of the instability (inconsistency) of the measurement results when two sizes are made on the same subject group [25]. The concept of reliability of measurement results is related to errors in sampling. However, both can affect the quality of the evaluation results obtained [26] [17].

Based on the explanation, consistency in the measurement results is essential. This study aims to estimate the reliability index of final examination items in the Educational Technology Department of the State University of Surabaya. The determination of this research topic is based on three things—first, the importance of the final examination function in the lecture process. Second, the high proportion of final examination is 3 or 30% of a course's final score. Third, the final exam material covers all lecture material. This research also aims to critically examine the reliability index of final exam questions as an indicator of educational quality in higher education. Through the application of rigorous statistical analyses and a comprehensive review of existing assessment methods, this study aims to evaluate the consistency, precision, and accuracy of final exam questions across multiple courses. By doing so, it seeks to contribute valuable insights into the effectiveness of current assessment practices, identify areas for improvement, and provide actionable recommendations for educators, policymakers, and institutions aiming to optimize educational outcomes

2 Method

This study is an evaluation study whose purpose is to evaluate the quality of the questions on the final exam. Assessment research is the process of applying research skills to determine the value or benefits of educational practice [2], [5], [27]. All of the courses offered by the department are included in this study's population. A stratified proportional random sample approach was utilized to gather the data for this study. The sample was chosen at random and proportionally, considering the number of classes that make up each subject area, to obtain a sample of seven different areas of study.

This research was conducted over six months in 2020 at the Faculty of Education in the Educational Technology Department at the State University of Surabaya. Of all the courses, seven chosen courses could represent the whole because of the large number of courses. Documentary research is the approach that was taken to obtain the data. The student's response sheet for the final exam of the semester, which the lecturer of the course has assessed is used for data collecting. The answer sheet that has already been scored for each item is the answer sheet that can be evaluated further.

In this particular research project, the single presentation approach was chosen to analyze the test's dependability. This strategy has been decided upon since the questions on the final examination cannot be repeated. The researcher is not authorized to request lecturers or a team of lecturers to organize parallel tests in any capacity. The alpha formula developed by Cronbach is the basis for the calculation technique used as a reference for the dependability score [17].

3 Results

This study examines the approximated index of exam questions' reliability. This study will characterize the reliability index of final exam questions for each course and define the average reliability index of final exam questions for the department courses. Before showing the final exam reliability index for each course, here is the index classification that will be used to determine the reliability of the analyzed questions.

Table 1. The coeficient clasification of reliability

Number	Coefficient reliability	Description
1	$0,80 \leq r < 1,00$	Very high
2	$0,60 \leq r < 0,80$	High
3	$0,40 \leq r < 0,60$	Moderate
4	$0,20 \leq r < 0,40$	Low
5	$r < 0,20$	Very low

The reliability index of the final examination's questions in each course can be seen in Table 2.

Table 2. The reliability index of the final examination's questions for each course

Number.	Course	Reliability index
1	Course 1	0,755
2	Course 2	-0,542
3	Course 3	0,566
4	Course 4	0,365
5	Course 5	-2,706
6	Course 6	-1,758
7	Course 7	0.270
Mean		-0.435

In accordance with the course lecturer, deliberate use of numbers is made in course names. To preserve the lecturer's credibility, this is done. As a result of Table 2, the following can be discussed.

In course 1, the estimation test was carried out on an essay test instrument of 7 questions with 13 respondents. The result of the estimated reliability score is 0.755. Based on the reference in the reliability test in Table 1., it can be concluded that the final examination's items for course 1 have a high level of reliability or reliable. In course 2, the estimation test was carried out on an essay test instrument of 3 questions with 42 respondents. The result of the estimated reliability score is 0.542. Based on the reference in the reliability test in Table 1, it can be concluded that the final examination's items for course 2 have a very low reliability level or are unreliable. In course 3, the estimation test was conducted on an essay test instrument of 5 questions with

34 respondents. The result of the estimated reliability score is 0.566. Based on the reference in the reliability test in Table 1, it can be concluded that the final examination's items for course 3 have a moderate level of reliability or reliable enough. In course 4, the estimation test was carried out on an essay test instrument of 3 questions with 81 respondents. The result of the estimated reliability score is 0.365. Based on the reference in the reliability test in Table 1, it can be concluded that the final examination's items for course 4 have a very low-reliability level or are unreliable. In course 5, the estimation test was carried out on an essay test instrument of 5 questions with 41 respondents. The result of the estimated reliability score is -2,706. Based on the reference in the reliability test in Table 1, it can be concluded that the final examination's items for course 5 have a very low-reliability level or are unreliable. In course 6, the estimation test was carried out on an essay test instrument of 5 questions with 44 respondents. The result of the estimated reliability score is -1,758. Based on the reference in the reliability test in Table 1, it can be concluded that the final examination's items for course 6 have a very low-reliability level or are unreliable. In course 7, the estimation test was carried out on an essay test instrument of 5 questions with 45 respondents. The result of the estimated reliability score is 0.270. Based on the reference in the reliability test in Table 1, it can be concluded that the final examination's items for course 7 have a low level of reliability or are less reliable.

Overall, the average reliability score on the final exam for all subjects in the sample shows a score of -0.435. Referring to the classification of scores in Table 1, it can be concluded that the final examination item of all subjects in the sample has a very low-reliability level or is not reliable.

This research calculated the average reliability score across all subjects in the study sample, which yielded a score of -0.435. This numerical finding is particularly significant because it falls below the threshold for what is generally considered an acceptable level of reliability in educational assessments. According to the categorizations presented in Table 1, this score corresponds to a 'very low-reliability level' or is deemed 'not reliable.' This classification is alarming and suggests that the final exams in the sampled subjects do not effectively measure what they intend to measure.

The average dependability index of the final exam questions is rather low, according to the study's findings. If examined, this results from the instructor not following the correct protocol for creating the question instruments. However, given the length of time, expense, and effort required to complete the treatment, this can be seen as justified. Lecturers and policymakers should be able to cooperate in allocating time, resources, and effort to be able to carry out procedures for creating final exam question instruments optimally, given the significance of the ideality of research outcomes.

Such low reliability could be due to a variety of factors, such as poor question design, ambiguous phrasing, or a mismatch between the exam content and the course objectives. This undermines the exams' utility as an assessment tool and calls into question the validity of the educational outcomes derived from these exams. Furthermore, it poses a serious concern for educators, policymakers, and academic institutions, as unreliable exams may lead to inaccurate evaluations of student performance, thereby affecting academic progression, policy formulation, and even accreditation.

Therefore, the findings signify an urgent need for reviewing and likely revising the assessment strategies currently in place.

4 Discussion

In this study, the reliability index of final exam questions is critically analyzed as an assessment of educational quality at the higher education level. Assessment is a systematic process inseparable from education that aims to evaluate the course of the teaching and learning process [28]. The evaluation's target was to measure student learning's progress and the success of educators and institutions as learning service providers [2].

Reviewing the timing of its implementation, the assessment can be divided into two forms: formative and summative. Formative assessment is carried out during the learning process, while summative assessment refers to the assessment carried out after the learning process is complete. Formative assessment can be done by observing their activity, judging from the assignments given, and giving tests or quizzes on the sidelines of the learning process. Summative research can be done by providing exam questions, recapping observations, and notes during learning, or a combination of several assignments for the final project [29].

The University has implemented formative and summative assessments. Both formative and summative assessments have an essential role in becoming a policy consideration as a follow-up to the evaluation results. This study aims to test the quality of the final exam questions in the Department. Based on the estimation test of the final exam reliability index in each course, it can be seen that only one out of seven courses has a high reliability score. At the same time, the rest are divided into moderate to very low categories. Even three of them have very low or unreliable reliability scores.

Reliability itself is an indicator of a measuring instrument that explains the extent to which the measurement results are consistent when measured twice on the same subject group [17]. The same thing is also explained by Allan [30], where reliability is a measurement characteristic related to the accuracy, precision, and consistency of a measuring instrument. In other words, the higher the constancy of the data, the better the quality of the measuring instrument used [17]. When viewed from the level of reliability, there are several classifications related to the level of reliability based on the characters. But in general, they divide the level of reliability into very low/unreliable, low/less reliable, moderate/fairly reliable, high/reliable, and very high/very reliable [17], [31][32], [33].

The constancy of the evaluation instrument is essential, considering the data from the evaluation can be used as a consideration for making decisions [34], [35]. The results of the assessment should be able to provide information that enhances and encourages learning [1]. Therefore, the ability of lecturers as educators to create good evaluation instruments is something that needs to be considered [36]. A good evaluation instrument is characterized by its reliability, ensuring consistent and stable results

over time, and its validity, meaning it accurately measures the specific subject or constructs it is designed to assess.

Practicality is also essential; the tool should be easy to administer, score, and interpret while also being cost-effective. Sensitivity in range and discrimination allows the instrument to measure a wide range of abilities or attitudes and effectively differentiate between closely matched subjects. Fairness is also crucial, ensuring the instrument is culturally sensitive and accessible to all individuals, including those with disabilities. Clarity and precision in question formulation and scoring procedures also minimize ambiguity and subjectivity. Flexibility in adaptability and versatility enables the instrument to be used in various contexts and for different assessments without losing effectiveness. Lastly, transparency in the instrument's procedures and scoring contributes to its credibility and acceptability among administrators and respondents. Overall, a well-designed evaluation instrument enhances the quality of assessments and contributes to more informed decision-making [16], [37].

In this study, the samples analyzed were questions in the form of descriptions of seven courses. The essay questions themselves are items whose answers are entirely freed from the participants' thoughts. The advantages of description questions include measuring learning outcomes that are quite complex, involving skills in integrating ideas and information into problem-solving and requiring students to express their thoughts in writing [38]. On the other hand, the reliability of the questions in the form of descriptions is low [38]. The scores obtained by participants may be inconsistent when the test is retested several times. These results can be possible because of the form of description questions that give students freedom to express their ideas. This condition can also explain the results of this study, which shows the low reliability of the final examination items.

Based on the causes of low reliability, it also can be caused by low methodological quality at the time of instrument preparation. Who conducts the assessment and assesses the performance of student assessments and the creation of questions that can only test a sample of all topics and levels of learning can affect the reliability of the assessment results. Therefore, the importance of teachers' understanding and use of statistical analysis of test materials to improve their teaching strategies and test construction also needs to be emphasized [18].

CDCAT, which stands for "Competence of Designing Competence Assessment Tool," is a methodology created to guide educators, researchers, and professionals through the process of developing successful assessment instruments. The model is arranged into several main components, each containing unique indicators that function as milestones or standards. There are four components, and between two and four indicators are included in each component. The first step is to determine the purposes and goals of the assessment. This component is made up of three different indicators: determining the objectives of the assessment task system, identifying the characteristics of the scenarios, and identifying the purposes for why the situations were created. The second part of the process is called "plan the development," and it consists of determining the kind of data to collect, how much of it to collect, and which assessment methodologies to use. The creation of assessment tools is the third step in the process. This part of the process involves determining the sort of information that will

be utilized, searching for the type of information that will be used, designing the assessment, and figuring out the specific actions that the evaluator will take. The next step is to practice using the evaluation tool and make any necessary adjustments. It includes putting the assessment tool through some tests, analyzing the results of those tests, and making adjustments to the assessment tool. Altogether, CDCAT provides a structured framework that aims to create an assessment tool that is not only reliable and valid but also aligned with its intended purpose [39]–[41].

In order relation to the types of questions used in the final examination of this particular department in the form of descriptions or essays, several things should be considered as for things that can be regarded as when preparing essay questions, among others, having a clear framework so that it does not give rise to multiple interpretations for participants who read test questions and formulate questions carefully [38]. Furthermore, another thing that must be considered is providing sufficient time for participants to complete the questions given.

In supporting the competence of educators related to the preparation of the right items, numerous activities can be considered. Things that must be considered include the provision of training as professional development, observing the impact of the provision of training and professional development on professional behavior for the implementation of teaching and learning, and observing the output and impact on students. Suppose the obstacles faced are in the form of limited capacity in the form of time. In that case, the higher education institution can consider special personnel in preparing the instrument to maximize it.

5 Conclusion

The study demonstrates conclusively that the average reliability index of final exam questions in the examined academic setting is depressingly low. This result is primarily the result of instructors not adhering to set norms and protocols for question creation, which has a negative impact on the quality of assessment tools. Even though this technique may be somewhat justified due to time, expense, and effort constraints, it jeopardizes the integrity and validity of the final examination questions and, by extension, the quality of education. Given that the ideality and rigor of academic evaluations are essential to both instructional efficacy and academic research, the current state of affairs requires quick correction.

A multifaceted approach involving educators and legislators must remedy this predicament. First and foremost, schools should prioritize the professional development of lecturers, with an emphasis on training them in the correct procedures for producing valid and reliable test questions. This may involve workshops, seminars, or even individual mentorship sessions. Second, authorities should devote sufficient financial, time, and human resources to support these educational initiatives. This may entail altering academic calendars or devoting cash to training programs. Implementing a structured approach such as CDCAT could provide a foundation for assuring the quality of future assessment instruments. Through collaborative efforts, it is feasible to

enhance the dependability of final test questions, raising the institution's educational and research standards.

REFERENCES

1. C. R. Reynolds, R. B. Livingston, and V. Willson, Measurement and Assessment in Education. London: Pearson Education, Inc., 2010.
2. D. W. Johnson and R. T. Johnson, Meaningful assessment. Boston: Allyn and Bacon, 2002.
3. H. De Wit, "Internationalisation of higher education in Europe and its assessment, trends and issues." NVAO The Hague, The Netherlands, 2010.
4. T. Kubiszyn and G. D. Borich, Educational testing and measurement. John Wiley & Sons, 2016.
5. R. L. Linn, Measurement and assessment in teaching. Pearson Education India, 2008.
6. A. Irons and S. Elkington, Enhancing learning through formative assessment and feedback. Routledge, 2021.
7. D. Boud, Enhancing learning through self-assessment. Routledge, 2013.
8. G. V Helden, V. Van Der Werf, G. N. Saunders-Smits, and M. M. Specht, "The Use of Digital Peer Assessment in Higher Education—An Umbrella Review of Literature," IEEE Access, vol. 11, pp. 22948–22960, 2023, doi: 10.1109/ACCESS.2023.3252914.
9. M. T. Flórez and P. Sammons, Assessment for Learning: Effects and Impact. ERIC, 2013.
10. F. M. Van der Kleij, J. A. Vermeulen, K. Schildkamp, and T. J. H. M. Eggen, "Integrating data-based decision making, assessment for learning and diagnostic testing in formative assessment," Assess Educ, vol. 22, no. 3, pp. 324–343, 2015.
11. K. Govindan, S. Rajendran, J. Sarkis, and P. Murugesan, "Multi criteria decision making approaches for green supplier evaluation and selection: a literature review," J Clean Prod, vol. 98, pp. 66–83, 2015.
12. X. Zeng et al., "The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review," J Evid Based Med, vol. 8, no. 1, pp. 2–10, 2015.
13. A. Irons and S. Elkington, Enhancing learning through formative assessment and feedback. Routledge, 2021.
14. M. S. Abou El-Seoud, I. A. T. F. Taj-Eddin, N. Seddiek, M. M. El-Khouly, and A. Nosseir, "E-learning and students' motivation: A research study on the effect of e-learning on higher education," International Journal of Emerging Technologies in Learning (Online), vol. 9, no. 4, p. 20, 2014.
15. X. Zeng et al., "The methodological quality assessment tools for preclinical and clinical studies, systematic review and meta-analysis, and clinical practice guideline: a systematic review," J Evid Based Med, vol. 8, no. 1, pp. 2–10, 2015.
16. C. A. Tomlinson and T. R. Moon, Assessment and student success in a differentiated classroom. ascd, 2013.
17. S. Azwar, Reliabilitas dan validitas. Yogyakarta: Pustaka Pelajar, 2012.
18. G. A. Brown, J. Bull, and M. Pendlebury, Assessing student learning in higher education. Routledge, 2013.
19. J.-A. Baird, D. Andrich, T. N. Hopfenbeck, and G. Stobart, "Assessment and learning: Fields apart?," Assess Educ, vol. 24, no. 3, pp. 317–350, 2017.

20. M. D. Hanus and J. Fox, "Assessing the effects of gamification in the classroom: A longitudinal study on intrinsic motivation, social comparison, satisfaction, effort, and academic performance," Comput Educ, vol. 80, pp. 152–161, 2015.
21. T. A. Angelo and K. P. Cross, "Classroom assessment techniques: A handbook for college teachers," (No Title), 2018.
22. D. Boud, Enhancing learning through self-assessment. Routledge, 2013.
23. J. McDavid and L. Hawthorn, "Program evaluation & performance measurement, chapter Applying qualitative evaluation methods." Sage Publications Inc, 2006.
24. L. Suskie, Assessing student learning: A common sense guide. John Wiley & Sons, 2018.
25. B. Noonan and C. R. Duncan, "Peer and self-assessment in high schools," Practical assessment, research, and evaluation, vol. 10, no. 1, p. 17, 2019.
26. M. Hussain, W. Zhu, W. Zhang, and S. M. R. Abidi, "Student engagement predictions in an e-learning system and their impact on student course assessment scores," Comput Intell Neurosci, vol. 2018, 2018.
27. P. W. Airasian, Classroom assessment. ERIC, 1997.
28. J. W. Gikandi, D. Morrow, and N. E. Davis, "Online formative assessment in higher education: A review of the literature," Comput Educ, vol. 57, no. 4, pp. 2333–2351, 2011, doi: https://doi.org/10.1016/j.compedu.2011.06.004.
29. D. D. Dixson and F. C. Worrell, "Formative and summative assessment in the classroom," Theory Pract, vol. 55, no. 2, pp. 153–159, 2016.
30. R. N. Allan, Reliability evaluation of power systems. Springer Science & Business Media, 2013.
31. J. P. Guilford, "Creativity: A quarter century of progress," in Perspectives in creativity, Routledge, 2017, pp. 37–59.
32. S. Arikunto, Dasar-Dasar Evaluasi Pendidikan Edisi 3. Bumi Aksara, 2021.
33. S. Arikunto, "Prosedur penelitian suatu pendekatan praktik," 2013.
34. D. Daryanto, Pendekatan pembelajaran saintifik Kurikulum 2013. Yogyakarta: Gava Media, 2014.
35. R. A. Sani, Pembelajaran saintifik untuk kurikulum 2013. Jakarta: Bumi Aksara, 2013.
36. H. Pagarra, P. Bundu, M. Irfan, Hartoto, and S. Raihan, "Peningkatan Kompetensi Guru Dalam Mengevaluasi Pembelajaran Daring Menggunakan Aplikasi Berbasis Tes Dan Penugasan Online," Publikasi Pendidikan, vol. 10, no. 3, pp. 260–265, 2020.
37. T. Kubiszyn and G. D. Borich, Educational testing and measurement. John Wiley & Sons, 2016.
38. Y. Astriani and I. Marzuki, "Pjj: Digital Transformasi Daring Pada Evaluasi Pendidikan Di Era Pandemi Covid -19," Rausyan Fikr : Jurnal Pemikiran dan Pencerahan, vol. 17, no. 1, pp. 76–83, 2021, doi: 10.31000/rf.v17i1.4205.
39. Y. Xu, B. Xia, Y. Wan, F. Zhang, J. Xu, and H. Ning, "CDCAT: A multi-language cross-document entity and event coreference annotation tool," Tsinghua Sci Technol, vol. 27, no. 3, pp. 589–598, 2021.
40. N. T. D. Linh, "Competency model of designing competency assessment tool: A pilot study with Vietnamese science pre-service teacher," in Journal of Physics: Conference Series, IOP Publishing, 2021, p. 012067.
41. N. T. D. Linh, "A competence model to assess and develop designing competence assessment tool," International Journal of Learning, Teaching and Educational Research, vol. 20, no. 2, pp. 81–103, 2021.