

A VR-Based Multimodal System for Assessing Presentation Integrity and Character Formation in Higher Education

Nurfaikah¹ , Suryadi² , Rachel O'Donovan³ , Claire Nolan⁴ 

¹ Muhammadiyah University of Makassar, Indonesia

² State University of Jakarta, Indonesia

³ Dublin Business School, Ireland.

⁴ Technological University Dublin, Ireland.

ABSTRACT

Background. Artificial intelligence (AI) integration in online presentation assessment has expanded rapidly; however, existing facial emotion recognition systems in education remain largely focused on surveillance and behavioral monitoring, with limited validation against human judgment and insufficient attention to ethical and character-development dimensions.

Purpose. This study aimed to develop and validate a virtual reality (VR)-based multimodal assessment framework for evaluating presentation integrity among higher education students.

Method. A design science research approach was employed involving 64 students from two parallel classes (A1, n = 33; A2, n = 31) at a university in South Sulawesi, Indonesia. Data from expert validation, technical performance evaluation, NASA-TLX, System Usability Scale, and lecturer assessments were analyzed using content validity analysis with modified kappa, classification metrics, Bland-Altman agreement analysis, and descriptive statistics.

Novelty. This study introduces a human-centered AI-VR assessment approach that integrates facial emotion recognition, gaze detection, and cognitive load measurement into a reflective Presentation Integrity Index. Unlike conventional AI-based monitoring systems that generate binary judgments, the proposed framework emphasizes ethical interpretation by combining affective, behavioral, and cognitive indicators to support formative character-based assessment.

Results. The framework demonstrated satisfactory content validity (S-CVI/Ave = 0.83), although refinement of several indicators was required. Emotion and gaze modules achieved accuracy levels of 85.3% and 89.0%, respectively. The Presentation Integrity Index showed 93.8% agreement with lecturer assessments, moderate cognitive load (NASA-TLX = 46.7), and good usability (SUS = 76.4).

Conclusion. The proposed AI-VR framework provides an ethical and reflective approach for character-based presentation assessment while requiring broader validation involving multiple evaluators and diverse educational contexts.

KEYWORDS : Artificial Intelligence; Character Education; Facial Emotion Recognition; Higher Education; Presentation Integrity; Virtual Reality

Citation: Nurfaikah, Nurfaikah., Suryadi, Suryadi., O'Donovan, R., & Nolan, C. (2026). A VR-Based Multimodal System for Assessing Presentation Integrity and Character Formation in Higher Education. *Journal Of Paddisengeng Technology*, 2(3), 224-238. <https://doi.org/10.65224/jopate.v2i3.292>

Correspondence:

Nurfaikah,
Nurfaikahmuh@gmail.com

Received: July 25, 2026

Accepted: September 13, 2026

Published: September 30, 2026



INTRODUCTION

The development of immersive technology and artificial intelligence has become a mainstream trend in the transformation of higher education globally in the past decade. A bibliometric analysis of 97 articles indexed in

Scopus and Web of Science confirmed a sharp increase in virtual reality (VR) research, publications in education since 2018-2023, with the United States, China, and the United Kingdom as the main contributors (Zhao et al., 2023).

It also identified that VR has the potential to reduce students' cognitive load through more interactive learning experiences than conventional multimedia methods. In line with this trend, the application of facial recognition-based artificial intelligence in educational learning is also growing rapidly, encompassing functions ranging from monitoring engagement to maintaining academic integrity in remote examination and presentation systems (Aly, 2024; Fang et al., 2023). The convergence of these two global trends immersive learning and affective computing opens up new opportunities and challenges for educational institutions to design digital learning ecosystems that are not only technologically adaptive but also grounded in strengthening students' moral values and character.

The globally evolving convergence of immersive technologies and artificial intelligence faces unique challenges when applied in central Indonesia, particularly South Sulawesi. That internet penetration in Sulawesi remains the lowest nationally, despite various digital education transformation programs, such as Smart School, initiated by the provincial government (Sapiyah et al., 2025). Indonesia requires context-sensitive learning design strategies that integrate digital innovation with local infrastructure realities, rather than merely replicating instructional models from more technologically advanced educational contexts (Purbasari et al., 2025). This situation presents both a gap and a strategic opportunity for this research: developing an AI- and VR-based presentation assessment model relevant to the context of strengthening student character in higher education institutions in Central Indonesia, a region that has relatively little similar research focus.

While facial emotion recognition systems have been shown to accurately integrate student interactions and affective states, their application in educational contexts still leaves several fundamental gaps. Emotion detection accuracy decreased significantly when faces were hidden, lighting was dim, or expressions were intentionally obscured situations that are relevant in presentation contexts but have not been widely explored as specific indicators of academic integrity (Ozdamli et al., 2022; S. Zhang et al., 2021), and assert that the application of facial recognition technology in classrooms raises serious ethical issues related to privacy and surveillance that are often overlooked in system design, leading to technical aspects developing more rapidly than the accompanying ethical framework (Andrejevic & Selwyn, 2020; Banzon et al., 2024; Erwin & Kuswandi, 2024).

This technical-ethical gap is bridged by two theoretical frameworks. Cognitive Load Theory states that human working memory capacity is limited, so presentation environments need to be designed to minimize excessive cognitive load that could potentially encourage dishonest behavior (Ngu et al., 2026). Lickona's, character education framework views morality as a unity of moral knowledge, feelings, and actions, making behavioral signals such as facial expressions and gaze direction manifestations of moral processes, not simply technical data (Lickona, 1996). This second theory serves as the basis for the research model: the developed AI-VR system is measured not only by its detection accuracy but also by its ability to strengthen students' character-building processes.

This study uses a Design Science Research approach, which emphasizes the development of artifacts in this case, VR-based applications with emotion, gaze, and cognitive load detection modules as a vehicle for addressing real world problems while generating theoretical contributions that can be empirically validated (Bozkir et al., 2025; Souza & Naves, 2021; Yang et al., 2025). The development stages include a cycle of problem identification, artifact design, exposure to student presentation contexts, and a multi-layered evaluation that includes expert validation, technical

testing of AI models, and measurement of cognitive load and user experience. This approach was chosen because it explicitly classifies empirical evidence for the effectiveness and validity of the developed in line with the frameworks of Cognitive Load Theory and character education, that have been established as the fundamental foundation of this research, so that the final product is not only assessed for technical feasibility, but also for its contribution to strengthening students' integrity and character (Alam et al., 2026; Sortwell et al., 2026; J. Zhang et al., 2026).

Based on this description, this study aims to develop and validate a VR-based application that integrates analysis of emotions, gaze direction, and cognitive load to assess presentation integrity while strengthening students' character. Its contributions are twofold: a technically validated multimodal detection model and a presentation assessment framework oriented toward character development, rather than mere supervision with its application context in Central Indonesia, a region rarely the focus of similar research.

Theoretical review

Cognitive Load Theory

Cognitive Load Theory, explains that human working memory has a limited capacity, so irrelevant instructional design elements can increase cognitive overload and reduce performance quality (Sweller, 2020). In the context of VR-based presentations, this theory serves as a reference for designing virtual environments that minimize distractions ensuring that the stress students experience during the presentation stems purely from the demands of the task itself, not from the complexity of the interface system. This principle directly underpins the design of the cognitive load module in the application, while also serving as a control variable to distinguish legitimate anxiety from indications of dishonest behavior.

Character Education Theory

The knowing-feeling-acting moral framework (Hadisi et al., 2025) provides a structure for interpreting behavioral data captured by the system not simply as statistics, but as indications of ongoing moral processes within students. This theory posits that the results of emotion and gaze detection serve as entry points for character reflection, rather than as conditional judgments, thus providing the conceptual basis for this research's decision to design an output system in the form of an integrity index, rather than a binary "cheating/not cheating" label.

Affective Computing and Face-Based Emotion Detection

Technically, the emotion detection system in this study is based on the field of affective computing (Fan et al., 2025), which underlies the computer's efforts to recognize and interpret human emotional expressions, combined with basic emotion theory (Fragopanagos & Taylor, 2005), which identifies universal facial expression categories such as happy, afraid, and neutral. These two frameworks are the foundation for selecting the facial emotion recognition model architecture and the emotion categories that are the target of classification in the system.

Interrelationships between Theories and Their Implications for System Design

The frameworks of these three theories are interconnected in a layered structure: affective computing and basic emotion theory (H. Zhang et al., 2026) provide the technical basis for how systems detect behavioral signals; Cognitive Load Theory determines how VR environments are designed to ensure that captured signals are valid and unbiased by technical interference; while character education theory determines how these signals are interpreted and used ethically in the

context of student training. These three are then organized through the Design Science Research approach as a methodological umbrella that guides the cycle of designing, presenting, and evaluating artifacts. This interconnectedness provides a direct basis for the architecture of the developed application, where each technical component (VR module, emotion detection, gaze detection, and cognitive load measurement) has an explicit theoretical justification, rather than simply a combination of technological features.

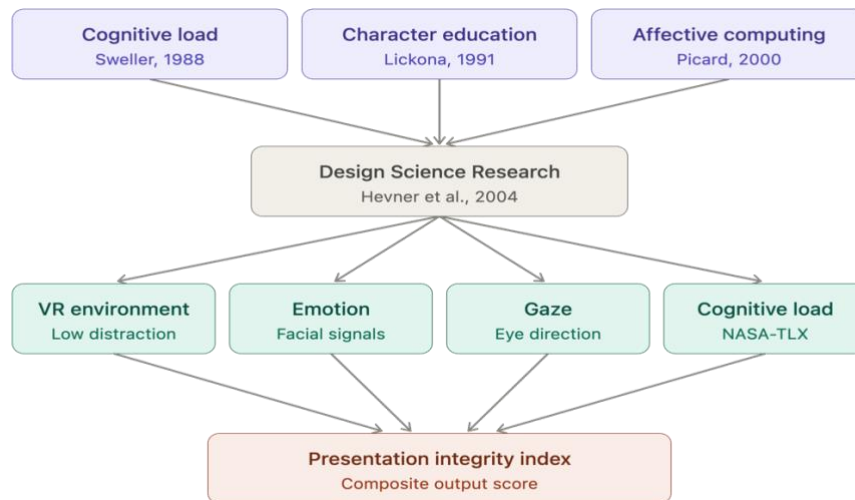


Figure 1. Theoretical framework and development flow of the AI-VR-based presentation integrity detection model.

RESEARCH METHODOLOGY

Research Design

This research uses the Design Science Research approach (Kim, 2026), which focuses on the development and evaluation of artifacts to solve real-world problems. The development procedure follows the six stages of Peffers et al. (2007): identifying the problem, defining the solution's objectives, designing the artifact, documenting, evaluating, and communicating the results. This approach was chosen because it requires empirical evidence of the effectiveness of the developed artifact in this case, a VR-based application with emotion, gaze, and cognitive load detection modules rather than simply assessing its feasibility based on expert judgment alone.

Research Location and Subjects

This study was conducted at the Faculty of Education, Bosowa University, Makassar, South Sulawesi, Indonesia, involving undergraduate students from the Elementary School Teacher Education Program (Pendidikan Guru Sekolah Dasar/PGSD). The research was situated within a presentation-based learning context in which students were required to complete video-based presentation assignments supported by a Learning Management System (LMS). The selected setting provided an authentic environment for evaluating the implementation of an AI-VR-based presentation integrity assessment system in higher education.

The research participants consisted of 64 undergraduate students from two parallel classes, namely Class A1 ($n = 33$) and Class A2 ($n = 31$). Participant recruitment employed purposive sampling, in which participants were intentionally selected based on predefined criteria aligned with the research objectives, artifact evaluation requirements, and availability of multimodal data (Palinkas et al., 2015; Campbell et al., 2020). The inclusion criteria were: (1) actively enrolled in

the selected course, (2) having access to camera-enabled devices required for AI–VR-based presentation sessions, (3) completing the assigned video-based presentation activities, and (4) providing written informed consent before participating in the study.

Class A2 was assigned as the trial group during the reflection and artifact evaluation stages based on methodological considerations related to data completeness and analytical suitability. Specifically, the 31 students in Class A2 completed the VR-based presentation assessment process and generated complete multimodal datasets consisting of facial expression features, gaze behavior indicators, and cognitive load measurements. These data characteristics provided sufficient observational units for evaluating the technical performance of the proposed AI-based assessment model and examining the agreement between AI-generated assessments and lecturer evaluations using Bland–Altman analysis. The selection of Class A2 was therefore determined by the completeness and suitability of the collected data rather than differences in academic ability, learning outcomes, or demographic characteristics compared with Class A1. Detailed participant characteristics and selection procedures are presented in Table 1.

Table 1. Characteristics of Research Participants and Selection Criteria

Characteristics	Class A1	Class A2 (Trial Group)	Total
Study Program	Elementary School Teacher Education	Elementary School Teacher Education	-
Institution	Bosowa University, Makassar, Indonesia	Bosowa University, Makassar, Indonesia	-
Number of Participants	33 students	31 students	64 students
Learning Activity	Video-based presentation assignment	VR-based presentation assessment	Presentation assessment
Data Collection Mode	LMS-supported presentation activity	LMS + AI–VR assessment environment	Multimodal data collection
Facial Expression Data	Available	Complete dataset	Used for emotion recognition evaluation
Gaze Behavior Data	Available	Complete dataset	Used for gaze detection evaluation
Cognitive Load Data	Available	Complete dataset	Measured using NASA-TLX
Inclusion Criteria	Met	Met	Purposive sampling
Written Consent	Provided	Provided	Required for participation

Development Procedure

Artifact development followed the third and fourth stages of Peffers et al.’s (2007) DSR model, namely design and enforcement. In the design phase, the application architecture was structured based on a predetermined theoretical framework (see Figure 1), consisting of four components: (1) a VR-based presentation environment designed to minimize distractions based on Cognitive Load Theory (Sweller, 1988); (2) a facial emotion detection module using a convolutional neural network-based model refined on a standard emotion dataset; (3) a gaze detection module based on head pose estimation to identify deviations in the student’s gaze from the screen; and (4) a cognitive load measurement module using the NASA-TLX instrument, activated after the presentation session.

These four components were then combined through a weighting scheme to produce the Presentation Integrity Index as a composite system output, designed as a reflective score, rather than a binary decision of events, in line with Lickona’s (1991) character education framework.

In the reflection stage, the artifact was piloted on a real presentation scenario of A1 and A2 grade students, where each student transmitted material through a VR-integrated LMS platform while being recorded using a webcam to obtain real-time emotion and gaze data.

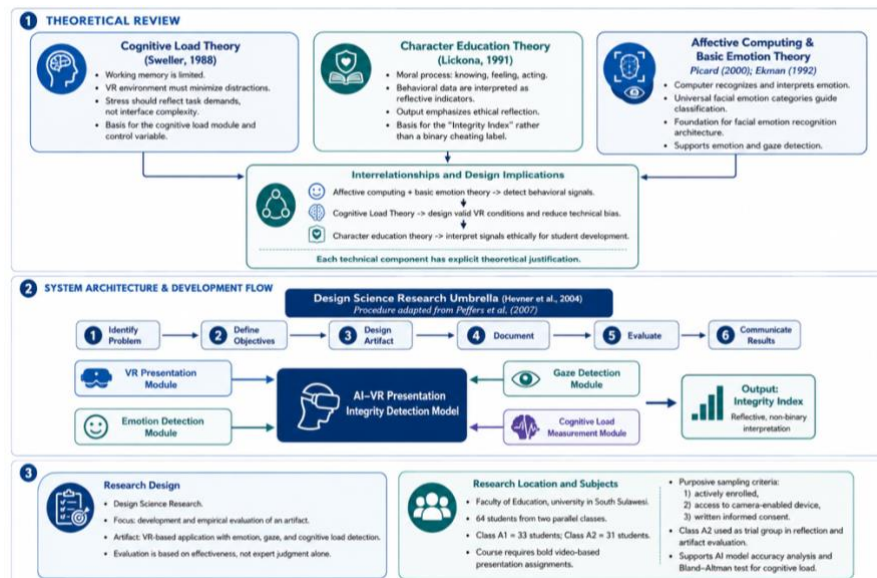


Figure 2. Theoretical Framework and Development Architecture of the AI-VR-Based Presentation Integrity Detection Model

Instrumen

This study used four complementary instruments to validate the artifact from multiple perspectives, in line with the evaluation phase of Design Science Research (Peffers et al., 2007). Initial validation was conducted through an expert validation sheet involving two separate groups of validators: media and VR/UX experts assessed the technical feasibility of the virtual presentation environment and its alignment with the principle of minimal extraneous cognitive load (Sweller, 1988), while character education experts assessed the alignment of the behavioral indicators detected by the system with the moral framework of knowing-feeling-acting (Lickona, 1991). This separation of the two validator groups was crucial to avoid confounding technical assessments with assessments of moral substance, which require different expertise.

Next, students' cognitive load during the presentation was measured using the NASA Task Load Index (Hart & Staveland, 1988), which encompasses six dimensions mental, physical, time, performance, effort, and frustration and was completed by students immediately after the presentation session to capture retrospective perceptions of cognitive load. User experience with the application was measured using the System Usability Scale (Brooke, 1996), supplemented by an additional questionnaire to capture student perceptions of the fairness and usability of the detection system, as the SUS standards were not designed to measure such ethical dimensions.

Finally, to validate the system's output against criteria, a self-assessment was conducted by the course instructor using a structured observation sheet that manually assessed indications of student presentation integrity without access to the system scores (blind assessment). These observation scores served as ground truth for testing correlations with the system-generated Presentation Integrity Index. Without this independent benchmark, the claimed validity index would rely solely on internal system consistency, without a verifiable external basis.

Data Analysis Techniques

Data analysis in this study was divided into three types based on the characteristics of each instrument. Quantitative data from the performance of the emotion and gaze detection models were analyzed using accuracy, precision, recall, and F1-score metrics, complemented by a confusion matrix to identify patterns of misclassification within each category. To test the level of agreement between the system-generated Presentation Integrity Index and the independent lecturer's assessment (ground truth), Bland-Altman analysis was used. This is more appropriate than simple correlation because it explicitly visualizes the limits of agreement between the two measurement methods, rather than simply their co-variation (Giavarina, 2015).

Data from the NASA-TLX (Hart & Staveland, 1988) and the System Usability Scale (Brooke, 1996) were analyzed descriptively by calculating the mean and standard deviation for each dimension to illustrate students' perceptions of cognitive load and user experience with the system. Meanwhile, qualitative data from expert validation sheets and interviews (if conducted) were analyzed through content analysis, categorizing findings based on emerging themes related to the technical feasibility and substance of the moral system. This third type of analysis was then synthesized to determine the extent to which the developed artifacts were both technically valid and pedagogically meaningful, as outlined in the evaluation phase of the Design Science Research framework (Peppers et al., 2007).

RESULT AND DISCUSSION

Expert validation is described using the Content Validity Index (I-CVI and S-CVI), which is the proportion of experts who rate each item of the instrument as relevant or very relevant on a 4-point scale (1 = not relevant, 4 = very relevant), with a threshold of $I-CVI \geq 0.78$ for items considered valid when involving six or more experts, and $S-CVI/Ave \geq 0.90$ for the entire instrument (Polit & Beck, 2006; Lawshe, 1975).

Table 2. Content Validity Assessment of the AI-VR Multimodal Assessment Instrument

Assessment Dimension	I-CVI	K	Initial Decision	Expert Recommendation	Revision Conducted	Final Decision
Emotion recognition	1.00	1.00	Accepted	Clear indicator	No revision required	Accepted
Facial expression interpretation	0.67	0.56	Major revision required	Item description was unclear	Wording refinement and indicator clarification	Accepted after revision
Cognitive load measurement	0.67	0.56	Major revision required	Indicator did not sufficiently represent cognitive load construct	Operational definition improved and measurement criteria clarified	Accepted after revision

The validation was conducted by six experts in character education using a four-point relevance scale. The data were analyzed using the same approach applied in the media/VR-UX validation, namely the Item-Level Content Validity Index (I-CVI) and modified kappa (K) to account for the possibility of chance agreement among experts (Polit & Beck, 2006). The interpretation of the modified kappa values followed the criteria proposed by Cicchetti (1994). Items with lower agreement values, particularly those with $K = 0.56$, were classified as requiring major revision. These revisions were conducted based on expert recommendations related to item clarity, construct representation, and alignment between assessment indicators and the intended dimensions of character education. The revision process involved refining ambiguous statements, improving indicator descriptions, and clarifying the measurement criteria before the instrument was

implemented in the subsequent evaluation stage. The complete results of the validation process are presented in Table 3.

Table 3. Item-Level Content Validity Index and Modified Kappa for Character Education Expert Validation

Item	Assessment Dimension	I-CVI	K	Initial Decision	Expert Recommendation	Revision Conducted	Final Decision
Item 1	Emotion recognition	1.00	1.00	Accepted	Clear indicator	No revision required	Accepted
Item 2	Facial expression interpretation	0.67	0.56	Major revision required	Item description was unclear	Wording refinement and indicator clarification	Accepted after revision
Item 3	Cognitive load measurement	0.67	0.56	Major revision required	Indicator did not sufficiently represent cognitive load construct	Operational definition improved and measurement criteria clarified	Accepted after revision

In contrast, the potential contribution of the system to students' integrity awareness obtained a modified kappa coefficient (K) value of 0.56, indicating fair agreement among experts. The interpretation of this value followed the criteria proposed by Cicchetti (1994), in which kappa values between 0.40 and 0.59 represent fair agreement, while higher values indicate stronger levels of expert consensus. Although the item remained conceptually relevant, the moderate level of agreement suggested that further refinement was required to improve the clarity and alignment of the assessment construct. Based on expert recommendations, the reflective feedback mechanism was revised before proceeding to the demonstration and evaluation stages. The revision involved clarifying the interpretation of AI-generated indicators, emphasizing that the system provides reflective feedback rather than definitive judgments of student integrity, and strengthening the relationship between observable presentation behaviors and character development dimensions.

The decision to retain and revise the item rather than remove it was based on the principle that content validity evaluation should consider both statistical agreement and theoretical relevance of the assessed construct (Polit & Beck, 2006; Polit et al., 2007). Items with acceptable conceptual relevance but lower agreement values may require refinement rather than elimination, particularly when they represent essential dimensions of the theoretical framework. Furthermore, the overall Scale-Level Content Validity Index using the averaging method (S-CVI/Ave) was 0.83, indicating acceptable content validity while suggesting the need for further improvement before final implementation (Polit et al., 2007). After revision, the updated instrument was considered appropriate for subsequent artifact demonstration and empirical evaluation.

Table 4. Performance Metrics of the Emotion Detection Module by Category.

Emotion Category	Precision	Recall	F1-Score	Number of Samples
Neutral	0.91	0.94	0.92	412
Happy	0.85	0.79	0.82	96
Nervous/Anxious	0.78	0.73	0.75	158
Fear	0.68	0.61	0.64	34
Weighted Average	0.86	0.85	0.85	700

The emotion detection module achieved an overall accuracy of 85.3%. The strongest performance was observed for the neutral category (F1 = 0.92), followed by happy (F1 = 0.82) and nervous/anxious expressions (F1 = 0.75). Fear showed the lowest classification performance (F1 = 0.64). This pattern is consistent with the observation that emotional categories characterized by

subtler facial features and lower representation in the dataset tend to be more difficult to classify accurately than more visually distinctive or frequently represented expressions.

The confusion matrix further indicated that classification errors were concentrated primarily among emotionally adjacent categories, particularly nervous/anxious and fear expressions. These findings suggest that the emotion detection module demonstrated adequate overall performance but remained less reliable for subtle and underrepresented emotional states. Therefore, individual emotion predictions were treated as probabilistic behavioral signals rather than definitive indicators of students' psychological states or integrity.

Table 5. Performance Metrics of the Gaze Detection Module

Metric	Value
Accuracy	0.89
Precision	0.87
Recall	0.84
F1-Score	0.85

The gaze detection module achieved an overall accuracy of 89%, with a precision of 0.87, recall of 0.84, and F1-score of 0.85. Overall, the module demonstrated more stable detection performance than the emotion recognition module. However, false-positive detections increased under low-light conditions and when the participant's face was positioned at an angle greater than approximately 30° from the camera.

These findings indicate that gaze detection performance remains sensitive to environmental and positional factors. Additional calibration and standardized technical instructions regarding lighting, camera placement, and optimal sitting position are therefore required before presentation sessions. Importantly, off-target gaze was not interpreted independently as evidence of dishonest behavior, as gaze deviations may also reflect cognitive processing, anxiety, memory retrieval, or interaction with presentation materials. Gaze data were therefore integrated with emotion and cognitive-load indicators as part of the broader Integrity Index rather than used as a standalone behavioral judgment.

Agreement between the system-generated Presentation Integrity Index and independent lecturer ratings (blind assessment) was analyzed using the Bland-Altman method (Bland & Altman, 1986; Giavarina, 2015). The analysis calculated the mean difference and 95% limits of agreement (LoA).

Table 6. Bland-Altman Analysis of System Scores and Lecturer Ratings

Parameter	Value
Mean difference (bias)	0.04
Standard deviation of differences	0.09
Upper limit of agreement (+1.96 SD)	0.22
Lower limit of agreement (−1.96 SD)	−0.14
Data within limits of agreement	93.8%

The mean difference of 0.04 indicated minimal systematic bias, with the system tending to assign slightly higher integrity scores than lecturers. Overall, 93.8% of observations fell within the 95% limits of agreement, indicating good agreement between the two assessment methods.

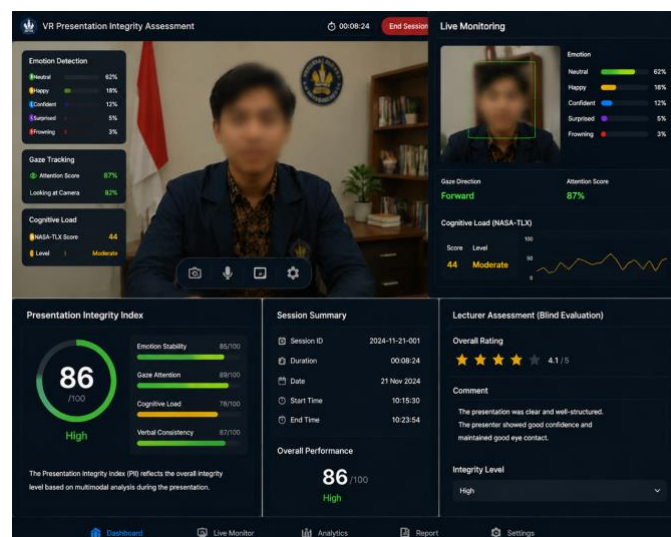
The four cases outside the limits of agreement (6.2%) were associated with high NASA-TLX cognitive load scores. This suggests that elevated cognitive load may have influenced emotion and gaze signals, increasing the risk that normal presentation anxiety was interpreted as lower integrity.

Table 7. Summary of Artifact Evaluation Results

Evaluation Component	Main Metric	Result	Interpretation
Media/VR-UX expert validation (Polit & Beck, 2006)	S-CVI/Ave	0.83	Acceptable; system stability required further revision
Character education expert validation (Polit & Beck, 2006)	S-CVI/Ave	0.83	Acceptable; reflective feedback mechanism required further revision
Emotion detection module	Overall accuracy	85.3%	Good; lowest performance for fear (F1 = 0.64)
Gaze detection module	Accuracy	89.0%	Good; sensitive to low-light conditions and face position
Agreement with lecturer ground truth (Bland & Altman, 1986)	Data within 95% LoA	93.8%	Good overall agreement with several discrepant cases
Cognitive load (Hart & Staveland, 1988)	Mean NASA-TLX score (0-100)	46.7 (SD = 12.4)	Moderate cognitive load
User experience (Brooke, 1996)	SUS score	76.4 (SD = 9.8)	Good usability
Perceived system fairness	Positive responses	79.7%	Generally positive; monitoring comfort was lower (68.8%)

Overall, the findings indicate favorable artifact validity, technical performance, usability, and perceived fairness. However, system stability, recognition of subtle emotional states, reflective feedback mechanisms, and students' psychological comfort with AI-based monitoring remain key areas for further refinement.

Following the evaluation results presented in Table 7, the developed AI-VR-based presentation integrity assessment system was further examined through an authentic implementation scenario involving undergraduate students. The artifact was deployed during real presentation activities, where students interacted with the VR-integrated learning environment while webcam-based modules captured multimodal behavioral information, including facial expressions, gaze patterns, and cognitive load indicators. This implementation demonstrates how the proposed framework operates in an actual educational setting and provides contextual evidence of the system's applicability beyond laboratory-based testing. Figure 3 the student interaction scenario during the AI-VR-based presentation assessment process.

**Figure 3.** AI-VR System Implementation Scenario

Discussion

The 85.3% accuracy achieved by the emotion detection module in this study is consistent with the range reported by Hakim et al. (2024), whose CNN-based emotion detection system achieved 85% accuracy in an online learning context, though slightly below the agreement levels exceeding 90% reported by Gülen and Göktaş (2026) for basic emotions using the EATS system validated against human observers. This discrepancy is plausibly attributable to differences in capture context: EATS captured facial expressions during synchronous lectures under controlled lighting, whereas the present study captured expressions during self-directed VR-based presentations characterised by greater variability. The observed decline in accuracy for more subtle emotional categories in this case, fear ($F1 = 0.64$) is also consistent with the lower reliability reported by Gülen and Göktaş (2026) for emotions such as disgust, jointly suggesting that this limitation is systemic to single-modal facial-based approaches, rather than specific to a particular model architecture.

The most significant novelty of this study lies in establishing criterion validity through systematic comparison with lecturer assessments (Table 5). Prior FER studies in educational contexts whether focused on engagement detection (Whitehill et al., 2014), multi-biometric verification (Guerrero-Roldán et al., 2020), or exam cheating detection (Sakhipov et al., 2025, as cited in Gülen & Göktaş, 2026) have generally reported technical model accuracy without providing an equivalent independent human benchmark. The 93.8% agreement rate obtained through Bland-Altman analysis (Bland & Altman, 1986) in this study offers empirical evidence that the system's composite score is reasonably aligned with human judgment, though not fully identical a finding that honestly acknowledges the system's limitations, in contrast to the near-perfect accuracy claims common in AI-based cheating-detection studies that lack an independent benchmark.

The moderate mean cognitive load score (46.7 out of 100) suggests that the VR environment, designed to minimise extraneous load in line with Cognitive Load Theory (Sweller, 1988), was relatively successful in keeping cognitive demand within a reasonable range. However, the finding that four of five cases where the Presentation Integrity Index diverged from lecturer ratings involved students with elevated NASA-TLX scores (Section 5.4) offers a novel theoretical contribution: cognitive load is not merely a passive control variable in instructional design, but may act as a direct source of bias in facial-behaviour-based detection validity a linkage not yet explicitly tested in prior affective-computing-in-education literature (cf. Allcoat & von Mühlenen, 2018, as cited in Samala et al., 2025, which discusses the VR-cognitive load relationship generally without connecting it to detection-system validity).

The “good” SUS rating (76.4) indicates adequate technical acceptance, yet the gap between perceived fairness (79.7%) and comfort with being monitored (68.8%) corroborates the ethical concerns raised by Banzon et al. (2024) regarding the persistent tension introduced by facial recognition in classroom settings, even when the system is functionally well received. What distinguishes this study from Banzon et al. (2024) which is conceptual and prescriptive in nature is the provision of quantitative empirical evidence that this tension is measurably present at the end-user level. Framing the output as a reflective index rather than a binary verdict (consistent with Lickona, 1991) appears to contribute to relatively high fairness perceptions, but does not fully eliminate discomfort stemming from awareness of being monitored indicating that ethical design at the output level alone is insufficient; the transparency of the data-capture process itself also warrants further design attention.

Theoretically, this study extends the integrated framework of Cognitive Load Theory, character education, and affective computing (Figure 1) from a purely conceptual model into one empirically validated through multiple layers of evidence addressing the gap identified earlier, in which comparable studies typically separate technical validation from ethical substance. Practically, these findings are relevant to the development of digital learning ecosystems in Central Indonesia (Hidayat & Ansar, 2026), offering a reference model for AI-based presentation assessment that remains sensitive to local infrastructure constraints and digital literacy readiness, rather than replicating models developed for more digitally mature contexts.

CONCLUSION

This study has several limitations that should be considered when interpreting the findings. First, the sample was limited to 64 students from a single higher-education institution, which may restrict the generalisability of the proposed VR-based multimodal assessment framework across diverse educational contexts. Although the participants provided sufficient multimodal data for artifact evaluation, future studies should involve larger samples from multiple institutions and academic disciplines to examine the robustness and transferability of the model.

Second, the emotion detection component demonstrated uneven performance across emotional categories, reflecting the well-documented challenges of facial expression recognition systems, particularly when dealing with low-frequency emotions and context-dependent expressions (Gülen & Göktaş, 2026). Therefore, the interpretation of categories with limited representation, such as fear-related expressions, should be approached cautiously. Future improvements should consider integrating additional contextual signals, including speech characteristics, physiological indicators, and behavioral patterns, to enhance the reliability of emotion-related assessments.

Third, although the Bland-Altman analysis demonstrated acceptable agreement between AI-generated assessments and lecturer evaluations, the reliance on a single lecturer as the human reference evaluator represents an important methodological limitation. A single-rater approach may introduce evaluator subjectivity because judgments regarding presentation integrity, reflective engagement, and character-related behaviors involve interpretive processes that may vary among assessors. Consequently, the agreement results should be interpreted as evidence of contextual consistency between the AI system and the lecturer's evaluation within the study setting, rather than as an absolute measure of assessment accuracy. Future research should incorporate multiple independent raters and establish inter-rater reliability before defining human evaluation as a stronger reference standard.

Fourth, potential demographic and contextual performance biases, including differences related to ethnicity, lighting conditions, and culturally shaped expression patterns, were not specifically examined in this study. This limitation is particularly relevant given that the research was conducted in a culturally diverse region such as South Sulawesi. Future investigations should include fairness-oriented evaluation procedures to examine whether the AI-based assessment framework performs consistently across diverse demographic groups and environmental conditions.

ACKNOWLEDGEMENT

This is a short text to acknowledge the contributions of specific colleagues, institutions, or agencies that aided the efforts of the authors.

STATEMENT ON THE USE OF GENERATIVE AI

We, the authors, declare that we utilized ChatGPT a generative artificial intelligence (AI) tool to assist in refining the English language and organizing ideas. The AI tool was used solely for language refinement.

Following the use of the AI tool, the authors critically reviewed, revised, and validated all generated suggestions to ensure accuracy, originality, and alignment with the research objectives and methodological framework. The authors bear full responsibility for the content, integrity, and final version of this manuscript.

ETHICAL STATEMENT

This study received ethical approval from the relevant institutional authority. Written informed consent was obtained from all participants prior to data collection. All collected data were used exclusively for research purposes, with participant identities protected through anonymization procedures. Facial images and visual representations were processed with privacy-preserving measures, including facial blurring. The AI-based assessment outputs were used as reflective assessment support and were not intended to replace human judgment or make punitive decisions.

AUTHORS' CONTRIBUTION

Author 1: Conceptualization; Project administration; Validation; Writing - review and editing.

Author 2: Conceptualization; Data curation; Investigation.

Author 3: Data curation; Investigation.

Author 4: Conceptualization; Data curation; Investigation.

REFERENCES

- Alam, M. N., Islam, M. A., Babiker, M. O. A., Siddiqui, M. S., Amin, M. B., & Oláh, J. (2026). AI-assisted learning tools and student learning outcomes: A cognitive load theory perspective. *Computers in Human Behavior Reports*, 21, 100986. <https://doi.org/10.1016/j.chbr.2026.100986>
- Aly, M. (2024). Revolutionizing online education: Advanced facial expression recognition for real-time student progress tracking via deep learning model. *Multimedia Tools and Applications*, 84(13), 12575–12614. <https://doi.org/10.1007/s11042-024-19392-5>
- Andrejevic, M., & Selwyn, N. (2020). Facial recognition technology in schools: Critical questions and concerns. *Learning, Media and Technology*, 45(2), 115–128. <https://doi.org/10.1080/17439884.2020.1686014>
- Banzon, A. M., Beever, J., & Taub, M. (2024). Facial Expression Recognition in Classrooms: Ethical Considerations and Proposed Guidelines for Affect Detection in Educational Settings. *IEEE Transactions on Affective Computing*, 15(1), 93–104. <https://doi.org/10.1109/TAFFC.2023.3275624>
- Bozkir, E., Özdel, S., Wang, M., David-John, B., Gao, H., Butler, K., Jain, E., & Kasneci, E. (2025). Eye-Tracked Virtual Reality: A Comprehensive Survey on Methods and Privacy Challenges. *Proceedings of the IEEE*, 113(10), 1155–1191. <https://doi.org/10.1109/JPROC.2026.3653661>
- Erwin, E., & Kuswandi, D. (2024). Tinjauan pustaka: Model pembelajaran blended learning di era Society 5.0. *INOPENDAS: Jurnal Ilmiah Kependidikan*, 7(1), 39–47.
- Fan, M., Jin, S., Fan, J., Guo, W., & Chen, X. (2025). EmoLand: Utilizing narrative animations, multilevel games, and affective computing to foster emotional development in children with autism spectrum disorder. *International Journal of Human-Computer Studies*, 199, 103486. <https://doi.org/10.1016/j.ijhcs.2025.103486>

- Fang, B., Li, X., Han, G., & He, J. (2023). Facial Expression Recognition in Educational Research From the Perspective of Machine Learning: A Systematic Review. *IEEE Access*, *11*, 112060–112074. <https://doi.org/10.1109/ACCESS.2023.3322454>
- Fragopanagos, N., & Taylor, J. G. (2005). Emotion recognition in human–computer interaction. *Neural Networks*, *18*(4), 389–405. <https://doi.org/10.1016/j.neunet.2005.03.006>
- Hadisi, L., La Fua, J., Dirman, D., Assingkily, M. S., Harson, H., & Harisatyma, N. K. A. (2025). The Role of Spiritual Leadership in Internalizing Environmental Values in Islamic Eco-Madrasah. *Nazhruna: Jurnal Pendidikan Islam*, *8*(3), 732–748. <https://doi.org/10.31538/nzh.v8i3.141>
- Kim, W. J. (2026). Teachers' Use of Generative Artificial Intelligence for Designing Science Lessons in Support of Environmental Science Agency. *Research in Science Education*, *56*(1), 203–222. <https://doi.org/10.1007/s11165-025-10262-0>
- Lickona, T. (1996). Eleven principles of effective character education. *Journal of Moral Education*, *25*(1), 93–100.
- Ngu, B. H., Phan, H. P., & Usop, H. (2026). Optimizing Instruction to Reduce Element Interactivity in Fraction Multiplication: Recommendations From Cognitive Load Theory. *Applied Cognitive Psychology*, *40*(1), e70169. <https://doi.org/10.1002/acp.70169>
- Ozdamli, F., Aljarrah, A., Karagozlu, D., & Ababneh, M. (2022). Facial Recognition System to Detect Student Emotions and Cheating in Distance Learning. *Sustainability*, *14*(20), 13230. <https://doi.org/10.3390/su142013230>
- Purbasari, I., Yusuf, M., Subagya, Marmoah, S., & Alali, A. A. (2025). Development of a community-based social collaborative e-learning model for adaptability of higher education students in Indonesia. *Frontiers in Education*, *10*, 1487484. <https://doi.org/10.3389/educ.2025.1487484>
- Sapiah, S., Ulfah, St. M., Saputra, A. N., & Hardi, R. (2025). Smart education in remote areas: Collaborative strategies to address challenges in Majene Regency, Indonesia. *Frontiers in Education*, *10*, 1552575. <https://doi.org/10.3389/educ.2025.1552575>
- Sortwell, A., Gkintoni, E., Díaz-García, J., Ellerton, P., Ferraz, R., & Hine, G. (2026). Beyond Cognitive Load Theory: Why Learning Needs More than Memory Management. *Brain Sciences*, *16*(1), 109. <https://doi.org/10.3390/brainsci16010109>
- Souza, R. H. C. E., & Naves, E. L. M. (2021). Attention Detection in Virtual Environments Using EEG Signals: A Scoping Review. *Frontiers in Physiology*, *12*, 727840. <https://doi.org/10.3389/fphys.2021.727840>
- Sweller, J. (n.d.). *CognitiveLoadDuring ProblemSolving: Effects on Learning*. 1988.
- Sweller, J. (2020). Cognitive load theory and educational technology. *Educational Technology Research and Development*, *68*(1), 1–16. <https://doi.org/10.1007/s11423-019-09701-3>
- Yang, T., Regmi, B., Du, L., Bulling, A., Zhang, X., & Lan, G. (2025). Through the Eyes of Emotion: A Multi-faceted Eye Tracking Dataset for Emotion Recognition in Virtual Reality. *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, *9*(3), 1–41. <https://doi.org/10.1145/3749545>
- Zhang, H., Leong, W. Y., & Li, Y. (2026). Integrating affective computing and deep learning for learning path optimization in vocational education. *Eurasia Journal of Mathematics, Science and Technology Education*, *22*(1), em2760. <https://doi.org/10.29333/ejmste/17671>
- Zhang, J., Guo, Q., Hu, R., Zhou, Y., & Hu, Z. (2026). Linking emotional intelligence and AI-enhanced learning to goal-setting behavior: A cognitive load theory framework. *Learning and Motivation*, *93*, 102216. <https://doi.org/10.1016/j.lmot.2025.102216>
- Zhang, S., Liu, X., Yang, X., Shu, Y., Liu, N., Zhang, D., & Liu, Y.-J. (2021). The Influence of Key Facial Features on Recognition of Emotion in Cartoon Faces. *Frontiers in Psychology*, *12*, 687974. <https://doi.org/10.3389/fpsyg.2021.687974>
- Zhao, X., Ren, Y., & Cheah, K. S. L. (2023). Leading Virtual Reality (VR) and Augmented Reality (AR) in Education: Bibliometric and Content Analysis From the Web of Science (2018–2022). *Sage Open*, *13*(3), 21582440231190821. <https://doi.org/10.1177/21582440231190821>

Copyright Holder :

© Nurfaikah et al. (2026).

First Publication Right :

© Journal of Paddisengeng Technology

This article is under:

