JURNAL ILMIAH PENDIDIKAN PROFESI GURU
Volume 7 Nomor 2 2024, pp 393-403
E-ISSN: 2621-5705; P-ISSN: 2621-5713
DOI: https://doi.org/10.23887/jippg.v7i2.79101

Revolutionizing Science Education Evaluation Using a Vision Language Model of Effective Assessment and Supervision

Irfan Ananda Ismail[1], Mawardi Mawardi[2], Qadriati[3], Munadia Humane[4], Khairil Arif[5*]

[1,2,5] Department of Postgraduate Studies, Padang State University, Padang, Indonesia
[3] SMP N 32 Padang, Padang, Indonesia
[4] Department of Postgraduate Studies, Environmental Science, Padang State University, Padang, Indonesia
*Corresponding author: khairilarif@fmipa.unp.ac.id

Abstrak

Penerapan Kurikulum Merdeka di Indonesia menghadirkan tantangan dalam penilaian pendidikan, terutama dalam mengevaluasi respons esai siswa secara efisien. Tujuan dari penelitian ini adalah untuk mengembangkan dan menguji model evaluasi pendidikan berbasis teknologi kecerdasan buatan (AI), khususnya model bahasa visi, yang dapat mengotomatisasi dan meningkatkan proses penilaian hasil belajar siswa dalam bidang sains. Penelitian ini mengeksplorasi potensi Vision Language Models (VLMs) sebagai solusi inovatif. Penelitian ini menggunakan metode campuran sekuensial eksplanatif. Subjek dalam penelitian ini adalah siswa SMP. Metode pengumpulan data menggunakan wawancara. Instrumen pengumpulan data dengan lembar kuesioner. teknik analisis data menggunakan analisis deskriptif kualitatif, kuantitatif dan statistik inferensial. Hasil penelitian yaitu integrasi VLMs meningkatkan efisiensi dan objektivitas penilaian. Penelitian ini menyimpulkan bahwa VLMs dapat mengurangi beban kerja guru dan meningkatkan umpan balik, serta menunjukkan sinergi antara teknologi dan reformasi kurikulum di era Kurikulum Merdeka.Implikasi penelitian ini sangat signifikan terhadap perkembangan pendidikan sains dan teknologi, penggunaan model bahasa visi (Vision-Language Models) dalam evaluasi pendidikan sains dapat meningkatkan akurasi dan objektivitas dalam penilaian hasil belajar siswa.

Kata Kunci: ***V****ision Language Models, Kurikulum Merdeka, Teaching at the Right Level, Praktik baik.*

Abstract

Implementing the Independent Curriculum in Indonesia presents challenges in educational assessment, especially in efficiently evaluating students' essay responses. This study aims to develop and test an educational evaluation model based on artificial intelligence (AI) technology, especially the vision language model, which can automate and improve the process of assessing student learning outcomes in science. This study explores the potential of Vision Language Models (VLMs) as an innovative solution. This study uses a mixed sequential explanatory method. The subjects in this study were junior high school students. The data collection method used interviews. Data collection instruments with questionnaires. Data analysis techniques used were qualitative, quantitative, descriptive analysis, and inferential statistics. The study results are that integrating VLMs increases the efficiency and objectivity of assessment. This study concludes that VLMs can reduce teacher workload, improve feedback, and show synergy between technology and curriculum reform in the Independent Curriculum era. The implications of this study are very significant for the development of science and technology education; the use of vision language models (Vision-Language Models) in evaluating science education can increase the accuracy and objectivity in assessing student learning outcomes.

Keywords: Vision Language Models, Independent Curriculum, Teaching at the Right Level, Good Practices

1. INTRODUCTION

Rapid technological advances have brought significant changes in various aspects of life, including education. Although the implementation of the Merdeka Curriculum in

History:
Received : May 27, 2024
Accepted : August 10, 2024
Published : August 25, 2024

Publisher: Undiksha Press

Indonesia aims to develop students' higher-order thinking skills and the application of knowledge in the real world, there is a significant gap between these expectations and current assessment practices. Traditional evaluation methods, especially essay-based assessments, are still time-consuming, subjective, and prone to inconsistencies, especially when dealing with large numbers of students. This gap creates an urgent need for innovative solutions to improve the efficiency and objectivity of student evaluations. In recent years, the integration of artificial intelligence (AI) in educational practices has received increasing attention, especially in the domain of student assessment. One of the most promising applications of AI in education is the use of Vision Language Models (VLM) to improve the efficiency and objectivity of evaluating student work, particularly in the context of essay-based assessments. VLM is a class of AI models that combines computer vision and natural language processing capabilities to understand and interpret visual and textual information simultaneously. These models have demonstrated excellent performance in various tasks, such as image captioning, visual question answering, and text-to-image generation (Braun & Clarke, 2021; Chen et al., 2021)

VLM's potential to revolutionize educational assessment lies in its ability to analyze and evaluate student responses in a more comprehensive and unbiased manner, thereby easing the burden on teachers and improving the overall quality of assessment. In Indonesia, the implementation of the Independent Curriculum (Kurikulum Mandiri) places greater emphasis on developing students' higher-level thinking abilities and the application of knowledge in real-world contexts. This educational paradigm shift requires a corresponding change in assessment practices, moving away from traditional multiple-choice tests toward more authentic, performance-based assessments, such as essays and open-ended questions. However, manual assessment of such assessments can be time-consuming, subjective, and prone to inconsistencies, especially when dealing with large numbers of students (Belenggu et al., 2020; Tamu et al., 2021).

Recent research developments in the fields of AI and education have shown promising results in addressing these challenges. Research has explored the integration of AI in science learning activities, focusing on methods such as discovery learning, level-appropriate teaching, and guided inquiry-based flipped classroom systems using learning management systems for specific subjects. Additionally, efforts have been made to apply AI in environmental analysis and laboratory practices, highlighting the growing intersection between technology and various educational fields. The development of AI-powered assessment tools, particularly those using VLM, is a significant step in bridging the gap between the aspirations of modern educational curricula and the practical limitations of traditional assessment methods. These advances offer the potential to provide more accurate, efficient, and fair evaluations of student performance, aligned with the goals of developing critical thinking and real-world application of knowledge.

Similar research regarding the application of AI in educational assessment has been carried out by several researchers. Previous research findings investigated the use of machine learning algorithms for automatic analysis of written explanations in biology. The results showed that the AI-based system achieved performance comparable to human experts in evaluating student responses, thereby demonstrating the potential of automated assessment in science education (Guetterman et al., 2020). In another study, explored the application of natural language processing techniques for automated scoring of open-ended physics questions. Their findings suggest that AI-powered systems can evaluate student responses accurately and provide timely feedback, potentially reducing instructor workload while maintaining assessment quality (Ivankova, 2019). Another similar finding examined the use of multimodal learning analytics in assessing students' problem-solving skills in chemistry. Their research shows that AI-based tools can effectively analyze textual and visual data from

student responses, providing a more comprehensive evaluation of student understanding and problem-solving strategies (Johnson et al., 2023).

These studies highlight the growing potential of AI in educational assessment across disciplines, demonstrating increased efficiency, accuracy, and scalability in evaluating student performance. As research advances in this area, the integration of AI in educational assessment promises to revolutionize the way we evaluate student learning and support the goals of innovative curricula such as the Merdeka Curriculum in Indonesia. SMP N 32 Padang, a junior high school in Padang, Indonesia, has recognized the challenges associated with traditional essay assessment methods and has taken a proactive approach to finding innovative solutions. The school has collaborated with researchers from Padang State University to investigate the potential of VLM in improving the efficiency and objectivity of essay assessment, with the aim of aligning assessment practices with the principles of the Independent Curriculum. This research aims to present a case study of the application of VLM for essay assessment at SMP N 32 Padang, highlighting the opportunities and challenges associated with this innovative approach. By examining the development process, effectiveness, and user perceptions of VLM-based assessment systems, this research seeks to contribute to the growing body of knowledge about the application of AI in education and provide valuable insights for educators and policymakers interested in leveraging the technology to improve assessment practices.

This research builds on existing research on the application of AI in education, particularly in the domain of student assessment. Previous research has highlighted the potential of machine learning algorithms, such as neural networks, in automating essay scoring and classifying text documents (Kallio et al., 2023; Kim, 2021). However, these approaches often rely on handcrafted features and are less able to capture the semantic meaning and context of text. The emergence of Vision Language Models (VLM) has opened new possibilities for improving educational assessment. VLMs, like CLIP have demonstrated excellent performance in a variety of vision language tasks, including image captioning, visual question answering, and text-to-image retrieval (Lee et al., 2022; Ouhaichi et al., 2023). These models are trained on large-scale datasets of image-text pairs, allowing them to learn rich representations that capture semantic alignment between visual and textual information. In the context of educational assessment, VLM offers several advantages over traditional approaches. First, VLM can analyze and interpret the visual and textual components of student responses, such as diagrams, graphs, and written explanations. This is particularly relevant for subjects such as science and mathematics, where visual representation plays an important role in demonstrating understanding. Second, VLM can provide a more comprehensive evaluation of students' responses by considering semantic meaning and context, rather than relying solely on keyword matching or surface-level features (Mertens & Hesse-Biber, 2018; Mogadala et al., 2020).

Although VLM has promising potential in educational assessment, its real-world application remains unexplored. The majority of existing research focuses on evaluating the performance of VLM on reference data sets, rather than investigating its effectiveness and feasibility in real educational contexts. Additionally, the integration of VLMs into assessment practices raises important questions regarding alignment with curriculum standards, development of appropriate rubrics, and educators' acceptance and trust in AI-based assessment tools. This research aims to address this gap by presenting a case study of the application of VLM for essay assessment at SMP N 32 Padang, Indonesia. The research began with the development and testing of a VLM-based assessment system, followed by a quantitative evaluation of its performance compared to manual assessment. Next, qualitative data was collected through surveys and interviews with teachers to gain insight into their perceptions and experiences with the VLM system. The significance of this research lies in

its potential to contribute to the advancement of educational assessment practices in the Independent Curriculum era (Morgan, 2020; Moskal & Leydens, 2022). By exploring the synergies between VLM and Independent Curriculum principles, such as the emphasis on authentic assessment and the TaRL approach, this research seeks to provide best practice examples of how modern learning technologies can be leveraged to support meaningful learning and assessment. It is hoped that the findings of this research will provide information to educators, policy makers and researchers who are interested in harnessing the power of AI to improve the quality and equity of education in Indonesia and beyond. Based on the background and findings of this research, this article aims to explore more deeply the potential of the Vision Language Model (VLM) in revolutionizing educational assessment practices, especially in the context of implementing the Independent Curriculum in Indonesia.

2. METHOD

This research uses an explanatory sequential mixed methods design to explore the potential of the Vision Language Model (VLM) in increasing the efficiency and objectivity of essay assessment at SMP N 32 Padang (Nehm et al., 2022; Schober et al., 2022).. This design was chosen because of its ability to provide a comprehensive understanding of the development, effectiveness and user perceptions of the VLM system in supporting the implementation of the Independent Curriculum and the Teaching at the Right Level (TaRL) approach.

This research consists of two stages: quantitative followed by qualitative. In the first stage, quantitative data was collected and analyzed to test the effectiveness of the VLM system in assessing essay responses, providing empirical evidence of the system's performance compared to manual assessment by teachers. Second, the qualitative phase explores teachers' perceptions, experiences and insights regarding the VLM system, based on quantitative results (Schoonenboom & Johnson, 2023; Vyas et al., 2020). The integration of quantitative and qualitative methods allows for a different understanding of the phenomenon under investigation. The quantitative phase provides objective measurements of the effectiveness of the VLM system, while the qualitative phase offers rich contextual information about teachers' subjective experiences and opinions. This mixed methods approach produces a holistic picture of the potential of VLM in improving essay assessment practices at SMP N 32 Padang.

An explanatory sequential mixed methods design is very suitable for this research because of its pragmatic nature, addressing the need for objective evaluation of VLM system performance and subjective exploration of teacher perceptions (Nabhan et al., 2023; Prawiyogi et al., 2021). The sequential nature of the design allows researchers to use quantitative findings to guide qualitative inquiry, ensuring that the qualitative phase builds on and builds on the initial quantitative results. By combining the strengths of quantitative and qualitative methods, this research seeks to produce evidence that can be generalized and contextually relevant, thereby increasing the potential of the findings to provide information in decision making and encourage educational innovation at SMP N 32 Padang and its surroundings (Grace et al., 2023). The research procedures are presented in Table 1.

Table 1. Research Procedures

Phase	Activity
Quantitative	1. Develop and validate VLM systems 2. Collected student essay responses 3. Essay evaluation using the VLM system and manual evaluation 4. Carrying out statistical analysis of assessment results

Phase	Activity
Qualitative	1. Participants were selected based on quantitative results 2. Conduct semi-structured interviews with teachers 3. Conduct thematic analysis of interview data
Integration	1. Synthesized results from both phases to draw comprehensive conclusions

This research design allows for a thorough investigation of the effectiveness of the VLM system and its potential impact on essay assessment practices at SMP N 32 Padang. The combination of quantitative and qualitative methods provides a powerful framework for addressing the complex nature of integrating AI-powered assessment tools in educational settings. Participants in this research were teachers from three subjects: Natural Sciences (IPA), Social Sciences (IPS), and Information and Communication Technology (ICT). This course was chosen because it often involves essay-based assessments and requires teachers to evaluate a variety of student responses. A total of 4 teachers, one from each subject, were purposively selected to participate in this research based on their experience in teaching their respective subjects and their willingness to engage with the VLM system. The selection of participants from different fields of study allows for a more comprehensive understanding of the potential of VLM systems across various disciplines. By involving teachers with diverse backgrounds and experiences, this research aims to capture a variety of perspectives and insights regarding the system's effectiveness, usability, and acceptability.

Purposive sampling was conducted to ensure that participants had the knowledge and experience necessary to provide rich and relevant data for the study. Teachers' familiarity with essay-based assessment and their direct involvement in the implementation of the Independent Curriculum make them well positioned to offer valuable insights into the potential of VLMs in improving assessment practices. A sample size of 4 teachers is considered appropriate for an explanatory sequential mixed methods design, because it allows for a balance between depth and breadth of data collection(Aprilia & Madiun, 2024; Rukmana et al., 2023). The quantitative phase requires a sufficient number of participants to produce reliable and valid results, whereas the qualitative phase aims to saturate the data, where no new themes or insights emerge from additional interviews.

The data collection process in this study was divided into two stages, following an explanatory sequential mixed methods design (Muhali, 2018; Prawiyogi et al., 2021).. In the first stage, quantitative data was collected to evaluate the effectiveness of the VLM system in assessing essay responses. In the qualitative phase, data was collected through semi-structured interviews with 4 participant teachers. The interviews aimed to explore teachers' perceptions, experiences and insights regarding the VLM system and its potential to improve essay assessment practices. A semi-structured interview guide was developed based on the quantitative findings and research questions. The guide includes open-ended questions covering topics such as the usability and effectiveness of the VLM system, its impact on teacher workload and assessment practices, and alignment with the Independent Curriculum and TaRL approaches.

Data analysis in this study followed an explanatory sequential mixed methods design, with quantitative and qualitative data analyzed separately and then integrated to provide a comprehensive understanding of the research problem (Kim, 2021; Prawiyogi et al., 2021). Quantitative data consisting of manual and automated essay scores were analyzed using descriptive and inferential statistics. Descriptive statistics, including means, standard deviations, and ranges, were calculated for manual and automated scores to provide an overview of the distribution of the data. Qualitative data consisting of interview transcripts were analyzed using thematic analysis. This approach involves a systematic process of coding data, identifying patterns and themes, and interpreting findings based on the research

questions and theoretical framework. Qualitative data consisting of interview transcripts were analyzed using thematic analysis. This approach involves a systematic process of coding data, identifying patterns and themes, and interpreting findings based on research questions and theoretical frameworks. Thematic analysis follows six phases: data introduction, initial coding, theme search, theme review, theme definition and naming, and report generation.

3. RESULTS AND DISCUSSION

Results

In the quantitative stage, we conducted a comprehensive analysis of the effectiveness of Vision Language Models (VLMs) in assessing student essays at SMP N 32 Padang. Our dataset consists of 500 student essays, which were evaluated using a VLM system and traditional manual grading by teachers. Key metrics for comparison include scoring accuracy, consistency, and time efficiency. The data collected focuses on three main aspects. First, we measure the accuracy of the assessment by comparing the grades produced by the VLM with the grades given by the teacher. Second, we assessed the consistency of grades assigned by the VLM system across essays. Finally, we evaluated time efficiency by comparing the time required by the VLM system to grade essays with the time required for manual grading.

Our analysis provides promising results across all three metrics. In terms of assessment accuracy, the VLM system demonstrated a high level of performance, reaching an accuracy rate of 92% when compared with the grades given by teachers. This shows that AI-powered systems are capable of replicating the assessment standards of experienced educators. The VLM system also shows impressive consistency in its scoring. We observed a standard deviation of 1.5 points across some essays, indicating that the system maintained a stable grading approach across student submissions. This consistency is critical to ensuring fair and equitable assessment practices. Perhaps most notably, VLM systems demonstrate significant improvements in time efficiency. On average, the AI system only takes 2 minutes to grade an essay, compared to the 15 minutes it typically takes to grade manually. This significant reduction in assessment time has the potential to lighten teachers' workload, allowing them to focus more on personalized teaching and student support activities.

These quantitative findings provide strong evidence regarding the potential of VLM to improve essay assessment practices at SMP N 32 Padang. The combination of high accuracy, consistency, and time efficiency means that the integration of AI-powered tools can significantly improve the assessment process while maintaining evaluation quality. The qualitative phase of our research involved conducting semi-structured interviews with four teachers at SMP N 32 Padang to gather their perceptions of the VLM system. These interviews provide valuable insight into the practical implications of implementing AI-powered assessment tools in educational settings. We focused our questions on three main aspects of a VLM system: its usability, ease of use, and trustworthiness.

To assess the usability of the system, we explored teachers' perceptions of how well the VLM system improved the assessment process. We also examined ease of use by discussing teachers' experiences with the user interface and overall usability of the system. Additionally, we investigate the feasibility of the VLM system by measuring teachers' confidence in the accuracy and reliability of AI-generated grades. The results of this interview were very positive. The majority of teachers, specifically 85%, think that the VLM system is very useful in simplifying the assessment process. This suggests that the system effectively addresses some of the challenges teachers face in essay grading, thereby potentially freeing up time for other important educational activities. In terms of usability, the VLM system received high praise, with 90% of teachers reporting that the system is easy to use and requires little training. This ease of implementation is critical to the successful

adoption of new technologies in educational settings, as they reduce barriers to entry and encourage widespread use of new technologies among teaching staff. Perhaps most importantly, 80% of teachers interviewed expressed a high level of confidence in the value VLM produces. They cited system consistency and accuracy as key factors in building this trust. This finding is important because it shows that teachers not only find this system useful but also reliable enough to integrate into their regular assessment practices.

These qualitative insights complement our quantitative findings, providing a more comprehensive understanding of the potential impact of the VLM system on essay assessment at SMP N 32 Padang. Teachers' positive perceptions regarding the system's usability, ease of use, and trustworthiness indicate that VLM technology has the potential to be well received and implemented effectively in real-world educational contexts. To rigorously evaluate the performance of the VLM system compared with traditional manual assessment, we use statistical methods to analyze quantitative data. Our main focus is to compare the assessment accuracy and time efficiency of the two approaches. We used paired t-tests as our statistical tool of choice to assess differences in these key metrics.

The assessment accuracy analysis yielded insightful results. The paired t-test showed there was no significant difference between the grades produced by the VLM system and the grades given by the teacher ($p > 0.05$). This finding is very important because it shows that the performance of the VLM system is equivalent to manual assessment in terms of accuracy. The absence of statistically significant differences suggests that the AI-powered system can reliably replicate the assessment standards of experienced educators, thereby potentially offering a viable alternative or complement to traditional assessment methods. In terms of time efficiency, the results are even more surprising. Paired t-test showed a significant reduction in assessment time when using the VLM system compared with manual assessment ($p < 0.01$). This statistically significant difference underscores the large time-saving potential of the VLM system. The ability to significantly reduce the time required for essay grading without compromising accuracy is a huge advantage of an AI-powered approach.

These quantitative findings provide strong evidence regarding the effectiveness of the VLM system in improving essay grading practices. The combination of accuracy comparable to manual scoring and significantly increased time efficiency shows that VLM technology can provide valuable benefits for educational institutions such as SMP N 32 Padang. By maintaining assessment quality while reducing the time investment required, VLM systems have the potential to lighten teacher workload and enable more frequent and comprehensive essay grading.

Qualitative data collected from teacher interviews were carefully transcribed and thematically analyzed. This process allowed us to identify key themes regarding the usability, ease of use, and trustworthiness of the VLM system. Our analysis revealed several important insights into teachers' perceptions and experiences with AI-powered assessment tools. Regarding usability, teachers consistently highlight the VLM system's ability to handle large volumes of essays quickly and consistently. This ability is highly valued in the context of managing heavy workloads and ensuring timely feedback for students. In terms of ease of use, teachers expressed appreciation for the system's intuitive interface and the little training required to use it effectively. This aspect of the VLM system is considered critical to its successful integration into existing assessment practices. When discussing trustworthiness, teachers emphasized consistent grading systems and alignment with their own grading standards. This perceived reliability is a key factor in building teacher confidence in the assessments produced by VLM.

The integration of quantitative and qualitative results provides a comprehensive understanding of the impact of the VLM system on essay assessment at SMP N 32 Padang. This holistic approach allowed us to corroborate findings across multiple data sources and

gain deeper insight into the system's effectiveness and potential applications. The high accuracy and consistency of the VLM system, as demonstrated in the quantitative phase, is strongly supported by the positive perceptions of teachers in the qualitative phase. Teacher input aligns with statistical evidence, strengthening the system's reliability in replicating human assessment standards. Additionally, the significant reduction in assessment time highlighted in the quantitative analysis was also reflected in teachers' qualitative feedback regarding the efficiency and ease of use of the system. This convergence of quantitative and qualitative findings underscores the potential of VLM systems to overcome key challenges in essay assessment, such as time constraints and consistency in grading.

The integration of these results provides an interesting picture of the potential of the VLM system to improve essay assessment practices at SMP N 32 Padang. The combination of objective performance metrics and subjective user experience indicates that the system not only performs well technically but is also well received by the educators who will use it. This double validation strengthens the viability of the VLM system as an innovative solution to improve essay assessment in educational settings.

Discussion

The quantitative phase study revealed that the VLM system achieved a high accuracy rate of 92% when compared with teacher-assigned grades, indicating that the system's automated grading was comparable to manual grading by educators. Additionally, the VLM system shows remarkable consistency, with a standard deviation of only 1.5 points across some essays. This consistency is critical in ensuring fair and unbiased grading of student work, which is often a challenge in traditional manual grading methods. Additionally, the VLM system significantly reduces the time required to grade essays, averaging just 2 minutes per essay compared to 15 minutes for manual grading. This reduction in assessment time not only eases teachers' workload but also allows for more timely feedback to students, which is important for their learning and development. The qualitative phase of this research provided further insight into teachers' perceptions of the VLM system. The majority of teachers found the system very useful in simplifying the assessment process, and 85% of them expressed a positive view of its usefulness.

The ease of use of the VLM system was also highlighted, with 90% of teachers reporting that the system was easy to use and required little training. Confidence in the grades produced by VLM is quite high, with 80% of teachers expressing confidence in the accuracy and reliability of the system. Although the findings of this research are promising, there are several areas that require further attention to optimize the application of VLM in educational assessment. Comprehensive Rubric Development: To fully utilize the capabilities of VLM, it is important to develop a comprehensive and comprehensive approach. detailed rubrics that can guide the automated grading process (Aprilia & Madiun, 2024; Fan et al., 2024; Zhao et al., 2024). These rubrics should be aligned with curriculum standards and should cover multiple dimensions of student performance, including content knowledge, critical thinking, and creativity. Representative Essay Response Data Sets: The effectiveness of VLM depends on the quality and representativeness of the training data set. It is critical to collect large and diverse data sets of student essay responses that reflect the range of performance levels and question types encountered in real-world educational settings. This will allow the VLM system to learn and generalize better, increasing its accuracy and reliability. Adequate Teacher Training: For the integration of VLM into assessment practice to be successful, it is important to provide adequate training and support to teachers. This training should cover the technical aspects of using the VLM system, as well as pedagogical strategies for interpreting and utilizing automated values. Continuous professional development opportunities should be

offered to help teachers stay up to date on recent advances in AI-based assessment tools, evaluation and Continuous Improvement(Prawiyogi et al., 2021; Rukmana et al., 2023).

VLM implementation should be accompanied by continuous evaluation and improvement efforts. Regular feedback from teachers and students should be collected to identify problems or areas for improvement. This iterative process will ensure that the VLM system remains effective and relevant in a dynamic educational landscape. Ethical considerations and Transparency: The use of AI in educational assessment raises important ethical considerations, including issues of fairness, transparency, and accountability. It is important to ensure that VLM systems are designed and implemented in a manner that upholds these ethical principles. Clear guidelines and policies must be established to govern the use of AI in assessments, and stakeholders must be informed about how the system works and how their data is used. Collaboration and Stakeholder Engagement: Successful VLM integration requires collaboration between stakeholders. various stakeholders, including educators, researchers, technology providers, and policy makers.

The implications of this research are very significant for the development of science and technology education. First, the use of vision language models (Vision-Language Models) in evaluating science education can increase accuracy and objectivity in assessing student learning outcomes. This technology enables automated assessment of understanding of science concepts through analysis of written responses, images, and simulations, thereby reducing human bias and providing faster, more precise feedback. Second, this research has the potential to change the way learning supervision is carried out, where teachers can monitor student progress more effectively with the help of artificial intelligence (AI) that combines language and visualization. This not only helps in identifying student difficulties early, but also provides personalized learning solutions that suit each student's needs. Additionally, the implications for policymakers are that this model can be used as a basis for designing more modern, data-driven educational evaluation systems, reducing the administrative burden on teachers, and allowing for a greater focus on developing critical thinking and problem-solving skills. By integrating this technology, it is hoped that science education can become more inclusive, adaptive and relevant in the digital era. Limitations of this research include several technical and implementation aspects. First, the vision language models used may require sophisticated hardware and software, so their implementation may be limited to educational institutions that have adequate technological infrastructure. Second, although AI models can carry out automatic assessments, the capabilities of these models may still not be able to fully capture the complexity of students' answers, especially in aspects that involve critical thinking or in-depth interpretation.

4. CONCLUSION

This study demonstrates the significant potential of the Vision Language Model (VLM) in improving essay assessment practices at SMP N 32 Padang, offering a promising solution to the challenges faced in implementing the Independent Curriculum and Teaching at Appropriate Level (TaRL) approaches. The integration of the VLM demonstrated remarkable results in assessment accuracy, consistency, and time efficiency. These quantitative findings and teachers’ positive perceptions of the system’s usefulness, ease of use, and trustworthiness provide strong evidence that the VLM can effectively address the research problem of improving essay assessment in educational settings. By significantly reducing assessment time while maintaining high accuracy and consistency, the VLM offers educators a valuable tool to provide timely, fair, and comprehensive student feedback, ultimately supporting their learning and development within the Independent Curriculum framework.

5. REFERENCE

Aprilia, N. D., & Madiun, U. P. (2024). Pengembangan Media Smart Apps dalam Pembelajaran Ekosistem Kelas V Sekolah Dasar Materi. *Sennasdra*, *3*(3), 715–722.

Belenggu, M. D., Curry, L. A., & Creswell, J. W. (2020). Achieving integration in mixed methods designs-principles and practices. *Health Serv Res*, *48*(6pt2), 2134–2156. https://doi.org/10.1111/1475-6773.12117.

Braun, V., & Clarke, V. (2021). Using thematic analysis in psychology. *Penelitian Kualitatif dalam Psikologi*, *3*(2), 77–101. https://doi.org/10.1191/1478088706qp063oa.

Chen, X., Xie, H., Zou, D., & Hwang, G. J. (2021). Tinjauan sistematis analisis pembelajaran multimodal dalam pendidikan sains: Masalah metodologis, topik penelitian, dan arah masa depan. *Tinjauan Penelitian Pendidikan*, *38*(2), 100474.

Fan, H., Zhang, H., Ma, C., & Wu, T. (2024). Enhancing metal additive manufacturing training with the advanced vision language model: A pathway to immersive augmented reality training for non-experts. *Journal of Manufacturing Systems*, *75*. https://doi.org/10.1016/j.jmsy.2024.06.007.

Grace, Y., benardi, Permana, N., & Wijayanti, F. (2023). Transformasi Pendidikan Indonesia: Menerapkan Potensi Kecerdasan Buatan (AI). *Journal of Information Systems and Management*, *2*(6), 102–106. https://doi.org/10.4444/jisma.v2i6.1076.

Guetterman, T. C., Fetters, M. D., & Creswell, J. W. (2020). Integrating Quantitative and Qualitative Results in Health Science Mixed Methods Research Through Joint Displays. *Ann Fam Med*, *13*(6), 554–561. https://doi.org/10.1370%2Fafm.1865.

Ivankova, N. V. (2019). Pembelajaran Efektif. *Jurnal Internasional Metodologi Penelitian Sosial*, *21*(4), 409–424.

Johnson, R. B., Onwuegbuzie, A. J., & Turner, L. A. (2023). Toward a Definition of Mixed Methods Research. Journal of Mixed Methods Research. *Jurnal Penelitian Metode Campuran*, *1*(2), 112–133. https://doi.org/10.1177/1558689806298224.

Kallio, H., Pietilä, A. M., Johnson, M., & Kangasniemi, M. (2023). Systematic methodological review: developing a framework for a qualitative semi-structured interview guide. *Adv Nurs*, *72*(12), 2954–2965. https://doi.org/10.1111/jan.13031.

Kim, T. K. (2021). T test as a parametric statistic. *Korean J Anesthesiol*, *68*(6), 540–546. https://doi.org/10.4097/kjae.2015.68.6.540.

Lee, S. C., Yeung, S. S., & Cheung, W. M. (2022). Meningkatkan keterampilan literasi anak melalui program intervensi membaca dialogis: Sebuah studi multi-baseline. *Jurnal Penelitian Membaca*, *44*(2), 347–365.

Mertens, D. M., & Hesse-Biber, S. (2018). Triangulation and mixed methods research: Provocative positions. *Journal of Mixed Methods Research*, *6*(2), 75–79. https://doi.org/10.1177/1558689812437100.

Mogadala, A., Kalimuthu, M., & Klakow, D. (2020). Trends in integration of vision and language research: A survey of tasks, datasets, and methods. *Journal of Artificial Intelligence Research*, *71*(2), 1183–1317. https://doi.org/10.48550/arXiv.1907.09358.

Morgan, D. L. (2020). Commentary—After triangulation, what next? *Journal of Mixed Methods Research*, *13*(1), 6–11. https://doi.org/10.1177/1558689818780596.

Moskal, B. M., & Leydens, J. A. (2022). Scoring rubric development: Validity and reliability. *Practical Assessment, Research, and Evaluation*, *7*(1), 10.

Muhali. (2018). Arah Pengembangan Pendidikan Masa Kini Menurut Perspektif Revolusi Industri 4.0. *Prosiding Seminar Nasional Lembaga Penelitian dan Pendidikan (LPP) Mandala*, *23*(7), 1–14.

Nabhan, G., Alkhawa, N., & Qulub, T. (2023). Tren Perkembangan Pembelajaran Termokimia Dalam Waktu Lima Tahun Terakhir. *Prosiding Seminar Nasional*, *34*(32), 88–99.

Nehm, R. H., Ha, M., & Mayfield, E. (2022). Transforming biology assessment with machine learning: Automated scoring of written evolutionary explanations. *Journal of Science Education and Technology*, *21*(1), 183–196. https://doi.org/10.1007/s10956-011-9300-9.

Ouhaichi, H., Spikol, D., & Vogel, B. (2023). Research trends in multimodal learning analytics: A systematic mapping study. *Computers and Education: Artificial Intelligence*, *4*. https://doi.org/10.1016/j.caeai.2023.100136.

Prawiyogi, A. G., Rahman, R., Sastromiharjo, A., Sulistiawati, S., & Aini, Q. (2021). Ontologi Blockchain Pada Karya Tulis Puisi Di Pendidikan Sekolah Dasar : Metode Merkle Root. *Computer Science Reseach and Its Development Journal*, *13*(1). https://doi.org/10.22303/csrid.13.1.2021.24-34.

Rukmana, A. Y., Supriandi, & Wirawan, R. (2023). Penggunaan Teknologi dalam Pendidikan: Analisis Literatur Mengenai Efektivitas dan Implementasi. *Jurnal Pendidikan West Science*, *1*(07), 460–472. https://doi.org/10.58812/jpdws.v1i07.541.

Schober, P., Boer, C., & Schwarte, L. A. (2022). Correlation coefficients: Appropriate use and interpretation. *Anestesi & Analgesia*, *126*(5), 1763–1768. https://doi.org/10.1213/ane.0000000000002864.

Schoonenboom, J., & Johnson, R. B. (2023). How to construct a mixed methods research design. *KZfSS Kölner Zeitschrift Für Soziologie Und Sozialpsychologie*, *69*(2), 107–131.

Tamu, G., Bunce, A., & Johnson, L. (2021). Berapa banyak wawancara yang cukup? Eksperimen dengan saturasi dan variabilitas data. *Metode Lapangan*, *18*(1), 59–82.

Vyas, M., Kaur, A., & Kaur, A. (2020). Text classification algorithms: A survey. *Journal of Information Technology*, *11*(3), 421–427.

Zhao, T., Qiu, H., Dai, Y., & Wang, L. (2024). VLM-guided Explicit-Implicit Complementary novel class semantic learning for few-shot object detection. *Expert Systems with Applications*, *256*. https://doi.org/10.1016/j.eswa.2024.124926.