Review of International Journal Education and School Leadership (RITEDUL)
Vol. 3 No. 1 November 2026, Page., 25-38
https://sagorvos.net/index.php/RITEDUL
e-ISSN: 3089-7491

Evaluation of Pupils' Mathematical Problem-Solving Skills: A Literature Review of HOTS-Based Instruments and Assessment Rubrics

Busnawir
Universitas Halu Oleo, Indonesia
busnawir@uho.ac.id

Abstract

Assessing pupils' mathematical problem-solving skills requires an instrument capable of revealing both the quality of their thought processes and the accuracy of their solutions. This article aims to examine the characteristics and quality of *Higher Order Thinking Skills* (HOTS)-based instruments, as well as the design and use of assessment rubrics to evaluate pupils' mathematical problem-solving skills. The method used was a literature review employing a descriptive-qualitative approach. The findings indicate that the characteristics of HOTS instruments need to be determined based on the demands of the cognitive process, rather than solely on the complexity of calculations or the use of real-life contexts. Instruments need to align ability indicators, task requirements and expected evidence of responses, and be supported by relevant evidence of validity and reliability. Assessment rubrics complement the instruments through criteria and descriptors that distinguish the quality of problem understanding, strategy selection, problem-solving execution, and interpretation of results. Analytical rubrics can provide detailed information regarding pupils' abilities, but consistent application requires clarity in descriptors, response examples, and alignment of assessors' perceptions. This study concludes that the integration of HOTS-based instruments and assessment rubrics forms the foundation for process- and outcome-oriented evaluation, whilst also providing information for feedback and the improvement of mathematics learning.

Keywords: mathematical problem-solving; HOTS; assessment tools; assessment rubrics; literature review.

Introduction

Mathematics education plays a vital role in developing pupils' ability to understand problems, reason about relationships between pieces of information, and make decisions based on sound judgement. This approach positions the learning of mathematics not merely as an activity involving the mastery of formulas and the application of calculation procedures, but as a process of building the ability to apply knowledge in a variety of situations (Melinda dkk., 2024). The Organisation for Economic Co-operation and Development (OECD, 2023) emphasises that mathematical literacy encompasses an individual's capacity to reason mathematically and to formulate, use and interpret mathematics in solving problems across a variety of real-life contexts. This framework highlights the interconnection between conceptual understanding, reasoning and problem-solving as mutually supportive elements of mathematical competence. In the context of learning, this interconnection provides the basis for the need for assessment that reveals how pupils utilise their mathematical knowledge, rather than merely how much material they can recall. Therefore, the assessment of mathematical problem-solving

skills must be regarded as an integral part of efforts to understand the quality of pupils' learning processes and outcomes more comprehensively.

Mathematical problem-solving skills should be understood as skills involving a number of interrelated processes, ranging from identifying relevant information to interpreting and evaluating the solutions obtained. Within the PISA 2022 framework, problem-solving activities are linked to the processes of formulating situations mathematically, applying mathematical concepts and procedures, and interpreting and evaluating the results of the solution, with mathematical reasoning occupying a central position (OECD, 2023). Consequently, a student's success is not adequately represented by the accuracy of the final answer alone, but must also be assessed through the accuracy of the model constructed, the appropriateness of the chosen strategy, and the reasonableness of the interpretation of the results (Gradini, 2022). For example, a pupil may demonstrate an understanding of the problem's structure despite making a calculation error during the execution stage, whilst another pupil may obtain the correct result without providing an adequate explanation of the rationale behind the procedure used (Budiman & Jailani, 2014). These differences in response characteristics provide a conceptual basis for distinguishing between the assessment of outcomes and the assessment of the problem-solving process. Consequently, assessments designed to measure problem-solving must provide opportunities for pupils to demonstrate their reasoning whilst also providing a basis for assessors to interpret this evidence accurately.

One relevant framework for designing such assessments is the *Higher Order Thinking Skills* (HOTS) approach (Ikhlas et al.., 2024). Krathwohl, (2002), which outlines the dimensions of cognitive processes comprising remembering, understanding, applying, analysing, evaluating and creating, with a focus on increasingly complex processes, particularly analysis, evaluation and creation. This framework can be used to examine whether a task merely requires pupils to follow a predetermined procedure or demands that they analyse relationships, consider alternatives, and produce solutions based on their own thinking. In the assessment of mathematical problem-solving, the requirement for analysis may focus on recognising the structure and relationships between the elements of the problem, whilst evaluation may be carried out by assessing the appropriateness of the strategy or the reasonableness of the solution. Meanwhile, the demand for creation can be formulated through tasks that require pupils to construct a model or devise a solution. However, such mapping must be based on the thought processes actually required by the question, not merely on the inclusion of specific action verbs. Thus, HOTS is used as a framework for examining the quality of an instrument's cognitive demands, not merely as a label for questions deemed complex.

The development of HOTS-based instruments further requires alignment between the measurement objectives, ability indicators, task characteristics, and the evidence of responses expected from pupils. Krathwohl et al. (2002)) demonstrate the usefulness of a taxonomic framework in analysing the relationship between objectives, learning activities and assessment, whilst the OECD (2023) organises mathematics assessment

through the interconnection between reasoning, problem-solving processes, content knowledge and the context of its application. Based on these two frameworks, the development of assessment instruments in this study is considered to need to begin with the identification of the competencies to be assessed, followed by the design of tasks that allow these competencies to be demonstrated. If the aim of the assessment is to evaluate pupils' ability to evaluate strategies, for example, test items should require pupils to provide their reasoning or justification regarding the strategy, rather than merely demanding the result of a calculation. Similarly, if the aim is to assess the ability to formulate models, the required response must allow students to articulate the mathematical relationships they have established. This alignment is crucial to ensure that the interpretation of scores does not go beyond the evidence actually generated by the instrument. Therefore, the review of HOTS-based instruments must cover the task requirements and their alignment with problem-solving indicators, rather than merely the format of presentation or the level of complexity of the material.

In addition to the quality of the work, the assessment of mathematical problem-solving skills requires marking guidelines that can clearly distinguish between different levels of response quality. Assessment rubrics are a relevant approach as they provide criteria and descriptions of performance levels as a reference for assessing students' work. Prasetyaningsih et al.., (2025) , in their discussion of the assessment of mathematical problem-solving tasks among university students, distinguish between analytical rubrics, which assess each aspect of performance, and holistic rubrics, which assess the overall quality of the response. The rubric they developed covers conceptual understanding and connections, strategies and reasoning, calculation or execution, and communication. Although the context of this study is higher education—meaning its application to school pupils requires adaptation—these dimensions provide examples of how to break down the quality of problem-solving into observable components. In this article, the focus on the rubric is on the clarity of the relationship between ability indicators and performance descriptors, including how partially correct responses or those using different strategies can be assessed. Thus, the rubric is not treated merely as a list of scoring categories, but as a framework for explaining the meaning of the quality of answers that underlies the scoring.

Nevertheless, the use of rubrics does not in itself guarantee that the assessment has produced valid and consistent results. Jonsson & Svingby, (2007) , through a review of 75 studies, concluded that scoring consistency can be improved through the use of rubrics, particularly when the rubrics are analytical, topic-specific, and accompanied by examples of performance or assessor training. However, the review also emphasised that the mere presence of a rubric does not automatically result in valid assessment. These findings highlight the need to distinguish between the clarity of scoring guidelines and the sufficiency of evidence supporting the interpretation of assessment results. In the evaluation of mathematical problem-solving, an examination of rubrics must therefore consider whether the criteria used truly represent the target competencies and whether

the descriptors can be understood in a relatively uniform manner by assessors. Jonsson & Svingby (2007)) also identified the potential of rubrics to support learning through the explanation of expectations and criteria that can be used in both feedback and self-assessment. Consequently, the quality of a rubric needs to be assessed from two interrelated perspectives: its ability to support scoring decisions and its usefulness in providing information for improving learning.

The development of rubrics that specifically integrate Higher-Order Thinking Skills (HOTS) into mathematics learning has also been undertaken in educational research in Indonesia (Budiman & Jailani, 2014). Gradini, (2022), for example, developed an assessment rubric by integrating the characteristics of critical thinking, problem-solving, mathematical understanding, modelling, proof and reasoning, representation, and mathematical communication. The study evaluated the quality of the product based on the aspects of validity, reliability and practicality, and reported that the rubric developed met the quality criteria used in the study. These findings provide a concrete example that the development of HOTS assessment does not stop at the formulation of test items, but also encompasses the formulation of criteria for interpreting the quality of pupils' performance. However, the breadth of dimensions within HOTS rubrics also raises the need for analytical scrutiny to distinguish the constructs being measured: a rubric covering various thinking skills cannot be directly treated as a specific problem-solving rubric without examining its indicators. When considered in relation to the findings of Jonsson & Svingby, (2007), the issue that requires further examination is the alignment between target competencies, evidence of response, and the basis for scoring. On this basis, the need for a literature review in this article lies in the development of a synthesis linking the characteristics of HOTS instruments with the quality of assessment rubrics, rather than on the assumption that research on both is not yet available.

Based on the above, this article aims to examine the assessment of pupils' mathematical problem-solving abilities through two main areas of discussion: the characteristics and quality of HOTS-based instruments, and the design and use of assessment rubrics. The first discussion focuses on analysing the relationship between ability indicators, cognitive process requirements, task formats, and evidence of the instrument's quality. The second discussion focuses on the dimensions of assessment, the clarity of performance level descriptors, the consistency of scoring, and the alignment of the rubric with the responses elicited by the questions. These two discussions are brought together through the principle of alignment, namely the connection between the ability to be measured, the tasks used to elicit it, the evidence evident in students' responses, and the criteria used to assess that evidence.

The focus of this study stems from the OECD's (2023) emphasis on reasoning in problem-solving and the work of Jonsson & Svingby (2007) on the quality of rubric-based assessment. Through a synthesis of the literature, this article aims to provide a conceptual foundation and practical considerations for teachers and researchers in selecting,

developing and evaluating instruments and their rubrics, whilst also identifying aspects that still require empirical testing.

Research Methodology

This study employed a literature review using a descriptive-qualitative approach to examine *Higher Order Thinking Skills* (HOTS)-based instruments and assessment rubrics in the evaluation of mathematical problem-solving abilities among students at (Eliyah & Aslan, 2025). The literature reviewed included journal articles, academic books and assessment framework documents relevant to the research focus. The literature was selected based on topic relevance, clarity of the source, and its relevance to the development of instruments and the scoring of mathematical problem-solving skills. Literature meeting the criteria was then examined by noting the research objectives, educational context, ability indicators, instrument characteristics, rubric dimensions, and evidence of validity and reliability as reported at . The analysis was conducted through grouping, comparison and thematic synthesis into two main areas of discussion, namely the characteristics and quality of HOTS-based instruments, and the design and use of assessment rubrics. The results of the synthesis were used to identify similarities, differences, limitations and alignment between the requirements of the questions, evidence of pupils' responses, and the scoring criteria in the evaluation of mathematical problem-solving skills (Walliman & Walliman, 2021) .

Results and Discussion

HOTS-Based Instruments for Assessing Mathematical Problem-Solving Skills

Higher Order Thinking Skills (HOTS)-based instruments for assessing mathematical problem-solving skills need to be designed to reveal the meaningful application of knowledge, rather than merely the ability to recall formulas or follow demonstrated procedures. The conceptual basis for this can be traced back to the revised Bloom's taxonomy, which distinguishes cognitive processes into remembering, understanding, applying, analysing, evaluating and creating (Krathwohl, 2002).

In this discussion, the last three processes are the main focus as they relate to the requirements of analysing relationships between pieces of information, making judgements based on criteria, and formulating solutions or mathematical models. Nevertheless, these skills do not stand apart from foundational knowledge, as pupils still require conceptual understanding and procedural mastery to construct sound reasoning (Brown, 2018). Therefore, HOTS instruments should not be understood as tools that disregard basic skills, but rather as tools that require pupils to organise and utilise these skills to address more complex demands. The implication for assessment is the need to ensure that higher-order thinking processes are genuinely required by the task and can be identified through students' responses. Thus, the classification of an item as a HOTS question must be based on an analysis of its cognitive demands, rather than merely on the use of verbs such as 'analyse' or 'evaluate' in the question (Krathwohl, 2002).

The link between HOTS instruments and mathematical problem-solving skills can be explained through the role of reasoning in connecting problem situations with mathematical concepts. The OECD (2023) identifies mathematical reasoning as a central element underpinning the processes of formulating problems, applying mathematics, and interpreting and evaluating the results of problem-solving. In the problem-formulation process, pupils need to identify relevant information, define variables, state assumptions, and construct a mathematical representation of the situation at hand. In the process of applying mathematics, pupils select and implement strategies, manipulate representations, and formulate explanations leading to a solution. Furthermore, in the process of interpreting and evaluating, pupils need to relate the mathematical results back to the original context and consider their plausibility. This framework can be used to develop more operational assessment indicators, such as the ability to explain the reasons for choosing a model, compare strategies, or identify the limitations of a solution. However, not every item needs to encompass the entire process at once; the OECD (2023) explains that an assessment task may assess a specific part of the modelling cycle. Therefore, the scope of competence must be ensured across the entire set of assessment instruments, so that conclusions regarding pupils' competence are not based solely on one type of task.

A key characteristic of HOTS instruments lies in the requirement for mathematical decision-making, rather than merely in the length of the questions, the magnitude of the numbers, or the complexity of the calculations. Seaman, (2011), provides a basis for distinguishing between analysis, evaluation and creation based on their cognitive processes, whilst the OECD (2023) explains that reasoning can be observed when pupils select strategies, draw logical conclusions, explain their solutions and defend their arguments. Based on these two frameworks, the level of difficulty of a question needs to be conceptually distinguished from the depth of the thinking process required. As an illustrative example, a lengthy calculation accompanied by a sequence of steps does not necessarily provide evidence of a student's ability to select strategies independently (Egodawatte, 2010). Conversely, a task involving simple operations can be used to assess evaluation skills if students are asked to compare two solutions and explain why one of them does not apply under certain conditions. The use of real-life contexts must also serve a clear mathematical purpose, rather than merely serving as a narrative backdrop. The context should encourage pupils to identify assumptions, select information, or interpret results (Lewis & Gilman, 2005). Consequently, the analysis of the quality of assessment instruments needs to focus on the relationship between the stimulus, the questions, and the mathematical decisions pupils must make, rather than merely on the appearance of the item as a contextual problem.

The development of HOTS-based instruments requires alignment between the construct being measured, ability indicators, content, task requirements and expected responses. The taxonomic framework discussed by Seaman, (2011) provides a basis for examining the relationship between learning objectives and assessment, whilst the OECD

framework (2023) highlights the importance of considering mathematical processes, content and context in an integrated manner. In practice, the design of an assessment matrix can be directed towards clarifying the specific skills that each item is intended to assess. If the indicator is the ability to evaluate strategies, questions should require pupils to provide an assessment along with their reasoning, rather than merely producing a numerical answer. If the indicator is the ability to formulate a model, the assessment instrument should provide space for pupils to state the variables, assumptions and mathematical relationships used. As an example of such a design, a problem comparing several purchasing options could be framed not merely to identify the cheapest price, but also to explain the conditions that make a particular choice more advantageous. In this type of design, the evidence assessed includes the comparison model, the reasons for the choice of strategy, and the accuracy of the interpretation. Clarification of the evidence for responses from the stage of developing the assessment framework forms the basis for ensuring that the questions and their marking schemes are aligned with the same competency.

The choice of assessment instrument format must also take into account the type of evidence required to interpret pupils' mathematical problem-solving abilities. Research by Budiman & Jailani, (2014) indicates that mathematics HOTS assessment instruments can be developed in the form of multiple-choice or open-ended questions. In this study, which focused on first-semester Year 8 secondary school mathematics, the resulting instrument comprised 24 multiple-choice items and 19 open-ended items, which were deemed valid and suitable for use based on an assessment of content, construction and language. These findings suggest that the measurement of HOTS need not be limited to a single response format. However, the existence of these two types of assessment instruments must be considered in conjunction with the purpose of the assessment: selecting an answer and writing an essay provide different types of response evidence. If the evaluation specifically aims to reveal how pupils construct models or maintain strategies, the design outlined in this article emphasises the need for requests for explanation that allow these reasons to be examined. Conversely, when multiple-choice questions are used, item analysis must ensure that pupils are indeed faced with reasoning demands that correspond to the indicators. Thus, the question format is not the sole determinant of HOTS quality; what is more important is the alignment between the response format, the thinking demands, and the basis for interpreting assessment results (Budiman & Jailani, 2014).

The quality of HOTS-based instruments must be supported by test evidence, so that their designation as higher-order thinking instruments does not rely solely on the conceptual considerations of their developers. Budiman & Jailani, (2014) reported reliability coefficients of 0.713 for the multiple-choice items and 0.920 for the open-ended items they developed. These figures should be understood as the results of testing specific test instruments and research contexts, rather than as evidence that all HOTS instruments or all essay questions possess the same measurement quality. In a literature review, such evidence must be examined alongside explanations regarding indicators, participant

characteristics, validation procedures, and item analysis. The review must also distinguish between expert assessments of content, construction and language, and the empirical evidence obtained from pupils' responses. Furthermore, analyses of difficulty level and discriminative power provide information on item performance within the tested group, whilst analyses of cognitive demands are used to examine alignment with HOTS objectives. Consequently, decisions regarding an instrument's suitability must take several types of evidence into account collectively. Good reliability findings must still be linked to fundamental questions regarding what abilities the task actually measures and how the results will be used (Budiman & Jailani, 2014) .

Based on this synthesis, HOTS-based instruments for assessing mathematical problem-solving skills should be understood as tools that link cognitive demands to verifiable evidence of performance. Seaman, (2011), provides a foundation for analysing cognitive processes; the OECD (2023) offers a framework for the relationship between reasoning and problem-solving processes; whilst Budiman & Jailani, (2014) provide examples of the development and testing of instruments within the context of mathematics education in Indonesia. These three sources serve different functions and should be used proportionately: the conceptual framework helps to formulate target competencies, whilst the development study demonstrates how a design is tested within a specific context. The implication for teachers and researchers is the need to avoid defining HOTS questions solely on the basis of their format, narrative context, or level of difficulty. Instruments must demonstrate a clear relationship between indicators, questions and expected responses, whilst being supported by relevant evidence of quality. Furthermore, as problem-solving responses may exhibit different qualities of reasoning even when yielding similar final answers, the discussion of assessment instruments must be followed by an examination of assessment rubrics. Such rubrics are essential in the design of an evaluation to translate the established evidence from responses into targeted assessment criteria, thereby ensuring that the measurement objectives and the basis for scoring remain consistent.

Assessment Rubrics for Evaluating the Process and Outcomes of Mathematical Problem-Solving

An assessment rubric is a scoring guide that links evaluation criteria to descriptions of the levels of quality in pupils' performance. In the assessment of mathematical problem-solving skills, a rubric is necessary to clarify the aspects being assessed and the basis for scoring pupils' responses, so that assessment does not rely solely on the correspondence of the final answer with the answer key.

Egodawatte, (2010) explains that a rubric sets out the criteria and assessment scales used to identify the quality of work, whilst also providing information on the characteristics of the expected responses. Using this framework, pupils' work can be assessed based on evidence of understanding, strategy, execution and mathematical communication evident in their solutions. However, the presence of a rubric does not

automatically eliminate the assessor's professional judgement, as the assessor must still interpret responses based on the available descriptors.

Jonsson & Svingby, (2007), emphasise that rubrics have the potential to support consistency in marking, but that this benefit depends on the characteristics of the rubric and how it is used. Therefore, rubrics in problem-solving assessment should be viewed as a structured decision-making framework, rather than merely a table for assigning marks. Their quality is determined by the clarity of the relationship between the target competencies, evidence of performance, and the criteria used to make the assessment.

Assessing the process and outcomes of problem-solving requires a distinction to be made between the quality of the approach taken and the accuracy of the final result. Egodawatte, (2010), emphasises the assessment of the process as a key consideration when using rubrics for mathematical problem-solving tasks, particularly so that assessment information can be utilised to improve learning (Egodawatte, 2010). Based on this approach, rubrics need to enable assessors to identify which parts of the solution demonstrate understanding and which parts still require improvement. To illustrate the design, a pupil who chooses the correct strategy but makes an arithmetic error should be distinguished in their performance profile from a pupil who uses an inappropriate concept, even though both arrive at an incorrect final answer. Conversely, a correct numerical answer does not always provide sufficient evidence regarding the rationale behind the choice of strategy if the task specifically requires an explanation. Consequently, criteria for both process and outcome must be explicitly formulated to ensure that marking does not treat all errors as equivalent or base assessment solely on the final product. In this context, the outcome of the task remains an important component, but it is considered alongside evidence of reasoning and the implementation of strategies (Egodawatte, 2010). Thus, a rubric can provide a more informative picture of a student's ability than a simple 'correct' or 'incorrect' judgement, provided the task allows students the opportunity to demonstrate this process.

Based on the method of scoring, rubrics can be categorised into holistic rubrics and analytical rubrics. Egodawatte, (2010) explains that holistic rubrics are used to assess the overall quality of a response, whilst analytical rubrics break down performance into several dimensions that are assessed separately. This distinction has implications for the information obtained from the assessment. Holistic rubrics are appropriate when the aim of the assessment is to provide a general overview of the quality of the work, whilst analytical rubrics provide more detailed information regarding strengths and weaknesses in each aspect. For assessments aimed at diagnosing problem-solving abilities, the separation of dimensions allows teachers to see whether the main obstacles lie in conceptual understanding, strategy selection, procedural execution, or mathematical reasoning. However, Egodawatte, (2010) also points out that analytical scoring requires more time as each component must be examined individually. The findings (2007), reinforce this relevance by demonstrating that analytical rubrics—particularly those specific to the topic and supported by examples or assessor training—can help improve

scoring consistency. Therefore, the choice of rubric type should take into account the objectives of the assessment, the characteristics of the task, the need for diagnostic information, and the available resources.

The definition of assessment dimensions is a crucial stage as it determines the aspects of competence that will be reflected in pupils' score profiles. In his discussion of mathematics assessment, Egodawatte, (2010) outlines the importance of mathematical knowledge, strategic knowledge, and explanations of the steps and reasoning behind solutions. These dimensions indicate that assessment should not only focus on the accuracy of concept application but also on how pupils organise information and communicate mathematical relationships. For the purposes of this article, the rubric dimensions can be designed to include problem understanding, strategy planning, solution implementation, and interpretation or verification of results. This grouping constitutes a synthetic framework that needs to be adapted to the competency framework and task requirements; it is not intended as a mandatory structure to be applied to all assessment instruments. If a question merely requires the evaluation of a pre-existing solution, for example, the assessment dimensions need not include the independent construction of a model (Zapf & Roesch, 2008). Conversely, modelling tasks require criteria that pay attention to the assumptions, variables and relationships constructed by the pupils. Thus, the dimensions of the rubric must be derived from the competencies actually required by the question, so that the assessor does not award marks for aspects that the pupil has not had the opportunity to demonstrate in their response (Wiggins, 1990).

Once the dimensions have been established, the rubric requires descriptors that explain the differences in performance quality at each score level. Sitopu et al.., (2024) emphasise the importance of clear and specific criteria, whilst also discussing the risks of assessment when criteria are too vague or difficult to understand. On that basis, labels such as 'very good', 'good' and 'poor' need to be accompanied by descriptions of the evidence that can be identified in the students' work. For example, in the 'strategy' dimension, the highest level of performance might be characterised by the selection of an appropriate strategy and an explanation of the reasons for its use, whilst the level below might describe a relevant strategy that has not yet been explained or fully implemented. These descriptors need to be meaningfully distinguished, not merely through the use of different adjectives that do not indicate a change in the quality of the response. A 0–4 scale is one option discussed by Sitopu et al.., (2024) , but the number of levels still needs to be adjusted to account for reasonably distinguishable variations in performance. Furthermore, the wording of descriptors should not restrict marking to a single sequence of steps if the task permits valid alternative strategies. Thus, the clarity of descriptors must go hand in hand with openness to diversity in mathematical solutions, so that the rubric assesses the quality of reasoning, rather than merely the similarity of responses to model answers (Warju, 2016) .

The quality of a rubric also needs to be assessed in terms of the validity and consistency of its scoring. Brookhart, (2013) , based on a review of 75 studies, concluded

that a rubric alone does not automatically guarantee the validity of assessment, although it can support more consistent scoring under certain conditions. This finding emphasises the need to check whether the dimensions and descriptors truly represent the problem-solving skills to be assessed, rather than merely being convenient for assigning marks. On the other hand, the use of rubrics by multiple assessors requires attention to differences in the interpretation of criteria, tendencies to over- or under-score, and the influence of overall impressions on the assessment of each dimension. Prasetyaningsih et al.., (2025) discuss these issues as sources of scoring discrepancies that need to be anticipated. In the development process, examples of responses at various performance levels can be used as discussion material to align assessors' understanding, whilst scoring trials help to identify parts of the rubric that still cause ambiguity. Jonsson and Svingby (2007) specifically highlight the benefits of performance examples and assessor training in supporting reliability. Therefore, reports on the quality of rubrics need to include procedures for their use, rather than merely presenting a table of criteria or stating that the rubric has been reviewed by an expert.

In the context of HOTS-based assessment, the development of rubrics must take into account the alignment between the cognitive demands of the questions and the aspects that are rewarded in the marking process. Gradini, (2022) provides an example of the development of a mathematics HOTS rubric that integrates critical thinking, problem-solving, mathematical understanding, modelling, proof and reasoning, representation, and mathematical communication. The study employed a development process comprising initial investigation, design, development, implementation and evaluation, with the quality of the product assessed in terms of validity, reliability and practicality. The authors reported a validity score of 0.82, a reliability score of 0.79 and a practicality score of 81.02, which were categorised as 'good' according to the study's criteria. However, these figures must be interpreted in accordance with the testing procedures and context: the research abstract mentions the use of assessment by two experts and a practicality review by five mathematics teachers; consequently, the results should not be directly equated with evidence of scoring consistency across a variety of student responses in all contexts. Furthermore, the breadth of the HOTS dimensions in the rubric highlights the need to distinguish between general HOTS rubrics and those specifically designed to assess problem-solving. The contribution of this study lies in its example of integrating various skills, whilst its application still requires adaptation to the indicators and characteristics of the tasks used (Gradini, 2022) .

Assessment rubrics can ultimately be utilised not only to assign marks, but also to guide feedback, self-assessment and peer assessment. Prasetyaningsih et al.., (2025) developed an assessment approach for mathematical problem-solving tasks that involves students in analysing their own performance and that of their peers, with a primary focus on improving learning. Although the context is higher education, meaning that its application to school pupils requires adjustments to language, the complexity of criteria and guidance, this approach demonstrates how rubrics can make performance

expectations more transparent. In lesson planning, score profiles for each dimension can be used to formulate specific feedback—for example, asking pupils to clarify the reasoning behind their strategies or to check the validity of their conclusions—rather than simply providing a total mark. However, to realise these benefits, pupils must be familiarised with the criteria and given practice in using the rubric, so that they do not merely match their work to a number.

Jonsson & Svingby, (2007) state that rubrics have the potential to support learning and the quality of teaching, but that this potential must be understood in relation to how they are applied. Based on this synthesis, rubrics that are relevant for the assessment of mathematical problem-solving are those that provide information on the quality of both the process and the outcome, have descriptors that can be applied consistently, and provide a basis for follow-up learning tailored to pupils' needs.

Conclusion

Assessing pupils' mathematical problem-solving abilities requires instruments based on *Higher Order Thinking Skills* (HOTS) that are capable of revealing the quality of their reasoning, rather than merely the accuracy of their final answers. Such instruments must provide pupils with the opportunity to analyse information, determine strategies, formulate solutions, and interpret and evaluate their results. The characteristics of HOTS are not determined solely by the complexity of the calculations, the length of the stimulus, or the use of real-life contexts, but by the demands of the thought processes that are genuinely required to complete the task. Therefore, the development of assessment instruments must ensure alignment between ability indicators, cognitive demands, question formats, and the expected evidence of response. The validity of the instrument must also be supported by relevant conceptual analysis and empirical testing, so that the interpretation of assessment results remains consistent with the abilities being measured.

The assessment rubric complements HOTS-based instruments by providing criteria and descriptors for assessing both the process and the outcomes of problem-solving in a targeted manner. Analytical rubrics can be used when an assessment requires detailed information regarding problem understanding, strategy selection, implementation of solutions, and interpretation of results, whilst adapting the assessment dimensions to the characteristics of the task. However, the presence of a rubric does not automatically guarantee the validity and consistency of scoring; clarity of descriptors, the use of sample responses, alignment of assessors' perceptions, and testing of the rubric's quality remain essential. Consequently, assessment instruments and rubrics need to be developed as an integrated assessment system that links target competencies, evidence of student performance, and the basis for scoring. This integration forms the foundation for generating assessment information that not only demonstrates students' achievements but also helps to identify difficulties and determine follow-up learning actions.

Bibliography

Brookhart, S. M. (2013). *How to Create and Use Rubrics for Formative Assessment and Grading*. ASCD.

Brown, T. M. (2018). *Playing to Win: Applying Cognitive Theory and Gamification to Augmented Reality for Enhanced Mathematical Outcomes in Underrepresented Student Populations*. http://hdl.handle.net/10919/97340

Budiman, A., & Jailani, J. (2014). PENGEMBANGAN INSTRUMEN ASESMEN HIGHER ORDER THINKING SKILL (HOTS) PADA MATA PELAJARAN MATEMATIKA SMP KELAS VIII SEMESTER 1. *Jurnal Riset Pendidikan Matematika*, *1*(2), 139–151. https://doi.org/10.21831/jrpm.v1i2.2671

Egodawatte, G. (2010). A Rubric to Self-Assess and Peer-Assess Mathematical Problem Solving Tasks of College Students. *Acta Didactica Napocensia*, *3*(1), 75–88.

Eliyah, E., & Aslan, A. (2025). STAKE'S EVALUATION MODEL: METODE PENELITIAN. *Prosiding Seminar Nasional Indonesia*, *3*(2), Article 2.

Gradini, E. (2022). Development of Rubric of Higher Order Thinking Skills Assessment on Mathematics Learning. *Jurnal Teori Dan Aplikasi Matematika*, *6*(1), 61–74. https://doi.org/10.31764/jtam.v6i1.5342

Ikhlas, I., Aslan, A., & Mutazam, M. (2024). IMPLEMENTASI STRATEGI INQUIRY GURU PAI DALAM MENINGKATKAN CRITICAL THINKING SISWA KELAS VI DENGAN KONSEP HIGHER ORDER THINKING SKILLS (HOTS) DI SDN 06 MENSERE TAHUN PELAJARAN 2023/2024. *ADIBA : JOURNAL OF EDUCATION*, *4*(4), 813–822.

Jonsson, A., & Svingby, G. (2007). The use of scoring rubrics: Reliability, validity and educational consequences. *Educational Research Review*, *2*(2), 130–144. https://doi.org/10.1016/j.edurev.2007.05.002

Krathwohl, D. R. (2002). A Revision of Bloom's Taxonomy: An Overview. *Theory Into Practice*, *41*(4), 212–218. https://doi.org/10.1207/s15430421tip4104_2

Lewis, C. W., & Gilman, S. (2005). *The ethics challenge in public service: A problem-solving guide* (2nd ed). Jossey-Bass.

Melinda, S., Novita, S., Rafiqoh, U. W., Sopariah, S., & Nugraha, D. A. (2024). Pemodelan Matematika dalam Efisiensi Pengisian Ulang Stok Bahan Bakar di SPBU Cikurubuk. *Cartesius : Jurnal Pendidikan Matematika*, 213–218.

Prasetyaningsih, P., Kaniawati, I., Riza, L. S., & Utama, J. A. (2025). The Role of Edutainment in Shaping Modern Education in Indonesia: A Literature Review. *Journal of Mathematics Science and Computer Education*, *5*(2), 171–187. https://doi.org/10.20527/jmscedu.v5i2.16823

Seaman, M. (2011). Bloom's Taxonomy: Its Evolution, Revision, and Use in the Field of Education. Dalam D. J. Flinders & P. B. Uhrmacher (Ed.), *Curriculum and Teaching Dialogue* (hlm. 0). Emerald Publishing Limited. https://doi.org/10.1108/978-1-61735-607-020251004

Sitopu, J. W., Khairani, M., Roza, M., Judijanto, L., & Aslan, A. (2024). THE IMPORTANCE OF INTEGRATING MATHEMATICAL LITERACY IN THE PRIMARY EDUCATION CURRICULUM: A LITERATURE REVIEW. *International Journal of Teaching and Learning*, *2*(1), Article 1.

Walliman, N., & Walliman, N. (2021). *Research Methods: The Basics* (3 ed.). Routledge. https://doi.org/10.4324/9781003141693

Warju, W. (2016). Educational Program Evaluation using CIPP Model. *INVOTEC*, *12*(1). https://doi.org/10.17509/invotec.v12i1.4502

Wiggins, G. (1990). The Case for Authentic Assessment. *Practical Assessment, Research, and Evaluation*, *2*(1). https://doi.org/10.7275/ffb1-mm19

Zapf, P., & Roesch, R. (2008). *Evaluation of Competence to Stand Trial*. Oxford University Press.

Organisation for Economic Co-operation and Development. (2023). *PISA 2022 assessment and analytical framework*. OECD Publishing. https://doi.org/10.1787/dfe0bf9c-en[oecd]