JRERE 13 (1) 2024 73-84

Journal of Research and Educational Research Evaluation

http://journal.unnes.ac.id/sju/index.php/jere

Development of Minimum Competency Evaluation Instrument: Analyzing Conceptual and Daily Life Skills Using Rasch Model

Muhammad Lutfi Amin[1]✉, Supriyadi Supriyadi[2], Endang Susilaningsih[3]

[1,2,3]Universitas Negeri Semarang, Indonesia

Article Info

History Articles
Received:
16 December 2023
Accepted:
18 January2024
Published:
30 March 2024

Keywords:
Minimum Competency Evaluation, Conceptual Abilities, Daily Life, Rasch Model

Abstract

Evaluating educational outcomes is essential for ensuring students develop both conceptual understanding and practical skills. However, current evaluation methods often lack comprehensive coverage of both areas. This study addresses this gap by developing an evaluation instrument using the Rasch Model to assess students' minimum competencies in conceptual understanding and real-life application. The research is developmental, involving content validity, reliability, and construct validity of the instrument. Respondents included 5 experts in the expert testing phase, 28 individuals in the small-scale test, and 132 individuals in the large-scale test. Data collection used both quantitative and qualitative approaches. The instrument provides a comprehensive assessment of minimum competencies. Reliability tests showed inter-rater reliability at 0.769, and Cronbach's Alpha reliability for the small-scale test at 0.86 for single-point questions and 0.83 for four-point questions. The large-scale test results indicated Cronbach's Alpha reliability of 0.87 for single-point questions and 0.83 for four-point questions. The findings suggest that the developed instrument demonstrates good validity and reliability. This study significantly contributes to a more holistic evaluation of education, ensuring a balanced assessment of both conceptual understanding and practical application.

✉Correspondence Address :
Postgraduate, Universitas Negeri Semarang
Jalan Kelud Utara III, Kel. Petompon Kec. Gajah Mungkur
Kota Semarang, Indonesia 50237
E-mail : lutfi190897@students.unnes.ac.id

p-ISSN 2252-6420
e-ISSN 2503-1732

INTRODUCTION

Evaluation is the process of gathering data or information, which is then compared against specific criteria to draw conclusions, known as evaluation outcomes (Arikunto, 2021). It can also be understood as a process of summarizing and interpreting facts, as well as making professional judgments to make decisions based on the collected information (Hayati et al., 2020). "Evaluation is a systematic process of determining the merit, value, and worth of information by applying scientific methods and principles to gather, analyze, and interpret data" (Matthes et al., 2023). Effective evaluation should not only measure learning outcomes but also provide feedback that guides future teaching and learning processes (Johnsen & VanTassel-Baska, 2022). Minimum Competency Assessment (AKM) is an assessment of the basic competencies that all students must possess to develop their capacities and contribute positively to society. "Minimum Competency Assessment (AKM) is designed to evaluate foundational skills that are essential for students to progress in their educational journey and to function effectively in society" (Wolf et al., 2023). The basic competencies measured in AKM include literacy in reading and mathematical literacy (numeracy) (Pursitasari et al., 2022). Both areas assess competencies that encompass logical and systematic thinking skills, the ability to reason using learned concepts and knowledge, and the skills to sift and process information. "Literacy in reading and mathematical literacy are critical components of educational assessments, as they reflect students' ability to process information and apply logical reasoning" (Sak, 2022). AKM aims to measure competencies in depth, not just content mastery. "AKM aims to assess deeper cognitive processes, focusing on students' ability to apply their knowledge in practical situations rather than merely recalling facts" (da Silva et al., 2023). This aligns with science literacy assessments, which not only evaluate students' knowledge but also emphasize science competencies in everyday life.

Science learning (IPA) focuses on exploring nature and is a comprehensive discovery process (Saputri et al., 2020). This learning invites students to use their thinking skills to solve various problems (Zahara et al., 2022). This aligns with one of the main goals of science learning, which is to create a society with science literacy (Rahayu, 2014). "Science literacy involves the integration of scientific knowledge, understanding of scientific processes, and the ability to apply this knowledge in real-world contexts" (Schutz & Muis, 2023). Science literacy is the ability of students to apply science concepts in everyday life (PISA, 2015). "The development of scientific literacy in students is crucial for preparing them to navigate and address complex issues in modern society" (Higgins, 2021). Science literacy includes various strategies used by students to solve various problems in society (Thomson & De Bortoli, 2008). According to Holbrook & Rannikmae (2009), science literacy encompasses three aspects: understanding of science in terms of concepts (what people know), moral or ethical values (what people value), and context (what people can do). Evaluations in science education should encompass not only knowledge acquisition but also the application of scientific reasoning and ethical considerations (Anderson et al., 2022).

According to the Program for International Student Assessment (PISA) survey in 2006, Indonesia's students' science literacy ranked 50 out of 57 countries. In 2009, it dropped to 60 out of 65 participating countries. In 2012, Indonesia's students' science literacy ranking was 64 out of 65 participating countries (OECD, 2017). The latest survey in 2015 shows Indonesia ranked 62 out of 70 participating countries (OECD, 2016). This data indicates that science literacy among Indonesian students is still low. "The persistent low ranking of Indonesian students in international assessments highlights the

urgent need for educational reforms that enhance science literacy" (Astuti et al., 2024). Hayat et al., (2014) research also reinforces this, stating that Indonesian students have not been able to apply the science concepts they understand in everyday life. Bybee (Soobard & Rannikmae, 2011; Hadromi et al., 2019) proposed that in certain activities or situations (for example, during tests), students' understanding often only comes from the textbooks they read. At higher levels, namely conceptual and procedural, students have understood the principles and theories of science, as well as how one concept is connected to another as a whole, understanding the scientific process, and having an understanding of inquiry. Students who can utilize various concepts and demonstrate the ability to relate them to everyday life, as well as understand that science, social studies, and technology are interconnected and influence each other, are at a multidimensional level (Odja & Payu, 2014). "Science literacy is a multifaceted competency that includes understanding scientific concepts, applying scientific methods, and evaluating scientific information critically" (Egalite & Kisida, 2016).

The implementation of the Merdeka Curriculum at all levels of basic and secondary education allows for greater flexibility, enabling teachers to adapt learning to students' abilities and the local context and content, as well as the teachers' own abilities. "The Merdeka Curriculum emphasizes the importance of contextualizing education to local needs while ensuring alignment with national and global educational standards" (Saranya et al., 2021). The Merdeka Curriculum also supports the abolition of National Exams, giving freedom to all schools in Indonesia without distinguishing between urban and rural schools (Mulyasa, 2023). Changes in the Merdeka Curriculum indicate that education is not only the responsibility of teachers but is a joint responsibility among teachers, principals, school supervisors, parents, and the community as a whole. "Flexibility in curriculum design and implementation allows educators to tailor instruction to the diverse needs and abilities of students" (Sodi, 2024). Teachers are also required to continuously refine and adjust evaluations to the development of science, technology, arts, as well as local, national, and global needs, so that evaluations in the school curriculum are truly relevant to students' needs, the development of the times, and the demands and burdens of tasks that will be faced after participating in learning evaluations. "Teachers play a critical role in continuously updating and adapting their evaluation methods to keep pace with advancements in knowledge and technology" (Braman et al., 2024).

Minimum competency evaluation is an effort to measure the basic abilities that every student must possess. These competencies include conceptual abilities and their application in everyday life. To produce valid and reliable evaluations, appropriate instruments are needed. "Effective evaluation instruments must be both valid and reliable, providing accurate and consistent measures of student competencies" (Huang & Hu, 2015). This research aims to develop a Rasch model-based minimum competency evaluation instrument capable of analyzing both aspects. "The Rasch model offers a robust framework for developing and validating assessment instruments, ensuring that they accurately reflect student abilities" (Khine, 2020). The Rasch model was chosen for its ability to provide in-depth and accurate analysis of student abilities. This model allows researchers to identify suitable and unsuitable items in the instrument, thus obtaining more valid and reliable data. "Using the Rasch model, educators can identify which test items are most effective in differentiating between different levels of student ability" (Ricciutti & Zhang, 2024). The developed evaluation instrument is expected to provide an accurate picture of students' abilities in understanding basic concepts and their application in real-life situations. This is important because

evaluation results can be used by teachers to improve the quality of learning and help students achieve the minimum competencies expected. "Evaluation results should inform instructional practices, guiding teachers in modifying and improving their teaching strategies to better support student learning" (Syafrial et al., 2023). Comprehensive evaluation systems should consider multiple aspects of student performance, including conceptual understanding, practical application, and critical thinking skills (Lai et al., 2022). "Educational policies and practices must be grounded in empirical research to ensure that evaluations are meaningful and beneficial for student development" (Zhang et al., 2022).

METHODS

The development procedure in this research follows the development guidelines laid out by Gall et al. (2003), which consist of ten steps in research and development implementation strategy, including: (1) preliminary analysis, (2) planning, (3) initial model or product development, (4) hypothetical model evaluation, (5) refinement, (6) small-scale testing, (7) adjustment based on test results, (8) large-scale testing, (9) final model revision, and (10) dissemination. The research participants comprise 5 individuals for expert testing, 28 individuals for small-scale testing, and 132 individuals for large-scale testing. Data collection is conducted through interviews, documentation, and questionnaire usage. Data analysis is performed using both quantitative and qualitative approaches.

Content validity is evaluated by five professional raters. Reliability is measured using the Inter-Class Correlation Coefficient (ICC) due to the number of raters being more than three. IBM SPSS 26 software is utilized for reliability analysis. Construct validity is measured using the construct validity method within the Rasch model, which can be observed through the Item Polarity results. A positive value of the Point Measure Correlation Coefficient (Pt. Mea-Corr) indicates the absence of conflicts between the item and the measured construct. A Mean Square Outfit value less than 1.5 signifies the productivity of the measurement value. Quantitative data analysis involves calculating validity and reliability. Additionally, the researchers evaluate item fit, item difficulty (item measure), individual fit (person fit), and DIF using Rasch Model analysis on small and large-scale tests conducted in eighth-grade classrooms at MTs Al Asror Patemon, Gunungpati, Semarang City.

RESULTS AND DISCUSSION

The needs analysis, as part of the preliminary study, was conducted by interviewing eighth-grade Science teachers. The purpose of the interviews was to understand the implementation of Science teaching and the process of creating questions used to assess students' Science abilities.

Planning for the development of a minimum competency evaluation instrument containing questions about conceptual and daily life skills has been undertaken. This planning involves creating a framework for the minimum competency evaluation instrument for Science subjects, emphasizing conceptual and daily life skills. A total of 30 questions have been prepared and divided according to the proportion of cognitive domains adjusted to the educational achievements in Science subjects. The items created include multiple-choice, complex true-false multiple-choice, complex yes-no multiple-choice, matching questions, fill-in-the-blank, and essay questions. The proportion of non-routine items adjusted to the educational achievements developed in the initial stages is described in Table 1.

Table 1. Cognitive Domain of Science Test Items in Conceptual and Daily Life Abilities.

No	Cognitive Domain		Item
1	Analyzing substances based on their properties and characteristics.	33.3	10
2	Analyzing physical and chemical changes in a substance and their relevance to the concept of simple mixture separation that can be applied in daily life.	40	12
3	Analyzing the influence of heat energy on temperature changes and its application to insulating and conducting objects in daily life	26.7	8
Total			30

According to Table 1, the minimum competency assessment instrument, designed to measure conceptual and daily life abilities, underwent testing to ensure its alignment with the desired skills. Content validity and reliability tests were conducted by a group of experts. The minimum competency assessment instrument, tailored for eighth-grade students to measure conceptual and daily life abilities, underwent content validity assessment by several experts. The content validity evaluation involved five experts assessing the suitability of the material, structure, and language used in the instrument. Experts evaluating the conceptual and daily life competency assessment instrument in Science questions include both academics and practitioners. Subsequently, the expert evaluation results of the instrument were processed using the Aiken V formula prepared in Excel format. The data from the content validity calculation using Aiken's V formula were then displayed in Table 2.

Table 2. Coefesient of Expert Agreement

Index Item	Aiken's V Index Item	Description	Index Item	Aiken's V Index Item	Description
1	0.8	Valid	16	0.85	Valid
2	0.9	Valid	17	0.8	Valid
3	0.85	Valid	18	0.85	Valid
4	0.85	Valid	19	0.85	Valid
5	0.95	Valid	20	0.85	Valid
6	0.9	Valid	21	0.8	Valid
7	0.85	Valid	22	0.8	Valid
8	0.8	Valid	23	0.8	Valid
9	0.85	Valid	24	0.85	Valid
10	0.85	Valid	25	0.8	Valid
11	0.85	Valid	26	0.85	Valid
12	0.85	Valid	27	0.9	Valid
13	0.85	Valid	28	0.9	Valid
14	0.85	Valid	29	0.9	Valid
15	0.85	Valid	30	0.9	Valid

Based on the data in Table 2, the agreement coefficients among experts obtained were then compared with the validity coefficient table. Aiken version items are deemed valid if the agreement coefficient among experts exceeds the value of V. The Aiken V table indicates that the V value for 5 assessors and 5 scale options is validity > 0.80. Conversely, the agreement coefficient among experts is considered invalid if it does not reach a value of 0.80 for an error probability of 0.04. The construct validity test proves that

the indicators developed accurately measure the variable abilities. Conceptual and daily life abilities involve understanding concepts through everyday life. The data used to test construct validity are the results of minimum competency evaluations on a small scale. Analysis was conducted using Winsteps software. The construct validity of the Rasch model can be observed through Item Polarity. A positive Point Measure Correlation (Pt. Mea Corr) value indicates no conflict between the measured item and construct. Unidimensionality testing is also depicted in Table 3.

Table 3. Unidimensionality Test

Index	Udimensionality Test	Empirical
question point one	Total raw variance in observations	100
	Raw variance explained by measures	53.8
	Raw variance explained by persons	16.4
	Raw varience explained by items	37.4
	Raw varience explained total	46.2
	Unexplned varience in 1st contrast	13.6
	Unexplned varience in 2nd contrast	10.4
	Unexplned varience in 3rd contrast	7.8
	Unexplned varience in 4th contast	5.7
	Unexplned varience in 5th contrast	3.8
question point four	Total raw variance in observations	100
	Raw variance explained by measures	54.0
	Raw variance explained by persons	28.8
	Raw varience explained by items	25.2
	Raw varience explained total	46.0
	Unexplned varience in 1st contrast	14.4
	Unexplned varience in 2nd contrast	9.4
	Unexplned varience in 3rd contrast	6.9
	Unexplned varience in 4th contast	6.4
	Unexplned varience in 5th contrast	5.3

Based on Table 3, the raw variance measurement results for the item with one point show a value of 53.8%. This is close to the expected value of 51.6%, indicating that the unidimensionality requirement of 20% has been met. A unidimensionality value of 40% is better, and a value of 60% is considered excellent. The unexplained variances are 13.6%, 10.4%, 7.8%, 5.7%, and 3.8%, respectively. This suggests that the unexplained variance should ideally not exceed 15%, classifying the instrument as good.

For the item with four points, the raw variance measurement results show a value of 54.0%, which is close to the expected value of 53.6%. This finding indicates that the unidimensionality criterion of 20% has been satisfied. A unidimensionality value of 40% is better, and 60% is exceptional (Sumintono, 2018). The unexplained variances are 14.4%, 9.8%, 6.9%, 6.4%, and 5.3%, respectively. This indicates that the unexplained variance should ideally not exceed 15%, placing the instrument in the good category.

Item Characteristics of Evaluation Instruments Analyzed with Item Response Theory or Rasch

The Rasch model analysis was applied to the small-scale data, which consisted of 30 questions. The item analysis included

evaluating item difficulty (item measure), item fit (item fit), individual abilities (person measure), and individual fit (person fit). Item Fit assesses whether the items function properly for measurement. The criteria used to evaluate item fit are outfit mean-square (MNSQ), z-standard (ZSTD), and point measure correlation (Pt. Mean Corr). Items that do not meet these three criteria are considered suboptimal and need to be revised or replaced. The results of the Item Fit analysis are detailed in Table 4.

Table 4. Output Item Fit

Item	*Outfit*		*Pt. Measure Corr*	Item	*Outfit*		*Pt. Measure Corr*
	MNSQ	ZSTD			MNSQ	ZSTD	
6	9.90	5.7	0.41	24	0.66	-0.5	0.58
15	4.51	4.1	0.45	19	0.49	-0.4	0.59
9	3.74	3.0	0.30	26	0.49	-0.4	0.59
8	1.36	0.8	0.36	2	0.07	-0.8	0.66
11	1.10	0.4	0.45	13	0.07	-0.8	0.66
20	1.10	0.4	0.45	22	0.07	-0.8	0.66
5	1.14	0.4	0.66	18	0.11	-1.1	0.76
1	0.67	0.1	0.53	25	0.11	-1.1	0.76
12	0.67	0.1	0.53	28	2.01	3.2	0.53
21	0.67	0.1	0.53	16	0.95	-0.1	0.66
7	1.05	0.3	0.61	27	0.93	-0.2	0.72
4	0.66	-0.5	0.58	10	0.89	-0.3	0.68
30	0.79	-0.8	0.81	29	0.77	-0.9	0.81
17	0.65	-1.5	0.74				

Based on Table 4, the accepted range for MNSQ is between 0.5 and 1.5. For ZSTD, the accepted range is between -2.0 and 2.0. As for Pt. Measure Corr, the accepted range of values is between 0.4 and 0.85 (Sumintono et al., 2015). Items are considered suitable or accepted if they meet at least one of these three criteria. The results of the Item Fit analysis indicate that out of a total of 30 test items administered to 28 students, 27 items were deemed acceptable, while 3 items were excluded. Items numbered 9, 28, and 30 were deemed unsuitable

In the Item Measure analysis, the standard deviation reaches 1.80. If the value exceeds 0.0 logits plus the standard deviation, then the question is considered difficult. If the value falls between 0.0 logits plus the standard deviation and 0.0 logits minus the standard deviation, then the question is categorized as moderate. Meanwhile, if the value is less than 0.0 logits minus the standard deviation, then the question is classified as easy. Details regarding the Item Measure results and the classification of difficulty levels can be seen in Table 5.

Table 5. Output Item Measure

Item	*Measure*	Conclusiom	Item	*Measure*	Conclusion
2	4.19	Difficult	19	-2.19	Moderate
13	4.19	Difficult	26	-2.19	Moderate
22	4.19	Difficult	1	-3.36	Easy
18	2.42	Moderate	12	-3.36	Easy
25	2.42	Moderate	21	-3.36	Easy

5	1.87	Moderate	3	-5.40	Easy
7	1.43	Moderate	14	-5.40	Easy
15	-0.42	Moderate	23	-5.40	Easy
4	-0.62	Moderate	27	1.07	Difficult
8	-0.62	Moderate	28	0.25	Moderate
11	-0.62	Moderate	30	0.13	Moderate
20	-0.62	Moderate	29	-0.29	Moderate
24	-0.62	Moderate	16	-0.35	Moderate
9	-0.82	Moderate	17	-0.35	Moderate
6	-1.92	Moderate	10	-0.47	Moderate

Based on Table 5 from the small-scale test, there are 4 items classified as difficult: items 2, 13, 22, and 27. A total of 20 items fall into the moderate difficulty category. The remaining 6 items are in the easy category: items 1, 3, 12, 14, 21, and 23. The reliability of the minimum competency evaluation instrument to measure conceptual and daily life skills for eighth-grade middle school students was tested based on a) interrater reliability; b) small-scale test reliability; and c) large-scale test reliability. The reliability values tested consecutively were 0.769; 0.86; 0.87 for the single-point items and 0.769; 0.83; 0.83 for the four-point items. Based on the analysis, the results are presented in Table 6.

Table 6. Estimated Of Reliability

The Experimental Phase	Reliability	Number of Item
Judges	0.769	30
Single-Point Questions		
Small-Scala Test	0.86	23
Large-Scala Test	0.87	18
Four-Point Questions		
Small-Scala Test	0.83	7
Large-Scala Test	0.83	5

Interrater Reliability

The results of the Intraclass Correlation Coefficient (ICC), which measure the reliability and consistency of the ratings provided by different judges or raters, are detailed in Table 7.

Table 7. Interclass Correlation Coefficient

95% Confidence Interval				F Test with True Value			
	Interclass Correlation	Low Bound	Upper Bond	Value	df1	df2	Sig
Single Measures	0.532	0.353	0.707	5.554	29	87	0.000
Average Measures	0.820	0.685	0.906	5.554	29	87	0.000

The analysis results from Table 7 indicate that the reliability value *rxy* is 0.769. According to reliability standards, this figure falls within the moderate range as it lies between 0.4 r < 0.6. However, under the same classification, this reliability can also be considered high as it falls between 0.6 ≤ r 0.8. Therefore, based on these criteria, the reliability value of 0.769 falls into the category that can be considered high. This pertains to the interclass correlation coefficient.

Small Scale Test Reliability

Person reliability scores are 0.86 for point one questions and 0.80 for point four questions, while item reliability scores are 0.89 for point one and 0.73 for point four. For point one questions, the consistency of student responses is quite good, and the item quality in this instrument demonstrates very high reliability. On the other hand, for point four questions, the consistency of student responses is good, and the item quality in this instrument shows adequate reliability. Cronbach's Alpha values are 0.86 for point one and 0.83 for point four. In the Rasch Model, the Cronbach's Alpha value indicates the overall interaction level between individuals and items. According to the criteria outlined by Sumintono et al., (2015), a Cronbach's Alpha value of less than 0.50 is categorized as poor, 0.50-0.60 as fair, 0.61-0.70 as sufficient, 0.71-0.80 as good, and greater than 0.80 as very good. Therefore, with Cronbach's Alpha values of 0.86 and 0.83, it can be concluded that this instrument falls into the very good category.

Large-Scale Test Reliability

The person and item reliability values in the large-scale test for point one questions indicate adequate consistency in student responses, with a value of 0.83 categorized as excellent. The quality of the items for point one in the instrument shows very high reliability with a value of 0.97. For point four questions, the consistency of student responses is adequate with a value of 0.76 categorized as good, while the quality of the items shows very high reliability with a value of 0.86. The Cronbach's Alpha values are 0.87 for point one and 0.83 for point four. These values reflect the overall interaction between students and the items within the Rasch Model.

Discussion

The instrument developed in this study is an evaluation tool for minimum competency in eighth-grade MTs students, consisting of 23 questions. The types of questions included are multiple choice, short answer, matching, and true/false. This evaluation instrument is designed based on the relevant basic competencies for eighth-grade students according to the independent curriculum, aiming to analyze conceptual and daily life skills.

The developed instrument has been tested for content validity and interrater reliability. The content validity of the instrument is proven with an expert agreement index ranging from 0.80 to 0.90. Additionally, the reliability tested, including interrater reliability, showed a value of 0.769.

Construct validity tested using the Rasch model shows that the raw variance for single-point questions is 48.4%, which is close to the expected value of 46.6%, with successive variances of 13.2%, 9.4%, 5.3%, 4.2%, and 4.0%. For four-point questions, the raw variance is 61.4%, also near the expected value of 61.0%, with successive variances of 14.4%, 12.0%, 6.5%, 3.8%, and 0%. These results indicate that the variance explained by the single-point and four-point questions does not exceed 15%, categorizing the instrument as good.

The implication of this research is that the developed minimum competency assessment instrument provides additional types of questions to evaluate students' abilities in concepts and daily life. It is hoped that these questions will help improve students' understanding of concepts and daily life. Teachers can utilize this guide as a tool to enhance students' conceptual and daily life skills, as well as to design questions that align with the taught material. The analysis of item characteristics based on the Rasch model provides in-depth information not only about the quality of the instrument and the abilities of the tested students but also about the relationship between the questions posed and the students' responses.

Research on educational assessment tools indicates that using a variety of question types, including multiple choice and constructed response, enhances the ability to

measure diverse cognitive skills (Zhuang et al., 2021). The validation and reliability testing methods employed in this study align with best practices for educational measurement, ensuring that the instruments are both reliable and valid for assessing student competency (Akib, 2015).

Furthermore, content validity indices, such as those used in this study, are critical for confirming that the instrument adequately covers the intended content areas (Brennan, 2023). The interrater reliability of 0.769 is within acceptable ranges, indicating consistent scoring among different raters (Garcia-Loro et al., 2020).

The use of the Rasch model for construct validity testing is supported by contemporary research as an effective method for evaluating the quality and functioning of assessment items (Tunç, 2023). The findings from the Rasch analysis in this study provide valuable insights into item characteristics and student performance, contributing to the instrument's overall validity (Silverstein et al., 2023).

This study's development and validation of a minimum competency assessment instrument contribute to the field of educational measurement by providing a robust tool for evaluating eighth-grade students' conceptual and daily life skills. The rigorous testing of content validity, reliability, and construct validity ensures that the instrument is both effective and reliable for educational use.

CONCLUSION

Based on the findings and research discussions, it can be concluded that the instrument developed to assess minimum proficiency in conceptual understanding and everyday life applications has demonstrated validity and reliability. The assessment results from this instrument can accurately measure students' cognitive abilities. This tool can be invaluable for teachers in evaluating the conceptual and everyday life abilities of eighth-grade students in middle school. The strengths of this instrument include comprehensive evaluation of conceptual abilities and practical applications, high content validity and reliability, as well as diverse question types. However, the instrument also presents weaknesses such as the potential need for item revision or replacement if found inadequate, and the necessity for continuous monitoring to ensure impartiality and effectiveness in implementation by teachers.

REFERENCES

Akib, E. (2015). The Validity and Reliability of Assessment for Learning (AfL). *Education Journal*, *4*(2), 64.

Anderson, K., Stern, M., Powell, R., Dayer, A., & Archibald, T. (2022). A culturally responsive evaluation framework and its application in environmental education. *Evaluation and Program Planning*, *92*, 102073.

Arikunto, S. (2021). *Dasar-dasar evaluasi pendidikan edisi 3*. Bumi aksara.

Astuti, E. P., Wijaya, A., & Hanum, F. (2024). Characteristics of junior high school teachers' beliefs in developing students' numeracy skills through ethnomathematics-based numeracy learning. *Journal of Pedagogical Research*, *8*(1), 244–268.

Braman, J., Brown, A., & Richards, M. J. (2024). Reshaping learning with next generation educational technologies. In *Reshaping Learning with Next Generation Educational Technologies* (Issue February).

Brennan, R. L. (2023). *Educational measurement*. Rowman & Littlefield.

da Silva, A. N., Matos, M., Faustino, B., Neto, D. D., & Roberto, M. S. (2023). Rethinking leahy's emotional schema scale (LESS): Results from the Portuguese adaptation of the LESS. *Journal of Rational-Emotive & Cognitive-Behavior Therapy*, *41*(1), 95–114.

Egalite, A., & Kisida, B. (2016). School size and student achievement: a longitudinal analysis. *School Effectiveness and School Improvement*, *27*, 1–12.

Gall, M., Borg, W., & Gall, J. (2003). Educational Research: An Introduction. *British Journal of Educational Studies*, *32*.

Garcia-Loro, F., Martin, S., Ruipérez-Valiente, J. A., Sancristobal, E., & Castro, M. (2020). Reviewing and analyzing peer review Inter-Rater Reliability in a MOOC platform. *Computers & Education*, *154*, 103894.

Hadromi, Yudiono, H., & Budiman, F. (2019, November). The Optimization of the Vocational School Teacher Candidates' Employability Skills. In *Journal of Physics: Conference Series* (Vol. 1273, No. 1, p. 012001). IOP Publishing.

Hayat, B., Suryadi, B., & Surata, W. (2014). Assesment for Quality Education. *Prosiding Konferensi Ilmiah Tahunan Himpunan Evaluasi Pendidikan Indonesia*.

Hayati, N., Wadi, H., & Suud, S. (2020). IMPLEMENTASI PENDEKATAN SAINTIFIK BERBASIS PENGUATAN PENDIDIKAN KARAKTER DALAM PEMBELAJARAN SOSIOLOGI KURIKULUM 2013. *Jurnal Pendidikan Sosial Keberagaman*, *7*.

Higgins, M. (2021). *Unsettling Responsibility in Science Education: Indigenous Science, Deconstruction, and the Multicultural Science Education Debate*.

Holbrook, J., & Rannikmae, M. (2009). The meaning of scientific literacy. *International Journal of Environmental and Science Education*, *4*(3), 275–288.

Huang, X., & Hu, Z. (2015). On the Validity of Educational Evaluation and its Construction. *Higher Education Studies*, *5*(4), 99–105.

Johnsen, S., & VanTassel-Baska, J. (2022). *Handbook on Assessments for Gifted Learners: Identification, Learning Progress, and Evaluation*.

Khine, M. S. (2020). Rasch measurement: Applications in quantitative educational research. *Rasch Measurement: Applications in Quantitative Educational Research*, *January*, 1–281.

Lai, J. W. M., De Nobile, J., Bower, M., & Breyer, Y. (2022). Comprehensive evaluation of the use of technology in education – validation with a cohort of global open online learners. In *Education and Information Technologies* (Vol. 27, Issue 7). Springer US.

Matthes, J., Schneider, M., & Preckel, F. (2023). The Relation Between Prior Knowledge and Learning in Regular and Gifted Classes: A Multigroup Latent Growth Curve Analysis. *Journal of Educational Psychology*, *116*(2), 278–296.

Mulyasa, E. (2023). Implementasi Kurikulum Merdeka (A. Ulinnuha (ed.). *PT Bumi Aksara*.

Odja, A. H., & Payu, C. S. (2014). Analisis kemampuan awal literasi sains siswa pada konsep IPA. *Prosiding Seminar Nasional Kimia*, *20*, 1–8.

OECD. (2016). Results from PISA 2015: Indonesia. *OECD Publishing*, 1–8.

OECD. (2017). PISA for development assessment and analytical framework: reading , mathematics and science. *OECD Publishing*, *1*(1), 1–198.

PISA, O. (2015). *Draft Science Framework. 2013*.

Pursitasari, I. D., Permanasari, A., & Jaenudin, D. (2022). Pelatihan penyusunan e-asesmen literasi sains berbasis akm bagi guru IPA SMP di kabupaten bogor. *Jurnal Pemberdayaan Masyarakat*, *1*(1), 26–33.

Rahayu, S. (2014). Revitalisasi Scientific Approach dalam Kurikulum 2013 untuk Meningkatkan Literasi Sains: Tantangan dan Harapan. *Seminar Nasional Kimia Dan Pembelajarannya*.

Ricciutti, N. M., & Zhang, S. (2024). A pilot study of the behavioral addictions knowledge survey: Ensuring students' knowledge about process/behavioral

addictions. *Journal of Addictions & Offender Counseling*, *n/a*(n/a).

Sak, M. (2022). The Routledge Handbook of the Psychology of Language Learning and Teaching: edited by Tammy Gregersen and Sarah Mercer, Oxford and New York, Routledge, 2021, Pp. 446, £190.00 / $250.00 (hbk), ISBN 978-0-36733-723-0. *Journal of Multilingual and Multicultural Development*, 1–2.

Saputri, D. F., Sari, I. N., & Fadillah, S. (2020). *Panduan Desain Pembelajaran Bermuatan Karakter bagi Guru IPA Sekolah Menengah Pertama (SMP) Kelas VIII*. CV. Pustaka One Indonesia.

Saranya, V., Kalyani, S., & Ramachandran, V. (2021). Fuzzy decision analysis for regional contextualization of global educational frameworks. *Sādhanā*, *46*(2), 92.

Schutz, P., & Muis, K. (2023). *Handbook of Educational Psychology*.

Silverstein, M. C., Bjälkebring, P., Shoots-Reinhard, B., & Peters, E. (2023). The numeric understanding measures: Developing and validating adaptive and nonadaptive numeracy scales. *Judgment and Decision Making*, *18*, e19.

Sodi, D. (2024). Assessing the Effectiveness of Education Policy in Addressing Skill Gaps and Enhancing Employability. *INTERANTIONAL JOURNAL OF SCIENTIFIC RESEARCH IN ENGINEERING AND MANAGEMENT*, *08*, 1–11.

Soobard, R., & Rannikmae, M. (2011). Assessing Student's Level of Scientific Literacy Using Interdisciplinary Scenarios. *Science Education International*, *22*(2), 133–144.

Sumintono, B. (2018). *Rasch Model Measurements as Tools in Assesment for Learning* (Issue October 2017). Trim Komunikata Publishing House.

Sumintono, B., Islam, U., Indonesia, I., Widhiarso, W., & Mada, U. G. (2015). *RascH* (Issue September). Trim Komunikata Publishing House.

Syafrial, Nopiyanto, Y. E., Pujianto, D., Insanistyo, B., & Cotton, W. (2023). Assessing the profesional competence of physical education teachers in Bengkulu Province: Examining the role of teacher characteristics. *Journal Sport Area*, *8*(2), 151–164.

Thomson, S., & De Bortoli, L. (2008). *Exploring Scientific Literacy: How Australia measures up. The PISA 2006 survey of students' scientific, reading and mathematical literacy skills*.

Tunç, E. B. (2023). ASSESSING THE PSYCHOMETRIC PROPERTIES OF THE BRIEF RESILIENCE SCALE: A RASCH MODELING APPROACH. *International Journal of Eurasian Education & Culture*, *8*(23).

Wolf, M. K., Lopez, A. A., & Lee, J. (2023). An investigation of the use of standardized and local assessments for young EAL students. In *EAL Research for the Classroom* (pp. 164–184). Routledge.

Zahara, S. R., Imanda, R., & Alvina, S. (2022). Development of assessment sheet for measuring students' scientific literacy level in the era of revolution 4.0. *Jurnal Penelitian Pendidikan IPA*, *8*(3), 1096–1101.

Zhang, L., Kirschner, P. A., Cobern, W. W., & Sweller, J. (2022). There is an evidence crisis in science educational policy. *Educational Psychology Review*, *34*(2), 1157–1176.

Zhuang, L., Yang, Y., & Gao, J. (2021). Cognitive assessment tools for mild cognitive impairment screening. *Journal of Neurology*, *268*(5), 1615–1622.