JMKSP (Jurnal Manajemen, Kepemimpinan, dan Supervisi Pendidikan)
Volume 10 (2) 2025, 1209-1218
E-ISSN 2614-8021, P-ISSN 2548-7094

How to Develop Hots Questions Based on Contextual Problems

Dwi Ajeng Safitri[1]*, Sulistyarini[1], Dede Suratman[1]
[1]Universitas Tanjungpura, Indonesia

Corresponding author e-mail: hellobibyofficial@gmail.com*;
sulistyarini@fkip.untan.ac.id; dede.suratman@fkip.untan.ac.id

Article History: Received on September 11, 2025, Revised on October 12, 2025,
Published on October 28, 2025

***Abstract*:** The necessity of Higher Order Thinking Skills (HOTS) in the 21st-century education context is increasingly evident. To assess this construct, it requires an instrument that aligns with HOTS criteria. This study aimed to develop and test HOTS-based question instruments derived from contextual problems in thematic learning for Grade V students at SDN 36 South Pontianak. The research employed the ADDIE development model, but the focus of this paper lies in the tested product results. The final instrument demonstrated excellent feasibility according to expert validation, with an overall average score of 96% (very feasible). Reliability testing yielded a Cronbach's Alpha value of 0.942, indicating very high internal consistency. Furthermore, item difficulty analysis showed that 36 items were in the medium category, making the instrument appropriate for further use in drilling and assessment activities.

***Keywords*:** Contextual, Develeopment, HOTS Questions, HOTS Thinking Skills

A. Introduction

Education is a fundamental human right that must be accessible, inclusive, and equitable for all citizens (Sujasan, S., & Wibowo, 2021). The purpose of education is to produce high-quality human resources capable of developing their potential, knowledge, and skills through meaningful learning experiences. The learning process significantly influences learning achievement, which reflects changes in cognitive, affective, and psychomotor domains. Therefore, an accurate assessment is needed to measure students' learning progress.

According to *Permendikbudristek No. 21 of 2022*, assessment is a process of collecting and processing information to measure students' learning needs and achievements. Commonly, teachers use written tests to identify students' level of understanding of the material (Farida, 2019). However, conventional tests often emphasize memorization rather than analytical or critical thinking.

Developing Higher Order Thinking Skills (HOTS)-oriented assessments is essential to foster students' critical and creative thinking abilities (Gunartha, 2024). HOTS enables students to connect new information with existing knowledge, organize it, and apply it to solve complex problems (Taufik, A., & Arsid, 2020). These skills align with Bloom's revised taxonomy analyzing, evaluating, and creating which are fundamental competencies in 21st-century education.

In Indonesia's 2013 Curriculum, thematic learning is designed to integrate knowledge and experience meaningfully. The contextual learning approach supports this goal by linking concepts with students' real-life situations (Haryono, I., & Hikmah, 2023). Contextual-based learning encourages active participation, collaborative discussion, and the exploration of local issues as learning stimuli (Syaifuddin, Bharata, H., 2017).

However, the implementation of HOTS-oriented learning in Indonesian schools remains limited. Teachers often lack sufficient understanding and skills in developing HOTS based assessments (Hasibuan, et.al, 2022). Interviews with fifth-grade teachers at SDN 36 South Pontianak revealed that existing practice tests primarily measured low-order thinking skills (LOTS), while semester exams included higher-level questions that were not aligned with students' contextual environment. As a result, students struggled to answer and demonstrated low achievement in thematic subjects.

This study introduces an innovative development of *contextual-based HOTS test instruments* that integrate thematic materials and *Core Competencies (Kompetensi Dasar/KD)* specific to the local context of Pontianak Selatan. The novelty lies in the contextualization of HOTS questions with environmental themes familiar to students such as river ecosystem management and local cultural practices so that the test measures analytical and problem-solving skills grounded in real-life situations.

Therefore, the objective of this study is to describe the quality, validity, reliability, and effectiveness of a contextual-based HOTS test instrument for fifth-grade thematic learning at SDN 36 South Pontianak.

B. Methods

This research employed the ADDIE development model (Analysis, Design, Development, Implementation, and Evaluation) as a systematic framework for developing a contextual-based HOTS question instrument. The product testing process was carried out in two stages: a Small-Scale Trial and a Field Trial. The small-scale trial involved 10 fifth-grade students to identify initial readability, clarity, and comprehension of the test items. The Field Trial, involving 20 students, was conducted to obtain empirical data for item validation and reliability testing.

The difference in the number of participants (N=10 vs. N=20) follows the general principle of product development research, where a small-scale trial is intended for preliminary testing, and a larger sample in the field trial provides more stable and representative statistical results for the validity and reliability analyses. Data were collected through interviews (to identify students' comprehension difficulties) and questionnaires/tests (to gather empirical responses for item analysis). Item analysis in this study was conducted using the Classical Test Theory (CTT) approach with the Pearson Product Moment Correlation to determine item validity. The formula used is as follows:

$$r_{xy} = \frac{N\sum XY - (\sum X)(\sum Y)}{\sqrt{[N\sum X^2 - (\sum X)^2][N\sum Y^2 - (\sum Y)^2]}}$$

where:

- r_{xy} = correlation coefficient of item validity
- X = item score
- Y = total test score
- N = number of respondents

An item is considered valid if the calculated r_{count} (Pearson correlation coefficient) > r_{table} at a significance level of 0.05. The reliability of the test items was analyzed using Cronbach's Alpha through SPSS version 26. The interpretation of reliability levels refers to Guilford's classification (Okayana, 2019), as shown in Table 1.

Table 1. Item Reliability Classification (Guilford, cited in Okayana, 2019)

No	Reliability Range	Category
1	$r_{11} < 0.20$	Very Low
2	$0.20 \leq r_{11} < 0.40$	Low
3	$0.40 \leq r_{11} < 0.70$	Moderate
4	$0.70 \leq r_{11} < 0.90$	High
5	$0.90 \leq r_{11} < 1.00$	Very High

In addition, the **item difficulty level** was analyzed using SPSS based on the proportion of correct responses. The criteria for interpreting item difficulty are presented in Table 2.

Table 2. Item Difficulty Level Criteria

No	Difficulty Index Range	Category
1	0.71 – 1.00	Easy
2	0.31 – 0.70	Medium
3	0.00 – 0.30	Difficult

C. Results and Discussion

Results

To determine the relevant HOTS-based question standards for students, a question requirements analysis was conducted. The results of the interviews to obtain information related to requirements are as follows.

Tabel 2. Jawaban Wawancara Bersama Guru Kelas VB

No	Daftar Pertanyaan
1.	Jumlah peserta didik Kelas VB ada 30 orang
2.	Proses belajar mengajar mengikuti Kurikulum 2013
3.	KKM yang ditetapkan di kelas V adalah 70 untuk tematik, 75 untuk muatan pelajaran Pendidikan Agama dan PJOK, serta 65 untuk pelajaran MTK.
4.	Beberapa butir soal *HOTS* sudah ada di soal PTS dan PAS namun tidak

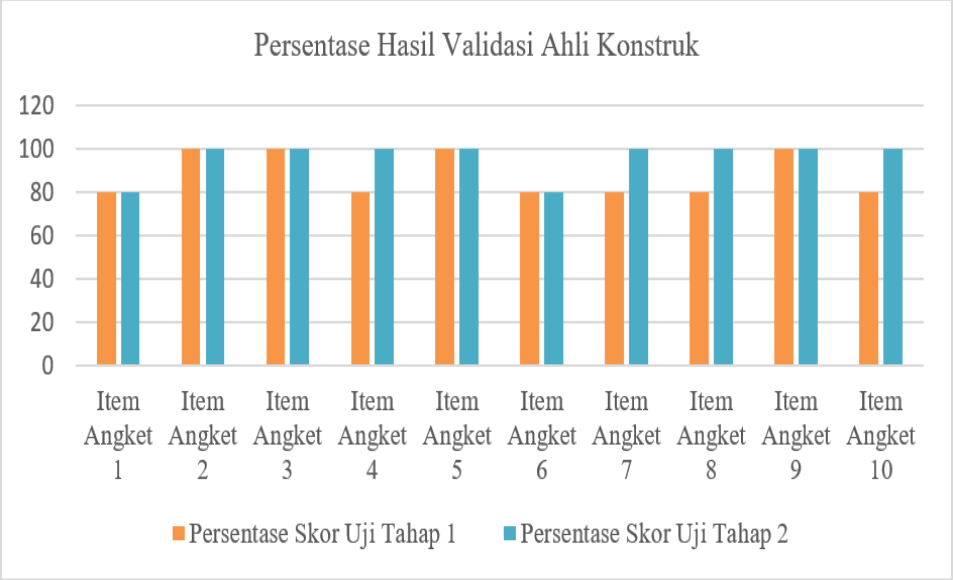
No	Daftar Pertanyaan
	pernah diberikan pada saat penilaian harian
5.	Guru kelas VB belum bisa memastikan semua kelas sudah menerapkan soal *HOTS* pada pembelajaran tematik, tetapi di bahan ajar dijumpai beberapa butir soal *HOTS*
6.	Guru kelas VB belum pernah menerapkan soal HOTS secara langsung ke peserta didik pada saat melakukan penilaian harian, namun peserta didik pernah mengerjakan soal HOTS pada saat PTS dan PAS
7.	Soal Latihan yang dibuat oleh guru kelas VB belum menggunakan indikator HOTS dan hanya mengukur kemampuan berpikir pada tingkatan dasar dan menghafal atau dikenal dengan *low-order thinking skills (LOTS)*
8.	Guru kelas VB memberikan soal latihan yang sudah ada di buku LKS, namun beberapa kali pernah membuat soal latihan sendiri
9.	Guru kelas VB mengalami kesulitan ketika menyusun soal HOTS dan memerlukan pelatihan khusus dalam membuat soal HOTS.
10.	Kelebihan soal HOTS menurut guru yaitu anak akan terlatih dan terukur dalam mengembangkan ide kreatifitasnya dan menerima ragam jenis informasi, dan kreatif dalam memikirkan pemecahan masalah
11.	Menurut guru kelas VB perlu dikembangkan soal HOTS agar anak dari dini sudah berlatih untuk berpikir kritis sehingga akan melahirkan peserta didik siswi yang cerdas, pintar dan kritis. Selain itu soal Latihan berbasis HOTS perlu dikembangkan agar anak terbiasa mengerjakan soal berbasis HOTS

In addition, curriculum analysis was also conducted to ensure that the questions would be relevant to the applicable curriculum. At this stage, interviews were also conducted to determine the learning resources used, the extent to which the learning material had been delivered in class, and the questions that had been used in learning to focus on basic skills that could be used in compiling HOTS questions, indicators and learning objectives to be achieved.

The design process begins by establishing a theme for thematic learning, namely theme 5, 'Ecosystems'. Next, an analysis of the KD is conducted to determine whether the KD contains actual, theoretical, systematic, or metacognitive knowledge. If a KD only contains actual knowledge, then HOTS questions cannot be used to measure that competency. After analysing the basic competencies, a grid of HOTS questions is compiled, taking into account the HOTS indicators of analysis, evaluation, and creation.

The product development stage is the stage at which HOTS questions based on contextual problems are compiled from the description of question indicators contained in the grid. After the questions have been compiled, answer options that are homogeneous, logical and have only one correct answer are created. In developing the questions, attention is also paid to creating engaging and contextual stimuli in the form of images, texts, case studies, tables, or data sets.

Once the questions have been developed, the answer keys and assessment rubrics are then developed. The scoring guidelines used are as follows: if the answer is correct, it receives a score of 1; if the answer is incorrect or not answered, it receives a score of 0.
In this development, a validation test or error correction was conducted with the assistance of a questionnaire. The testing was carried out in two stages with the answer scores shown in the following graph.

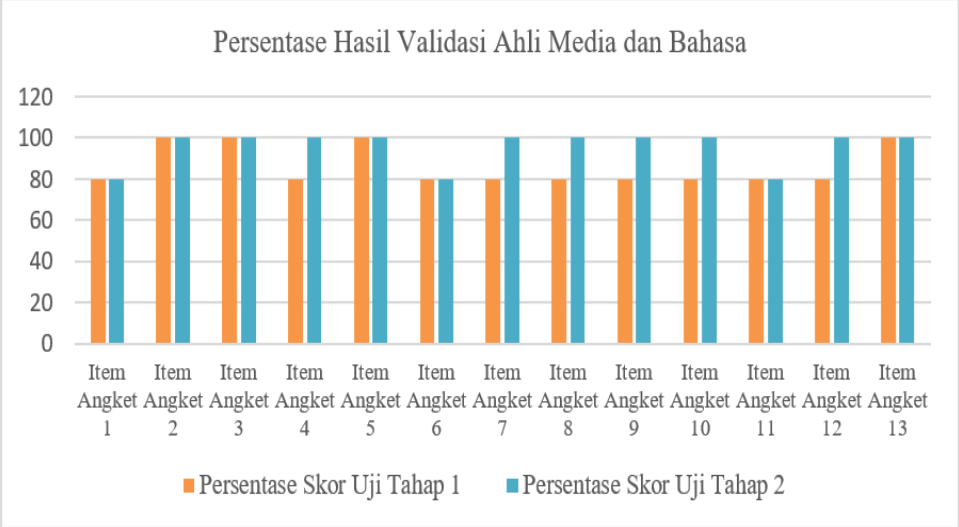


Gambar 1. Persentase Hasil Validasi Ahli Konstruk

After obtaining the results of the first stage of feasibility testing from construction experts, HOTS-based questions were declared to have high feasibility with an average percentage of 88% and were deemed very feasible. Researchers obtained several recommendations to improve the quality of the questions, namely: 1) Include sources for each image, 2) Include sources in the reading text, 3) If you want to break up a paragraph because it is too long, do not just take a sentence from it. Try to add sentences to make it a little longer or write the paragraph in its entirety, 4) Make the answer choices or options more diverse and evenly distributed so that the answer options are homogeneous

After making improvements based on the suggestions and critiques from the experts, a second-stage construct validity test was conducted, and the results indicated that the questions had a high level of validity with an average percentage of 96%, meeting the criteria for 'very valid.' With such results, the questions were deemed suitable for pilot testing.

The first material and language feasibility test was conducted by material and language experts with the help of a questionnaire. The testing was carried out in two stages with the answer scores shown in the following graph.



Gambar 2. Persentase Hasil Validasi Ahli Materi dan Bahasa

After obtaining the results of the first stage of feasibility testing from subject matter and language experts, the questions were declared to be highly feasible with an average percentage of 88% and very feasible criteria. The experts made several recommendations to improve the quality of the questions, including:
1) The questions were almost relevant to the rules for compiling HOTS questions. However, the wording of the questions needed to be reviewed so that students would have to think first. Therefore, the questions were made longer.
2) The distribution of distractors needed to be reviewed so that the correct and incorrect options were not too obvious.

After making improvements based on the suggestions and critiques from the experts, a second stage of material and language validation testing was conducted, the results of which showed that the questions were deemed highly feasible with an average percentage of 95% and met the criteria for being highly feasible. Therefore, the product was deemed feasible for testing.

The product testing process was carried out in two stages: a Small-Scale Trial and a Field Trial. The Small-Scale Trial involved 10 students, serving as an initial pilot to review the readability, clarity, and contextual relevance of the questions. The Field Trial involved 20 students, representing a larger and more reliable sample for statistical analysis. The difference in the number of respondents (N = 10 and N = 20) affected the *r* table value used for the validity test, which was 0.632 in the first trial and 0.444 in the second. A larger number of participants decreases the *r* table threshold, thus increasing the stability and generalizability of the statistical results. Therefore, the analysis in this section focuses on the final product results obtained from the Field Trial (N = 20).

Based on the Field Trial results, out of the 50 developed questions, 36 items were declared valid because their *r* calculated values exceeded the *r* table value (0.444). The reliability analysis using Cronbach's Alpha resulted in a coefficient of $\alpha = 0.942$, categorized as very high, indicating strong internal consistency among the items. These

results demonstrate that the developed HOTS-based instrument can consistently measure the intended construct of higher-order thinking skills.

In addition to the validity and reliability analysis, the item difficulty level was also calculated. The findings showed that 13 items (26%) were categorized as *easy*, 36 items (72%) as *medium*, and 1 item (2%) as *difficult*. The predominance of medium-difficulty items indicates that the developed instrument is balanced in measuring students' abilities it is neither too easy nor too difficult, and therefore appropriate for assessing elementary students' higher-order thinking skills.

The results of the expert validation also confirmed that the instrument was feasible for use in the classroom. The construct validation achieved an average score of 96% (very feasible), while the material and language validation obtained 95% (very feasible). These high percentages indicate that the instrument's content, structure, and language quality meet the standards required for HOTS based assessments.

Overall, the final product of this research consists of 36 valid, reliable, and feasible HOTS based multiple-choice questions. The instrument shows a high level of consistency (α = 0.942) and balanced item difficulty distribution, making it suitable for use in learning evaluations, formative assessments, and drilling exercises for Grade V students in thematic learning contexts.

Discussion

The purpose of this study was to develop and test a Higher Order Thinking Skills (HOTS)-based question instrument derived from contextual problems in thematic learning for Grade V elementary school students. The discussion focuses on interpreting the results of expert validation, reliability testing, and item analysis, as well as connecting these findings with relevant theories and previous studies.

The expert validation results revealed that the developed HOTS-based instrument obtained a very high feasibility score of 96%, indicating that the instrument met the standards of validity in terms of content, construct, and language. This finding signifies that the developed questions accurately represent the cognitive processes defined within Bloom's revised taxonomy namely, analyzing, evaluating, and creating(Oktaviana & Susiaty, 2020). The high validation result also demonstrates that the contextual problems embedded in the items were relevant to the thematic learning materials, thus allowing students to connect academic knowledge with real-life situations (Oktaviana & Susiaty, 2020). Such integration aligns with the principles of contextual teaching and learning, which emphasize learning through meaningful experiences that foster critical and reflective thinking (Fitri Ana et al., 2025) .

Furthermore, the instrument achieved a Cronbach's Alpha coefficient of 0.942, indicating an excellent level of reliability. According to Saw et al.,(2025), a reliability value exceeding 0.9 demonstrates very high internal consistency among items within a test. This means that the questions measure the same construct consistently across different conditions, suggesting that the contextual approach used in this study successfully generated coherent items. The high reliability value also implies that students' responses

were stable and dependable, reflecting the clarity of the question wording, the balance in item structure, and the appropriateness of the contextual scenarios used. These results are consistent with research by Jiang & Zhang, (2025), who emphasized that integrating real-world contexts in assessment can increase both the reliability and interpretability of students' cognitive outcomes.

The results of item analysis further support the robustness of the developed instrument. The distribution of difficulty levels showed that 36 items were categorized as "medium," meaning that the instrument was neither too easy nor too difficult for the target learners. In educational measurement, a balanced proportion of moderate-level items is desirable because it allows for a fair discrimination of students' abilities (Baudin, 2025). The predominance of medium-level items in this study indicates that the test can effectively measure variations in students' higher-order thinking without causing excessive cognitive load or demotivation. Moreover, this balance supports the cognitive development of elementary students, who are in the transition phase from concrete to abstract thinking, as described by Piaget's theory of cognitive development.

Another significant aspect of the findings is the effectiveness of contextual-based assessment in promoting student engagement and deeper thinking. Contextual problems presented in the instrument encouraged students to analyze, compare, and make decisions based on real-life scenarios familiar to them. This aligns with the constructivist view that knowledge is best acquired when learners are actively involved in constructing meaning through authentic experiences (Vygotsky, 1978). When students are able to relate test items to their environment and prior experiences, their motivation and cognitive processing improve, resulting in more meaningful learning outcomes (Akram & Abdelrady, 2025). In this sense, the developed HOTS instrument not only serves as a measurement tool but also as a learning stimulus that enhances metacognitive awareness.

From a pedagogical perspective, the integration of contextual problems into HOTS assessment also supports the implementation of the 2013 Curriculum in Indonesia, which emphasizes thematic and competency-based learning. By aligning the assessment with real-world contexts, teachers can evaluate not only students' cognitive mastery but also their ability to apply knowledge, reason critically, and solve problems creatively. This outcome corresponds with the goals of 21st-century education, which prioritizes critical thinking, problem-solving, and collaboration as essential competencies for lifelong learning (Malviya, 2024) .

Furthermore, the findings highlight the importance of developing localized assessment instruments tailored to specific learning contexts. The contextual materials used in this study were adapted from students' daily environments in Pontianak, ensuring cultural and linguistic relevance. Such localization is essential to enhance fairness and inclusivity in assessment, particularly in diverse educational settings like Indonesia. Instruments developed without considering local contexts often fail to capture students' actual abilities because they present abstract or unfamiliar problems. Therefore, this study contributes to the broader effort of creating culturally responsive assessment tools that bridge the gap between global competencies and local realities.

Overall, the results demonstrate that the developed contextual HOTS-based instrument is theoretically sound, empirically valid, and practically useful. It can serve multiple purposes: as a diagnostic tool to identify students' levels of higher-order thinking, as a formative assessment to guide instructional improvement, and as a model for teachers who wish to integrate contextual elements into their evaluation strategies. Future studies could expand on this work by testing the instrument in different regions, subjects, and grade levels, as well as exploring digital adaptations that align with current trends in educational technology (Rosani et al., 2025; Yusmiono et al., 2020).

In conclusion, the discussion confirms that developing HOTS based instruments through contextual problem design is not only feasible but also highly effective in assessing and fostering students' higher-order thinking skills in thematic learning. This finding reinforces the notion that assessment is an integral part of the learning process one that, when designed meaningfully, can both measure and cultivate students' intellectual growth.

D. Conclusions

Based on the research findings, it can be concluded that the HOTS-based questions developed through the ADDIE model successfully met the criteria of context-based assessment instruments. The items are aligned with contemporary environmental and social issues that are relevant to students' daily experiences, making them suitable for measuring cognitive processes at levels C4 (analyzing), C5 (evaluating), and C6 (creating). This alignment enables the instrument to effectively assess and foster students' higher-order thinking skills within the framework of 21st-century learning, particularly critical thinking, problem-solving, and creativity.

Moreover, the use of contextual problems within the questions encourages students to connect classroom knowledge with real-world situations. This approach not only supports deeper understanding and reflection but also promotes the development of essential life skills such as reasoning, decision-making, and collaboration. Thus, the instrument has practical value as both an assessment and a learning tool for teachers aiming to cultivate students' higher cognitive abilities in thematic learning.

For future research, it is recommended to conduct further testing to examine the effectiveness of the developed HOTS-based questions on students' learning outcomes and cognitive growth. Expanding the implementation to different subjects, grade levels, and digital formats would also provide broader evidence of the instrument's applicability and its potential contribution to enhancing meaningful learning in diverse educational contexts.

E. Acknowledgement

We would like to thank you for to the supervising lecturer who has provided direction and guidance during the research process, as well as to the Universitas Tanjungpura who has provided permission and facilities. The author also thanks all participants who have been willing to be respondents in this study.

References

Akram, H., & Abdelrady, A. H. (2025). Examining the role of ClassPoint tool in shaping EFL students' perceived E-learning experiences: A social cognitive theory perspective. *Acta Psychologica*, *254*(February), 104775. https://doi.org/10.1016/j.actpsy.2025.104775

Baudin, J. S. P. (2025). *Assessing the psychometric properties of AI-generated multiple-choice exams in a psychology subject*. *7*(3).

Farida, I. (2019). *Evaluasi Pembelajaran Berdasarkan Kurikulum Nasional (E. Kuswandi (ed.)*. Remaja Rosdakarya.

Fitri Ana, M., Syahri, M., & Tinus, A. (2025). The Contextual Teaching and Learning Using Media for Student Character Formation. *Academia Open*, *10*(1), 1–16. https://doi.org/10.21070/acopen.10.2025.11077

Gunartha, I. W. (2024). engembangan Penilaian Berorientasi Hots: Upaya Peningkatan Kemampuan Berpikir Kritis Siswa Di Era Global Abad Ke-21. *Widyadari, 25*(1), 133–147.

Haryono, I., & Hikmah, K. (2023). The Application Of The Contextual Teaching And Learning (CTL) Model In Arabic Language Learning To Improve The Learning Outcomes. *Buana Pendidikan Jurnal Fakultas Keguruan Dan Ilmu Pendidikan*, *19*(1), 45–60.

Hasibuan, M., Damayanti, R., & A. (2022). Upaya Peningkatan Pemahaman Pada Mata Pelajaran Fiqih Melalui Model Pembelajaran Student Teams Achievement Divisions Di Kelas VIII MTS Negeri 2 Langkat. *Ability: Journal of Education And Social Analysis*, *3*(2), 140–150.

Jiang, Z., & Zhang, Z. (2025). From black box to transparency: Enhancing automated interpreting assessment with explainable AI in college classrooms. *Research Methods in Applied Linguistics*, *4*(3). https://doi.org/10.1016/j.rmal.2025.100237

Malviya, S. B. (2024). 21st Century Skills in Education: A Review of Frameworks and Implementation. *National Education Policy 2020- The Key To Development In India (Volume-1)*, *1*, 45–55.

Oktaviana, D., & Susiaty, U. D. (2020). Development of Test Instruments Based on Revision of Bloom's Taxonomy to Measure the Students' Higher Order Thinking Skills. *JIPM (Jurnal Ilmiah Pendidikan Matematika)*, *9*(1), 21. https://doi.org/10.25273/jipm.v9i1.5638

Rosani, M., Lestari, N. D., & Valianti, R. M. (2025). Transformation of Education to Welcome the Golden Generation of Indonesia 2045. *JMKSP (Jurnal Manajemen, Kepemimpinan, dan Supervisi Pendidikan)*, *10*(1), 407-427. https://doi.org/10.31851/jmksp.v10i1.18895

Saw, Z. K., Yuen, J. J. X., Ashari, A., Bahemia, F. I., Low, Y. X., Mustapha, N. M. N., & Lau, M. N. (2025). Forward-backward translation, content validity, face validity, construct validity, criterion validity, test-retest reliability, and internal consistency of a questionnaire on patient acceptance of orthodontic retainer. *PLoS ONE*, *20*(1 January), 1–14. https://doi.org/10.1371/journal.pone.0314853

Sujasan, S., & Wibowo, U. B. (2021). The survival of school financing management in COVID-19 pandemic. *Journal of Education and Learning (EduLearn)*, *15*(4), 563–570.

Syaifuddin, Bharata, H., & C. (2017). Jurnal Pendidikan Matematika Universitas Lampung Pengembangan LKPD Berbasis Kontekstual untuk Meningkatkan Kemampuan Pemecahan Masalah dan Self-Efficacy Matematis. *Jurnal Pendidikan Matematika Universitas Lampung, 5*, *11*.

Taufik, A., & Arsid, I. (2020). Kemampuan pemecahan masalah matematis siswa dalam menyelesaikan soal hots. *Jurnal Pendidikan Matematika*, *4*(2), 581–589.

Yusmiono, B. A., Lestari, N. D., & Januardi, J. (2020). Analysis of Social Economic Life of Transmigrant Communities on Musi River Banks Muara Medak Village. *International Journal Of Scientific & Technology Research*, *9*(3), 1340–1345. https://www.ijstr.org/final-print/mar2020/Analysis-Of-Social-Economic-Life-Of-Transmigrant-Communities-On-Musi-River-Banks-Muara-Medak-Village.pdf.