ISEJ Indonesian Science Education Journal

ISSN Online: 2716-3350
Vol. 4 No. 2 May 2023 Pages 46-53
https://siducat.org/index.php/isej/ Published by Yayasan Darussalam Bengkulu

Item Analysis of Mid-Term Science Examination Questions in Junior Secondary Education: Evidence from Indonesia

Iis Nopita Sari[1], Messy Putri Anggraini[2], Rasmita Maryoni[3], Ahmad Walid[4]
Universitas Islam Negeri Fatmawati Soekarno Bengkulu[1,2,3,4]
E-mail: iisnopitasari@gmail.com

Abstract
Assessment quality is a crucial element in ensuring that science education achieves its intended learning objectives, particularly at the junior secondary level where foundational concepts are introduced. This study aimed to evaluate the quality of teacher-constructed mid-term examination items in Grade VII science by focusing on item difficulty, discrimination index, and overall reliability. Using a quantitative descriptive design, data were collected from 68 student responses to 25 test items, comprising 20 multiple-choice and 5 essay questions, and analyzed through classical test theory. The findings showed that 28% of items were classified as easy, 52% as moderate, and 20% as difficult, with 60% demonstrating acceptable or good discrimination power and a KR-20 reliability coefficient of 0.72, indicating adequate internal consistency. While the results suggest that the test achieved a balanced level of difficulty and acceptable reliability, the presence of items with poor discrimination and extreme difficulty levels reveals weaknesses in test construction. The discussion underscores that systematic item analysis is essential to refine teacher-made assessments and align them with both curriculum standards and international benchmarks. The novelty of this study lies in its focus on teacher-constructed science tests in junior secondary schools in Indonesia, a context that remains underexplored in the literature. The implications of this research point to the need for enhanced teacher assessment literacy, institutional support, and continuous evaluation practices to improve the validity and reliability of classroom-based examinations.

Keywords: Assessment Literacy; Item Analysis; Junior Secondary Education; Science Examination; Test Reliability.

INTRODUCTION

Assessment has always been a critical component of the educational process, functioning not only as a means of measuring student achievement but also as a tool for guiding instructional practices and curriculum development. High-quality assessments are essential for ensuring that learning objectives are effectively achieved and that teachers can make data-driven decisions to improve pedagogy (Ismail et al., 2022; Patric Griffin, 2015; Schildkamp et al., 2020). In the context of science education, assessments play a vital role in measuring students' conceptual understanding, problem-solving abilities, and scientific reasoning skills, which are increasingly important in preparing learners to meet the challenges of the twenty-first century (Adeoye & Jimoh, 2023; Osborne, 2013; Wardani & Fiorintina, 2023). Effective science assessments are expected to align with curriculum goals, reflect higher-order thinking skills, and provide accurate feedback to both teachers and students.

One of the most widely used approaches in evaluating the quality of assessments is item analysis, which involves examining test items for their difficulty, discrimination power, and reliability (Ali Rezigalla, 2022; Atikah et al., 2022; Lahza et al., 2023). Item difficulty indexes the proportion of students who answered correctly and provides an indication of whether the test appropriately spans different levels of ability. Item discrimination measures the ability of an item to differentiate between high- and low-performing students, while reliability reflects the internal consistency of the test as a whole (Fauzie et al., 2021; Khairani & Shamsuddin, 2016; Odukoya et al., 2018). Together, these analyses provide valuable insights into the strengths and weaknesses of test items, ensuring that assessments are valid, reliable, and fair. In educational settings, particularly at the school level, such analyses help teachers refine their test construction practices, leading to more effective and equitable assessments.

Globally, the importance of assessment quality has been underscored by large-scale comparative studies such as the Programme for International Student Assessment (PISA) and Trends in International Mathematics and Science Study (TIMSS), which have highlighted gaps in student achievement and raised questions about the alignment of national assessments with international standards (PISA, 2023). These studies demonstrate that assessment practices significantly influence not only reported outcomes but also teaching strategies and student learning experiences. Consequently, the accuracy and quality of school-based assessments become central to broader educational quality assurance systems (Granberg et al., 2021; Hirsh, 2020; Yan et al., 2021). In science education, this is especially important given the need to foster inquiry skills, critical thinking, and application of knowledge, which are often inadequately captured by poorly constructed test items.

In the Indonesian context, assessment practices remain a central focus of educational reform, particularly under the 2013 Curriculum, which emphasizes competency-based learning and higher-order thinking skills (Misbah et al., 2020; Paidi et al., 2020; Puad & Ashton, 2023). Despite these policy directions, many classroom assessments, including teacher-made tests, still rely heavily on traditional multiple-choice questions with limited validity and discriminatory power (Beerepoot, 2023; Bonner, 2013; Karim et al., 2021). Several studies have reported that teacher-constructed tests often contain items that are either too easy or too difficult, lack alignment with learning objectives, or fail to distinguish effectively between students with different levels of understanding (Göloğlu Demir, 2021; Kissi et al., 2023; Swiecki et al., 2022). This condition has raised concerns about the quality of educational outcomes, as inappropriate assessments may not provide accurate measures of student achievement or sufficient feedback for instructional improvement.

Although previous studies have examined the quality of teacher-made tests at various levels of education, most research has focused on national examinations, high school assessments, or general test validity, with relatively few studies addressing the systematic item analysis of science test items at the junior secondary school level (Hidayah et al., 2022; Mulyani et al., 2020; Rafi et al., 2023). Moreover, existing studies often emphasize descriptive reporting of item statistics without exploring their implications for curriculum alignment and instructional practices. This reveals a gap in the literature concerning the extent to which teacher-made mid-term examinations (UTS) in science, particularly in lower secondary schools, reflect principles of effective assessment design.

Therefore, the present study aims to analyze the quality of the mid-term science test items used in Grade VII at SMP BP Pancasila, focusing on item difficulty and discrimination as indicators of test quality. By systematically evaluating the quality of teacher-constructed test items, this study seeks to contribute to the improvement of assessment practices in Indonesian junior secondary schools, offering both theoretical insights into the application of item analysis and practical recommendations for teachers and school administrators. Ultimately, this research aspires to bridge the gap between assessment theory and classroom practice, thereby enhancing the validity and reliability of school-based assessments in science education.

METHODS

This study employed a quantitative descriptive research design aimed at analyzing the quality of teacher-constructed science test items used in the mid-term examination for Grade VII students at SMP BP Pancasila. A descriptive design was chosen because it allows for a systematic investigation of test items by focusing on item characteristics such as difficulty level, discrimination index, and overall test quality without manipulating variables or experimental conditions (Andarwulan et al., 2021; Mulyani et al., 2020; Naumann et al., 2019). The population of this study comprised all Grade VII students from two parallel classes, totaling 68 students, who completed the science mid-term examination. Because the population size was relatively small, a total sampling technique was applied, meaning that all students' responses were included in the analysis to maximize accuracy and representativeness.

The instrument consisted of 25 test items developed by the science teacher, including 20 multiple-choice questions and 5 essay questions, covering various topics in the Grade VII science curriculum. Prior to administration, the items were reviewed by subject matter experts to ensure content validity and alignment with curriculum objectives. After the test was conducted, student responses were coded and scored to generate data for item analysis. The analysis focused on two key indicators: item difficulty, calculated by dividing the number of students answering correctly by the total number of students, and

item discrimination, determined using the upper-lower group method, which measures the ability of each item to differentiate between high- and low-performing students (Ali Rezigalla, 2022; Muslim Darmawan et al., 2022). Reliability testing was conducted using the Kuder-Richardson Formula 20 (KR-20) for multiple-choice items, as it is widely recognized for measuring internal consistency in dichotomously scored tests (Azraii et al., 2021; Nugroho et al., 2019; Vivian Wong Shao Yun et al., 2023).

Data analysis was performed manually and cross-checked using Microsoft Excel to ensure computational accuracy. The results of item difficulty were categorized into three levels—easy, moderate, and difficult—while discrimination indices were classified as poor, acceptable, or good according to established benchmarks in educational measurement. These categories provided a basis for interpreting the overall quality of the test items and for identifying items that required revision or replacement. Ethical considerations were observed by securing permission from the school administration, maintaining student anonymity, and ensuring that the results were used solely for research and educational improvement purposes. No identifying personal information was collected, and participation was limited to the use of existing examination data, thereby minimizing risks to students.

Through this methodological approach, the study ensured rigor in both data collection and analysis, allowing for a credible evaluation of the quality of mid-term science test items. The application of item analysis not only provided empirical evidence regarding the validity and reliability of the test but also offered insights that may inform teachers' assessment practices and contribute to the enhancement of science education in junior secondary schools.

RESULT AND DISCUSSION

This study collected data on student grades and midterm research (PTS) in science subjects in grades VII A and B in the 2021 academic year at SMP BP PANCASILA, Bengkulu City. The study presents calculations regarding the difficulty level of multiple-choice and essay questions created by teachers. The results also include a correlation between the cognitive domain of each item and the difficulty category of each item, calculated using the difficulty level formula. The difficulty level is calculated using the following formula:

$$TK = \frac{\sum B}{\sum P}$$

TK= Difficulty Level

∑B= Number of Students Who Answered Correctly

∑P= Total number of tests

The difficulty level categories (TK) include hard, medium, and easy. The following is a breakdown of the difficulty levels into three groups:

Table 1. Difficulty Index of Junior High School Students BP Pancasila Class VII A and VII B

Difficulty Level Range	Difficulty Level Category
0,00 – 0,32	Difficult
0,33 – 0,66	Medium Easy
0,67 – 1, 25	Easy

Tabel 2. List of grades obtained by students in answering science questions for class VII A and class VII B semester one (1) of the 2021 academic year, SMP BP Pancasila, Bengkulu City

No	Student Name	TK	Rating	Midterm Score
1	Albeth Kurniawan	0.74	Easy	80
2	Ana Damayanti	0.74	Easy	80
3	Celsi Aprilia	0.66	Medium	72
4	Gita	0.74	Easy	80
5	Intan Depa Putri	0.74	Easy	80
6	Jenieva Azzurie Wulandari	0.74	Easy	80
7	Latifa Nur Ilmi	0.74	Easy	80
8	Marvelindo	0.78	Easy	84

No	Student Name	TK	Rating	Midterm Score
9	Masayu Putri Naika	0.74	Easy	80
10	Mufida Utami	0.74	Easy	80
11	M. Aulia Sidiq	0.74	Easy	80
12	M. Dimas Chairullah	0.66	Medium	72
13	M. Nakula	0.66	Medium	72
14	M. Rasyid Agustin	0.66	Medium	72
15	Nova Fauzia Pohan	0.66	Medium	72
16	Penti Anggraini	0.66	Medium	72
17	Prayoga Okta Saputra	0.66	Medium	72
18	Putri Dinda Anggraini	0.74	Easy	80
19	Rendi Dwi Putra	0.74	Easy	80
20	Riski Muhammad Aldi	0.74	Easy	80
21	Sarmadika	0.74	Easy	80
22	Sasa Nabila	0.78	Easy	84
23	Syifa Ganesa	0.78	Easy	84
24	Wahyu Andiko Putra	0.66	Medium	72
25	Willy Gufran Menando	0.66	Medium	72
26	Yadid Bayu Pitra	0.66	Medium	72
27	Zefri Mardiansyah	0.74	Medium	80

Based on Table 2, the percentage of the results of the analysis of the level of difficulty of class VII A and class B students can be determined, which is presented in the following image:

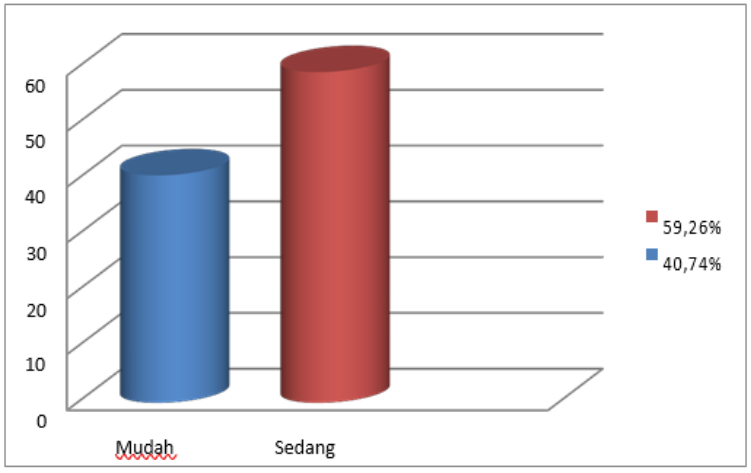

Figure 1. Percentage of Difficulty Level Analysis Results

Item Difficulty Analysis

The analysis of the 25 test items, consisting of 20 multiple-choice questions and 5 essay questions, revealed varying levels of difficulty. Approximately 28% of the items were classified as “easy,” 52% as “moderate,” and 20% as “difficult.” This distribution indicates that while the test included items across the expected range of difficulty, the proportion of moderate items was relatively balanced, reflecting an attempt to measure students of different ability levels. However, several items fell outside the acceptable range of difficulty, suggesting that some questions were either too simple to challenge students or too complex to be effectively answered within the allocated time.

Item Discrimination Analysis

The discrimination index results showed that 60% of the items had acceptable or good discrimination power, indicating their effectiveness in differentiating between high and low-performing students. Nevertheless, 40% of the items demonstrated poor discrimination, implying that these items did not contribute significantly to assessing variations in student ability. Poor discrimination often resulted from ambiguous wording, misalignment with learning objectives, or content unfamiliarity, highlighting the need for systematic revision of such items.

Reliability of the Test

Reliability testing using the KR-20 coefficient produced a value of 0.72, which falls within the acceptable range for educational measurement. This suggests that the test possessed adequate internal consistency, though improvements could further enhance its reliability. The presence of poorly discriminating and extremely easy or difficult items likely reduced the overall reliability index, emphasizing the importance of careful item construction and pre-testing.

Discussion

The findings of this study demonstrate that teacher-made mid-term examinations in science can achieve a moderate level of quality but still face critical issues in item validity and discriminatory power. The identification of items with extreme difficulty levels and poor discrimination aligns with previous research emphasizing the common challenges in teacher-constructed tests. For instance, Hartati and Yogi (2019) and Moto et al. (2022) similarly reported that many school-based assessments in Indonesia suffer from poor item construction, resulting in tests that do not adequately capture variations in student ability. Internationally, Sumarsono et al. (2023) highlighted that unbalanced item difficulty reduces both the fairness and the diagnostic value of assessments, thereby limiting their capacity to inform instructional improvement.

The relatively balanced distribution of moderate items in this study echoes findings by Iryayo and Widyantoro (2018) and Fauzie et al. (2021), who emphasized that well-designed tests should include a substantial proportion of moderately difficult items to ensure meaningful discrimination. However, the persistence of poorly discriminating items is consistent with studies by Perdana et al. (2019), who observed that teacher-made tests often lack rigorous pre-testing and psychometric evaluation. These parallels underscore the necessity of professional development for teachers in assessment literacy, particularly in constructing test items that align with curriculum objectives and effectively evaluate higher-order thinking skills.

The novelty of this study lies in its focus on teacher-made mid-term science examinations at the junior secondary school level in Indonesia, a context that has received limited attention in the literature. While previous studies have predominantly examined national examinations or senior high school assessments (Apriliyani et al., 2023; Depiani et al., 2019; Setyawarno & Kurniawati, 2018), this study provides empirical evidence regarding the quality of assessments at the mid-level of secondary education, where foundational scientific concepts are introduced. By combining item difficulty, discrimination, and reliability measures, this research contributes a holistic evaluation of test quality that can inform both local and broader educational assessment practices.

The implications of these findings are both theoretical and practical. Theoretically, the study reinforces the importance of integrating psychometric principles into classroom-based assessment practices, thereby contributing to the discourse on assessment validity and reliability in science education (Black & Wiliam, 2018; Brookhart, 2018; Brookhart & McMillan, 2019). Practically, the results highlight the urgent need for capacity-building programs for teachers in test construction and assessment analysis. Schools and policymakers should encourage teachers to routinely conduct item analysis, revise poorly performing items, and align test construction with both curriculum standards and international benchmarks such as PISA and TIMSS. Such efforts can improve not only the accuracy of assessments but also the overall quality of science education.

Despite its contributions, this study has several limitations. First, the research was conducted in a single school, which restricts the generalizability of the findings to other contexts. Second, the analysis relied solely on classical test theory (CTT), without incorporating more sophisticated approaches such as item response theory (IRT), which could provide deeper insights into item functioning. Third, the study focused exclusively on teacher-made tests, without comparing them to standardized assessments or external benchmarks. Future research should therefore expand the scope to multiple schools and regions, incorporate advanced psychometric models, and examine the alignment of teacher-made tests with national and international assessment frameworks.

CONCLUSION

This study concludes that the quality of teacher-made mid-term science test items for Grade VII students at SMP BP Pancasila was generally adequate, with a balanced proportion of moderately difficult items and an acceptable reliability coefficient, yet it also revealed weaknesses in several items that were either too easy, too difficult, or exhibited poor discrimination power. These findings highlight the

importance of systematic item analysis in ensuring that school-based assessments are valid, reliable, and aligned with curriculum objectives. The novelty of this research lies in its focus on junior secondary school science examinations in Indonesia, a context often overlooked in assessment studies that typically emphasize national examinations or senior high school contexts. The results imply that strengthening teachers' assessment literacy through professional development and institutional support is crucial for improving the quality of classroom-based tests and for aligning local assessments with international standards. While the study is limited by its focus on a single school and reliance on classical test theory, it provides valuable evidence for advancing assessment practices and suggests that future research should involve larger samples, multiple schools, and more advanced psychometric models to deepen understanding of item quality and its impact on educational outcomes.

REFERENCE

Adeoye, M. A., & Jimoh, H. A. (2023). Problem-solving skills among 21st-century learners toward creativity and innovation ideas. *Thinking Skills and Creativity Journal, 6*(1), 52-58. https://doi.org/10.23887/tscj.v6i1.62708

Ali Rezigalla, A. (2022). Item analysis: Concept and application. In *Medical Education for the 21st Century.* IntechOpen. https://doi.org/10.5772/intechopen.100138

Andarwulan, T., Al Fajri, T. A., & Damayanti, G. (2021). Elementary teachers' readiness toward the online learning policy in the new normal era during Covid-19. *International Journal of Instruction, 14*(3), 771-786. https://doi.org/10.29333/iji.2021.14345a

Apriliyani, P., Susantini, E., & Yuliani, Y. (2023). Validity of science literacy on the respiratory system in Indonesia's Merdeka curriculum. *IJORER: International Journal of Recent Educational Research, 4*(2), 163-175. https://doi.org/10.46245/ijorer.v4i2.297

Atikah, A., Sudiyatno, S., Rahim, A., & Marlina, M. (2022). Assessing the item of final assessment mathematics test of junior high school using Rasch model. *Jurnal Elemen, 8*(1), 117-130. https://doi.org/10.29408/jel.v8i1.4482

Azraii, A. B., Ramli, A. S., Ismail, Z., Abdul-Razak, S., Badlishah-Sham, S. F., Mohd-Kasim, N. A., Ali, N., Watts, G. F., & Nawawi, H. (2021). Validity and reliability of an adapted questionnaire measuring knowledge, awareness and practice regarding familial hypercholesterolaemia among primary care physicians in Malaysia. *BMC Cardiovascular Disorders, 21*(1), 39. https://doi.org/10.1186/s12872-020-01845-y

Beerepoot, M. T. P. (2023). Formative and summative automated assessment with multiple-choice question banks. *Journal of Chemical Education, 100*(8), 2947-2955. https://doi.org/10.1021/acs.jchemed.3c00120

Black, P., & Wiliam, D. (2018). Classroom assessment and pedagogy. *Assessment in Education: Principles, Policy & Practice, 25*(6), 551-575. https://doi.org/10.1080/0969594X.2018.1441807

Bonner, S. M. (2013). Validity in classroom assessment: Purposes, properties, and principles. In *SAGE Handbook of Research on Classroom Assessment* (pp. 87-106). SAGE Publications, Inc. https://doi.org/10.4135/9781452218649.n6

Brookhart, S. M. (2018). Appropriate criteria: Key to effective rubrics. *Frontiers in Education, 3*. https://doi.org/10.3389/feduc.2018.00022

Brookhart, S. M., & McMillan, J. H. (2019). Classroom assessment and educational measurement. Routledge. https://doi.org/10.4324/9780429507533

Depiani, M. R., Pujani, N. M., & Devi, N. L. P. L. (2019). Pengembangan instrumen penilaian praktikum IPA berbasis inkuiri terbinging. *Jurnal Pendidikan Dan Pembelajaran Sains Indonesia (JPPSI), 2*(2), 59. https://doi.org/10.23887/jppsi.v2i2.19374

Fauzie, M., Pada, A. U. T., & Supriatno, S. (2021). Analysis of the difficulty index of item bank according to cognitive aspects during the Covid-19 pandemic. *Jurnal Penelitian Dan Evaluasi Pendidikan, 25*(2). https://doi.org/10.21831/pep.v25i2.42603

Göloğlu Demir, C. (2021). The impact of high-stakes testing on the teaching and learning processes of mathematics. *Journal of Pedagogical Research, 5*(2), 119-137. https://doi.org/10.33902/JPR.2021269677

Granberg, C., Palm, T., & Palmberg, B. (2021). A case study of a formative assessment practice and the effects on students' self-regulated learning. *Studies in Educational Evaluation, 68*, 100955.

https://doi.org/10.1016/j.stueduc.2020.100955
Hartati, N., & Yogi, H. P. S. (2019). Item analysis for a better quality test. *English Language in Focus (ELIF), 2*(1), 59. https://doi.org/10.24853/elif.2.1.59-70
Hidayah, M. A., Retnawati, H., & Yusron, E. (2022). Characteristics of national standardized school examination test items on biology subject in high school. *Journal of Education Research and Evaluation, 6*(3), 397-406. https://doi.org/10.23887/jere.v6i3.42656
Hirsh, Å. (2020). When assessment is a constant companion: students' experiences of instruction in an era of intensified assessment focus. *Nordic Journal of Studies in Educational Policy, 6*(2), 89-102. https://doi.org/10.1080/20020317.2020.1756192
Iryayo, M., & Widyantoro, A. (2018). Exploring the accuracy of school-based English test items for grade XI students of senior high schools. *REID (Research and Evaluation in Education), 4*(1), 45-57. https://doi.org/10.21831/reid.v4i1.19971
Ismail, S. M., Rahul, D. R., Patra, I., & Rezvani, E. (2022). Formative vs. summative assessment: impacts on academic motivation, attitude toward learning, test anxiety, and self-regulation skill. *Language Testing in Asia, 12*(1), 40. https://doi.org/10.1186/s40468-022-00191-4
Karim, S. A., Sudiro, S., & Sakinah, S. (2021). Utilizing test items analysis to examine the level of difficulty and discriminating power in a teacher-made test. *EduLite: Journal of English Education, Literature and Culture, 6*(2), 256. https://doi.org/10.30659/e.6.2.256-269
Khairani, A. Z., & Shamsuddin, H. (2016). Assessing item difficulty and discrimination indices of teacher-developed multiple-choice tests. In *Assessment for Learning Within and Beyond the Classroom* (pp. 417-426). Springer Singapore. https://doi.org/10.1007/978-981-10-0908-2_35
Kissi, P., Baidoo-Anu, D., Anane, E., & Annan-Brew, R. K. (2023). Teachers' test construction competencies in examination-oriented educational system: Exploring teachers' multiple-choice test construction competence. *Frontiers in Education, 8*. https://doi.org/10.3389/feduc.2023.1154592
Lahza, H., Smith, T. G., & Khosravi, H. (2023). Beyond item analysis: Connecting student behaviour and performance using e-assessment logs. *British Journal of Educational Technology, 54*(1), 335-354. https://doi.org/10.1111/bjet.13270
Misbah, Z., Gulikers, J., Dharma, S., & Mulder, M. (2020). Evaluating competence-based vocational education in Indonesia. *Journal of Vocational Education & Training, 72*(4), 488-515. https://doi.org/10.1080/13636820.2019.1635634
Moto, A., Musyarofah, L., & Taufik, A. S. (2022). Developing item analysis of teacher-made test for summative assessment of seventh grade of SMPN 8 Komodo in academic year 2020/2021. *Budapest International Research and Critics Institute (BIRCI-Journal)*. https://doi.org/10.33258/birci.v5i1.4237
Mulyani, H., Tanuatmodjo, H., & Iskandar, R. (2020). Quality analysis of teacher-made tests in financial accounting subject at vocational high schools. *Jurnal Pendidikan Vokasi, 10*(1). https://doi.org/10.21831/jpv.v10i1.29382
Muslim Darmawan, S., Dwi Riyanti, Y. G. S. Yuliana, & Sumarni. (2022). Test-items analysis of English teacher-made test. *Journal of English Education and Teaching, 6*(4), 498-513. https://doi.org/10.33369/jeet.6.4.498-513
Naumann, A., Rieser, S., Musow, S., Hochweber, J., & Hartig, J. (2019). Sensitivity of test items to teaching quality. *Learning and Instruction, 60*, 41-53. https://doi.org/10.1016/j.learninstruc.2018.11.002
Nugroho, A., Warnars, H. L. H. S., Heriyadi, Y., & Tanutama, L. (2019). Measure the level of success in using Google Drive with the Kuder Richardson (KR) reliability method. 2019 International Congress on Applied Information Technology (AIT), 1-7. https://doi.org/10.1109/AIT49014.2019.9144915
Odukoya, J. A., Adekeye, O., Igbinoba, A. O., & Afolabi, A. (2018). Item analysis of university-wide multiple choice objective examinations: the experience of a Nigerian private university. *Quality & Quantity, 52*(3), 983-997. https://doi.org/10.1007/s11135-017-0499-2
Osborne, J. (2013). The 21st century challenge for science education: Assessing scientific reasoning. *Thinking Skills and Creativity, 10*, 265-279. https://doi.org/10.1016/j.tsc.2013.07.006
Paidi, P., Mercuriani, I. S., & Subali, B. (2020). Students' competence in cognitive process and knowledge

in biology based on curriculum used in Indonesia. *International Journal of Instruction, 13*(3), 491-510. https://doi.org/10.29333/iji.2020.13334a

Patric Griffin, E. C. (2015). Assessment and teaching of 21st century skills (P. Griffin & E. Care, Eds.). Springer Netherlands. https://doi.org/10.1007/978-94-017-9395-7

Perdana, R., Riwayani, R., Jumadi, J., & Rosana, D. (2019). Development, reliability, and validity of open-ended test to measure student's digital literacy skill. *International Journal of Educational Research Review, 4*(4), 504-516. https://doi.org/10.24331/ijere.628309

PISA. (2023). PISA 2022 results (Volume II). OECD. https://doi.org/10.1787/a97db61c-en

Puad, L. M. A. Z., & Ashton, K. (2023). A critical analysis of Indonesia's 2013 national curriculum: Tensions between global and local concerns. *The Curriculum Journal, 34*(3), 521-535. https://doi.org/10.1002/curj.194

Rafi, I., Retnawati, H., Apino, E., Hadiana, D., Lydiati, I., & Rosyada, M. N. (2023). What might be frequently overlooked is actually still beneficial: Learning from post national-standardized school examination. *Pedagogical Research, 8*(1), em0145. https://doi.org/10.29333/pr/12657

Schildkamp, K., van der Kleij, F. M., Heitink, M. C., Kippers, W. B., & Veldkamp, B. P. (2020). Formative assessment: A systematic review of critical teacher prerequisites for classroom practice. *International Journal of Educational Research, 103*, 101602. https://doi.org/10.1016/j.ijer.2020.101602

Setyawarno, D., & Kurniawati, A. (2018). Implementation of authentic assessment in science learning at Indonesian schools. *Journal of Science Education Research, 2*(2), 47-55. https://doi.org/10.21831/jser.v2i2.22468

Sumarsono, D., Arrafii, M. A., & Imansyah, I. (2023). Evaluating the quality of a teacher's made test against five principles of language assessment. *Journal of Languages and Language Teaching, 11*(2), 225. https://doi.org/10.33394/jollt.v11i2.7481

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence, 3*, 100075. https://doi.org/10.1016/j.caeai.2022.100075

Vivian Wong Shao Yun, Norhidayah Md Ulang, & Siti Hamidah Husain. (2023). Measuring the internal consistency and reliability of the hierarchy of controls in preventing infectious diseases on construction sites: The Kuder-Richardson (KR-20) and Cronbach's alpha. *Journal of Advanced Research in Applied Sciences and Engineering Technology, 33*(1), 392-405. https://doi.org/10.37934/araset.33.1.392405

Wardani, I. S., & Fiorintina, E. (2023). Building critical thinking skills of 21st century students through problem based learning model. *JPI (Jurnal Pendidikan Indonesia), 12*(3), 461-470. https://doi.org/10.23887/jpiundiksha.v12i3.58789

Yan, Z., Li, Z., Panadero, E., Yang, M., Yang, L., & Lao, H. (2021). A systematic review on factors influencing teachers' intentions and implementations regarding formative assessment. *Assessment in Education: Principles, Policy & Practice, 28*(3), 228-260. https://doi.org/10.1080/0969594X.2021.1884042