Indonesian Journal of Educational Development (IJED)
Volume 7, Issue 2, 2026, pp. 598-613
ISSN: 2722-1059 (Online); ISSN: 2722-3671 (Print)
DOI: https://doi.org/10.59672/ijed.v7i2.6999

ViLabS: Deep learning AI and computer vision-based smart website for improving basic laboratory techniques

Muhammad Farhan Ivan Phalosa[1], Rusly Hidayah*)[2], Muhammad Zainur Rifai[3]
[1]Universitas Negeri Surabaya, Surabaya, Indonesia; muhammad.23032@mhs.unesa.ac.id
[2]Universitas Negeri Surabaya, Surabaya, Indonesia; ruslyhidayah@unesa.ac.id
[3]Thursina International Islamic Boarding School, Malang, Indonesia; zainurrifaim@thursinaiibs.sch.id
*)Corresponding author: Rusly Hidayah; E-mail addresses: ruslyhidayah@unesa.ac.id

Article Info

Article history:
Received July 15, 2026
Revised July 25, 2026
Accepted August 08, 2026
Available online August 17, 2026

Keywords: Basic laboratory technique, Computer vision, Deep learning AI, Smart website

Abstract. To address excessive cognitive load and the lack of subject-based psychomotor evaluation in chemistry education, this study evaluates ViLabS, a deep-learning and computer vision website, for improving students' cognitive and psychomotor abilities. Using an explanatory sequential mixed-methods design, a quasi-experimental study involved 40 Grade X students selected via purposive sampling. Data were collected through cognitive tests, AI system logs, video-observation rubrics, and practicality questionnaires, followed by in-depth interviews. Results revealed ViLabS is highly feasible and practical (>92%). Cognitively, the experimental group achieved a significantly higher N-Gain (0.71) than the control (0.41) ($p<0.05$). Autonomously monitored psychomotor accuracy also significantly outperformed classical manual demonstrations ($p<0.05$). Thematic analysis confirmed the AI's instant corrective feedback minimized cognitive load and fostered precise muscle memory. Despite technical constraints such as internet and lighting dependency, ViLabS shows strong efficacy in accelerating theoretical understanding and kinesthetic proficiency. Future research should expand datasets and develop offline mobile applications. Ultimately, this study contributes a proactive, multimodal AI framework that advances the paradigm of AI-assisted laboratory learning.

Introduction

Science and technology education faces a sharp transformation alongside the massive penetration of Artificial Intelligence (AI) in the Society 5.0 era (Fukuda, 2020; Hwang et al., 2020; Popenici, 2017). This digital transformation is no longer limited to automating academic administration systems; it is slowly entering the pedagogical ecosystem to create a more meaningful learning experience. Various studies confirm that integrating AI into the curriculum can improve students' learning abilities (Celik, 2023; Zawacki-Richter et al., 2019).

In STEM education, artificial intelligence offers personalized learning paths, intelligent feedback, and automated assessment tools that facilitate a more efficient learning process (Chen et al., 2020; Crompton & Burke, 2023; Widana et al., 2021). AI can analyze users' digital footprints, predict potential misconceptions, and present adaptive instructional paths based on students' needs (Hariyono, Hidayati, & Suharto, 2026). Although this technology has been widely implemented in cognitive assessment instruments, AI use in psychomotor evaluation remains relatively limited

(Holmes & Tuomi, 2022; Chiu et al., 2023). This gap in technological development shows that conceptual understanding is largely automated, while observing physical movement skills still lacks adequate supervisory support.

At the secondary education level, chemistry learning relies on mastery of precise basic laboratory techniques as an absolute prerequisite for achieving complete scientific literacy (Seery, 2020). Practicum activities are not just a means of proving theories, but a space to develop critical thinking frameworks, analytical inquiry, and a strict scientific safety culture (Williams et al., 2018; Weaver et al., 2016; Mahsup et al., 2026). However, empirical evidence shows that students often experience fundamental difficulties in the psychomotor domain when manipulating glass instruments. Recent studies reveal that procedural mistakes such as inaccurate pouring angles or incorrect hand postures when handling pipettes occur repeatedly (Kyynäräinen et al., 2024; Schechtel & Bongers, 2026). These errors not only corrupt the precision of experimental data but also trigger cognitive anxiety that hinders the science assimilation process. Monitoring the psychomotor proficiency of dozens of students simultaneously under these conditions presents a highly complex managerial challenge for laboratory instructors. The limited ratio of educators to the student population often leaves the skill transfer process stuck in a one-way classical demonstration approach Cheung (2011); I Made Elia Cahaya et al. (2024), causing many learners to struggle when articulating theoretical procedures into targeted manipulative gestures in the field (Abrahams & Millar, 2008).

To address the challenge of observing practicum performance, educational institutions generally distribute conventional multimedia, such as interactive modules and demonstration videos, to accommodate various learning styles. Website-based platforms and dual-channel (visual and auditory) information presentation have been shown to accelerate the assimilation of chemical concepts into students' working memory (Hidayah & Nuraini, 2024; Mayer, 2021). However, these conventional flat-screen media have structural limitations because of their passive communication mechanisms. They cannot validate the accuracy of a user's hand movements; if a learner operates a pipette at an incorrect angle, the interface cannot interrupt to provide instant corrective feedback. Without this real-time intervention, physical motion misconceptions can become permanent motor habits (Wisniewski et al., 2020).

The combination of deep learning and computer vision offers a revolutionary solution to these mechanical evaluation limits. Architecturally, ViLabS uses a Convolutional Neural Network (CNN) algorithm in the TensorFlow framework, enabling real-time, browser-based inference. This system utilizes pose estimation modeling integrated with a cloud-based Teachable Machine ecosystem (Carney et al., 2020). The standard webcam acts as a proactive visual sensor, extracting and tracking the user's anatomical keypoints in milliseconds. By synchronizing the user's actual hand maneuvers with ideal posture datasets pre-trained and hosted in cloud infrastructure, this computational acceleration can achieve strong validation accuracy under strictly controlled laboratory lighting and background conditions. Ultimately, these AI sensors transform passive webcams into proactive, fatigue-free surveillance agents that autonomously validate compliance with Standard Operating Procedures (SOP), an essential layer of supervision previously lacking in experimental chemistry education.

Despite its promising technical capabilities, the downstream integration of computer vision into the pedagogical landscape remains highly fragmented (Zhai et al., 2021). While recent original experimental studies have successfully deployed human pose estimation for automated psychomotor assessments in physical education Wang & Chen (2023); Bakti et al. (2026) and surgical medical training, its specific application in school-level science laboratories remains critically underexplored. Previous experimental research in chemistry contexts, such as the prominent study conducted by Sasaki et al. (2024), has successfully implemented object detection algorithms to automatically recognize chemical experiment procedures and equipment. However, their system is strictly limited to detecting non-living experimental objects and cannot analyze human anatomical execution. The fundamental novelty of ViLabS lies in its paradigm shift from 'object' supervision to 'subject' psychomotor evaluation. ViLabS specifically evaluates the user's kinesthetic precision through continuous temporal pose estimation. By consolidating theoretical modules, interactive assessments, and real-time motion-accuracy validation into a unified smart website platform that acts as an intelligent laboratory assistant, ViLabS bridges the critical gap between passive multimedia and interactive motor skill calibration.

Responding to the limitations of conventional evaluation methods and the urgency for smart educational technology, this study formulates the main research problem: How feasible, practical, and effective is the integration of the deep learning AI and computer vision-based ViLabS platform in improving students' cognitive abilities and psychomotor precision in basic laboratory techniques compared to conventional demonstration methods? Specifically, this research aims to evaluate the architectural feasibility of ViLabS and statistically measure its impact on users. Grounded in the theoretical framework of multimedia learning and AI-assisted motor calibration, the proposed hypotheses are: (H_1) The experimental class utilizing ViLabS demonstrates a significantly higher improvement in cognitive understanding compared to the control class; and (H_2) The autonomous computer vision-based monitoring intervention is significantly more effective in training students' psychomotor precision compared to manual classical demonstrations.

Method

Research Method and Design

This study uses a mixed-methods approach with the Explanatory Sequential Design model Creswell & Clark (2018) to comprehensively integrate quantitative and qualitative analysis. In the quantitative phase (main phase), this study adopted a quasi-experimental non-equivalent control group design (Gopalan et al., 2020). This design was chosen because the study subjects were a whole class group, so individual randomization was not possible. Furthermore, the qualitative (supporting) phase used in-depth interviews to confirm, explain, and triangulate the quantitative metrics from the intervention.

The research design systematically bridges quantitative and qualitative methodologies as detailed in Image 1 below. The procedure initiated with a preliminary phase focused on ViLabS development and expert validation. Next, the quantitative phase used a quasi-experimental approach to measure cognitive shifts with pretest-posttest instruments and evaluate psychomotor precision through AI-generated accuracy logs (experimental) and video-based observer assessments (control). To provide a deeper behavioral context for these numerical findings, the final explanatory qualitative phase deployed semi-structured interviews. This sequential connection ensures that the qualitative thematic analysis directly triangulates and explains the underlying pedagogical mechanisms driving the quantitative cognitive and psychomotor achievements.

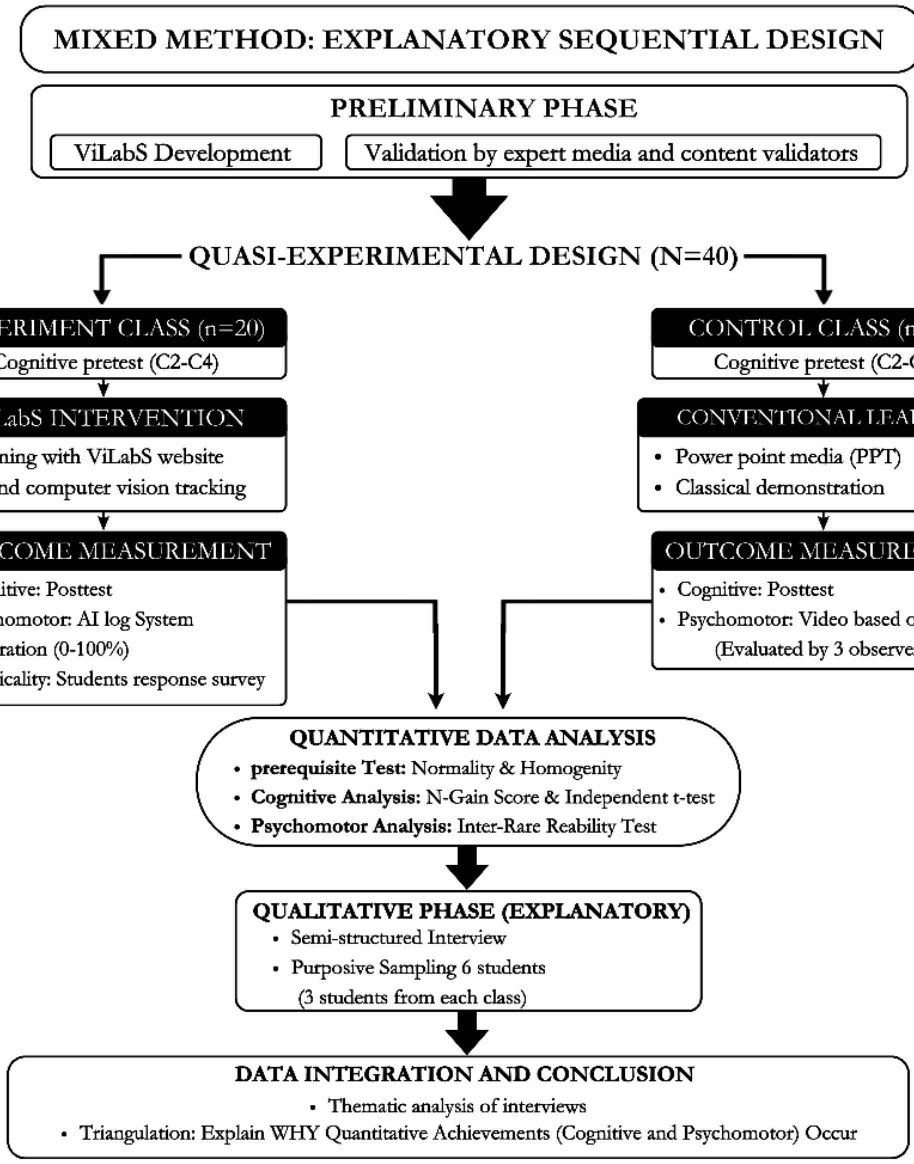

Image 1. Research Design

Research Location and Timeline

This research was conducted during the even semester of the 2025/2026 academic year, spanning from February to June 2026. The comprehensive research timeline encompassed media development, expert validation, field implementation, and data analysis. The limited field trial and primary pedagogical interventions were specifically executed at Thursina International Islamic Boarding School on June 9 and 11, 2026.

Participants and Sampling Techniques

This study involved 40 Grade X students as research subjects. The researchers used purposive sampling based on the equivalence of the two groups' initial academic abilities. The sample was divided into two intact classes: the experimental class (n=20), which received the ViLabS intelligent media intervention, and the control class (n=20), which followed conventional PowerPoint-based learning. Furthermore, for the qualitative explanatory phase, six informants were purposively selected based on their cognitive pretest scores to proportionally represent students with low, medium, and high academic proficiencies.

Instruments Validity and Reliability Test

The feasibility of the ViLabS product is measured through three main aspects: validity, practicality, and effectiveness. Initially, three expert validators (material and educational technology experts) evaluated the media's content and construct validity using a questionnaire sheet, and the results were analyzed descriptively using a Likert Scale. Beyond this human expert appraisal, the technical validity of the AI-based observation instrument was strictly calibrated before the experimental phase. Operating on a MobileNet convolutional neural network architecture a standard robust framework for real-time human pose estimation (Zheng et al., 2023; Suryaningsih & Cahaya, 2026). We developed distinct models in TensorFlow.js for four fundamental laboratory procedures: pipette handling, Erlenmeyer flask handling, beaker handling, and measuring cylinder meniscus reading. To ensure algorithmic reliability, we constructed a perfectly balanced custom dataset for each procedure. This dataset comprised 300 image samples per tool (100 correct posture, 100 incorrect posture, and 100 undetected objects), culminating in a total of 1,200 training images. Following standard machine learning evaluation protocols Tharwat (2021), the training phase utilized specific hyperparameters (50 epochs, a batch size of 16, and a learning rate of 0.001) and an autonomous train-test split that reserved exactly 45 images per model exclusively for objective validation. We then evaluated model performance using standard classification metrics, including accuracy, precision, recall, and F1-score, to empirically confirm the robustness of the automated psychomotor corrective feedback before the field intervention.

In parallel to the AI validation for the experimental class, the psychomotor dimension test for the control class required manual assessment by three independent observers. We tested the reliability of these human observation instruments using inter-rater reliability via Fleiss' Kappa to ensure assessment consistency and reduce inter-observer subjectivity. To measure the media's practicality, we used a student response questionnaire based on the Guttman Scale to objectively capture users' affective responses, which were then aggregated into percentages. Meanwhile, the instrument's effectiveness in the cognitive realm was evaluated through written tests (pretest and posttest). The study calculated students' cognitive improvement using the Normalized Gain (N-Gain) metric. Because the cognitive data were abnormally distributed, the study used the non-parametric Mann-Whitney U test to test for differences in learning-outcome improvement between the experimental and control classes.

Data Collection Techniques and Research Instruments

Data collection proceeded systematically. Following expert validation, both groups completed a cognitive pretest. During the 2-hour intervention, the experimental group utilized ViLabS with real-time AI guidance, while the control group received conventional PowerPoint demonstrations. Post-intervention, cognitive posttests were administered. Psychomotor proficiency was evaluated concurrently: the AI logged the experimental group's accuracy autonomously, whereas the control group submitted video demonstrations evaluated by three independent observers. Finally, practicality questionnaires were distributed, followed by in-depth interviews with six purposively selected students (representing low, medium, and high cognitive abilities) to triangulate and deepen the operational rationale behind their quantitative achievements.

To support these data-collection stages, the research used several meticulously designed instruments. Cognitive outcomes were measured using a 20-item multiple-choice pretest and posttest targeting C2-C4 levels. These questions covered the use and handling of laboratory tools (e.g., dropper pipettes, Erlenmeyer flasks, beakers, measuring cylinders, and test tubes), along with laboratory safety case studies. Both experimental and control classes received identical cognitive evaluation instruments. For the psychomotor domain, the control class submitted video demonstrations of tool use, which three independent observers assessed manually using a dichotomous observation rubric and then converted to ground-truth data by taking the mode across the three observers. Conversely, the log system evaluated the experimental class autonomously. Both manual and AI-based psychomotor assessments strictly focused on four identical assessment indicators: body posture, finger position, the tilt angle of the apparatus, and finger movement accuracy. Table 1 summarizes the research instruments used across all stages.

Table 1. Research Instrument

Instrument	Indicators	Scale
Validation sheet	Truth and suitability of the material, media design, and AI functionality	Likert 1–5
Student response questionnaire	Media display, ease of features, and usefulness	Guttman (Yes/No)
Observation sheet & Log system	Body posture, finger position, tilt angle of the tool, and finger movement accuracy	Dichotomous / Score 0–100
Pre-test & Post-test (multiple choice)	Utility of lab tools, handling procedures, lab safety case studies (Cognitive levels C2–C4)	Score 0–100

Data Analysis Techniques and Criteria

Data analysis techniques integrate statistical computation and qualitative synthesis. Expert validation data are converted and analyzed descriptively as percentages. To quantify these validation results, the experts' assessment scores were measured using a 5-point Likert scale to determine the media's feasibility (Joshi et al., 2015). Table 2 presents the conversion criteria for the validation scores.

Table 2. Likert Scale

Score	Criteria
5	Very good
4	Good
3	Good enough
2	Less good
1	Not good

For cognitive data, we applied the Shapiro-Wilk normality test and the Levene homogeneity test as prerequisites. We measured cognitive ability improvement using the Normalized Gain (N-Gain) formula formulated by Hake (1998) in Equation 1, with interpretation criteria in Table 3.

$$g=\frac{\text{Spost-Spre}}{\text{Smax-Spre}}\times 100\% \quad (1)$$

Table 3. N-Gain score criteria

N-gain Score	Criteria
$g \geq 0.7$	High
$0.7 > g \geq 0.3$	Medium
$g < 0.3$	Low

Since the results of the prerequisite test showed abnormal data distribution, the significance of the mean difference in the N-Gain score between the experimental and control classes was tested using Mann-Whitney U non-parametric statistics (Field, 2018; Schober & Vetter, 2020). Furthermore, practicality responses were measured using the Guttman Scale DeVellis & Thorpe (2021) to capture the objectivity of affection without ambiguity.

Table 4. Guttman Scale

Respons	Answer	Score
Negative	Yes	0
	No	1
Positive	Yes	1
	No	0

$$\text{Practicality Percentage}=\frac{\text{total score for each statement}}{\text{number of respondent}}\times 100\% \quad (2)$$

Table 5. Practicality Criteria Percentage

Percentage (%)	Criteria
81-100	Very good
61-80	Good
41-60	Good enough
21-40	Less good
0-20	Not good

For the control class psychomotor evaluation, we assessed the stability of the three observers using inter-rater reliability via Fleiss' Kappa (Gwet, 2014). To overcome the granularity mismatch between the human dichotomous rubric and the continuous decimal percentage output of computer vision, this study applied the Item-Rater Pooling composite technique (Campagner et al., 2021; Seery et al., 2017). The three observers' scores on the four instruments were combined into a single ground-truth dataset using the mode across the three observers, which was then compared (apple-to-apple) with the ViLabS deep learning agent's accuracy metrics using the Mann-Whitney U Test (Sasaki et al., 2024). Finally, the selected informants' interview transcripts were reduced through thematic analysis. To ensure the trustworthiness of the qualitative findings, this study rigorously employed methodological triangulation and member checking. The interview transcripts were cross-referenced with the empirical psychomotor AI logs to verify behavioral consistency, and the summarized thematic findings were presented back to the informants to confirm the credibility and confirmability of their articulated learning experiences.

Product Specifications

ViLabS is engineered as a comprehensive digital workspace encompassing theoretical modules, interactive evaluations, and laboratory simulations. Built on pure HTML, the platform seamlessly integrates the CNN-based computer vision architecture and is powered by the Teachable Machine ecosystem. This design enables standard device webcams to autonomously detect and evaluate users' anatomical movements in real-time, eliminating the need for high-end computing hardware.

Results and Discussion

The primary objective of this study was to evaluate the feasibility, practicality, and effectiveness of integrating the ViLabS smart website into basic chemistry laboratory instruction. To provide a systematic and robust analysis, the research findings are presented sequentially, aligned with the explanatory mixed-methods design. The evaluation began with rigorous technical calibration of the AI models, followed by pedagogical feasibility appraisals by human experts and practicality assessments by target users. Subsequently, the quantitative cognitive and psychomotor impacts of the intervention were statistically analyzed, culminating in a qualitative thematic triangulation to elucidate the underlying pedagogical mechanisms. Table 6 comprehensively summarizes the initial findings, including the machine's technical performance (AI) alongside appraisals from human experts and target users.

Table 6. Summary of Feasibility and Practicality Assessments

Assessment Aspect	Sub-Indicators	Score / Percentage	Category
Content Validity	Curriculum Alignment, Concept Accuracy, Quality of Evaluation, Language	5	Very Good
Construct Validity	Interface Design, Ease of Use, Interactivity, Learning Appeal	4	Good
Technical Validity (AI Performance)	Confusion Matrix	Zero Misclassification	Flawless
Practicality (User Response)	Accuracy, Precision, Recall, F1-Score Validation Loss	1.00 (100%) < 0.17	Flawless Converged
	Media View	92.50%	Very Good
	Ease of Use Benefits	92.50% 96.25%	Very Good Very Good

As shown in Table 6, the AI's internal evaluation metrics demonstrated extreme convergence. Across all four procedural models, the confusion matrices showed zero misclassifications, yielding a stable 1.00 (100%) across accuracy, precision, recall, and F1-score, with validation losses dropping below 0.17. From a computer vision standpoint, this 100% metric must be interpreted with high caution. It reflects an artificial ceiling effect symptomatic of overfitting, driven by the highly controlled, uniform lighting and static backgrounds of the data acquisition environment. While these metrics confirm that the system functioned robustly as an intervention tool in this specific, constrained quasi-experiment, future deployments in dynamic classroom environments require significant dataset diversification to achieve true real-world generalizability.

Following technical calibration, human experts and target users evaluated the pedagogical feasibility and practicality of the ViLabS media. Table 6 summarizes the assessment results. Expert validation yielded a mode score of 5 (Very Good) for content validity, indicating that the procedural visualizations and cognitive questions align precisely with chemical science principles. Construct validity had a mode score of 4 (Good), confirming intuitive interface navigation. Furthermore, the user practicality survey showed excellent technological acceptance (>92%), indicating that the user-centered design minimized technical barriers and allowed students to operate the AI features independently (Plomp & Nieveen, 2007).

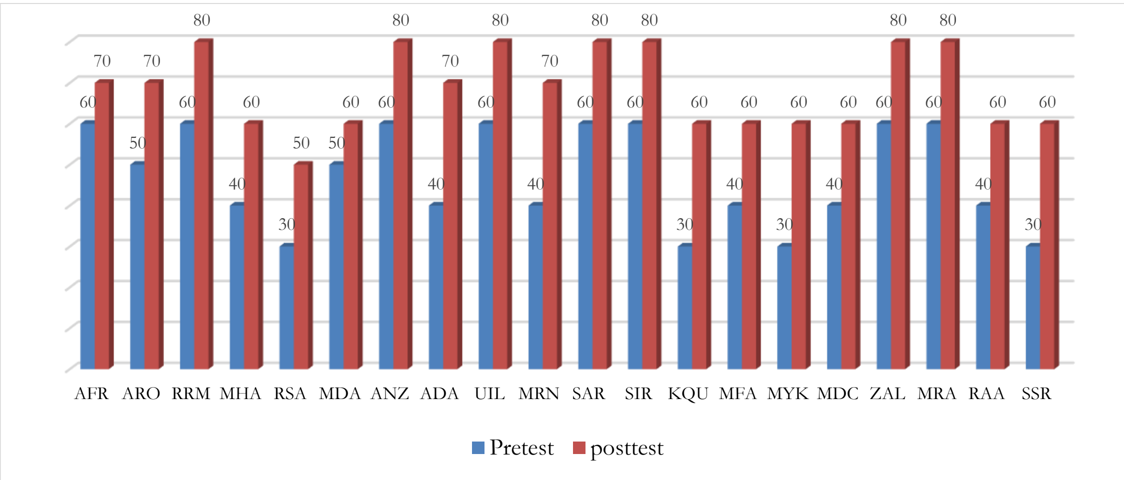

Image 2. Pretest and Posttest Scores of the Control Class

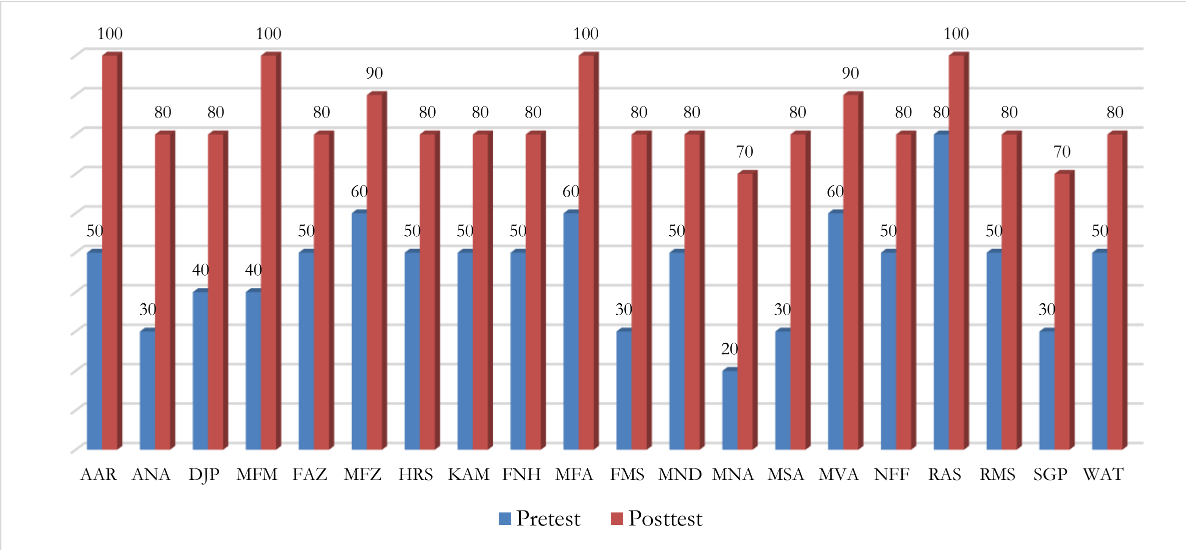

Image 3. Pretest and Posttest Scores of the Experimental Class

To map whether each learning method meaningfully increases understanding, Images 2 and 3 visually represent the distribution of individual pretest and posttest scores in both classes.

Moving from product feasibility to pedagogical effectiveness, the first step in the quasi-experimental design was to ensure no sample selection bias. This visual representation reveals a stark contrast in data distribution. While the control class showed moderate, varied improvement, the experimental class showed a consistent, significant spike, with most learners reaching the optimal cognitive range (80-100) post-ViLabS. Alongside the cognitive evaluation, we meticulously recorded students' psychomotor proficiency in basic laboratory techniques. For full transparency, Image 4 and Image 5 show the distribution of final psychomotor scores for both classes.

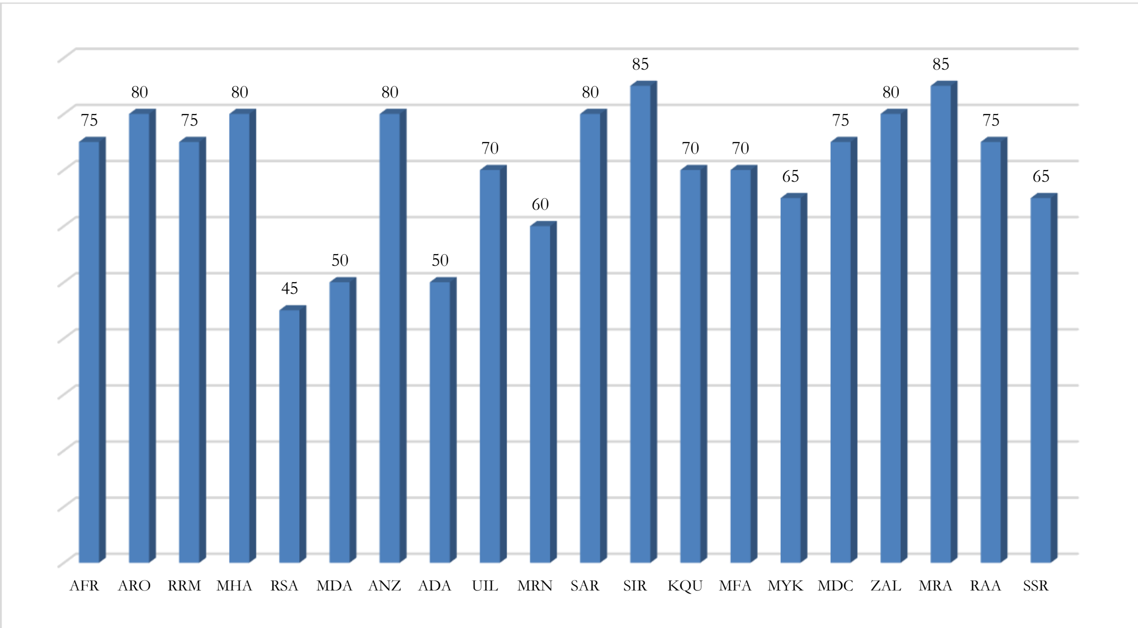

Image 4. Psychomotor Score Distribution of the Control Class

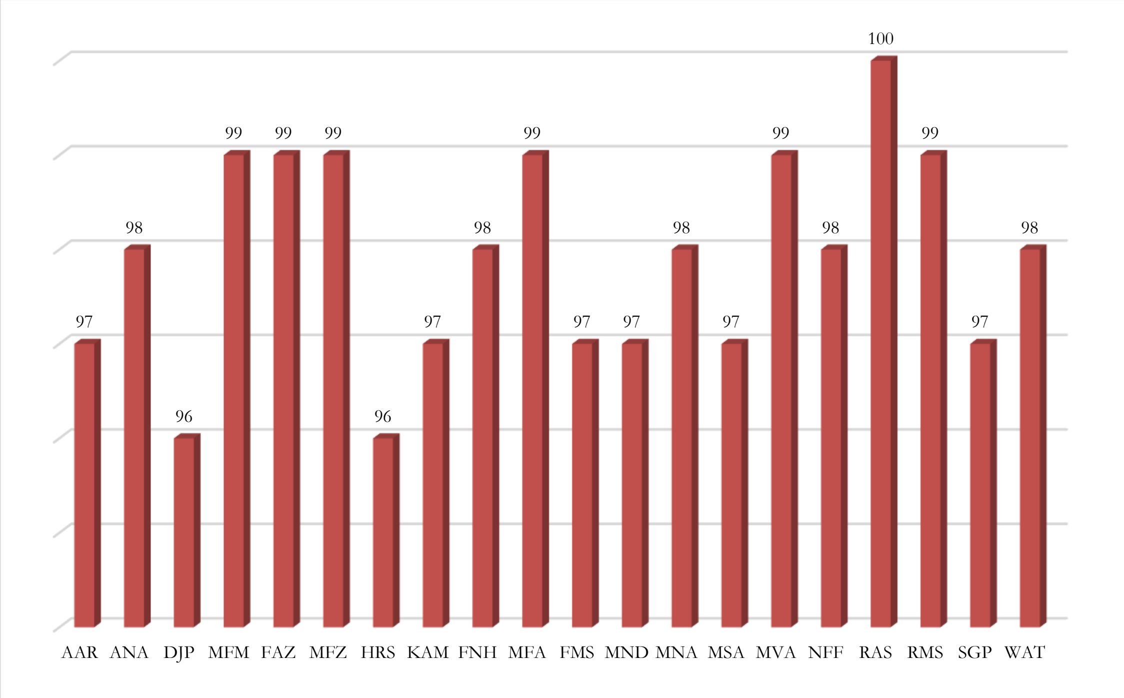

Image 5. Psychomotor Score Distribution of the Experimental Class

The psychomotor data for the experimental class were automatically extracted from the ViLabS system logs, recorded in real time during their practical sessions. Conversely, the psychomotor scores for the control class were derived through manual video-based observations. Three independent observers evaluated the students' movements using a strict dichotomous observation rubric. To reduce human subjectivity and establish a robust baseline, the final scores for the control class were aggregated by taking the mode (majority agreement) of the three observers' ratings, serving as the ground-truth data (Campagner et al., 2021; Seery et al., 2017). To rigorously quantify these visual trends and present the data scientifically, Table 7 systematically consolidates the comparative magnitude of the cognitive increase (N-Gain) and the prerequisite inferential statistical tests for both cognitive and psychomotor domains.

Table 7. Summary of Statistical Analysis for Cognitive and Psychomotor Domains

Variables & Tests	Control Class (PPT)	Experimental Class (ViLabS)	Test Statistic	p-value	Conclusion
Prerequisite Test					
Normality (Shapiro-Wilk)	Not Normal	Normal	-	-	Non-parametric required
Pretest Equivalence					
Mann-Whitney U Test	Mean Rank: 21.05	Mean Rank: 19.95	Z = -0.236	0.813	No significant initial difference
Cognitive Improvement					
Pre-Post Paired (Wilcoxon)	Pre: 47 → Post: 68.5	Pre: 46.5 → Post: 84	-	< 0.001	Significant improvement in both

Variables & Tests	Control Class (PPT)	Experimental Class (ViLabS)	Test Statistic	p-value	Conclusion
Average N-Gain Score	0.41 (Medium)	0.71 (High)	-	-	-
N-Gain (Mann-Whitney U Test)	Mean Rank: 10.50	Mean Rank: 30.50	Z = -5.481	< 0.001	H_1 Accepted (Exp > Control)
Psychomotor Evaluation					
Observer Reliability	Fleiss' Kappa: 0.887	N/A (AI Monitored)	Z = 30.7	< 0.001	'Very Strong' Agreement
Score (Mann-Whitney U)	Mean Rank: 10.50	Mean Rank: 30.50	Z = -5.443	< 0.001	H_2 Accepted (Exp > Control)

As detailed in Table 7, the initial cognitive equivalence between the two groups was established ($p = 0.813$). Post-intervention, the experimental class achieved a High N-Gain category (0.71), significantly outperforming the control group (0.41) with $p < 0.001$. This superiority shows that the intelligent laboratory assistant effectively helps learners solve high-level analytical evaluation instruments (Chiu & Cheng, 2023). This significant cognitive spike perfectly aligns with the theoretical implications of information processing. In the control class, reliance on PowerPoint media triggered passive observation, leading to cognitive overload as learners struggled to bridge spatial gaps without interactive intermediaries (Mayer, 2021). In contrast, ViLabS overcomes these static limitations by simultaneously representing the three essential chemical domains (Liu et al., 2023; Yaman, 2020). This integration activates dual memory channels, facilitating the seamless transfer of concepts into long-term memory (Slavin, 2014).

Furthermore, in the psychomotor domain, after establishing a 'Very Strong' ground truth via inter-rater reliability (Fleiss' Kappa = 0.887), the Mann-Whitney test confirmed that ViLabS autonomous monitoring was significantly more effective than manual demonstrations ($p < 0.001$). These advantages stem from precise anatomical tracking. While conventional practices lack a mirror reference to validate maneuvers, the ViLabS computer vision agent provides instantaneous feedback, emitting an instant alert when it detects a posture anomaly (Sasaki et al., 2024). This machine-supervised experience guides students to autonomously perform motor calibration repeatedly, leading to the formation of precise muscle memory and strict adherence to standard operating procedures.

To fully justify the experimental class's quantitative superiority, we conducted a qualitative phase. We analyzed semi-structured interview transcripts from six students purposively selected to represent low (R), medium (S), and high (T) cognitive abilities from both the experimental (SE) and control (SK) groups. This data reduction identified two central themes validating the intervention's success.

The first theme highlights the severe limitations of static representation and motor impasse in conventional learning. Although control-class students recognized the urgency of practical skills ("so that later when practicum in the lab we are not confused," SK-S), they experienced acute constraints of abstraction. All cognitive levels heavily criticized PowerPoint's inability to dynamically visualize kinesthetic procedures. Informant SK-R (low cognitive) stated, "*It is impossible to understand because we must know how it is practiced*," while SK-S (medium cognitive) added, "*Theory* *without practice will not be maximized.*" Strikingly, even the high-cognitive informant (SK-T) admitted to fundamental difficulties, specifically in "*reading the measuring cylinder and holding the dropper pipette,*" because they only received flat visual inputs. This phenomenon aligns with Mayer's (2021) cognitive theory of multimedia learning, which argues that static media fails to activate dual memory channels effectively for complex spatial-motor tasks, explaining the cognitive overload and lower N-Gain in the control class.

Conversely, the second theme uncovers the central role of real-time AI validation as a robust pedagogical scaffold. In the experimental class, students autonomously eliminated difficulties with instrument manipulation across all student stratifications. Informant SE-R (low cognitive) explicitly affirmed, "*The difficulty in holding the tools correctly is overcome because there is a demonstration video and an AI that judges right or wrong.*" This technological novelty was heavily favored, with informant SE-T (high cognitive) highlighting the "AI scanner" as the ultimate feature for detecting posture accuracy. Most importantly, this sensory-motor adaptation led to a significant affective transformation: self-efficacy. Informant SE-S (medium cognitive) projected mature psychological readiness, stating, "*I am more capable... because I have been assessed with the AI scanner so I know whether it is right or wrong.*" This confirms Wisniewski et al. (2020) meta-analysis on educational feedback, proving that instantaneous, temporal corrective intervention is the most critical variable in forming precise muscle memory and breaking down abstraction barriers. Image 6 presents photographic evidence of the students' practical activities using the ViLabS autonomous tracking system, alongside the in-depth interview sessions.

Image 6. Photographic evidence of students' practical activities using ViLabS and the in-depth interview sessions

Through this cross-methodological triangulation, the statistical divergence is phenomenologically confirmed. The integration of ViLabS succeeds not merely as a digital interface replacement, but as an active supervisory agent capable of independently orchestrating cognitive mastery, psychomotor reliability, and student self-efficacy in chemical laboratory techniques.

Synthesizing these findings, the fundamental novelty of this research lies in its paradigm shift from conventional 'object-based' supervision to interactive 'subject-based' psychomotor evaluation. While prominent prior studies, such as Sasaki et al. (2024), used computer vision to recognize chemical equipment, their systems were limited to non-living objects and could not evaluate human anatomical execution. ViLabS bridges this gap by directly evaluating the user's kinesthetic precision through continuous pose estimation. This paradigm perfectly aligns with recent discourse in Artificial Intelligence in Education (AIED), which emphasizes the urgent transition from passive automated assessments to proactive, multimodal pedagogical AI agents (Chiu et al., 2023). Furthermore, from a computer vision perspective, the real-time autonomous correction executed by ViLabS successfully extends current human action recognition frameworks Kong & Fu (2022) into practical, automated educational scaffolding. Theoretically, these findings solidify Mayer's (2021) Cognitive Theory of Multimedia Learning by demonstrating that AI-driven interactive interfaces significantly reduce extraneous cognitive load compared to passive media. Practically, integrating this intelligent agent has a major impact on the field: it autonomously alleviates teachers' supervisory burden in large science classes, mitigates the risk of laboratory accidents due to procedural errors, and gives students a safe, corrective environment to calibrate their motor skills before physical experiments.

Despite the highly significant outcomes, this study acknowledges several methodological limitations. First, the current ViLabS model accommodates only four fundamental laboratory instruments (e.g., dropper pipettes, Erlenmeyer flasks, beakers, measuring cylinders, and test tubes). Second, as a cloud-based web platform, its real-time inference relies heavily on a stable internet connection. Third, the system's flawless validation accuracy was achieved under controlled laboratory lighting, making it potentially susceptible to false positives in dimly lit or highly heterogeneous environments. Finally, the intervention was limited to a sample of 40 students. Therefore, to support broader pedagogical development, future research should expand the AI dataset to include more complex laboratory techniques. Technologically, transitioning this web-based architecture into an offline mobile application (Android/iOS) is highly recommended to resolve internet dependency. Additionally, augmenting the machine learning training dataset with diverse lighting conditions will significantly enhance the model's detection sensitivity for everyday classroom environments.

Conclusion

This study concludes that the deep learning-based Virtual Laboratory Assistant (ViLabS) is highly feasible, practical, and effective in significantly improving students' cognitive understanding and psychomotor precision in basic chemistry laboratory techniques. The autonomous AI corrective feedback minimizes extraneous cognitive load, eliminates micro-movement anomalies, and fosters students' self-efficacy before actual practicums. However, the system's current efficacy is limited by specific technical and methodological constraints, including dependence on stable internet connectivity, standard webcam positioning, uniform laboratory lighting, and a small sample size, which may constrain the model's generalizability across diverse users. Based on these critical findings, it is highly recommended that future researchers address these technical boundaries by augmenting training datasets with diverse lighting conditions and camera angles, transitioning the web-based architecture into offline mobile applications to resolve hardware and connectivity dependencies, and expanding the evaluation to larger, more heterogeneous demographics with more complex chemical procedures.

Bibliography

Abrahams, I., & Millar, R. (2008). Does practical work really work? A study of the effectiveness of practical work as a teaching and learning method in school science. *International Journal of Science Education*, *30*(14), 1945–1969. https://doi.org/10.1080/09500690701749305

Bakti, I., Winarti, A., Mirah Surya Dewi, N. P. P., Mahdian, & Eka Sari, S. M. (2026). Profiling readiness for deep learning: A study of chemistry students' multiple intelligences and learning reflections. *Indonesian Journal of Educational Development (IJED)*, *6*(4), 1526–1541. https://doi.org/10.59672/ijed.v6i4.5400

Campagner, A., Ciucci, D., & Cabitza, F. (2021). Ground truthing from multi-rater labeling with three-way decision and possibility theory. *Information Sciences*, *545*, 771–790. https://doi.org/10.1016/j.ins.2020.09.049

Carney, M., Webster, B., Alvarado, I., Phillips, K., Howell, N., Griffis, J., Roberts, J., Barman, P. R., & Chen, S. Y. (2020). Teachable machine: Approachable web-based machine learning.

Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems, 1–8. https://doi.org/10.1145/3334480.3382839
Celik, I. (2023). Towards intelligent-TPACK: An empirical study on teachers' professional knowledge to ethically integrate artificial intelligence (AI)-based tools into education. *Computers in Human Behavior*, *138*, 107468. https://doi.org/10.1016/j.chb.2022.107468
Chen, L., Chen, P., & Lin, Z. (2020). Artificial intelligence in education: A review. *IEEE Access*, *8*, 75264–75278. https://doi.org/10.1109/ACCESS.2020.2988510
Cheung, D. (2011). Teacher beliefs about implementing guided-inquiry laboratory experiments for secondary school chemistry. *Journal of Chemical Education*, *88*(11), 1462–1468. https://doi.org/10.1021/ed1008409
Chiu, W. K. (2021). Pedagogy of emerging technologies in chemical education. *Education Sciences*, *11*(11), 709. https://doi.org/10.3390/educsci11110709
Chiu, T. K. F., Xia, Q., Zhou, X., Chai, C. S., & Cheng, M. (2023). Systematic literature review on opportunities, challenges, and future research recommendations of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *4*, 100118. https://doi.org/10.1016/j.caeai.2022.100118
Creswell, J. W., & Plano Clark, V. L. (2018). *Designing and conducting mixed methods research* (3rd ed.). SAGE.
Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. *International Journal of Educational Technology in Higher Education*, *20*(1), 22. https://doi.org/10.1186/s41239-023-00392-8
DeVellis, R. F., & Thorpe, C. T. (2021). *Scale development: Theory and applications* (5th ed.). SAGE Publications.
Field, A. (2018). *Discovering statistics using IBM SPSS Statistics* (5th ed.). SAGE Publications.
Fukuda, K. (2020). Science, technology and innovation policy for Society 5.0. *Journal of Open Innovation: Technology, Market, and Complexity*, *6*(2), 47. https://doi.org/10.1016/j.ijpe.2019.07.033
Gopalan, M., Rosinger, K., & Ahn, J. B. (2020). Use of quasi-experimental research designs in education research: Growth, promise, and challenges. *Review of Research in Education*, *44*(1), 218–243. https://doi.org/10.3102/0091732X20903302
Gwet, K. L. (2014). *Handbook of inter-rater reliability: The definitive guide to measuring the extent of agreement among raters* (4th ed.). Advanced Analytics, LLC.
Hake, R. R. (1998). Interactive-engagement versus traditional methods: A six-thousand-student survey of mechanics test data for introductory physics courses. *American Journal of Physics*, *66*(1), 64–74. https://doi.org/10.1119/1.18809
Hariyono, H., Hidayati, U., & Suharto. (2026). The effect of AI-optimized project-based learning on digital pedagogical readiness. *Indonesian Journal of Educational Development*, *7*(1), 49–64. https://doi.org/10.59672/ijed.v7i1.6145
Hidayah, R., & Nuraini, S. (2024). The effectiveness of website on improving students' creative thinking skills in chemical bonding. *AIP Conference Proceedings*, *3116*(1), 100014. https://doi.org/10.1063/5.0210532
Holmes, W., & Tuomi, I. (2022). State of the art and practice in AI in education. *European Journal of Education*, *57*(4), 542–570. https://doi.org/10.1111/ejed.12533
Hwang, G. J., Xie, H., Wah, B. W., & Gašević, D. (2020). Vision, challenges, roles and research issues of artificial intelligence in education. *Computers and Education: Artificial Intelligence*, *1*, 100001. https://doi.org/10.1016/j.caeai.2020.100001
I Made Elia Cahaya, I. M. E., Suryaningsih, N. M. A., Parwata, I. M. Y., Poerwati, C. E. (2024). The influence of a guided inquiry learning model in improving students' creative thinking abilities: A meta-analysis study. (2024). *Indonesian Journal of Educational Development (IJED)*, *5*(3), 376–384. https://doi.org/10.59672/ijed.v5i3.4211

Joshi, A., Kale, S., Chandel, S., & Pal, D. K. (2015). Likert scale: Explored and explained. *British Journal of Applied Science & Technology*, *7*(4), 396–403. https://doi.org/10.9734/BJAST/2015/14975

Kyynäräinen, R., Vilhunen, E., Li, P.-H., Laakso, M.-J., & Vesterinen, V.-M. (2024). How making mistakes shapes students' situational engagement in chemistry laboratory?. *International Journal of Science Education*, 1-21. https://doi.org/10.1080/09500693.2024.2439142

Kong, Y., & Fu, Y. (2022). Human action recognition and prediction: A survey. *International Journal of Computer Vision*, *130*(5), 1366–1401. https://doi.org/10.1007/s11263-022-01594-9

Liu, Q., Ma, J., Yu, S., Wang, Q., & Xu, S. (2023). Effects of an augmented reality-based chemistry experiential application on student knowledge gains, learning motivation, and technology perception. *Journal of Science Education and Technology*, *32*(2), 153–167. https://doi.org/10.1007/s10956-022-10014-z

Mahsup, Syaharudin, Nepomuceno, M., & Jali, H. (2026). Effectiveness of guided inquiry learning and self-regulated learning on students' numeracy skills. *Indonesian Journal of Educational Development (IJED)*, *6*(4), 1427–1440. https://doi.org/10.59672/ijed.v6i4.5756

Mayer, R. E. (2021a). Evidence-based principles for how to design effective instructional videos. *Journal of Applied Research in Memory and Cognition*, *10*(2), 229–240. https://doi.org/10.1016/j.jarmac.2021.03.007

Mayer, R. E. (2021b). *Multimedia learning* (3rd ed.). Cambridge University Press. https://doi.org/10.1017/9781316941355

Ouyang, F., & Jiao, P. (2021). Artificial intelligence in education: The three paradigms. *Computers and Education: Artificial Intelligence*, *2*, 100020. https://doi.org/10.1016/j.caeai.2021.100020

Plomp, T., & Nieveen, N. (2007). *An introduction to educational design research*. SLO.

Popenici, S. A. D., & Kerr, S. J. (2017). Exploring the impact of artificial intelligence on teaching and learning in higher education. *Research and Practice in Technology Enhanced Learning*, *12*(1), 22. https://doi.org/10.1186/s41039-017-0062-8

Sasaki, R., Fujinami, M., & Nakai, H. (2024). Application of object detection and action recognition toward automated recognition of chemical experiments. *Digital Discovery*, *3*(1), 2458–2469. https://doi.org/10.1039/D4DD00015C

Schechtel, S., & Bongers, A. (2026). The faces of failure: Understanding students' affective experiences with failure in introductory chemistry laboratory learning activities. *Chemistry Education Research and Practice*, *27*(1), 479–493. https://doi.org/10.1039/d5rp00297d

Schober, P., & Vetter, T. R. (2020). Nonparametric statistical methods in medical research. *Anesthesia & Analgesia*, *131*(6), 1862–1863. https://doi.org/10.1213/ANE.0000000000005101

Seery, M. K., Agustian, H. Y., Doidge, E. D., Kucharski, M. M., Owen, S., & Shirran, A. (2017). Developing laboratory skills by incorporating peer-review and digital badges. *Chemistry Education Research and Practice*, *18*(2), 403–419. https://doi.org/10.1039/C7RP00003K

Seery, M. K. (2020). Establishing the laboratory as the place to learn how to do chemistry. *Journal of Chemical Education*, *97*(3), 666–672. https://doi.org/10.1021/acs.jchemed.9b00764

Slavin, R. E. (2014). *Educational psychology: Theory and practice* (11th ed.). Pearson.

Suryaningsih, N. M. A. ., & Cahaya, I. M. E. . (2026). Enhancing early childhood science literacy through deep learning–based instruction using simple coding games . *Indonesian Journal of Educational Development (IJED)*, *7*(1), 230–243. https://doi.org/10.59672/ijed.v7i1.5924

Tharwat, A. (2021). Classification assessment methods. *Applied Computing and Informatics*, *17*(1), 168-192. https://doi.org/10.1016/j.aci.2018.08.003

Wang, X., & Chen, Y. (2023). Deep learning-based human pose estimation for real-time action recognition in physical education. *Computers & Education*, *198*, 104770. https://doi.org/10.1016/j.engappai.2022.105813

Weaver, M. G., Samoshin, A. V., Lewis, R. B., & Gainer, M. J. (2016). Developing students' critical thinking, problem solving, and analysis skills in an inquiry-based synthetic organic

laboratory course. *Journal of Chemical Education*, *93*(5), 847–851. https://doi.org/10.1021/acs.jchemed.5b00678
Widana, I. W., Sopandi, A. T., Suwardika, I. G. (2021). Development of an authentic assessment model in mathematics learning: A science, technology, engineering, and mathematics (STEM) approach. *Indonesian Research Journal in Education, 5*(1), 192-209. https://doi.org/10.22437/irje.v5i1.12992
Williams, S. R., Ziebell, A. L., Kitson, R. R. A., Coppo, P., Thompson, C. D., & Overton, T. L. (2018). 'What do you think the aims of doing a practical chemistry course are?' A comparison of the views of students and teaching staff across three universities. *Chemistry Education Research and Practice*, *19*(2), 463–473. https://doi.org/10.1039/C7RP00177K
Wisniewski, B., Zierer, K., & Hattie, J. (2020). The power of feedback revisited: A meta-analysis of educational feedback research. *Frontiers in Psychology*, *10*, 3087. https://doi.org/10.3389/fpsyg.2019.03087
Yaman, F. (2020). Pre-service science teachers' development and use of multiple levels of representation and written arguments in general chemistry laboratory courses. *Research in Science Education*, *50*(6), 2331–2362. https://doi.org/10.1007/s11165-018-9781-0
Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of AI applications in higher education. *International Journal of Educational Technology in Higher Education*, *16*(39), 1–27. https://doi.org/10.1186/s41239-019-0171-0
Zhai, X., Chu, X., Chai, C. S., Jong, M. S. Y., Idrus, A., & Li, Y. (2021). A review of artificial intelligence (AI) in education from 2010 to 2020. *Complexity*, *2021*, 1–18. https://doi.org/10.1155/2021/8812542
Zheng, C., Wu, W., Yang, T., Zhu, S., Chen, C., Liu, R., Yuan, J., Zou, C., & Shah, M. (2023). Deep learning-based human pose estimation: A survey. *ACM Computing Surveys*, *56*(1), 1–37. https://doi.org/10.1145/3603618