pISSN 2442-532X | eISSN 2528-4541
https://journal.unnes.ac.id/journals/jone

An In-Depth Analysis of the Students' Perceptions of How the Final Exam Administered in Package C Program

Evy Ratna Kartika Waty[1*], Yanti Karmila Nengsih[2], Ciptro Handrianto[3], Shahid Rasool[4]

[1,2]Universitas Sriwijaya, Indonesia
[3]Universitas Negeri Padang, Indonesia
[4]Florida Gulf Coast University, USA

Email: evyrkwaty@gmail.com

Submitted: 2024-06-19. Revised: 2024-07-10. Accepted: 2024-07-28.

Abstract. Assessments are essential in influencing students' academic results, but there are still worries about the equity and thoroughness of tests created by tutors for different subjects. This research examines how students view these assessments in Citizenship Education, Islamic Religious Education, and History Education. The main goal is to investigate how Package C Program students view the distribution and quality of final exam questions and evaluate their preparedness and performance. This qualitative study used thematic analysis on interviews with 20 randomly chosen Package C Program students in Palembang. The information was examined to find common themes on how feel about their readiness and the tutors' involvement in creating exams. The results suggest that students believe the exam questions do not accurately represent their understanding or what they have learned. Numerous individuals are unhappy with mistakes in the test questions, which hinder their success. The study also identifies differences between students' anticipations and the real exam material. Prejudiced evaluation methods and varying test creation by educators can greatly harm student learning and the development of future educational policies. Improvements are necessary in exam design to guarantee fairness and accuracy. This research offers fresh perspectives on how the quality of exams affects student performance and highlights the necessity of inclusive test development with input from various parties to prevent biases.

Keywords**:** students' perceptions; test development; test administration; test evaluation; final examination

How to Cite: Kartika Waty, E. R., Nengsih, Y. K., Handrianto, C., & Rasool, S. . (2024). An In-Depth Analysis of the Students' Perceptions of How the Final Exam Administered in Package C Program. *Journal of Nonformal Education*, *10*(2), 435-445. https://doi.org/10.15294/jone.v10i2.9960

INTRODUCTION

Assessment is commonly perceived as collecting data on individuals' acquired knowledge to facilitate decision-making processes (Mohan, 2023; Yan & Boud, 2021). The conventional understanding of assessments has experienced a significant transformation in the last ten years following the emergence of critical pedagogy. The democratic assessment approach incorporates test takers' perspectives, experiences, and expectations and validates them within the educational system. Additionally, it should be noted that learners are active participants in educational systems and possess their rights, which are duly acknowledged and upheld (Meisuri et al., 2023).

Kubiszyn and Borich (2024) hold that assessments serve as a mechanism to enforce tacit notions regarding achievement, expertise, and aptitude. The author observes that how assessments are conducted can significantly impact individuals' career trajectories, either facilitating or impeding their success. Specifically, while assessments can serve as a valuable tool for career advancement, they can also have detrimental effects, such as imposing undue stress on individuals and creating unnecessary obstacles to their achievement. According to Van Groen and Eggen (2019), an examination of tests within the framework of social, educational, and political contexts situates the field of testing within the realm of critical testing. The approach to language testing is commonly known as critical language testing.

Rimfeld et al. (2019) assert that tutors employ assessments to gauge their students' performance relative to established national standards. The tutor devised this task and disseminated it among the students. As per the findings of Pratiwi et al. (2022), the collected data can aid educational institutions in assessing their students' progress and ascertaining their preparedness for progression to the subsequent academic level. Conversely, scholars in education and specialized institutions (Abdillah et al., 2023) concur that the ultimate assessment administered to students in educational institutions is beset with significant issues and necessitates modification. Multiple domains exist in which tutors' existing testing and evaluation protocols exhibit potential for enhancement.

According to Rose and Johnson (2020), the creation of the exam must be designed more effectively, make better judgments, and take productive actions; for the exam's excellence in assessments, it is crucial to incorporate solid quantitative traits such as validity and reliability. The efficacy of the test construction requires improvement. According to Liu et al. (2020), there is a possibility of enhancing the classroom examinations that have been utilized so far because certain tutors lack proficiency in test development. Furthermore, scant evidence suggests that the standardized tests administered in educational institutions are precise (Kilag et al., 2023).

Furthermore, it should be noted that the scope of test scores and item scores may differ across various aptitude tests, as indicated by Adom (2020). Diverse assessments are employed to evaluate the aptitudes of individual students. According to Trigueros et al. (2020), the results of a student's examinations can offer valuable information regarding their overall academic achievement. Reading oneself for and undertaking final assessments is a crucial aspect of the educational experience in academic institutions. In this regard, it is not uncommon for tutors to relocate to different Package C Programs, where they may encounter unfamiliar colleagues developing their pedagogical practices and evaluating their students' academic advancement.

According to Tanjung et al. (2021), it is customary to establish test teams at the onset of every academic year, comprising tutors from diverse disciplinary domains. As mentioned above, the committee is responsible for overseeing the administration of Package C Program examinations, and its evaluations are utilized as a framework for determining the outcome. The subsequent stage involves the creation of examination items by the academic institution, which are subsequently utilized to generate replicated test queries. According to Hidayat et al. (2023), the majority of seasoned tutors possess the ability to create more demanding assessments. Regular implementation of formative assessments by tutors enables them to administer comprehensive summative evaluations that accurately reflect their students' comprehension of the subject matter under consideration. The test score obtained by a student is a significant determinant of the degree to which it reflects their cognitive aptitude. The role of the tutor in the capacity of a test developer is of utmost importance. Provided that tutors can produce precise formative assessments throughout the academic year, it stands to reason that they should possess the ability to construct valuable summative evaluations at the culmination of the school year.

Adom et al. (2020) contend that the examination has been criticized due to its utilization of objective testing methods, which may not accurately represent students' knowledge levels. Furthermore, it is imperative to adopt a modern approach to assessment that surpasses the sole measurement of intelligence quotient (IQ). As per Marzuki's (2024) findings, a majority of tutors exhibit a preference for employing objective assessments that comprise multiple-choice questions as a means of conducting their concluding evaluations. Assistance may be necessary for students to effectively demonstrate their genuine abilities on objective assessments, as they are solely tasked with selecting the most appropriate response. The escalating demand for students to undergo testing at supplementary educational institutions may lead to a decline in overall proficiency.

The creation of a superior examination necessitates meticulous consideration and the utilization of diverse resources, as suggested by Bhat et al. (2023), Wang et al. (2023), Wilson (2023), and Rusyid et al. (2024). Brennan (2023) posits that a dependable test should encompass various attributes, such as but not limited to relevance, specificity, equilibrium, objectivity, efficiency, equal opportunity, difficulty level, discrimination index, and dependability. The efficacy of a test, as determined by the quality of the problems presented, is contingent upon the level of expertise and familiarity possessed by the tutor. The efficacy of exams in accurately assessing students' actual levels of accomplishment is contingent upon their appropriate construction. Marzuki (2024) asserts that many tutors attach great significance to delivering superior education to their students in the context of both learning and examination.

The quality of the final evaluation is significantly influenced by internal and external factors relevant to the tutor. The sources cited in the text include works by Bowman et al. (2024), LaBelle and Wozniak (2023), Luo et al. (2024), Nguyen et al. (2020), McHaney (2023), and Roche et al. (2023). Kubiszyn and Borich (2024) posit that the quality of a tutor's evaluation is contingent upon various factors, such as their level of knowledge, emotional disposition, and socioeconomic status. There is a widely held belief among individuals that it is of utmost significance to examine the educational qualifications of tutors, their accumulated years of professional experience, the frequency of their test administration, and the extent of their training within the preceding three-year period. It is imperative to offer satisfactory solutions to these concerns in order to ensure the continued success of the ultimate academic examination. Internal factors can also influence the tutor's capacity to offer constructive feedback. In order to enhance the assessment of the caliber of assessments crafted for tutors, it is advantageous to delve into their viewpoints (Weng, 2023).

In addition, the impact of assessments holds considerable weight on students' academic performance. Students must receive adequate attention in order to optimize their performance in assessments. If such a scenario arises, objectively evaluating any reliable test's results would not be easy. Consequently, the difficulties students encounter present a captivating subject matter for discussion. This, together with the rising complexity of today's educational setting and the high-stake nature of final examinations, calls for urgent reconsideration of how structure and administration are executed. Poor or biased assessments violate the academic achievements of students and add to disparities in the system. The involvement and contribution of tutors in developing examinations call for critical investigation; poorly conducted testing will result in negative consequences for the students in their academic and future development. Understanding these perceptions and being informed by educational reforms that aim for equity and inclusion is important.

This study explores the perception of Package C Program students in Singapore regarding the format and neutrality of final exam questions in Civic Education, Islamic Religious Education, and History Education. This helpful research investigates students' perceptions of the adequacy of exam questions, assessment challenges, and tutor involvement in exam construction. These findings underscore the need to improve exam construction and further train tutors in assessment literacy to make exams more reliable and valid. These observations can inform and lead to changes in education towards building a more equitable system where assessments will be aligned with students' skills to help them achieve academic success. Taking these considerations into account, the researchers sought to ascertain a solution to describe the perspective of Package C Program students regarding the incorporation of final exams as a prerequisite for graduation.

METHODS

Research Design

The study aimed to examine students' perspectives regarding the distribution of final exam questions in Citizenship Education, Islamic Religious Education, and History Education, as prepared by their respective tutors. This research employed a qualitative design to gain in-depth insights into students' perceptions of exam preparation, fairness, and tutor involvement. This qualitative research entailed conducting interviews with a cohort of 20 Package C Program students in Palembang, selected at random (Leavy, 2022).

Data Collection Techniques

To procure pertinent data for this inquiry, the investigator devised an interview protocol to be administered to the study's participants (Nasution, 2023). The interview protocol was pre-tested to ensure clarity, and participants were encouraged to share their experiences freely within the structure of the semi-structured interview format. When the interviewer intends to extract only the necessary information, they may employ a structured interview form to administer it. This form comprises pre-planned questions designed to elicit specific responses. In addition, the interviews were audio-recorded for accuracy and complemented with notes taken during the sessions. This investigation aimed to validate the accuracy of the students' answers when they were directly interrogated using the dialogue mentioned above. Patton (2014) employed notes gathered from student interviews to comprehensively examine their perspectives during the preparation and execution of the final examination. This study aimed to gain insight into students' perspectives regarding creating test items for the final examination in the subjects of Citizenship Education, Islamic Religious Education, and Historical Education, as developed by their tutors. The findings of this research may inform future regulatory efforts in this area.

Data Validity

Data validity was ensured through several mechanisms. First, triangulation was employed by cross-referencing the interview data with classroom observations and informal discussions with tutors. This ensured a more comprehensive understanding of the context. Additionally, member checking was conducted, allowing participants to review their interview transcripts to verify their responses had been accurately captured. This process helped reduce researcher bias and ensure the authenticity of the data.

Data Analysis

The thematic analysis (Braun et al., 2022) of the collected data pertains to students' perspectives regarding their preparation and execution of test items, as well as their opinions on the correlation between the endeavors of tutors and the Ministry of Education in addressing the challenges faced by students arranged following the questions in the interview protocol.

Coding the transcribed interviews was conducted using a comparative analysis methodology recommended by Deterding and Waters (2021). The study employed an iterative examination methodology across the dataset to identify commonalities and recurring themes in the interview transcripts. Examining these resemblances and regularities progressively developed a coding system for the classifications. Through this coding process, the key themes identified included exam fairness, student preparation, and tutor involvement in test design. The study involved the development of units of analysis and coding schemes. The codes were converted into categorical labels or themes that were observed as recurring patterns in the interviews, per Patton's (2014) methodology. As per Cloutier and Ravasi's (2021) findings, this methodology aids researchers in creating a feasible classification or coding system, which serves as the initial stage of analysis and subsequently organizing it into a table. The analysis process was iterative, meaning codes were revisited and refined as new patterns emerged, ensuring that the final thematic structure represented the full range of student perspectives. Data analysis was conducted step-by-step, and only after achieving data coherence were conclusions derived from the analyzed data.

RESULTS AND DISCUSSION

The open-ended interview allowed freedom of expression, enabling participants to articulate their viewpoints more effectively. The interviews will be addressed sequentially in the following sections.

Q1: Does the test accurately reflect and indicate your genuine level of knowledge?

The inquiry into the students' reactions to this query revealed that most acknowledged that the examination failed to offer an accurate and authentic representation and manifestation of their genuine understanding. The rationale behind their refusals was based on the assertion that it is unfeasible to assess the knowledge of every individual through such examinations, as the scope of materials is confined to a finite set of inquiries. The statement is consistent with Park and Cho's (2023) findings. Which individual asserted that faculty members could potentially increase their ratings on teaching evaluations by restricting the range of student grades? In contrast, prior studies conducted by Beleche et al. (2012) and Hernon et al. (2023) have demonstrated a strong positive correlation, which is also statistically significant, between the measure of student learning and course evaluations.

The study participants believed that examinations of this nature could encompass only a restricted subset of the essential sources for the examination. The participants believed that a two-hour multiple-choice examination was inadequate in measuring their actual knowledge due to the potential for debilitating and detrimental stress among all participants, which may impede their ability to think clearly. Santisteban and Egues (2022) conducted a literature review on system evaluation tests to elucidate the known and unknown sources of bias associated with these evaluative tools. This paper contends that biased evaluation tests can have detrimental effects in the current era of advancing educational methodologies and a nationwide push toward inclusivity. This can result in the permeation of such biases into the education, policymaking, practice, and research aspects, ultimately hindering its progress, according to Clayson (2020). The present research reveals that students who perform better than their anticipated grades benefit their tutors psychologically by eliciting a more favorable evaluation. In contrast, the converse is true for students who receive grades lower than their expected performance. The empirical findings indicate that a prior understanding regarding the impact of student grades on the quality of teaching in a given course may need to be reconsidered to some extent.

Q2: Is it possible to enhance any facets of the concluding assessment?

The respondents primarily focused on the sources of the test and expressed their belief that a thorough and consistent survey and consensus on the primary exam materials could enhance various aspects of the exam. The participants expressed dissatisfaction with the lack of specific sources, particularly for the general section of the test. Alternative viewpoints were put forth by some participants, who proposed that the examination would be more effective if it assessed their comprehension of practical application and theoretical knowledge. According to Adom et al. (2020), there is an increasing necessity to consistently reassess educational assessment due to its dynamic nature over time. Despite the abundance of scholarly literature on measurement, testing, and evaluation in education, these concepts continue to pose challenges for tutors tasked with test creation. Tutors must comprehensively understand the consequences of their assessment methodologies and their impact on practical applications in real-world scenarios.

Conversely, most participants reported experiencing irritation and distraction due to the inaccuracies in the questions, which hindered their ability to maintain concentration.

> *"We, the students, feel upset because the tutor's mistake in making the question items incorrect greatly interferes with our concentration in doing the exam. Although some friends are happy with this because it adds a mark without working on asking question...."* (Mila).

Scholars have noted that students actively pursue high-quality questions in their academic pursuits (Hirai et al., 2022; Manfaat et al., 2021; Susongko et al., 2022). They aim to avoid inquiries that have the potential to result in misinterpretations. A significant proportion of the student participants reported frequent annoyance due to the continued inaccuracies in the questions. The individuals expressed apprehension regarding their ability to attain the objective due to inadequate posing or presenting the inquiries. This constituted the primary source of their concerns. Maldonado and De Witte (2022) have arrived at a conclusion based on an analysis of Package C Program characteristics, standardized testing in Grade 10, and the fixed influences of Package C Programs. The observed effects indicate that Package C Program closures lead to a concomitant loss of learning progress and learning. It has been observed that there is an increase of 7% in the discrepancy of mathematics instruction and 8% in the discrepancy of Dutch language instruction. Furthermore, Purwanto's (2020) study has indicated that online education confers advantages to students by allowing them to access courses from their homes, unrestricted by geographical boundaries, and listen to courses at their convenience.

Additionally, online education allows students to engage with familiar educational content without constraints of time and space constraints, such as environmental conditions, idle time, network disruptions, the incongruity between the voice of tutors and educational materials, and the unavailability of Wi-Fi leading to missed classes, have negatively impacted students' focus and attendance. Various proposals have been put forth to enhance the network's performance, including inducing greater instability, fostering engagement by amplifying one-way communication, and conducting in-person training sessions for practical application. Research has demonstrated a positive correlation between learning losses and observable Package C Program characteristics, indicating that Package C Program with economically disadvantaged students experience more significant learning losses. As expressed by Maya, *"We, students, the tutor's mistake in writing incorrect item questions greatly disturbs the student's concentration in answering the test...".*

Q3: Has the Ministry of Education adequately furnished you with comprehensive details concerning the conditions of various educational institutions, including the qualifications of their tutors and the quality of their educational facilities?

The participants asserted that they were not provided with any data on the final examination of the educational institutions during the decision-making process. The individuals indicated that while certain Package C Program websites may have assisted, obtaining access to crucial essential information was unattainable. Moreover, the Ministry of Education has been reported to conceal specific inadequacies, opting to remain silent rather than acknowledge them. It has been suggested that they would instead promote the excellence of educational institutions rather than admit to any shortcomings.

The findings indicate that a significant proportion of the student population favors diverse stakeholders' participation in the final examination's administration. Daud asserted that

> *"...there should be some correlation between the information taught in the curriculum and the test material, even though it might not be possible to involve all students in a complete assessment of this scope. I continued to expand on this idea...."*

As an illustration, within the context of our course, we examined various literary works on pedagogy and assessment. However, in completing Package C Program, the students were required to engage with alternative scholarly materials. Package C Program tutors can commence instructing the courses that will prove significant for the ultimate examination, at the very least. In the Package C Program curriculum, it is common for tutors to possess knowledge regarding the required reading material for the final examination while concurrently providing supplementary resources. It is suggested that a correlation between the students, tutors, and examiners should exist. Daud suggested that "*...the implementation of a certain action could potentially result in the conservation of time and energy....*"

Vina supported Daud's suggestion that *"...the Ministry of Education should conduct an annual workshop to gather feedback from students and tutors regarding their expectations and suggestions for future examinations..."*

Fatimah and Rinawati (2022) have supported the effectiveness of a particular training program. The authors assert that this training can enhance tutors' comprehension and proficiency in constructing questions that require high-level thinking skills. Consequently, such questions can be utilized to accurately assess students' cognitive development in line with desired learning outcomes. The ultimate goal of this training is to equip students with critical and creative thinking abilities that can be applied to identify and resolve real-world problems.

Q4: Did you know about the test makers and the test creation process?

Regarding inquiries about the respondents' knowledge of test creators, the test's creation timeline, and primary exam sources, the data analysis results indicate that the respondents' study plans in preparation for the Package C Program final exam were inadequately uncertain. The primary reason for the lack of a secure foundation in the interviewees' study plans was the difficulty in determining the technical and general topics that should be included in the exams, as reported by 80% of the interviewees. One of the students, Nina, asserted that *"...prior to taking the exam, I lacked access to the sources, and instead resorted to obtaining materials from the internet or from peers who had previously taken the exam..."* The students expressed uncertainty regarding the test makers' identity but speculated that certain tutors within the Package C Program might have formulated the questions. All participants lacked knowledge regarding the time the test was administered.

Marjanovic-Shane et al. (2023) posited the notion of democratic education, highlighting that We must foster an open and inclusive learning environment for students within our education system. Students must be allowed to voice their needs and concerns regarding the subject matter they are learning. Therefore, it is inappropriate for the tutor to unilaterally create questions without soliciting student input. The test administration should be preceded by collaborative efforts involving all relevant parties, including tutors, learners, and other stakeholders. At a minimum, the students know the content and purpose of the test they are scheduled to take. The assertion is corroborated by Yan et al.'s (2023) research, which indicates that involving students in evaluating their self-efficacy for assessment tasks or creating particular queries positively impacts the phase. This is because it guarantees that tutors and students receive feedback on learning outcomes and processes from themselves and their peers. Yan et al. (2023) posits that self-assessment is crucial in regulating learning. Engaging in this practice can enhance students' autonomy in the learning process. It enhances their ability as an evaluator of the assessment and as a creator.

Q5: Is there any potential for improvement in any aspect of the final examination?

Several interviewees presented their perspectives regarding specific areas of the Package C Program curriculum that require enhancement. One of the facets highlighted by the students pertained to the imperative of evaluating all competencies during the examination. The assessment is intended to evaluate the practical application of students' knowledge in real-world scenarios. However, it prioritizes rote memorization of answers over posing practical inquiries. According to Daud, *"...the assessment did not effectively evaluate the students' capacity to apply their knowledge in practical situations, as it relied solely on their memorization and reading abilities...."*

Consistent with Hadi's (2020) research, it was determined that incorporating practice tests into the learning process yields superior academic performance outcomes compared to traditional methods. The heightened level of engagement among students and the livelier classroom learning environment indicates this phenomenon. In the context of civics education, it is imperative that students not only possess knowledge of citizenship theory but also demonstrate proficiency in the practical application of said theory by providing apt examples. Similarly, in Islamic religious education, the primary objective is to cultivate ethical values in the context of daily living. The memorization of hadiths about cleanliness by students would prove futile if their conduct in daily life continues to be characterized by littering. In history instruction, assessments should evaluate students' comprehension of significant historical occurrences and gauge their ability to apply lessons from the past to enhance prospects.

Q6: Do you believe the final examination will impact your ability to complete it?

According to Purwanto's (2020) research, the level of concentration exhibited by students was influenced by the quality of questions posed to them. Individuals expected to focus on understanding the inquiries and furnishing precise answers are sidetracked by inaccurate questions. This phenomenon challenges one's ability to maintain focus and may impede one's capacity to address other complex matters of significance. Consequently, inquiries intended to evaluate students' proficiency objectively may inadvertently reflect values unsuitably linked to their aptitude levels.

> *"... although some individuals express contentment with this phenomenon as it provides an added value bonus without requiring any additional effort towards the items in question. As a student in Package C Program, I think the tutor should check the test before administering it."*

Conversely, there is a phenomenon wherein students prefer challenging questions. This is because if the tutor acknowledges the difficulty of the question, it is often elevated to the status of a bonus question, thereby enabling all students to be credited with a correct response. The principal obstacles presented by remote evaluation encompassed academic dishonesty, technological resources, congruity with educational goals, and compliance of students with submission guidelines. The most efficacious approach in mitigating occurrences of academic dishonesty was found to be the formulation of particular questions for every individual student. It has been found that delivering presentations through online platforms is a viable option for mitigating violations of academic honesty. The implementation of diverse evaluation methods, such as the submission of a written document alongside a virtual demonstration, can effectively mitigate occurrences of academic misconduct. This is because the examiner would be allowed to authenticate whether or not the submitted work was genuinely the student's work.

Many students tend to favor this approach despite harboring doubts regarding the legitimacy and reliability of the question. Initially, the inquiry is subjected to an assessment without imperfections. As an illustration, if 50 inquiries exist, and a pair of them are answered incorrectly, the number of available high-quality questions would be reduced to 48. While many tutors may still assign a score of 50, it is recommended that the accurate overall score be evaluated as 48 by the tutor. Hence, the accuracy of students' self-perceived capabilities is questionable. In the hypothetical scenario where a more significant number of erroneous inquiries exist, the gravity of the situation would be exacerbated in the given instance. It is possible to observe a deviation in the psychological impact on students' perception of competence when they answer more than five questions incorrectly.

Another consequence is that students tend to expect a higher frequency of incorrect questions. The pedagogical approach involves directing students to prioritize identifying errors over focusing solely on generating responses to posed inquiries. This phenomenon substantially influences the overall quality level of the evaluation. The research findings suggest that minimizing errors in the questioning process should be prioritized to the fullest extent possible. Consequently, tutors must assess the final examination prior to dissemination.

CONCLUSION

This paper argues that biased system evaluation test data can sabotage Package C Program so to permeate into and handicap the education, policy making, practice, and research of the profession itself. This qualitative research project aimed to investigate the students' views on administering the final examination questions of citizenship education, Islamic religious education, and history education, which the tutors made. The results of the data analysis revealed that the study plans they had mapped out in light of the Package C Program final exam were defectively dubious. Furthermore, the writers suggest that biased evaluation tests can have detrimental effects in the current era of advancing educational methodologies and a nationwide push toward inclusivity. tutors must assess the final examination prior to dissemination. The findings of this research conclude that the number of mistakes introduced into the questioning process should be reduced to the greatest extent feasible. As a result, the writers also suggest minimizing errors in the questioning process should be prioritized to the fullest extent possible. The implementation of diverse evaluation methods, such as the submission of a written document alongside a virtual demonstration, can effectively mitigate occurrences of academic dishonesty. Because of the study's limitations, such as the small sample size, it is essential to interpret these research findings cautiously to ensure reliable results. In the future, similar studies should be conducted with a larger sample size so that findings can be confirmed.

REFERENCES

Abdillah, F., Azmi, K., Hafizah, C. V., Anisha, D., Bintang, N. D., & Mulyani, S. (2023). Strategi Pelaksanaan Evaluasi Program Pendidikan Terhadap Kualitas Belajar Siswa di Sekolah. *Jurnal Bintang Pendidikan Indonesia*, *1*(2), 13-23. https://doi.org/10.55606/jubpi.v1i2.1190

Adom, D., Mensah, J. A., & Dake, D. A. (2020). Test, Measurement, and Evaluation: Understanding and Use of the Concepts in Education. *International Journal of Evaluation and Research in Education*, *9*(1), 109-119. https://doi.org/ 10.11591/ijere.v9i1.20457

Bhat, N., Deo, S. K., & Gurung, S. (2023). Assessing the quality of multiple-choice questions in allied health science summative exams: A retrospective analysis. *Journal of Gandaki Medical College-Nepal*, *16*(2), 111-117. https://doi.org/10.3126/jgmcn.v16i2.55893

Bowman, T. G., Thrasher, A. B., Kasamatsu, T. M., & Lyons, S. M. (2024). Multistakeholder perceptions of young professionals' integration during role transition. *Journal of Athletic Training*, *59*(1), 99–110. https://doi.org/10.4085/1062-6050-0505.22

Braun, V., Clarke, V., Hayfield, N., Davey, L., Jenkinson, E. (2022). Doing Reflexive Thematic Analysis. In: Bager-Charleson, S., McBeath, A. (eds) Supporting Research in Counselling and Psychotherapy. Palgrave Macmillan, Cham. https://doi.org/10.1007/978-3-031-13942-0_2

Brennan, R. L. (Ed.). (2023). *Educational measurement*. Rowman & Littlefield. https://eduq.info/xmlui/handle/11515/34503

Clayson, D. E. (2020). *A comprehensive critique of student evaluation of teaching: Critical perspectives on validity, reliability, and impartiality*. New York: Routledge. https://doi.org/10.4324/9781003091462

Cloutier, C., & Ravasi, D. (2021). Using tables to enhance trustworthiness in qualitative research. *Strategic Organization*, *19*(1), 113–133. https://doi.org/10.1177/1476127020979329

Deterding, N. M., & Waters, M. C. (2021). Flexible coding of in-depth interviews: A twenty-first-century approach. *Sociological methods & research*, *50*(2), 708-739. https://doi.org/10.1177/0049124118799377

Fatimah, S., & Rinawati, A. (2022). Pelatihan Penyusunan Instrumen Evaluasi Berbasis Higher Order Thinking Skills (HOTs) Untuk Guru MI di Kebumen. *BERNAS: Jurnal Pengabdian Kepada Masyarakat*, *3*(2), 152-161. https://doi.org/10.31949/jb.v3i2.2190

Hirai, A., Oka, H., Kato, T., & Maeda, H. (2022). Development and validation of an English test measuring EFL learners' critical thinking skills. *Language Testing in Asia*, *12*(1), 45. https://doi.org/10.1186/s40468-022-00193-2

Hadi, D. A. (2020). Implementasi Model Pembelajaran Discovery Learning Berorientasi HOTS Pada Mata Pelajaran Matematika di SMK Negeri 7 Mataram. *SUPERMAT: Jurnal Pendidikan Matematika*, *4*(1), 22-32. https://doi.org/10.33627/sm.v4i1.356

Hernon, O., McSharry, E., MacLaren, I., & Carr, P. J. (2023). The use of educational technology in teaching and assessing clinical psychomotor skills in nursing and midwifery education: A state-of-the-art literature review. *Journal of Professional Nursing*, *45*, 35–50. https://doi.org/10.1016/j.profnurs.2023.01.005

Hidayat, M. S., Fitra, D., Susetyo, A. M., Amarulloh, R. R., & Ardiansyah, R. (2023). *Pengantar Evaluasi Pendidikan*. Penerbit Widina. https://books.google.hu/books?id=EgvYEAAAQBAJ&lpg=PA6&ots=yWjpygo_5h&lr&hl=id&pg=PA6#v=onepage&q&f=false

Kilag, O. K. T., Evangelista, T. P., Sasan, J. M., Librea, A. M., Zamora, R. M. C., Ymas, S. B., & Alestre, N. A. P. (2023). Promising Practices for a Better Tomorrow: A Qualitative Study of Successful Practices in Senior High School Education. *Journal of Elementary and Secondary School*, *1*(1), 16-28. https://doi.org/10.31098/jess.v1i1.1379

Kubiszyn, T., & Borich, G. D. (2024). *Educational testing and measurement*. John Wiley & Sons. https://www.wiley.com/en-ae/Educational+Testing+and+Measurement%2C+11th+Edition-p-9781119228097

LaBelle, S., & Wozniak, T. (2023). Academic beliefs and prescription stimulant misuse among college students: Investigating academic locus of control, grade orientation, and academic entitlement. *Journal of American College Health*, *71*(8), 2370-2379. https://doi.org/10.1080/07448481.2021.1968408

Leavy, P. (2022). *Research design: Quantitative, qualitative, mixed methods, arts-based, and community-based participatory research approaches*. Guilford Publications. https://www.guilford.com/books/Research-Design/Patricia-Leavy/9781462548972

Liu, Z. Y., Lomovtseva, N., & Korobeynikova, E. (2020). Online learning platforms: Reconstructing modern higher education. *International Journal of Emerging Technologies in Learning (iJET)*, *15*(13), 4-21. Retrieved September 13, 2024 from https://www.learntechlib.org/p/217605/.

Luo, W., & Lim, S. Q. W. (2024). Perceived formative assessment and student motivational beliefs and self-regulation strategies: a multilevel analysis. *Educational Psychology*, 1–19. https://doi.org/10.1080/01443410.2024.2354686

Maldonado, J. E., & De Witte, K. (2022). The effect of school closures on standardised student test outcomes. *British Educational Research Journal*, *48*(1), 49-94. https://doi.org/10.1002/berj.3754

Manfaat, B., Nurazizah, A., & Misri, M. A. (2021). Analysis of mathematics test items quality for high school. *Jurnal Penelitian dan Evaluasi Pendidikan*, *25*(1), 108-117. http://dx.doi.org/10.21831/pep.v25i1.39174

Marjanovic-Shane, A., Kullenberg, T., & Gradovski, M. (2023). Scandinavian Experiments in Democratic Education. *Dialogic Pedagogy*, *11*(2). 94-132. https://doi.org/10.5195/dpj.2023.477

Marzuki, I. (2024). Implementasi prinsip-prinsip evaluasi pembelajaran pada mata pelajaran Pendidikan Agama Islam. *Tadarus Tarbawy: Jurnal Kajian Islam dan Pendidikan*, *6*(1). 91–97. http://dx.doi.org/10.31000/jkip.v6i1.11821

McHaney, R. (2023). *The new digital shoreline: How Web 2.0 and millennials are revolutionising higher education*. Taylor & Francis. https://dl.acm.org/doi/abs/10.5555/2018755

Meisuri, M., Nuswantoro, P., Mardikawati, B., & Judijanto, L. (2023). Technology Revolution in Learning: Building the Future of Education. *Journal of Social Science Utilizing Technology*, *1*(4), 214-226. https://doi.org/10.70177/jssut.v1i4.660

Mohan, R. (2023). *Measurement, evaluation and assessment in education*. PHI Learning Pvt. Ltd.. https://www.alibris.com/Measurement-Evaluation-and-Assessment-in-Education-Radha-Mohan/book/35257945

Nartin, N., Faturrahman, F., Deni, A., Santoso, Y. H., Paharuddin, P., Suacana, I., ... & Eliyah, E. (2024). *Metode penelitian kualitatif*. Cendikia Mulia Mandiri. http://eprints2.ipdn.ac.id/1373/1/Metode%20Penelitian%20Kualitatif%20%28cover%20dan%20daftar%20isi%29.pdf

Nguyen, D., Harris, A., & Ng, D. (2020). A review of the empirical research on teacher leadership (2003–2017) Evidence, patterns and implications. *Journal of educational administration*, *58*(1), 60–80. https://doi.org/10.1108/JEA-02-2018-0023

Park, B., & Cho, J. (2023). How does grade inflation affect student evaluation of teaching?. *Assessment & Evaluation in Higher Education*, *48*(5), 723-735. https://doi.org/10.1080/02602938.2022.2126429

Patton, M. Q. (2014). *Qualitative research & evaluation methods: Integrating theory and practice*. Sage publications. https://cir.nii.ac.jp/crid/1130282272245558272

Pratiwi, S. N., Prasetia, I., & Gajah, N. (2022). Literacy Culture in Elementary Schools: The Impact of the Literacy Movement Program and Library Facilities. *Jurnal Kependidikan: Jurnal Hasil Penelitian Dan Kajian Kepustakaan Di Bidang Pendidikan, Pengajaran Dan Pembelajaran*, *8*(3), 786-794. https://doi.org/10.33394/jk.v8i3.5559

Purwanto, A. (2020). University students online learning system during Covid-19 pandemic: Advantages, constraints and solutions. *Sys Rev Pharm*, *11*(7), 570–576. Available at SSRN: https://ssrn.com/abstract=3986850

Rimfeld, K., Malanchini, M., Hannigan, L. J., Dale, P. S., Allen, R., Hart, S. A., & Plomin, R. (2019). Teacher assessments during compulsory education are as reliable, stable, and heritable as standardised test scores. *Journal of Child Psychology and Psychiatry*, *60*(12), 1278-1288. https://doi.org/10.1111/jcpp.13070

Roche, A., Gervasoni, A., & Kalogeropoulos, P. (2023). Factors that promote interest and engagement in learning mathematics for low-achieving primary students across three learning settings. *Mathematics Education Research Journal*, *35*(3), 525–556. https://doi.org/10.1007/s13394-021-00402-w

Rose, J., & Johnson, C. W. (2020). Contextualising reliability and validity in qualitative research: Toward more rigorous and trustworthy qualitative social science in leisure research. *Journal of leisure research*, *51*(4), 432–451. https://doi.org/10.1080/00222216.2020.1722042

Rusyid, H. K., Suryadi, D., Herman, T., Adnan, M., Lutfi, A., & Mukhibin, A. (2024). Rasch modelling approach to measure the quality of algebraic thinking test item for junior high school students. *Beta: Jurnal Tadris Matematika*, *17*(1), 44-58. https://doi.org/10.20414/betajtm.v17i1.652

Santisteban, L., & Egues, A. L. (2022). How do student evaluations of teaching contribute to the hindrance of faculty diversity? *Teaching and Learning in Nursing*. https://doi.org/10.1016/j.teln.2022.04.007

Susongko, P., Sunu, H. W. D., & Arfiani, Y. (2022, January). Student Moral Quality Measurement Framework Based on Ancient Javanese Philosophy. In *2nd International Conference on Social Science, Humanities, Education and Society Development (ICONS 2021)* (pp. 195-201). Atlantis Press. https://doi.org/10.2991/assehr.k.220101.029

Tanjung, E. F., Harfiani, R., & Sampedro Hartanto, H. (2021). Formation Of Soul Leadership Model In Indonesian Middle Schools. *Kuram ve Uygulamada Eğitim Bilimleri/Educational Sciences: Theory And Practice*, *21*(1), 84-97. https://psycnet.apa.org/record/2021-77803-007

Trigueros, R., Padilla, A., Aguilar-Parra, J. M., Lirola, M. J., García-Luengo, A. V., Rocamora-Pérez, P., & López-Liria, R. (2020). The influence of teachers on motivation and academic stress and their effect on the learning strategies of university students. International Journal of Environmental Research and Public Health, 17(23), 9089. https://doi.org/10.3390/ijerph17239089

Van Groen, M. M., & Eggen, T. J. (2020). Educational test approaches: The suitability of computer-based test types for assessment and evaluation in formative and summative contexts. *Journal of Applied Testing Technology*, *21*(1), 12-24. *Retrieved from* http://www.jattjournal.net/index.php/atp/article/view/146484

Wang, Y., Derakhshan, A., Pan, Z., & Ghiasvand, F. (2023). Chinese EFL teachers' writing assessment feedback literacy: A scale development and validation study. *Assessing Writing*, *56*, 100726. https://doi.org/10.1016/j.asw.2023.100726

Weng, F. (2023). EFL teachers' writing assessment literacy: Surveying teachers' knowledge, beliefs, and practises in China. *Porta Linguarum Revista Interuniversitaria de Didáctica de las Lenguas Extranjeras*, (40), 57-74. https://doi.org/10.30827/portalin.vi40.23812

Wilson, M. (2023). *Constructing measures: An item response modeling approach*. New York: Routledge. https://doi.org/10.4324/9781003286929

Yan, Z., & Boud, D. (2021). Conceptualising assessment-as-learning. In *Assessment as learning* (pp. 11-24). Routledge. https://www.taylorfrancis.com/chapters/edit/10.4324/9781003052081-2/conceptualising-assessment-learning-zi-yan-david-boud

Yan, Z., Panadero, E., Wang, X., & Zhan, Y. (2023). A systematic review on students' perceptions of self-assessment: usefulness and factors influencing implementation. *Educational Psychology Review*, *35*(3), 81. https://doi.org/10.1007/s10648-023-09799-1