International Journal of Learning, Teaching and Educational Research
Vol. 25, No. 5, pp. 227-251, May 2026
https://doi.org/10.26803/ijlter.25.5.11
Received Feb 28, 2026; Revised Apr 20, 2026; Accepted Apr 22, 2026

AI-Supported Authentic Assessment in Science Education: Overcoming Logistical Barriers and Enhancing Outcomes

Marwan Abualrob* **and Deema Ghannam**
Arab American University,
Palestine

Abstract. As modern education shifts toward 21st-century skills, practical application often struggles to keep pace. To support the implementation of Authentic Assessment (AA), this study investigates its impact on grade 9 science students' achievement and engagement in Palestine, relative to Traditional Assessment (TA), while exploring the role of generative AI as a teacher-support tool. Adopting an explanatory sequential mixed-methods design with a purposive sample of 59 female students, the study utilized quantitative testing (Mann-Whitney *U*) alongside qualitative thematic analysis (interviews, focus groups, and structured teacher reflections). To ensure rigor and replicability, a prompt engineering strategy was used alongside blind grading to reduce teacher bias. The results indicated that the AA group significantly outperformed the TA group in academic achievement ($p = 0.007$) while also displaying higher cognitive, behavioral, and emotional engagement. Qualitatively, the findings revealed that students learned concepts more deeply by creating tangible products to present to the class rather than through memorization. In addition, teacher reflections revealed that implementing AA posed significant logistical and time-related challenges, particularly in rubric construction and instructional planning. Importantly, Artificial Intelligence (AI) helped overcome these obstacles by simplifying the design of rubrics, clarifying performance criteria, and supporting lesson preparation. The study concludes that teacher-mediated, AI-supported AA offers a practical model for enhancing educational quality and student agency, proposing a scalable solution for settings with logistical constraints, although further studies are needed to extend these findings beyond this teacher- and gender-specific context.

Keywords: authentic assessment; generative AI; student engagement; academic achievement; instructional design; teacher-support tools

Citation:
Abualrob, M., & Ghannam, D. (2026). AI-Supported Authentic Assessment in Science Education: Overcoming Logistical Barriers and Enhancing Outcomes. *International Journal of Learning, Teaching and Educational Research, 25(5)*, 227–251. https://doi.org/10.26803/ijlter.25.5.11

*Corresponding author: Marwan Abualrob; *marwan.abualrob@aaup.edu*

1. Introduction

It is widely accepted that AA plays a key role in preparing students for future learning that aligns with the demands of work and everyday life (Villarroel et al., 2018; Sokhanvar, Salehi, & Sokhanvar, 2021). Historically, this concept emerged in the early 1990s as a necessary counterbalance to standardized testing in U.S. secondary schools, with the central aim of "directly examining student performance on worthy intellectual tasks" (Wiggins, 1990, p. 1). While these principles provide the pedagogical justification for AA, recent research has further highlighted its importance in fostering 21st-century skills (Vlachopoulos & Makri, 2024). Nevertheless, for decades, performance measurement within educational systems has largely centered on standardized tests, assignments, and grades. Despite their practical advantages, these assessment tools have been widely criticized for their limited ability to reflect students' overall growth and actual abilities (Joseph, 2025).

TA systems have historically favored memorization practices and standardization, to the detriment of problem-solving, critical thinking, and the application of knowledge in real-world settings (Joseph, 2025). Studies have shown that traditional paper-and-pencil tests have not been effective in determining how well learners have internalized knowledge (Colley, 2008), nor have they provided a fair representation of learners' mastery of complex skills like problem-solving (Amrein & Berliner, 2002; Herrera, Murry, & Cabral, 2013; Volante, 2004). As a result, "teaching to the test" has become a common practice in schools, narrowing students' learning experiences to lower-level skills (Resnick, 1996). In the Palestinian context, studies have indicated a similar tendency toward TA practices, with science teachers often emphasizing rote learning and basic cognitive objectives (Diab, 2005).

In this context, moving beyond TA becomes essential. This need has been further underscored by the growing availability of AI systems among students (Papanastasiou et al., 2025). As educators who witness the widening gap between TA methods and students' abilities, we believe that this is vital for promoting genuine learning. While this technological accessibility poses a challenge to traditional testing, this study shifts the focus by investigating how teachers can use generative AI to overcome the historical barriers of implementing AA. Specifically, the study focuses on AI as a pedagogical support tool for teachers in designing rigorous AA tasks, rather than examining students' direct use of AI to complete these tasks.

Moreover, this shift requires a greater focus on the skills students need to operate effectively within a "more digital world, where technology not only enhances the assessment process but also affects the students' online identity and future employability " (Nieminen et al., 2023). AI has thus become a major driver that extends beyond TA boundaries, enabling more dynamic, personalized, and inclusive approaches to evaluation (Joseph, 2025). Prior research has indicated that generative AI can support teachers in developing instructional materials, assessments, and recommendations (Chen et al., 2020a), as well as facilitating the evaluation of students' work (Lo, 2023) and creating AA rubrics (Abualrob, 2025).

These tools have helped address long-standing challenges faced by teachers, including the substantial time and effort required to design assessment criteria and construct evaluation tools (Burden & Byrd, 2018). By reducing these practical barriers, AI serves as an effective aid in translating pedagogical practices from theory into practice.

Moreover, the integration of Performance-Based Assessment (PBA) represents an approach that embeds problem-solving processes within the assessment methodology itself (Messick, 1994; Torrance, 1995). PBA requires students to complete tangible tasks—such as producing a product or responding to an open-ended question—from which measurable indicators can be derived to assess their progress (Aladini et al., 2024). In parallel, self-assessment (SA) and peer assessment (PA) practices offer valuable learning opportunities for developing self-regulated and co-regulated learning, and they may contribute to improved academic performance (Yan & Boud, 2021; Yan et al., 2022). Most importantly, the efficiency of such sophisticated evaluation methods relies heavily on clearly defined guidelines and criteria—a task that can be effectively supported by the advanced capabilities of generative AI. By engaging students in active and reflective roles in understanding and applying assessment criteria, these approaches support the achievement of important educational goals (Andrade & Heritage, 2018; Harris & Brown, 2018).

Although the potential benefits of AA have been identified, its widespread application is still hampered by implementation challenges. To help close this research gap, this paper presents a new perspective on AI by shifting the perspective from the student to the teacher. Regarding the logistical facilitation of AA, there remains a distinct lack of studies examining the application of AI exclusively as a pre-instructional design tool. Therefore, the present study seeks to compare students' science performance on hands-on and TA tasks (Chi et al., 2021), while exploring how teacher-mediated, AI-supported AA may provide personalized learning experiences that respond to students' needs and prepare them for an uncertain future (Bearman et al., 2023; van den Berg & du Plessis, 2023).

2. Literature Review

2.1 Theoretical Foundations of AA

The alignment between Social Constructivism (SC) and Self-Determination Theory (SDT) forms the foundation of AA. Specifically, from a social constructivist perspective, educators do not view learning as the passive reception of information, but rather as an active process in which students construct knowledge by engaging in tasks situated within a realistic social context (Vygotsky, 1978; Vlachopoulos & Makri, 2024). In this framework, assessment transforms from a tool for measuring memorization into "Assessment for Learning," where tangible products serve as mediating tools to deepen understanding. In parallel, SDT provides the psychological framework for understanding student motivation; Ryan and Deci (2000) posit that learning environments that satisfy students' needs for autonomy and competence significantly enhance engagement and intrinsic motivation. Accordingly, AA

represents a practical application of these theories, offering students the opportunity for choice and creativity (autonomy) within tasks that challenge their abilities and allow them to demonstrate mastery (competence) in a collaborative environment (Zhan et al., 2025). However, applying such ideals in daily classroom practice requires sophisticated task design and feedback, which pose considerable logistical challenges for educators.

2.2 The Shift toward AA and PBA

AA differs significantly from traditional forms of assessment (Fawns et al., 2025). It aims to move students beyond repeating isolated facts by encouraging deeper understanding, knowledge construction, and connections between learning and real-life situations. Furthermore, scholars have heavily stressed the development of 21st-century skills—such as critical thinking, problem-solving, and communication (Villarroel et al., 2018; Vlachopoulos & Makri, 2024). As a result, educators outline the cognitive demands of AA along the lines of knowledge construction, professional skills, and 21st-century skills.

Educators view PBA as a practical method within AA rather than a separate evaluative concept. Through PBA, assessors recognize problem-solving as an important aspect of the evaluation process. (Messick, 1994; Torrance, 1995). In this mode of assessment, students' complete tasks to produce tangible outcomes, such as a product, or to provide an open-ended response to a question. This is based on the notion that evaluators can infer certain attributes from the task performed to measure students' development (Aladini et al., 2024). These assessments provide an effective method for assessing students' deep knowledge compared with more traditional evaluative processes (Frederiksen, 1984).

Furthermore, Mirzaei et al. (2024) highlight the additional benefits of AA, especially its value in terms of construct validity and consequential validity. In the same manner, PBA also enhances learners' abilities by promoting higher-order thinking in problem-solving (Espinosa, 2015). Despite these significant advantages, the practical challenge of developing effective grading rubrics for open-ended, project-based assessments often hinders the implementation of AA and PBA. While traditional forms of assessment are easy to administer, they lack the necessary ability to assess advanced, higher-order skills (Joseph, 2025; Colley, 2008).

On the other hand, alternative methods, which include portfolios, dynamic assessments, and manually implemented AAs, ensure high levels of educational validity but create heavy workloads for teachers (Papanastasiou et al., 2025; Burden & Byrd, 2018). Analytically comparing these paradigms reveals a strict trade-off between administrative convenience and pedagogical depth; standard tests scale easily but fail to capture authentic skills, whereas portfolios and dynamic assessments capture depth but are rarely scalable due to rigorous manual grading demands. AI-supported AA offers a practical compromise, combining the educational benefits of authentic tasks with the administrative efficiency typically associated with traditional testing (Ilieva et al., 2025; Abualrob, 2025).

2.3 Motivation, Student Agency, and Social Dimensions

Motivation influences the processes involved in learning, engagement levels, persistence in achieving goals, and students' approaches to learning (Chiu, 2021a, 2021b, 2022). Proponents argue that PBA is more motivating for students (Abualrob & Al-Saadi, 2019), as it evaluates achievement by engaging learners in authentic, real-life situations through tasks that function as learning activities.

Motivation in this framework closely aligns with the concept of student agency, which involves the capability to set personal goals, develop learning strategies, and engage in independent learning (Zhan et al., 2025). Through autonomy, teachers create environments that allow students to take responsibility for the learning process they undertake during AA tasks (Zhan et al., 2025). Furthermore, social collaboration serves as a founding principle of AA, as it is grounded in social interactions (Ajjawi et al., 2024; Boud & Bearman, 2024; Gravett, 2025; Timperley & Schick, 2025). Zhan et al. (2025) note that peer learning and social collaboration have been found to act as effective strategies for building social relations among learners.

Additionally, from a sociocultural perspective, Vygotsky's (1978) theory of social development suggests that peer assessment (PA) provides valuable opportunities for interaction that promote learning and development. Consequently, PA offers students insight into others' perspectives in ways that are often more accessible and meaningful. However, the need for clear, pre-established criteria and instructions further complicates the facilitation of this social collaboration, thereby increasing the logistical preparation required by the teacher.

2.4 AI in Assessment Design and Efficiency

Integrating AI technology has become increasingly significant in professional activities and practice, with some scholars suggesting that authentic engagement with AI is inherent to contemporary assessment discourse (Salinas-Navarro et al., 2024). Educators can employ ChatGPT to generate open-ended questioning prompts that, as suggested by Trust et al. (2023), align with success criteria and content knowledge for a particular instructional unit. Additionally, through prompt engineering, teachers can use ChatGPT to construct high-quality rubrics that clearly explain criteria for students at different levels (Trust et al., 2023).

While the broader literature explores AI's potential for student-facing adaptive testing or gamified assessments (Joseph, 2025; Boucher et al., 2021), the present study intentionally narrows its scope, both in theory and practice, to AI as a teacher-facing instructional design tool. Specifically, the literature most pertinent to this research highlights how generative AI addresses the logistical burdens of AA by automating the creation of comprehensive rubrics, refining task instructions, and suggesting evaluation scenarios (Ilieva et al., 2025; Abualrob, 2025). By explicitly restricting AI use to the pre-instructional design phase, educators can maintain pedagogical control while leveraging technology to overcome the time and workload constraints traditionally associated with implementing AA. This strategic use of AI does not position it as an evaluator, but rather as a practical tool for translating the theoretical ideals of AA into manageable classroom practices.

While recent systematic reviews highlight the extensive interest in AI technologies aimed at students (Bond et al., 2024), as well as more general technologies such as AI-based text detection software, they specifically call for more research into the application of generative AI to assist in designing educational assessments (Ogunleye et al., 2024). A major gap in current research lies in the lack of empirical studies examining the use of AI to overcome the logistical hurdles involved in AA rubric development and to reduce teachers' burden when designing authentic activities. The importance of closing this gap lies in the fact that the main problem with the widespread implementation of AA is not theoretical, but logistical; without adequate tools to help develop complex rubrics, the benefits of AA may remain less accessible in regular classrooms. Hence, overcoming such a problem provides an evident rationale for this research. The current study makes a new contribution by focusing explicitly on AI as a design tool for teachers.

Accordingly, this study examines the theoretical and practical implications of implementing such integration in Palestine, assessing its ability to facilitate learning gains and engagement among students and transform challenges into viable educational practices. Despite the benefits of using AI in education, researchers raise serious concerns regarding hallucination, algorithmic bias, and student data privacy (García-Carreño, 2025; Al Umri et al., 2025). Therefore, the educational community must make a fundamental distinction between using AI as a student-oriented assessment tool, which carries major ethical risks, and utilizing it exclusively as a teacher-oriented instructional design tool.

2.5 Research Questions

In light of the theoretical framework and the identified gap in the literature regarding AI-augmented AA, this study addresses the following research questions:

1. How does the academic achievement of students assessed via traditional methods compare to that of those assessed via authentic tasks designed with generative AI support?
2. What differences in student engagement (cognitive, behavioural, emotional, and social) are observed between the AA group and the TA group?
3. How do students describe their lived experiences during the AA process, and how do these qualitative insights explain the quantitative differences observed in their engagement and conceptual understanding?
4. How does the teacher interpret the transition from traditional to AA, in terms of implementation challenges, student performance, and perceived workload?
5. How did generative AI assist the teacher exclusively in the design phase (e.g., rubric construction, task selection) to overcome logistical barriers, and what opportunities or concerns emerged from this teacher-led integration?

3. Methodology

3.1 Research Design

This research employed a quasi-experimental sequential explanatory mixed-methods design to examine the phenomenon comprehensively. Specifically, the quantitative phase established the overall effect of the intervention, while the qualitative phase explained these outcomes. The independent variable was the

assessment technique used (AI-supported AA versus TA, while the dependent variables were students' academic performance and student engagement. In the quantitative phase, intact classes were assigned to either a control or an experimental group. A pre-test confirmed baseline equivalence ($p > 0.05$); however, the main statistical conclusions relied on post-tests. Subsequently, the qualitative phase explored participants' personal experiences and the efficacy of generative AI in supporting the logistical design of assessments.

3.2 Participants

3.2.1 Quantitative Phase Participants

The study took place at the Japanese Elementary School for Girls in Palestine between September 2025 and January 2026. The quantitative sample included 59 Grade 9 female students (classified as elementary students under the Palestinian educational system, which comprises grades 1–9). Due to the quasi-experimental design, convenience sampling was utilized. This involved the purposeful selection of intact classes from an accessible school to reflect the country's gender-segregated educational system. These participants were divided into an experimental group ($n = 29$) and a control group ($n = 30$). To assess the difference in achievement, the non-parametric Mann-Whitney *U* test was used since the data distribution deviated from normality according to the Shapiro-Wilk test ($p < 0.05$). Given the limited sample ($N=59$) and the use of only one classroom for each experimental condition, the findings were not highly generalizable. Accordingly, this research was considered a valuable pilot study or case study.

3.2.2 Qualitative Phase Participants

For the qualitative phase, participants were selected using purposive sampling:

1. Six experimental group students participated in individual semi-structured interviews, representing a range of achievement levels. Additionally, fifteen students participated in three focus group discussions (FGDs), stratified evenly into high-, medium-, and low-achievement groups (five students each).

2. The lead educator tasked with conducting the experiment participated in a semi-structured interview. Additionally, to broaden the evaluation of the technology and to mitigate the limitations of a single-teacher perspective, responses were sought from six science teachers who utilized generative AI to design assessment instruments such as rubrics, criteria, and scales. These teachers were purposively selected not to represent all science disciplines, but to provide rich, qualitative insights into use of AI for assessment design.

3.3 Instruments and Data Collection

The study utilized four main instruments:

3.3.1 Academic Achievement Test

This test was based on the Grade 9 science curriculum and instructional objectives. It consisted of 20 items, including 15 multiple-choice questions and 5 open-ended problem-solving tasks. The instrument demonstrated high internal reliability, with a Cronbach's alpha of 0.82. Importantly, to ensure accurate evaluation of the skills developed during the intervention, the test was designed to assess not only factual recall but also deep conceptual understanding, critical thinking, and

the application of knowledge —skills developed directly through the completion of AA products and performance tasks. To ensure content validity, the test was reviewed by a panel of experts to verify alignment with instructional objectives. It was administered post-intervention to measure learning outcomes. To ensure objectivity, control for experimenter bias, and minimize scorer bias, the tests were graded by an independent neutral evaluator (blind grading) using a standardized answer key.

3.3.2 Student Engagement Questionnaire

Adapted from Wang et al. (2016) and previously validated in the Palestinian context by the researcher (Abualrob, 2022), this 33-item instrument measured four dimensions: cognitive, behavioral, emotional, and social engagement, using a Likert-type scale. The instrument's validity was confirmed through previous factor analysis and a pilot test to ensure linguistic clarity. The questionnaire was administered four months after the intervention and was timed to assess sustained engagement rather than initial excitement over a new method.

3.3.3 Interviews and Focus Groups

Semi-structured techniques were used to explore students' views about the AA experience and the learning that resulted from it. Although a standard interview guide was followed, questions were kept open so that key themes could emerge directly from the students' responses. Data from these sessions were transcribed verbatim and analyzed using thematic coding (Creswell, 2018), whereby responses were open-coded, categorized, and grouped into main themes. To ensure qualitative validity (credibility), member checking and peer debriefing were utilized to verify the accuracy of the themes.

3.3.4 Teacher Reflections & Interviews

Data were collected through the primary teacher's daily logs and post-implementation interview, alongside the structured reflections of the six supporting teachers. These instruments focused on evaluating the efficiency of AI in reducing the teacher's workload related to the practical implementation of AA (e.g., rubric construction), ensuring that the "AI support" variable was measured at the instructional design level, rather than at the student-performance level.

3.4 Procedure

The experiment lasted for four months. To control for teacher-related confounding variables, (e.g., variations in teaching style or content delivery), both the experimental and control groups were taught by the same science teacher. To carefully control the risk of bias arising from this single-instructor setup, blind grading was rigorously applied to all assessments.

The control group received traditional instruction and testing, while the experimental group engaged in performance-based tasks (e.g., creating brochures and models and defending them orally) supported by rubrics designed via AI. To isolate the impact of the AA strategy and ensure the replicability of the technological intervention, the intervention was structured as follows: a generative AI tool (ChatGPT-4.0) was utilized exclusively by the teacher during the pre-instructional phase to design scenarios, refine instructions, and generate

assessment rubrics. Students were neither required nor instructed to use AI tools to produce their final artifacts; instead, they relied on manual creation, textbook synthesis, and collaborative group work to demonstrate their understanding. These AI-generated rubrics clearly outlined criteria such as scientific accuracy, creativity, and collaboration. All qualitative data were collected in private, quiet settings to ensure confidentiality and were subsequently audio-recorded and transcribed for thematic analysis.

3.5 AI Prompt Engineering and Iterative Refinement

To ensure the replicability of the AI intervention, a "Role and Context" framework was followed. The instructor used explicit pedagogical parameters to interact with the generative AI model, as follows: "Act as an expert in educational assessment. Design a 4-level rubric for evaluating a science project for 9th-grade students in Palestine, aligned with national standards in science." In addition, a step-by-step review process was also followed. In this process, the output of the AI was not copied directly, but was edited to match the students' age and level of understanding and to adapt the tasks to the Palestinian context, for instance, by making references to water pollution issues in the Palestinian community. Table 1 below shows some examples of the AI prompts used and the objectives set for the intervention tasks:

Table 1: Sample AI Prompts for AA Tasks

Educational Task	Sample Prompt	Goal of Intervention
Rubric Construction	"Design a four-level rubric for evaluating a presentation for a ninth-grade project on Water Pollution, with a focus on scientific inquiry skills."	To save time and ensure comprehensive criteria for AA.
Instructional Planning	"Design three authentic learning activities for a unit on genetics that connect concepts with real-world issues faced by Palestinian communities."	To transform instruction from memorization to practical application.

3.6 Data Analysis and Ethical Considerations

Data analysis was conducted using SPSS for the quantitative part. As noted in Section 3.2.1, the academic achievement data violated normal distribution assumptions, which required the use of the non-parametric Mann-Whitney *U* test. Conversely, the student engagement data met the assumptions of normality, allowing for the use of the parametric Independent-Samples *t*-test. The qualitative transcripts were analyzed thematically. Regarding ethical considerations, all participants and their guardians provided informed consent. Complete anonymity, confidentiality, and the right to withdraw at any time were ensured.

3.7 Objective Assessment and Bias Mitigation

To reduce possible teacher bias in the results associated with the single-instructor design, a strict and fair grading process was followed. Standardized statistical measures and achievement tests were used to assess both groups. In the case of

the experimental group, the rubrics were provided to the students in advance to ensure that the assessment was not based on the teacher's impressions, but on tangible evidence of performance. Moreover, triangulation was used through independent interviews and focus groups with the students, so that their experiences could be captured without the influence of the teacher. This helped indicate that improvements in performance were not due to the teacher, but to the level of understanding the students had gained through creating tangible products.

4. Research Results

4.1 AI-Supported AA Improves Academic Achievement

What differences in academic achievement were observed between Grade 9 students who were assessed using teacher-designed authentic tasks and those assessed using traditional methods? To answer this question, the Mann-Whitney *U* test was used, and Table 2 shows the test results.

Table 2: Results of the Mann-Whitney *U* Test for differences in post-intervention academic achievement between the AA group and the TA group

Variable	Group	*n*	Mean Rank	Sum of Ranks	Significance (*p*)	Effect Size (*r*)
Post aca-demic achievement	TA	30	24.12	723.50	.007	.35
	AA	29	36.09	1046.50		

The results from the Mann-Whitney *U* test showed significant differences in post-intervention academic achievement between the AA and TA groups. This was supported by a calculated *p*-value of .007, which was statistically significant at the $a = .05$ level. Additionally, the rank information indicated that the mean rank for the AA group was 36.09, which was substantially higher than the mean rank for the TA group (24.12), indicating that the AI-supported AA intervention measurably improved students' mastery of science concepts. To evaluate the practical significance of these findings, the effect size (Field, 2024) was calculated. The result ($r = 0.35$) indicated a moderate practical effect (Cohen, 1988) of the intervention. The mean ranks further showed that students assessed with the AI-supported method demonstrated a meaningful relative improvement over those evaluated traditionally.

4.2 AA Enhances Student Engagement

What differences in student engagement (cognitive, behavioral, emotional, and social) were observed between the AA group and the TA group? To answer this, arithmetic means and standard deviations were used for the sample's ratings on the scale items. In addition, an independent-samples *t*-test was used to compare the two groups, as shown in Table 3.

Table 3: Results of the independent samples *t*-test for the significance of differences between the AA group and the TA group across engagement dimensions

Type of Engagement	TA (n = 30)		AA (n = 29)		t	p	Effect Size (Cohen's d)
	M	*SD*	*M*	*SD*			
Cognitive engagement	3.26	.402	3.91	.657	4.594	.001*	1.2
Behavioral engagement	3.22	.399	3.65	.451	3.915	.001*	1.01
Emotional engagement	2.60	.846	3.08	.686	2.386	.020*	.62
Social engagement	3.24	.373	3.30	.410	0.565	.574	-
Overall score	3.19	.254	3.40	.302	2.836	.006*	.74

** Significance level: a ≤ 0.05*

Table 3 presents the results of the independent-samples *t*-test used to examine the significance of differences between the mean scores of the sample across various engagement dimensions at the significance level ($a \leq .05$), for a sample of 59 participants. The results indicated statistically significant differences between the AA group and the TA group in cognitive engagement, where the computed *t*-value was t(57) = 4.594, $p = .001$. The results also showed statistically significant differences in behavioral engagement, where the computed *t*-value was t(57) = 3.915, p = .001.

The results further revealed statistically significant differences in emotional engagement between the AA group and the TA group, where the computed *t*-value was t(57) = 2.386, p = .020, in favor of the AA group. Compared with the AA group (SD = .686), the traditional group showed a higher *SD* (.846) in terms of emotional responses. This suggested that the traditional method of testing may have produced more varied emotional responses among the students because of differences in test anxiety, while the AA method appeared to promote greater consistency.

The results did not show statistically significant differences between the AA group and the TA group in social engagement, where t(57) = 0.565, p = .574. As for the overall engagement score, the results showed statistically significant differences between the AA group and the TA group in favor of the AA group, where t(57) = 2.836, p = .006, indicating higher overall engagement associated with the AA approach. To determine whether the findings have any practical importance, effect sizes for Cohen's d were calculated (Cohen, 1988). The AI-supported method had strong effects on cognitive engagement ($d = 1.20$) and behavioral engagement ($d = 1.01$), as well as a moderate effect on emotional engagement ($d = 0.62$). Notably, the effect size on the composite variable of engagement was also moderate to strong ($d = 0.74$).

4.3 Student Perspectives: Deeper Understanding and Engagement

(Note: "Excellent achievement" reflects prior official school records, not a specific rubric score).

How did students describe their learning experience within the AA context, specifically regarding their engagement and conceptual understanding, based on individual interviews and focus group discussions? The thematic analysis of the individual interviews and focus groups revealed four main themes regarding how students experienced AA and its impact on their learning and engagement:

4.3.1 Theme 1: AA deepens understanding because it builds meaning rather than retrieving memorization

The students reported that producing AA products (a brochure, poster, PowerPoint, model/heart worksheet, and an oral presentation) helped clarify scientific concepts and build a deeper understanding of the relationships between parts and functions. Students learned more deeply by turning textbook content into visual or hands-on projects that required summarizing, organizing, tracing, and researching, which helped consolidate concepts and distinguish them more clearly than relying on memorization.

Illustrative quotes:

> *"I used to mix them up … but when we made the brochure, it helped me tell them apart and made studying them easier." (Salma, excellent achievement)*
>
> *"I couldn't connect the sequence of the digestion process, but when we made the product, I understood more through presenting and applying it." (Nora, very good achievement)*
> *"When we colored it, labeled the parts, and traced the blood pathway in the product, I understood it more." (Rama, good achievement)*
>
> *"The heart product helped me understand the blood pathway…. where it is oxygenated and where it is deoxygenated." (Focus group 1)*

This theme showed that students experienced AA as a learning mechanism that led to deeper conceptual understanding, thereby clarifying the direct link between the experience and their learning, and aligning closely with the principles of SC.

4.3.2 Theme2: AA increases engagement because it is enjoyable and breaks the routine of "class-textbook-exam"

The students described AA as a more engaging and effective learning experience than traditional lessons, associated with enjoyment, breaking boredom, and increasing interest in science. They demonstrated this engagement through active participation, increased confidence in presenting, improved focus during class, and extending discussions about science topics beyond the classroom.

Illustrative quotes:

> *"The work is nice and entertaining… and it changes the approach we're used to." (Sounbla, excellent achievement)*
>
> *"After the things we made, I started to like the lesson more and engage in it… and I even talk about science outside the lesson." (Reem, excellent achievement)*
>
> *"We used to color and understand in a fun way." (Focus group 1)*

> *"It's not like the usual test… when we're worried and scared about the grade." (Focus group 3)*

This theme reflected that AA was linked to increased engagement and motivation as a meaningful and enjoyable learning experience, highlighting its direct relationship to students' engagement, which strongly supported SDT regarding autonomy and motivation.

4.3.3 Theme3: AA develops multiple skills and reveals abilities that are less visible in paper-based assessment.

The students indicated that AA enabled them to practice technological, organizational, and communication skills, in addition to creativity, drawing, summarizing, and teamwork. This type of assessment also revealed personal abilities that were not previously clear, such as confidence in public speaking, self-confidence, and leadership, and contributed to discovering both their own talents and those of their classmates.

Illustrative quotes:

"I used technological applications like Google Slides and Microsoft Publisher… and learned to do new things." (Maram, excellent achievement)

"I used to be shy… but when we presented… my confidence increased and I learned how to stand and explain." (Lona, good achievement)

"I didn't know how to make a PowerPoint… so I learned this skill as something new." (Focus group 1)

"There were girls whose abilities we didn't know about… so when she stood up and explained, we recognized that in her." (Focus group 1)

This theme clarified that AA not only supported conceptual learning, but also contributed to developing multiple skills and expanding opportunities to express understanding, strengthening the relationship between AA experience and students' learning and engagement.

4.3.4 Theme4: The experience is positive, but it requires managing technical, group, and time/psychological challenges

While the experience was mostly positive, challenges emerged that might affect the quality of participation and learning, especially difficulties in designing PowerPoint presentations and using computers, differences in students' prior teamwork experience, as well as time pressure and stress associated with oral presentations. The students indicated that successful AA required supportive organization that reduced stress and ensured fairness in contributions. These issues expressed by the students highlighted the need for clear and highly structured rubrics and instructions, which the teacher had efficiently developed with the assistance of AI.

Illustrative quotes:

> *"Making the PowerPoint… requires computer skills, and that was the hardest thing." (Dalal, good achievement)*

> *"Designing the PowerPoint… where should I put the picture, and how do I arrange the information appropriately?" (Focus group 1)*
>
> *"Teamwork is a bit difficult… there's difficulty in taking responsibility and in leadership." (Afnan, excellent achievement)*
>
> *"The hardest situation was when I stood up present for the first time… I wasn't used to it." (Lama, low achievement)*

This theme highlighted that the AA experience is positively related to learning and engagement, but that this positive impact depends on providing technical and organizational support, a safe environment for presentations, and effective management of time and group work, which helped explain variations in students' experiences.

4.4 Teacher Perspectives: Transitioning to AA

How did the teacher interpret the transition from traditional to AA, in terms of implementation challenges, student performance, and perceived workload?

Analysis of the teacher interview and reflections (using thematic coding) showed that the teacher's interpretation of AA centered on five primary themes.

In the first theme, "Deep Understanding and Achievement," the teacher felt that the AA group performed better because students had multiple ways to demonstrate what they knew (visual, oral and technical) and to link concepts, which revealed "depth of comprehension" rather than mere memorization. She also mentioned a clear difference in average grades in favor of the AA group. She highlighted this by stating that "[t]he role of AA is not limited… rather, it provides multiple spaces to demonstrate understanding," and confirmed the impact on grades by noting that "[t]he difference… was large and clear."

Regarding the second theme, "The Nature of Errors as an Indicator of the Type of Learning," the teacher believed that errors in AA were mostly related to performance and organization (planning, time and teamwork) and decreased as the experience was repeated, whereas in traditional tests, errors were more cognitive in nature (forgetting and weak conceptual linking), indicating less deep understanding. Supporting this observation, she noted that performance errors "began to decrease with repeated experience…," while TAs exposed "[c]ognitive errors related to forgetting information."

Moving to the third theme, "Engagement and Motivation," the teacher described higher student engagement in AA due to the way it broke routine, activated students' talents, and gave them space to present knowledge in an appealing way, alongside "positive stress" associated with challenge and time pressure. She noted "[g]reater engagement and higher enthusiasm." In contrast, traditional tests were linked to stress about grades, boredom, and weaker motivation, a situation she explicitly described as being dominated by "[s] tress linked to grades… [and] signs of boredom."

In the fourth theme, "Skills Development and Fairness," the teacher reported that AA enabled the development of multiple skills (creativity, communication, technology, collaboration and organization) and considered it fairer because it allowed for diverse products and for individual and group work in ways that reflected students' actual abilities. She reinforced this observation by stating that "[i]t provided opportunities to develop multiple skills…," and emphasized that accommodating diverse workflows is the key factor "which made it fairer and more inclusive."

Concluding with the fifth theme, "Implementation Challenges and Accuracy of Scoring," despite the positive effect, the teacher emphasized that AA required greater time and effort for planning, developing criteria, and follow-up. She specifically pointed out that "[i]t required more time and effort… [along with] the necessity of accuracy in assessment…", especially when compared to the ease of preparing and grading traditional tests. This theme clearly showed the main logistical problem with AA: the administrative burden associated with rubric development and task planning is a major limitation on its adoption.

Ultimately, while the teacher believed that AA improved achievement and engagement because it transformed assessment into an active, multimodal learning experience and activated the teacher's supportive role, she acknowledged that it needed an organizational structure (time, criteria and group management) to ensure feasibility and fairness. This burden on the teacher's cognitive resources provided the rationale for the introduction of generative AI.

4.5 Generative AI as a Solution to Logistical Barriers

How did generative AI assist the teacher exclusively in the design phase (e.g., rubric construction, task selection) to overcome logistical barriers, and what opportunities or concerns emerged from this teacher-led integration?

The analysis of the teacher's narratives and the corroborating reflections from the six external science teachers showed that AI support contributed to designing AA in several key ways. These benefits were conditional upon critical use and careful verification of output validity. Specifically, the analysis of this AI integration centered on four primary themes.

The first theme focused on "Supporting Instructional Planning Through the Selection and Diversification of Products." In this context, AI functioned as a decision-support tool during AA planning: it recommended the most suitable product based on lesson constraints and provided alternatives that accommodated multiple learning styles. The teacher illustrated this support by explaining, "I used to hesitate between more than one product… and it would choose the most appropriate based on time, students' capabilities, and lesson objectives." Furthermore, she noted that the tool "[g]ave me ideas for different products that accommodate learning styles… and helped me build a collection of products." Ultimately, this addressed the method-selection component and demonstrated its planning impact by expanding options and enabling faster, more appropriate decision-making.

Regarding the second theme, "Improving Implementation Quality by Refining Product Design (Instructions/Time/Work Mode)," AI did not stop at "choosing the product," but contributed to making the product feasible to implement by clarifying instructions, ensuring realistic timelines, correcting errors, and specifying whether it should be individual or group work. The teacher highlighted this refinement process, stating that the AI assisted in "[m] odifying some products in terms of wording the instructions… or changing the submission time… and alerting me to some mistake and correcting them." She also relied on it for "[d]etermining the appropriate work type for the product: individual [or] group." This demonstrated AI's implementation impact by reducing student confusion and increasing feasibility.

Addressing the third theme, "Strengthening AA Through the Selection of Assessment Tools and the Development of More Comprehensive and Fairer Criteria," the most prominent contribution was at the "core" of assessment: choosing the appropriate assessment tool (rubric, checklist/rating scale) and specifying the type of assessment (peers, teacher, or blended), then building and refining criteria to measure multiple aspects and increase fairness. The teacher detailed this process, mentioning the AI's role in "[d]etermining the most appropriate assessment tool: rubric, checklist/rating scale… and the assessment type (peers/teacher/blended)." She further explained that the tool assisted by "[s]uggesting evaluation criteria or modifying them to be more comprehensive and precise… so we can measure all aspects… making the evaluation process fairer and more equitable." Additionally, she pointed out that "[p]reparing rubric templates… saved me time and effort."

In addition, feedback from the external instructors verified the effectiveness of the AI-assisted drafting in reducing the time required to build the rubric from hours to mere minutes, which greatly lowered the threshold for implementing AA. Importantly, the teacher noted that the AI-generated content showed considerable pedagogical validity. Instead of producing rating scales that could apply to anything, they effectively mapped the assessment criteria onto higher-level cognitive abilities, making sure that the rubrics would be academically sound and consistent with learning goals rather than bureaucratic expedience. These findings directly addressed the tool-selection and criteria/rubric-building components and highlighted opportunities such as fairness, comprehensiveness, and time savings.

Finally, addressing the theme of "Opportunities and Cautions as Indicated by the Teacher," the data reveal a comprehensive view of the AI experience. Regarding the opportunities, the teacher highlighted how the tool served to support decision-making when choosing among multiple assessment products within constraints such as time, resources, and objectives. Furthermore, it played a role in enhancing differentiation and accommodating learning styles through varied product suggestions. On a practical level, the AI contributed to improving design and implementation quality by refining instructions, estimating submission time, and reducing errors. It also aided in improving assessment tools and criteria toward greater comprehensiveness so as to effectively measure diverse aspects —

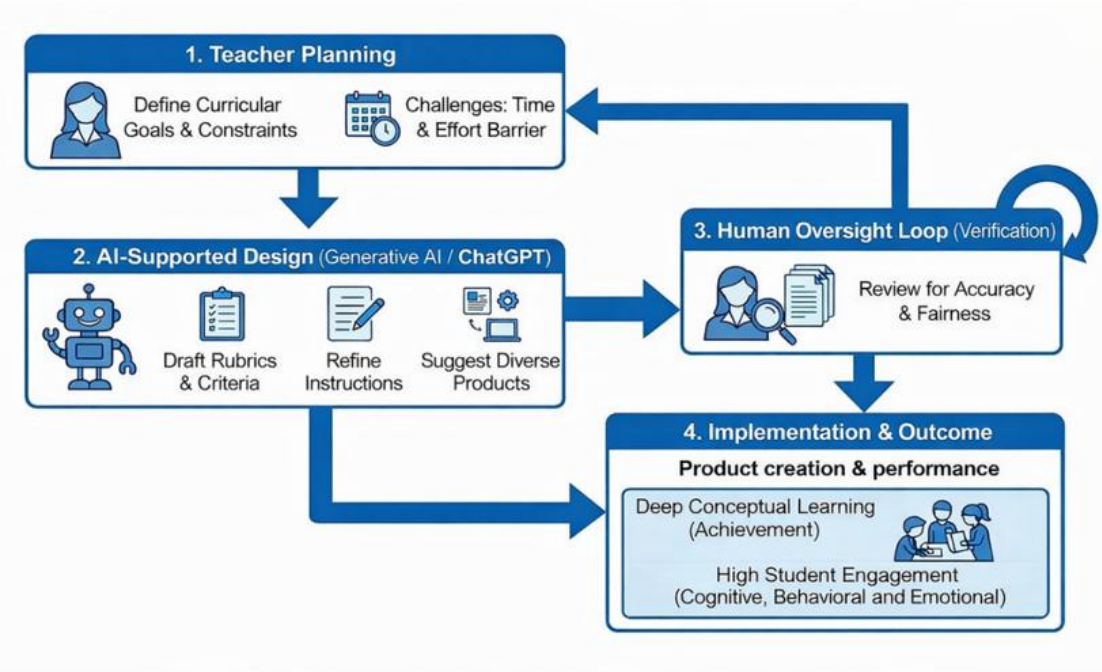
whether academic, personal, or creative. Beyond merely saving time and effort through ready-made rubric templates, the teacher emphasized a dimension of professional growth, noting that expanding thinking and developing critical capacity through dialogue with the tool was a significant outcome.

However, these benefits were carefully balanced by critical caution. The teacher strictly warned against accepting AI outputs "as they are." Instead, she stressed the necessity of review and verification, particularly regarding the suitability and comprehensiveness of the criteria for what the teacher intends to measure, as well as the correctness of the proposed modifications. To avoid irrelevant results, she also pointed out the need for precision in writing prompts (prompt engineering), because the tool may sometimes generate outputs that are off-target if the context or objectives are vague. Ultimately, these cautions served to reinforce the teacher's role as the final decision-maker, which she emphatically summarized by describing the AI as "a tool for support and enhancement… not a tool that takes my place."

4.6 The AI-Augmented AA Framework

Drawing on the findings from the teacher's reflections regarding the integration of AI, the operational process of implementing AA in this study is visualized in the AI-Augmented AA Framework (AAAA Model) shown in Figure 1. This figure summarizes the workflow described by the teacher, illustrating how AI tools were utilized to overcome planning barriers through a structured cycle of design, human verification, and implementation.

As illustrated in the figure, the cycle structures the assessment process into four interconnected stages. It begins with the teacher defining curricular goals and facing logistical constraints (Stage 1). This is followed by the AI-supported design phase, where tool like ChatGPT assist in drafting rubrics, refining instructions, and suggesting diverse products (Stage 2). Crucially, the model highlights the necessity of a human oversight loop for reviewing and verifying AI outputs to ensure accuracy and fairness (Stage 3) before implementation. Finally, the cycle concludes with students engaging in tangible products and performance tasks, leading to deep conceptual learning and high engagement (Stage 4).



Source: Developed by the researchers based on the study findings (visual layout assisted by AI

Figure 1: The AI-Augmented AA Framework (AAAA Model)

5. Discussion

The study investigated the impact of AA compared with TA on Grade 9 science students in Palestine and explored the role of generative AI in supporting the logistical application of AA. The results showed that the AI-supported AA group substantially outperformed the TA group in achievement and engagement. While AA encouraged deeper learning, it brought about practical challenges that could be addressed efficiently by AI tools.

5.1 Impact on Achievement and Deep Learning

The AA group demonstrated significantly higher post-test academic achievement compared to the TA group. These findings align with SC, suggesting that students shifted from rote memorization to "meaning-making" by creating tangible products, like brochures and models. Unlike traditional methods that emphasize basic cognitive objectives (Diab, 2005), AA allows students to consolidate concepts through active construction. This confirms that knowledge is best retained when students transform textbook information into visual and practical representations, fostering deep conceptual change and moving beyond superficial learning (Villarroel et al., 2018; Mirzaei et al., 2024). Most significantly, however, this level of learning became practically feasible because AI support reduced the teacher's traditional administrative burden in designing such complex tasks.

5.2 Dimensions of Student Engagement

The study found significant differences in favor of the AA group in cognitive, behavioral, and emotional engagement. This improvement can be explained by SDT; the authentic tasks transformed the high-pressure classroom into a supportive environment. The freedom to choose presentation formats enhanced students' autonomy, while successful task completion bolstered their competence.

In this way, the creation of the "product" fostered cognitive engagement through inquiry and synthesis, while the "performance" aspect (oral presentation) fostered emotional and behavioral engagement by building self-confidence and ownership, despite introducing some anxiety related to public speaking, in line with the competence and autonomy needs described in SDT.

Conversely, social engagement showed no significant difference (p = .574). Although students practiced teamwork, they reported challenges in leadership and responsibility. This finding aligns with research suggesting that placing students in groups does not automatically result in effective social engagement without explicit instruction in collaborative skills (Boud & Bearman, 2024). The students likely viewed group work as a logistical division of labor rather than a social learning opportunity, creating friction that offset the potential social benefits. This is an indication that real social interaction goes beyond merely being physically near each other or sharing tasks; it requires interventions within co-regulation and peer assessment frameworks, which were not implemented during this trial.

This reflects the complexity of collaborative assessments, which require careful management to be effective (Zhan et al., 2025). However, online technology can facilitate these interactions by reducing logistical burdens (Yan et al., 2022), and future AI integrations could specifically target the generation of structured peer-collaboration protocols.

5.3 Teacher's Challenges and the AI Solution

A critical tension emerged between the educational value of AA and its logistical burden. The teacher noted that AA "required more time and effort" regarding planning and the construction of criteria. This observation resonates with recent research indicating that "secondary teachers, operating under greater curricular pressure and TA mandates, viewed AA as more time-consuming and difficult to implement" (Papanastasiou et al., 2025, p. 14).

Generative AI served as a practical bridge to overcome this barrier. The teacher and the external validating teachers reported that AI "saved time and effort" by generating comprehensive rubric templates and suggesting diverse evaluation criteria. This corroborates findings that AI helps teachers produce high-quality rubrics and select assessment criteria efficiently (Abualrob, 2025; Baidoo-Anu & Owusu Ansah, 2023; Ilieva et al., 2025). By refining product instructions, generating rapid prompt-based iterations, and ensuring feasibility, AI allowed the teacher to focus on the pedagogical benefits rather than administrative hurdles.

5.4 Cautions and Limitations of AI Integration

Despite these benefits, the teacher emphasized that AI outputs must not be accepted "as they are." Critical human oversight is essential to verify the accuracy of the criteria and avoid potential biases or errors inherent in AI tools. This supports the view that AI should serve as a support tool that enhances—rather than replaces— the teacher's academic judgment and decision-making capacity (Joseph, 2025; Baidoo-Anu & Owusu Ansah, 2023).

Moreover, it is vital to recognize the limitations associated with the current research. Though experimenter bias was controlled through blind grading of the assessment tools, the role of a single implementing teacher poses a major threat to validity. There is always the possibility that the enthusiasm of the educator for the AA method and how strongly she advocated it may unconsciously affect the outcome of the experiment. Thus, the results might not be due only to the assessment exercise, since the sole teacher's enthusiasm might have directly influenced students' motivation and performance. Further research should consider involving more teachers to minimize possible bias. Moreover, considering the time lag between the engagement test and the assessment exercise (four months), there is a possibility of recall bias.

Therefore, the engagement findings should be interpreted with caution, as they may reflect retrospective rather than immediate engagement. There is a serious danger of overdependence on technology. Relying too heavily on AI for creating rubrics without proper analysis and reflection can result in the loss of essential teaching skills. This, in turn, means that teacher training programs should not only focus on assessment but also include AI literacy among the skills taught. At an international level, as education systems around the world increasingly adapt to the incorporation of AI, this study offers a framework for policymakers seeking to update their assessment strategies without overwhelming teachers.

Moreover, the small sample size (N = 59) and the focus on a single gender group in a specific geographic location (Palestine) with a particular academic discipline (Grade 9 science) must be considered limitations. Further research is needed to replicate this AI-enhanced AA framework with a larger sample size and in different academic disciplines. Ultimately, through blind grading and methodological triangulation, the observed benefits may be confidently interpreted as being associated with the AA process; however, teacher-related bias cannot be fully ruled out, particularly given the single-teacher design.

6. Research Implications

6.1 Theoretical Implications

Regarding the contributions to theory, the results support the relevance of SC and SDT in today's context. First, the study demonstrates that authentic tasks shift the learning environment from a source of test anxiety to a space that fosters autonomy and competence. Second, the research highlights that meaningful social engagement requires explicit instruction in collaboration.

6.2 Practical Implications

Accordingly, the practical significance of the research is that the AAAA model proposed in the study serves as a clear roadmap for implementation by educators. By positioning AI as a pre-instructional support infrastructure rather than a replacement, teachers can drastically reduce the logistical workload of rubric design from several hours to mere minutes while maintaining human pedagogical oversight to guard against AI hallucinations.

6.3 Policy Implications

From a policy perspective, especially in low-resource contexts, the study offers a scalable framework that supports investment in teacher training in "AI literacy for instructional design."

7. Conclusion

The study presents empirical data showing that the change from TA to AA positively affects the learning outcomes and motivation of students in Palestinian science classes. By constructing their knowledge through tangible products, students were able to gain a deeper conceptual understanding rather than merely memorizing the subject matter. However, the lack of a meaningful increase in social engagement indicates the complexity of organizing collaborative activities without direct support.

In terms of technology, AI plays an important role in implementing assessment innovations in the modern classroom. AI successfully reduces the challenges of implementing AA by substantially shortening teachers' preparation time. At the same time, the involvement of AI is only beneficial when the technology is controlled and monitored by humans who ensure its accuracy and equity. Ultimately, this study's main contribution is a scalable framework that pairs AA with teacher-guided AI. By democratizing access to innovative evaluations, this approach helps ensure that all students are well-prepared for the digital age while also building a highly resilient education system capable of meeting future global learning demands.

Conflict of Interest

The authors declare no conflict of interest.

Declaration of AI Use

AI-assisted tools were used in a limited capacity to support language refinement and editing of the manuscript. All intellectual content, data analysis, interpretations, and conclusions were developed and verified solely by the authors

Acknowledgments

We would like to acknowledge the generosity and insights of the study participants.

8. References

Abualrob, M. M. A., & Al-Saadi, S. H. (2019). Performance-based assessment: Approach and obstacles by higher-elementary science teachers in Palestine. *Journal of Education and Learning*, *8*(2), 1–11. https://doi.org/10.5539/jel.v8n2p1

Abualrob, M. (2022). Fifth and ninth grade students' engagement in science classes in Palestine. *South African Journal of Education*, *42*(2), Article 2070, 1–11. https://doi.org/10.15700/saje.v42n2a2070

Abualrob, M. M. (2025). Innovative teaching: How pre-service teachers use artificial intelligence to teach science to fourth graders. *Contemporary Educational Technology*, *17*(1), Article ep547. https://doi.org/10.30935/cedtech/15686

Ajjawi, R., Tai, J., Dollinger, M., Dawson, P., Boud, D., & Bearman, M. (2024). From authentic assessment to authenticity in assessment: Broadening perspectives. *Assessment & Evaluation in Higher Education*, *49*(4), 499–510. https://doi.org/10.1080/02602938.2023.2271193

Aladini, A., Bayat, S., & Abdellatif, M. S. (2024). Performance-based assessment in virtual versus non-virtual classes: Impacts on academic resilience, motivation, teacher support, and personal best goals. *Asian-Pacific Journal of Second and Foreign Language Education*, *9*, Article 5. https://doi.org/10.1186/s40862-023-00230-4

Al Umri, N., Karnyoto, A. S., & Pardamean, B. (2025, December). *The impact of AI-generated hallucinations in educational settings: Trends, gaps, and future directions* [Paper presentation]. 2025 International Conference on Information Technology, Information Systems, and Electrical Engineering, Purwokerto, Indonesia. https://doi.org/10.1109/ICITISEE68184.2025.11355145

Amrein, A. L., & Berliner, D. C. (2002). High-stakes testing & student learning. *Education Policy Analysis Archives*, *10*, Article 18. https://doi.org/10.14507/epaa.v10n18.2002

Andrade, H. L., & Heritage, M. (2018). *Using formative assessment to enhance learning, achievement, and academic self-regulation*. Routledge. https://doi.org/10.4324/9781315623856

Baidoo-anu, D., & Owusu Ansah, L. (2023). Education in the era of generative artificial intelligence (AI): Understanding the potential benefits of ChatGPT in promoting teaching and learning. *Journal of AI*, *7*(1), 52–62. https://doi.org/10.61969/jai.1337500

Bearman, M., Nieminen, J. H., & Ajjawi, R. (2023). Designing assessment in a digital world: An organising framework. *Assessment & Evaluation in Higher Education*, *48*(3), 291–304. https://doi.org/10.1080/02602938.2022.2069674

Boucher, E. M., Harake, N. R., Ward, H. E., Stoeckl, S. E., Vargas, J., Minkel, J., Parks, A. C., & Zilca, R. (2021). Artificially intelligent chatbots in digital mental health interventions: A review. *Expert Review of Medical Devices*, *18*(Suppl. 1), 37–49. https://doi.org/10.1080/17434440.2021.2013200

Boud, D., & Bearman, M. (2024). The assessment challenge of social and collaborative learning in higher education. *Educational Philosophy and Theory*, *56*(5), 459–468. https://doi.org/10.1080/00131857.2022.2114346

Bond, M., Khosravi, H., & De Laat, M. (2024). A meta systematic review of artificial intelligence in higher education: A call for increased ethics, collaboration, and rigour. *International Journal of Educational Technology in Higher Education*, *21*, Article 4. https://doi.org/10.1186/s41239-023-00436-z

Burden, P. R., & Byrd, D. M. (2018). *Methods for effective teaching: Meeting the needs of all students* (8th ed.). Pearson.

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

Chi, S., Liu, X., & Wang, Z. (2021). Comparing student science performance between hands-on and traditional item types: A many-facet Rasch analysis. *Studies in Educational Evaluation,* 70, Article 100998. https://doi.org/10.1016/j.stueduc.2021.100998

Chiu, T. K. F. (2021a). Digital support for student engagement in blended learning is based on self-determination theory. *Computers in Human Behavior,* 124, 106909. https://doi.org/10.1016/j.chb.2021.106909

Chiu, T. K. F. (2021b). Student engagement in K-12 online learning amid COVID-19: A qualitative approach from a self-determination theory perspective. *Interactive Learning Environments.* https://doi.org/10.1080/10494820.2021.1926289

Chiu, T. K. F. (2022). Applying the self-determination theory (SDT) to explain student engagement in online learning during the COVID-19 pandemic. *Journal of Research on Technology in Education*, *54*(Suppl. 1), S14–S30. https://doi.org/10.1080/15391523.2021.1891998

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates.

Colley, K. (2008). Performance-based assessment. *Science Teacher*, *75*(8), 68–72.

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approach* (5th ed.). SAGE Publications.

Diab, M. (2005). *The impact of using portfolios in the development of science thinking and retention for grade students* [Unpublished master's thesis]. Islamic University.

European Commission. (2024). *Regulation (EU) 2024/1689 of the European Parliament and of the Council of 13 June 2024 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act)*. Official Journal of the European Union. http://data.europa.eu/eli/reg/2024/1689/oj

Espinosa, L. F. (2015). Effective use of performance-based assessments to identify English knowledge and skills of EFL students in Ecuador. *Theory and Practice in Language Studies*, *5*(12), 2441–2447. https://doi.org/10.17507/tpls.0512.02

Fawns, T., Bearman, M., Dawson, P., Nieminen, J. H., Ashford-Rowe, K., Willey, K., Jensen, L. X., Damşa, C., & Press, N. (2025). Authentic assessment: From panacea to criticality. *Assessment & Evaluation in Higher Education*, *50*(3), 396–408. https://doi.org/10.1080/02602938.2024.2404634

Field, A. (2024). *Discovering statistics using IBM SPSS Statistics* (6th ed.). SAGE Publications.

Frederiksen, J. R. (1984). The real test bias: Influences of testing on teaching and learning. *American Psychologist*, *39*, 193–202. https://doi.org/10.1037/0003-066X.39.3.193

García-Carreño, I. (2025). A systematic review of emerging trends in education: Exploring the risks and benefits of generative artificial intelligence applications. *European Educational Researcher*, *8*(3), 1–20. https://doi.org/10.31757/euer.834

Gravett, K. (2025). Authentic assessment as relational pedagogy. *Teaching in Higher Education*, *30*(3), 608–622. https://doi.org/10.1080/13562517.2024.2380997

Harris, L. R., & Brown, G. T. L. (2018). *Using self-assessment to improve student learning*. Routledge. https://doi.org/10.4324/9781351036979

Herrera, S. G., Murry, K., & Cabral, R. M. (2013). *Assessment accommodations for classroom teachers of culturally and linguistically diverse students* (2nd ed.). Allyn & Bacon.

Ilieva, G., Yankova, T., Ruseva, M., & Kabaivanov, S. (2025). A framework for generative AI-driven assessment in higher education. *Information*, *16*(6), Article 472. https://doi.org/10.3390/info16060472

Joseph, S. (2025). Rethinking assessment: How AI is changing the way we measure student success? *AI & Society*, *40*, 5543–5545. https://doi.org/10.1007/s00146-025-02255-4

Lo, C. K. (2023). What is the impact of ChatGPT on education? A rapid review of the literature. *Education Sciences*, *13*(4), Article 410. https://doi.org/10.3390/educsci13040410

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational Researcher*, *23*(2), 13–23. https://doi.org/10.3102/0013189X023002013

Mirzaei, M., Hebblethwaite, D., & Yates, A. (2024). Exploring business students' perceptions of authentic project-based and work integrated assessments.

International Journal of Work-Integrated Learning, *25*(2), 183–199. https://hdl.handle.net/10652/6615

Nieminen, J. H., Bearman, M., & Ajjawi, R. (2023). Designing the digital in authentic assessment: Is it fit for purpose? *Assessment & Evaluation in Higher Education*, *48*(4), 529–543. https://doi.org/10.1080/02602938.2022.2089627

Ogunleye, B., Zakariyyah, K. I., Ajao, O., Olayinka, O., & Sharma, H. (2024). A systematic review of generative AI for teaching and learning practice. *Education Sciences*, *14*(6), Article 636. https://doi.org/10.3390/educsci14060636

Papanastasiou, E. C., Giallousi, M., & Pitri, E. (2025). Re-introducing authentic assessment in classroom assessment courses: Finding its place in the 21st century. *Education Sciences*, *15*(11), Article 1564. https://doi.org/10.3390/educsci15111564

Resnick, L. B. (1996). *Performance puzzles: Issues in measuring capabilities and certifying accomplishments*. CRESST/University of Pittsburgh, LRDC. https://doi.org/10.1037/e652072011-001

Ryan, R. M., & Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, *25*(1), 54–67. https://doi.org/10.1006/ceps.1999.1020

Salinas-Navarro, D. E., Vilalta-Perdomo, E., Michel-Villarreal, R., & Montesinos, L. (2024). Designing experiential learning activities with generative artificial intelligence tools for authentic assessment. *Interactive Technology and Smart Education*, *21*(4), 708–734. https://doi.org/10.1108/ITSE-12-2023-0236

Sokhanvar, Z., Salehi, K., & Sokhanvar, F. (2021). Advantages of authentic assessment for improving the learning experience and employability skills of higher education students: A systematic literature review. *Studies in Educational Evaluation*, *70*, Article 101030. https://doi.org/10.1016/j.stueduc.2021.101030

Timperley, C., & Schick, K. (2025). Assessment as pedagogy: Inviting authenticity through relationality, vulnerability and wonder. *Teaching in Higher Education*, *30*(3), 592–607. https://doi.org/10.1080/13562517.2024.2367662

Torrance, H. (Ed.). (1995). *Evaluating authentic assessment: Problems and possibilities in new approaches to assessment*. Open University Press.

Trust, T., Whalen, J., & Mouza, C. (2023). Editorial: ChatGPT: Challenges, opportunities, and implications for teacher education. *Contemporary Issues in Technology and Teacher Education*, *23*(1), 1–23.

Van den Berg, G., & du Plessis, E. (2023). ChatGPT and generative AI: Possibilities for its contribution to lesson planning, critical thinking and openness in teacher education. *Education Sciences*, *13*(10), Article 998. https://doi.org/10.3390/educsci13100998

Villarroel, V., Bloxham, S., Bruna, D., Bruna, C., & Herrera-Seda, C. (2018). Authentic assessment: Creating a blueprint for course design. *Assessment & Evaluation in Higher Education*, *43*(5), 840–854. https://doi.org/10.1080/02602938.2017.1412396

Vlachopoulos, D., & Makri, A. (2024). A systematic literature review on authentic assessment in higher education: Best practices for the development of 21st century skills, and policy considerations. *Studies in Educational Evaluation*, *83*, Article 101425. https://doi.org/10.1016/j.stueduc.2024.101425

Volante, L. (2004). Teaching to the test: What every teacher and policy maker should know. *Canadian Journal of Administration and Policy*, *35*(1), 1–6.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Wang, M. T., Fredricks, J. A., Ye, F., Hofkens, T. L., & Linn, J. S. (2016). The math and science engagement scales: Scale development, validation, and psychometric

properties. *Learning and Instruction*, *43*, 16–26. https://doi.org/10.1016/j.learninstruc.2016.01.008

Wiggins, G. (1990). The case for authentic assessment. *Practical Assessment Research and Evaluation*, *2*(2). https://doi.org/10.7275/FFB1-MM19

Yan, Z., Lao, H., Panadero, E., Fernández-Castilla, B., Yang, L., & Yang, M. (2022). Effects of self-assessment and peer-assessment interventions on academic performance: A meta-analysis. *Educational Research Review*, *37*, Article 100484. https://doi.org/10.1016/j.edurev.2022.100484

Zhan, Y., Boud, D., & Du, Z. (2025). Designing for authentic assessment: A scoping review. *Higher Education*. https://doi.org/10.1007/s10734-025-01588-9

Zhang, K., & Aslan, A. B. (2021). AI technologies for education: Recent research & future directions. *Computers and Education: Artificial Intelligence*, *2*, Article 100025. https://doi.org/10.1016/j.caeai.2021.100025