Issue VIII, 22 November 2025
e-ISSN 2707-9481 Institute of Metallurgy and Ore Beneficiation JSC, Satbayev University, Almaty, Kazakhstan
ISBN 978-601-323-547-9 https://doi.org/10.31643/2025.01

Zafrullah Zafrullah
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
E-mail: zafrullah.2022@student.uny.ac.id
ORCID ID: https://orcid.org/0009-0008-3752-8841

Siti Nurjanah
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
E-mail: siti960pasca.2023@student.uny.ac.id
ORCID ID: https://orcid.org/0009-0006-0727-0830

Ima Aprilia Fitri
Sekolah Menengah Pertama Negeri 1 Banguntapan,
Yogyakarta, Indonesia
E-mail: ima.aprilia48@gmail.com

Era Mutiara
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
E-mail: eramutiara.2023@student.uny.ac.id
ORCID ID: https://orcid.org/0009-0004-0955-4795

Resky Nuralisa Gunawan
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
E-mail: reskynuralisa.2022@student.uny.ac.id
ORCID ID: https://orcid.org/0009-0009-1805-9928

Ghany Desti Laksita
Universitas Negeri Yogyakarta, Yogyakarta, Indonesia
E-mail: ghanydesti.2023@student.uny.ac.id
ORCID ID: https://orcid.org/0009-0007-8475-0966

Analysis of Numeracy Literacy Question Items in Junior High School Mathematics: Analysis using Classical Test Theory

Abstract: Item analysis is an important part of evaluating student performance because it can identify the extent to which the questions can measure students' abilities accurately and fairly. This research aims to analyse Numeracy Literacy questions for 31 students from Class VIII at one of the Junior High Schools in Bantul, Special Region of Yogyakarta, using interview methods and analysis techniques for levels of difficulty and differences in questions based on gender and overall, with the help of the AnBuSo 8.0 application. The results showed that although the majority of questions were in the "Easy" category and were effective in measuring students' overall abilities, there were differences in the level of difficulty and differential ability of questions between genders, where questions were more effective in differentiating the abilities of male students compared to female students. These findings indicate the need to improve and develop questions to be more fair and equitable in measuring the abilities of students from both gender groups. By taking these differences into account, better question design can improve the accuracy and consistency of evaluations, providing a more accurate picture of students' overall abilities.

Keywords: numeracy literacy, question items, classical theory test, education.

Introduction

Education is an important foundation for the progress of a country (Reimers, 2024; Zafrullah et al., 2024). Developed countries always place education as a top priority in development. The existence of quality education determines the future of the nation, because, through education, the young generation is prepared to face global challenges and contribute to national development (Adeniyi et al., 2024; Ramadhani et al., 2024). Therefore, education must be a main concern in state policy, starting from improving the curriculum to providing adequate facilities and infrastructure (Valencia, 2024; Zafrullah & Zetriuslita, 2021). All of this must be pursued to create a conducive learning environment, which will ultimately lead to improving the quality of human resources. So, quality education is not just a dream, but a reality that can be felt by every child of the nation (Izzulhaq et al., 2024; Sanyal, 2024). All of these efforts started from one institution that has a central role in the education process, namely the school.

School is the main foundation in the process of forming the knowledge and character of the nation's children (Sakhiyya & Rahmawati, 2024). Schools act as places where moral, ethical and scientific values are taught and applied in everyday life (Qazi et al., 2024). The character of each individual is greatly influenced by the school environment that guides them during their growth and development (Alajmi, 2024). Schools make students ready to face life's challenges with adequate knowledge and skills (Ramdas et al., 2024). Thus, schools must be able to create a conducive and inspiring learning environment, where every student feels valued and motivated to continue to develop in the classroom.

The classroom is a space where educational interaction between teachers and students takes place intensively (Barker, 2024). Classes are challenged to always create an atmosphere that supports students' intellectual, emotional and social growth. Classes are an important element in education that can facilitate the process of critical thinking, creativity and collaboration (Pamuji & Mulyadi, 2024). In class, every student is given the opportunity to explore their potential and develop skills that are relevant to future needs (Bleazby, 2020). Thus, classes must be designed and managed well to be able to accommodate various effective and comprehensive teaching methods to achieve the ultimate goal of the learning process.

Learning is a crucial process in supporting student development and progress (Puspitasari et al., 2021). Learning encompasses a variety of components, including strategies, resources, and interactions, all of which work to advance students' overall knowledge and proficiency (Maree, 2022). Different requirements and learning styles are taken into consideration while designing instruction, together with the academic objectives to be met (Kennedy & Sundberg, 2020). As a consequence, studying helps pupils comprehend the subject matter better and prepares them to handle obstacles in the future. (Cebrián et al., 2020). Learning must always be supported by an objective and thorough evaluation procedure to guarantee the effectiveness of it.

Evaluation is a piece of crucial equipment for continuously assessing and tracking students' progress (Huang et al., 2021). Evaluation is crucial for determining how well learning objectives have been accomplished and indicating areas in need of development (Thornhill-Miller et al., 2023). Through evaluation, both teachers and students may determine the learning process's advantages and disadvantages and create plans for future development (Thornhill-Miller et al., 2023). Therefore, evaluation serves as both a measuring instrument and a source of constructive feedback during the teaching and learning process (Tang et al., 2020). Effective and targeted learning can be achieved by appropriate evaluation, as outlined in Classical Test Theory.

Classical Test Theory is a foundational method in educational and psychometric assessment that emphasizes using tests to measure a person's unique skills and traits (MacPhail et al., 2024; Minkos & Gelbar, 2021). With the assumption that test outcomes are composed of two primary components, true scores and measurement error, classical test theory offers a framework for comprehending how tests can be assessed and examined. (Darling-Hammond et al., 2020). The use of classical test theory is beneficial for the creation and assessment of test instruments as well as for the interpretation of test results to provide accurate feedback (El-Sabagh, 2021). Thus, in a variety of educational and research situations, the application of this theory enables the construction of more quantifiable and efficient evaluation systems.

Previous research yielded some notable conclusions about the quality of daily test questions and scientific literacy ability exams. The first study revealed that the validity of the questions remained poor, the reliability was insufficient, the level of difficulty and discrimination was pretty good, and the efficiency of the distractors was less than optimal. (Hassan et al., 2021). Meanwhile, the second study compared the quality of the questions using the classical test theory approach and the Rasch Model. The validity of the questions differed between the two approaches, with the classical test theory classifying 3 questions as valid and 12 as invalid, while the Rasch Model classified 6 as valid and 9 as invalid (Ahmad, 2020). The reliability score in the classical test theory is 0.40 (medium), while the Rasch Model shows 0.43 (medium), with a person reliability value in the Rasch Model of 0.54 (bad) and product reliability of 0.91 (very good). The level of difficulty of the questions is also different, with classical test theory grouping questions into easy, medium and difficult categories, while the Rasch Model has four categories of difficulty. In the aspect of differential power, both approaches show similar results, namely the majority of questions are classified as bad with only a few questions considered good.

Based on the background and previous research, the author is interested in analyzing the items on the mathematical Numeracy Literacy questions using classical test theory.

Research Methods

This research is descriptive quantitative research that analyzes question items to evaluate the quality of the questions in the context of Numeracy Literacy. Quantitative research is an approach that focuses on collecting and analyzing numerical data to obtain objective and measurable results (Hooda et al., 2022). This research was conducted at one of the junior high schools in Bantul, Daerah Istimewa Yogyakarta, Indonesia, involving 31 grades from Class 8 (Table 1). The questions used are 9 essay questions on Numeracy Literacy in Algebra Form material following the Kurikulum Merdeka. Data collection was carried out through interviews with teachers to obtain relevant information about the implementation and effectiveness of the questions. In the analysis of these questions, AnBuSo 8.0 software was used which takes into account the level of difficulty and differentiation to provide an in-depth picture of the quality of the questions used in the learning process. The criteria for levels of difficulty and different strengths can be seen in Table 2 and Table 3.

Table 1. Details of Respondents in Research

Gender	Male	16 (51,61%)
	Female	15 (48,39%)
	Total	**31 (100%)**
Age	13	20 (64,52%)
	14	11 (35,48%)
	Total	**31 (100%)**

Source: Data from Researcher

Table 2. Difficulty Level Criteria

Difficulty Level	**Description**
0,00-0,30	Hard
0,31-0,70	Medium
0,71-1,00	Easy

Source: (Istiyono, 2020)

Table 3. Differential Power Criteria

Criteria	**Description**
$D \leq 0{,}199$	Bad (Rejected)
$0{,}200 - 0{,}299$	Good enough (Needs revision)
$0{,}300 - 0{,}399$	Medium (No Revision Required)
$D \geq 0{,}400$	Very Good

Source: (Istiyono, 2020)

Research Results

This research focuses on analyzing test items using Classical Test Theory on Algebra Form material for grade 8 at one of the state schools in Bantul, Special Region of Yogyakarta, Indonesia. This research involved 31 students in grade 8 as research subjects. Details of the questions used and scoring in this study are in Table 4. Researchers will analyze based on gender and overall.

Table 4. Details of Material in Question Item Analysis using Classical Test Theory

Number	Learning Materials	Value Details	Total
1	Understand and know the structure of algebraic forms	8.5	20
2		11.5	
3	Simplify Forms of Many Tribes	7	20
4		13	
5	Understanding Explanations Using Algebraic Forms	20	20
6	Understand how Changing Eq	20	20
7	Understanding Multiplication and Division of Tribal Forms Single	3	20
8		3	
9		14	
Total			100

Source: Data from Researcher

Difficulty Level

The level of difficulty is an important indicator in evaluating the quality of questions, which measures the extent to which students can answer questions correctly (Owan et al., 2023). In this analysis, researchers focused on difficulty levels based on gender and the overall student population to see whether there were significant differences in ability to answer questions between these groups. Thus, the results of this analysis can provide deeper insight into how gender factors influence question difficulty and help in designing questions that are fairer and more effective for all students.

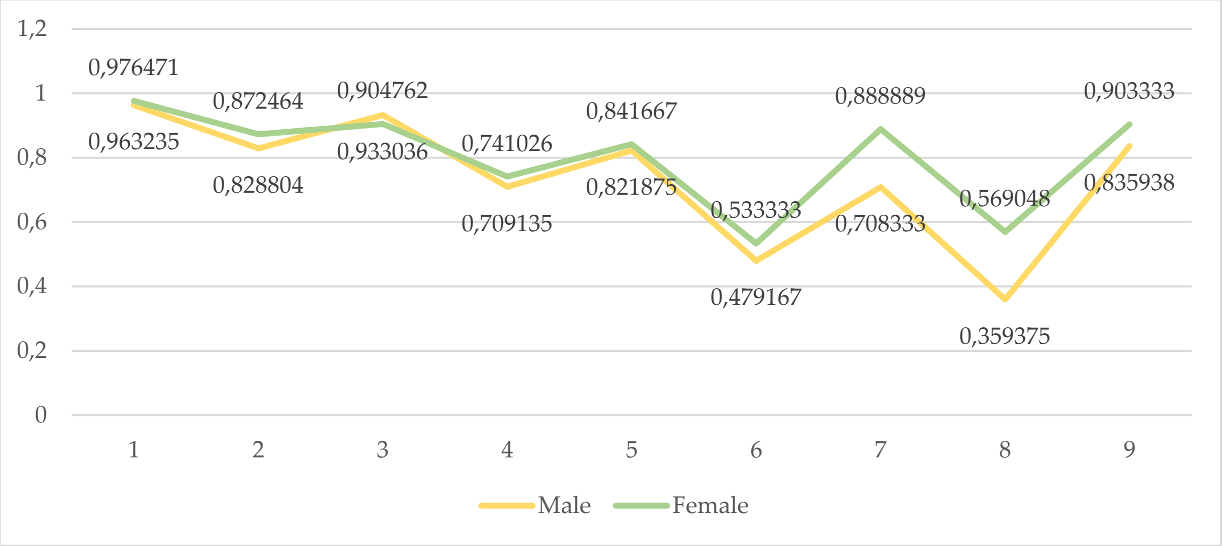

Fig. 1. Results of Difficulty Levels Based on Gender Analyzed with AnBuSo 8.0

Based on Figure 1, the level of difficulty of the questions for male students shows that of the nine question numbers, most of them fall into the "Easy" category with scores between 0.71 to 1.00. Question numbers 1, 3, 5, and 9 fall into this category, with difficulty values of 0.963, 0.933, 0.822, and 0.836 respectively, indicating that most male students can answer these questions correctly. Meanwhile, question number 6 with a value of 0.479 and question number 8 with a value of 0.359 are in the "Medium" category, indicating a medium level of difficulty. Question number 2 with a value of 0.829 and question number 4 with a value of 0.709 are also included in the "Easy" category, although the value of number 4 is close to the lower limit of this category.

In contrast, data for female students shows a slightly different pattern. Most of the question numbers also fall into the "Easy" category, with question numbers 1, 2, 3, 5, 7, and 9 showing a level of difficulty above 0.71, meaning most female students were able to answer these questions correctly. However, two question numbers indicate a "Medium" level of difficulty, namely number 6 with a value of 0.533 and number 8 with a value of 0.569, which shows that these questions are more challenging for female students than the other questions. There were no question numbers in the "Hard" category for female students, indicating that overall, these questions tended to be more accessible to female students than to male students, although this difference was not significant.

Based on an analysis of the level of difficulty of the questions between male and female students, it can be seen that most of the questions fall into the "Easy" category for both groups, although female students tend to find these questions slightly easier than male students. Questions that fall into the "Medium" category show that some questions are more challenging for both groups, with a slightly larger difference for female students. There were no questions classified as "Hard" for either of them, indicating that overall, the questions tended to be easier and more accessible for female students compared to male students, although the difference was not significant.

Apart from analyzing gender, researchers also analyzed it as a whole which can be seen in Figure 2.

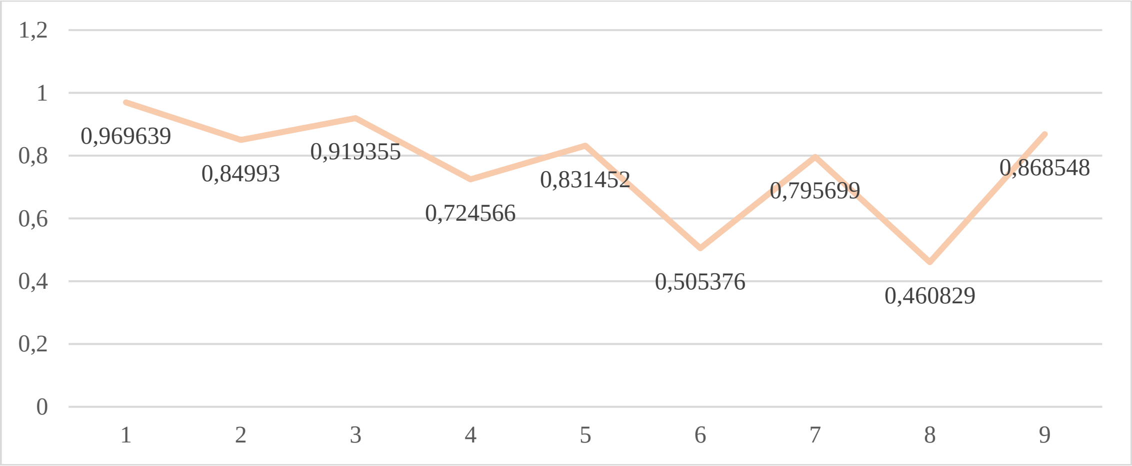

Fig. 2. Results of Difficulty Levels Overall Analyzed with AnBuSo 8.0

Based on data on the overall difficulty level of the questions, the majority of the questions fall into the "Easy" category, indicating that the majority of students were able to answer correctly. Question numbers 1, 3, 4, 5, and 9 are all in this category, with difficulty ratings indicating that the questions are relatively easy for students to access. This indicates that these questions, in general, succeeded in measuring students' abilities without causing excessive difficulty.

However, several questions fall into the "Medium" category, namely numbers 6 and 8, indicating that these questions are more challenging for students overall. This moderate level of difficulty may indicate that the questions require a deeper understanding or a more complex solution strategy. There are no questions that fall into the "Hard" category, which means all questions have a level of difficulty that students can reach, although with varying levels of success. This shows a fairly good balance in the distribution of question difficulty levels, although there is room for improvement in designing questions that are challenging but still accessible to all students.

Overall, analysis of the difficulty levels of the questions showed that the majority of the questions were in the "Easy" category, with only a few questions in the "Medium" category, and none in the "Hard" category. This indicates that these questions can generally be answered well by students, although some offer greater challenges. In conclusion, these questions were successful in measuring students' understanding of the material tested with a balanced level of difficulty, but there is an opportunity to design questions that are more varied in difficulty to be more effective in identifying differences in levels of understanding between students.

Different Power

Just like the level of difficulty, this research will focus on analyzing the power of differences based on gender and overall to see whether there are differences in the effectiveness of the questions in differentiating the abilities of male and female students, as well as to reflect the general performance of the questions in measuring variations in ability across the student population.

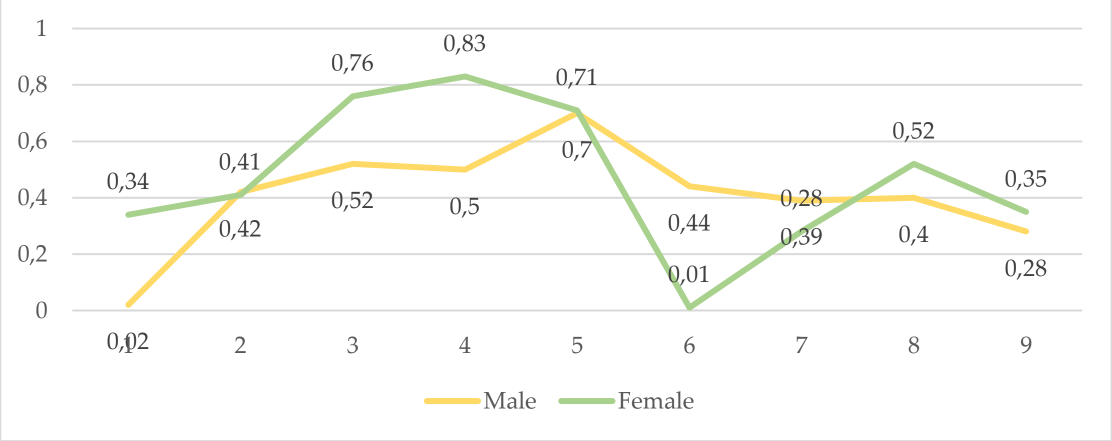

Fig. 3. Results of Different Power Based on Gender Analyzed with AnBuSo 8.0

Based on the analysis of the differential power of questions for male students, the results show that the majority of questions have a differential power that is in the "Very Good" category with a value of more than 0.400. Question numbers 2, 3, 4, 5, 6, and 8 all fall into this category, indicating that the questions can differentiate well between students with high and low ability. However, numerous questions do not match the good criterion, such as questions 1 and 9, which are in the "Bad" category with a power difference value of less than 0.200, indicating that they are ineffective in discriminating student abilities and may need to be revised or replaced.

Meanwhile, the analytical results suggest that differential power varies more widely among female students. Question numbers 3, 4, and 5 exhibit very good discrimination, with scores in the "Very Good" category, indicating that the questions are successful at discriminating female students' talents. However, numerous questions demonstrate little distinguishing power, such as question number 6, which is in the "Bad" category, and question number 1, which is only in the "Medium" category. This demonstrates that, while some questions perform well, others require adjustment to be more successful in assessing differences in female students' abilities. Female students generally have a greater ability to differentiate questions than male pupils.

Based on the examination of the questions' distinguishing power, it appears that the questions are more effective in differentiating the abilities of male students than female students, with the bulk of male-specific questions falling into the "Very Good" category. Meanwhile, the questions for female students revealed a higher variety in discriminating power, with some having good discrimination power and others performing badly. This shows that the questions for male students are generally more consistent in measuring differences in ability, so it can be concluded that in terms of differential power, the questions are more effective and consistent in differentiating the abilities of male students compared to female students.

Apart from analyzing gender, researchers also analyzed it as a whole which can be seen in Figure 4.

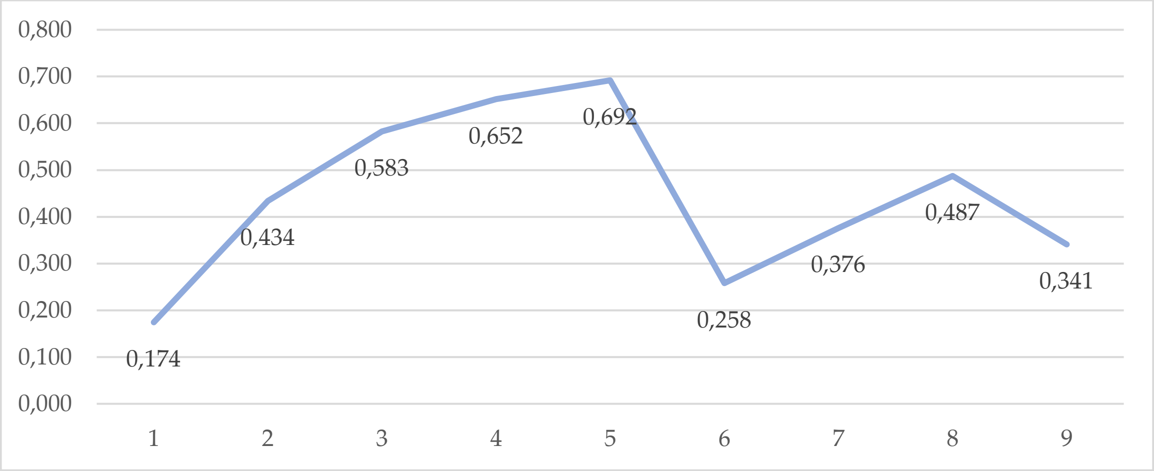

Fig. 4. Results of Different Power Overall Analyzed with AnBuSo 8.0

Based on the overall discriminative power analysis, the results show that the majority of questions have quite good discriminative power in differentiating between students with high and low abilities. Question numbers 2, 3, 4, 5, and 8 are all in the "Very Good" category with different power values above 0.400. This indicates that the questions are very effective in differentiating students' overall ability levels, so they do not require further revision.

However, there are several questions that have lower discrimination power, such as question numbers 1 and 6 which are in the "Bad" category with scores below 0.200 and 0.300, which means these questions are less effective in differentiating students' abilities and may require revision. or repair. Question numbers 7 and 9 are in the "Medium" category with sufficient differential power, but there is still room for improvement to be more optimal in evaluating students' abilities. Overall, although most of the questions have good discrimination, some require more attention to improve measurement quality.

Overall, the discriminating power analysis showed that the majority of questions had good qualities in discriminating students' abilities, with the majority being in the "Very Good" category. However, several questions require revision because they have low distinguishing power, especially questions numbers 1 and

6, which are less effective in measuring differences in student abilities. Although in general, the quality of the questions is quite adequate, improvements to some questions will further optimize the test's ability to evaluate differences in student abilities more accurately and consistently.

Research Discussion

Education creates a strong foundation for classroom development (Cobb et al., 2020). The classroom is the main forum for an effective learning process. Learning that takes place in the classroom requires appropriate evaluation to ensure the achievement of educational goals (Schildkamp et al., 2020). This evaluation then becomes an important tool in assessing student learning outcomes, and one method that is often used is classical test theory. Classical test theory plays an important role in providing an analytical framework for measuring the validity, reliability and distinguishability of questions so that evaluation results can be used to improve the overall quality of education.

This research focuses on analyzing test items using Classical Test Theory on Algebra Form material for grade 8 students at one of the state schools in Bantul, Daerah Istimewa Yogyakarta, Indonesia, involving 31 students as research subjects. Based on Table 4, the material tested covers various aspects of algebra, such as understanding the structure of algebraic forms, simplifying multi-term forms, as well as understanding the multiplication and division of single-term forms. Each question item is given a certain value which reflects its weight in the total assessment. Analysis was carried out both overall and based on gender, to identify the quality of the test items and their conformity with the principles of Classical Test Theory.

From the results of the difficulty level by gender, the analysis shows that the majority of questions are in the "Easy" category for both groups of students, although female students generally find these questions slightly easier than male students. Although most of the questions were easy for both groups to answer, there were a few questions that indicated a difficulty level of "Medium", meaning some questions presented a greater challenge for students, especially for female students. There were no questions classified as "Hard" for either group, indicating that overall, these questions tended to be more accessible to female students. Therefore, even if the difference in difficulty is not significant, adjustments in designing more challenging questions can help create tests that are more effective in measuring students' abilities equally across both genders.

As for the overall difficulty level results, the analysis shows that the majority of questions fall into the "Easy" category, which means the majority of students were able to answer correctly. This demonstrates that the questions were effective in testing students' abilities without incurring significant difficulty. Several problems, however, fall into the "Medium" category, indicating a larger challenge for students, which may necessitate deeper comprehension or more advanced problem-solving techniques. There were no questions in the "Hard" category, showing that the difficulty level of the questions remained within students' reach, albeit with varying degrees of success. According to Classical Test Theory, a balanced distribution of difficulty levels is needed to ensure that tests can accurately measure students' diverse abilities (Rajagukguk & Naibaho, 2023). As a result, while these questions have been useful in measuring student knowledge, there is room to construct questions with varied levels of difficulty to be more effective in finding disparities in levels of understanding among students.

In terms of gender differentiation outcomes, the analysis shows that the questions are more effective in differentiating the abilities of male students than female students, with the majority of the questions falling into the "Very Good" category. Meanwhile, the questions for female students showed a higher variance in discriminating power, with some questions having excellent discrimination power and others doing poorly. This demonstrates that questions for male students are more consistent in measuring differences in ability. Thus, while the findings varied between the two genders, the questions are more successful and consistent in distinguishing the talents of male students from female students.

Meanwhile, in terms of overall discriminating power results, the study shows that the majority of questions are of good quality in separating students' skills, with the majority falling into the "Very Good" group. This indicates that the questions effectively differentiate students' overall ability levels, so they do not require further revision. However, several questions have lower discrimination power, which indicates that these questions are less effective in differentiating students' abilities and may require improvement. In this context, Classical Test theory emphasizes the importance of good discrimination to ensure that each question can accurately measure differences in ability between students (Vincent & Shanmugam, 2020). Therefore, although in general the quality of the questions is quite adequate, improvements to some questions that have low differential power will further optimize the test's ability to evaluate differences in student abilities more accurately and consistently.

So, from all the analysis above, it can be concluded that although the majority of the questions are in the “Easy” category and can measure students' abilities without causing excessive difficulty, there are differences in the level of difficulty and the different power of questions based on gender, where the questions are more effective and consistent in differentiating the abilities of male students compared to female students. This shows that there needs to be improvements to several questions to increase the validity of the measurement, especially in ensuring that the test can evaluate students' abilities fairly and evenly between the two gender groups. Develop questions that not only measure the material in a representative manner but also take into account various levels of student ability, so that the test can be more effective in identifying differences in levels of understanding among students as a whole (Sistyawati & Apriani, 2024). As a consequence, designing more hard questions with more evenly distributed power will increase the quality of evaluating student abilities..

Conclusion

To sum up, this study is a lighthouse that illuminates the way to a better comprehension of how we assess prospective teachers' digital literacy abilities. The results paint a picture of a world full of growing curiosity and a spirit of cooperation, where scholars, organizations, and countries are working together to understand the complexities of digital literacy evaluation in teacher preparation. Spatial mapping becomes a potent tool as the scholarly conversation on this subject develops, not just for visualization but also for promoting cross-border interdisciplinary cooperation Researchers are better able to negotiate the difficulties of digital literacy assessment research with fresh focus and direction thanks to this approach, which enables them to identify research objectives, collaborative opportunities, and emerging trends. In the end, this study not only advances our knowledge of digital literacy assessment but also establishes a solid framework for further research in this crucial field of education, which could influence practice, policy, and instructional strategies for teacher preparation globally.

CRediT author statement: **Z. Zafrullah**: Conceptualization, Methodology, Software, Validation, Writing draft preparation; I.A. **Fitri**: Supervision, Data curation; **R.N. Gunawan**: Visualization, Investigation; **S. Nurjanah:** Reviewing; **E. Mutiara & G.E. Laksita**: Editing.

Acknowledgements. The authors would like to thank anonymous reviewers and the conference editors for their comments on earlier versions to improve this study’s quality.

Cite this article: Zafrullah, Z., FitriI, A., Gunawan, R.N., Nurjanah, S., Mutiara, E., Laksita, G.E. (2025). Analysis of Numeracy Literacy Question Items in Junior High School Mathematics: Analysis using Classical Test Theory. Materials of International Scientific-Practical Internet Conference “Challenges of Science”. Issue VIII, pp. 5-13. https://doi.org/10.31643/2025.01

References

Adeniyi, I. S., Al Hamad, N. M., Adewusi, O. E., Unachukwu, C. C., Osawaru, B., Onyebuchi, C. N., Omolawal, S. A., Aliu, A. O., & David, I. O. (2024). Educational reforms and their impact on student performance: A review in African Countries. *World Journal of Advanced Research and Reviews*, *21*(2), 750–762.

Ahmad, T. (2020). Scenario based approach to re-imagining future of higher education which prepares students for the future of work. *Higher Education, Skills and Work-Based Learning*, *10*(1), 217–238.

Alajmi, M. (2024). Promoting equity and equality in student learning: principals as social justice leaders in Kuwaiti schools. *International Journal of Educational Management*.

Barker, E. (2024). *National Character: and the factors in its formation*. Taylor & Francis.

Bleazby, J. (2020). Fostering moral understanding, moral inquiry & moral habits through philosophy in schools: a Deweyian analysis of Australia’s Ethical Understanding curriculum. *Journal of Curriculum Studies*, *52*(1), 84–100.

Cebrián, G., Palau, R., & Mogas, J. (2020). The smart classroom as a means to the development of ESD methodologies. *Sustainability*, *12*(7), 3010.

Cobb, P., Jackson, K., Henrick, E., & Smith, T. M. (2020). *Systems for instructional improvement: Creating coherence from the classroom to the district office*. Harvard Education Press.

Darling-Hammond, L., Flook, L., Cook-Harvey, C., Barron, B., & Osher, D. (2020). Implications for educational practice of the science of learning and development. *Applied Developmental Science*, *24*(2), 97–140.

El-Sabagh, H. A. (2021). Adaptive e-learning environment based on learning styles and its impact on development students' engagement. *International Journal of Educational Technology in Higher Education*, *18*(1), 53.

Hassan, M. A., Habiba, U., Majeed, F., & Shoaib, M. (2021). Adaptive gamification in e-learning based on students' learning styles. *Interactive Learning Environments*, *29*(4), 545–565.

Hooda, M., Rana, C., Dahiya, O., Rizwan, A., & Hossain, M. S. (2022). Artificial intelligence for assessment and feedback to enhance student success in higher education. *Mathematical Problems in Engineering*, *2022*(1), 5215722.

Huang, Y., Richter, E., Kleickmann, T., Wiepke, A., & Richter, D. (2021). Classroom complexity affects student teachers' behavior in a VR classroom. *Computers & Education*, *163*, 104100.

Istiyono, E. (2020). Pengembangan instrumen penilaian dan analisis hasil belajar fisika dengan teori tes klasik dan modern. *Yogyakarta: UNY Press. L, I.(2019). Evaluasi Dalam Proses Pembelajaran. Jurnal Manajemen Pendidikan Islam*, *9*, 478–492.

Izzulhaq, B. D., Gunawan, R. N., Zafrullah, Z., Ayuni, R. T., Ramadhani, A. M., & Fitria, R. L. (2024). Research Trends on Leadership in Indonesian Schools: Bibliometric Analysis (2008-2024). *Elementaria: Journal of Educational Research*, *2*(1), 19–38.

Kennedy, T. J., & Sundberg, C. W. (2020). 21st century skills. *Science Education in Theory and Practice: An Introductory Guide to Learning Theory*, 479–496.

MacPhail, A., Tannehill, D., & Ataman, R. (2024). The role of the critical friend in supporting and enhancing professional learning and development. *Professional Development in Education*, *50*(4), 597–610.

Maree, J. G. (2022). The psychosocial development theory of Erik Erikson: critical overview. *The Influence of Theorists and Pioneers on Early Childhood Education*, 119–133.

Minkos, M. L., & Gelbar, N. W. (2021). Considerations for educators in supporting student learning in the midst of COVID-19. *Psychology in the Schools*, *58*(2), 416–426.

Owan, V. J., Abang, K. B., Idika, D. O., Etta, E. O., & Bassey, B. A. (2023). Exploring the potential of artificial intelligence tools in educational measurement and assessment. *Eurasia Journal of Mathematics, Science and Technology Education*, *19*(8), em2307.

Pamuji, S., & Mulyadi, Y. (2024). Formation Of Students' Character Through Islamic Education. *International Journal of Islamic Thought and Humanities*, *3*(1), 26–35.

Puspitasari, D., Widodo, H. P., Widyaningrum, L., Allamnakhrah, A., & Lestariyana, R. P. D. (2021). How do primary school English textbooks teach moral values? A critical discourse analysis. *Studies in Educational Evaluation*, *70*, 101044.

Qazi, M. A., Sharif, M. A., & Akhlaq, A. (2024). Barriers and facilitators to adoption of e-learning in higher education institutions of Pakistan during COVID-19: Perspectives from an emerging economy. *Journal of Science and Technology Policy Management*, *15*(1), 31–52.

Rajagukguk, M. J. T., & Naibaho, D. (2023). Mampu Memilih Soal Berdasarkan Tingkat Kesukaran. *Jurnal Pendidikan Sosial Dan Humaniora*, *2*(4), 12736–12747.

Ramadhani, A. M., Yakob, N. Y. B., Ayuni, R. T., Zafrullah, Z., & Bakti, A. A. (2024). Trends in implementation of game use as learning at primary schools level in scopus database: a bibliometric analysis. *Jurnal Penyelidikan Sains Sosial*, *7*(23).

Ramdas, B., Rao, M. V., & Chandrika Reddy, D. N. (2024). Analysis On New National Education Policy 2020–School Education In India. *Turkish Online Journal of Qualitative Inquiry*, *15*(3).

Reimers, F. M. (2024). The sustainable development goals and education, achievements and opportunities. *International Journal of Educational Development*, *104*, 102965.

Sakhiyya, Z., & Rahmawati, Y. (2024). Overview of education in Indonesia. In *International Handbook on Education in South East Asia,* 1–25. Springer.

Sanyal, B. C. (2024). *Higher education and employment: An international comparative analysis*. Taylor & Francis.

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.

Sistyawati, R. I., & Apriani, D. (2024). Pengembangan Soal PISA Materi Bangun Ruang Kubus Untuk SMP. *Jurnal Pendidikan Matematika Sebelas April*, *3*(1), 48–56.

Tang, T., Vezzani, V., & Eriksson, V. (2020). Developing critical thinking, collective creativity skills and problem solving through playful design jams. *Thinking Skills and Creativity*, *37*, 100696.

Thornhill-Miller, B., Camarda, A., Mercier, M., Burkhardt, J.-M., Morisseau, T., Bourgeois-Bougrine, S., Vinchon, F., El Hayek, S., Augereau-Landais, M., & Mourey, F. (2023). Creativity, critical thinking, communication, and collaboration: assessment, certification, and promotion of 21st century skills for the future of work and education. *Journal of Intelligence*, *11*(3), 54.

Valencia, A. (2024). The Future of Higher Education in Latin America and the Caribbean: A Foresight Reflection. *The Bloomsbury Handbook of Context and Transformative Leadership in Higher Education, Edited by M. Drinkwater and P. Deane*, 302-332.

Vincent, W., & Shanmugam, S. K. S. (2020). The role of classical test theory to determine the quality of classroom teaching test items. *Pedagogia: Jurnal Pendidikan*, *9*(1), 5–34.

Zafrullah, Z., Ramadhani, A. M., Awliya, D., & Ayuni, R. T. (2024). Implementasi Project-based Learning di Sekolah: Analisis Bibliometrik (1998-2023): Indonesia. *Ciencias: Jurnal Penelitian Dan Pengembangan Pendidikan*, *7*(2), 11–23.

Zafrullah, Z., & Zetriuslita, Z. (2021). Minat belajar siswa kelas VII terhadap media pembelajaran matematika berbantuan Adobe Flash CS6. *Math Didactic: Jurnal Pendidikan Matematika*, *4*(2), 114–123.