DOI: https://doi.org/10.52131/jer.2025.v6i2.2925

*i*RASD Journal of Educational Research

Volume 6, Number 2, 2025, Pages 09 - 22

Journal Homepage:
https://journals.internationalrasd.org/index.php/jer

Assessing the Consistency of AI, Peer, Self, and Expert Assessments in Evaluating Graduation Projects

Haneen Taqatqa[1], Doaa Hakawati[2], Suad Qurini[3], Hussam Daood[4], Abdalkarim Ayyoub[5]

[1] An-Najah National University, Palistine, Email: haneenmtaqatqa@gmail.com
[2] An-Najah National University, Palistine, Email: S12370537@stu.najah.edu
[3] An-Najah National University, Palistine, Email: s12370526@stu.najah.edu
[4] An-Najah National University, Palistine, Email: S12370511@stu.najah.edu
[5] An-Najah National University, Palistine, Email: ayyoub@najah.edu

ARTICLE INFO

Article History:
Received: July 23, 2025
Revised: August 21, 2025
Accepted: August 23, 2025
Available Online: September 02, 2025

Keywords:
AI in Assessment
Self-Assessment
Peer Assessment
Authentic Assessment
Rubric
Experts' Assessment
Educational Technology

Funding:
This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

ABSTRACT

The rapid spread of artificial intelligence and modern trends towards using its applications in assessment processes that have become directed towards authentic assessment have prompted researchers to pay attention to investigating the consistency of the results of the authentic assessment of experts as a criterion with the results of several other assessments (artificial intelligence, self, and peer assessment) and determining the assessment method that is most consistent with expert assessments. This research also aimed to study the differences between the results of the authentic assessment of students' graduation projects using different assessment methods. To examine this, the researchers used a rubric that was prepared in a way that suits the graduation project standards set by a group of specialists, and based on the descriptive quantitative approach with a cross-sectional design, where a group of graduation projects of students from the Faculty of Education at An-Najah National University in Palestine were evaluated as a random cluster sample by experts as a criterion and assessment using Chat-GPT, peer assessment, and self-assessment. The results showed that there was a strong consistency between the expert assessment and Chat-GPT assessment, with a Pearson correlation coefficient of R=0.897. Compared to the peer assessment, the results show no meaningful alignment with the expert assessment, where the correlation coefficient is R=0.380. Unlike the peer assessment, the results show weak alignment to the expert evaluation with R = 0.380. The self-assessment, in contrast, shows a moderate alignment with the expert assessment, with a Pearson correlation of R = 0.484. From the above, it can be inferred that, among the other methods, the AI-based assessment had the highest alignment with expert judgment.

Corresponding Author's Email: haneenmtaqatqa@gmail.com

1. Introduction

In the educational realm, assessments are considered as an indispensable component. It serves in determining the level of following and mastering the material, the level of achievement, and the adaptability of the set goals. It gives a teacher a 'snapshot' of how the pupil progresses, what the pupil's strengths and weaknesses are, and what needs to be done to assist the pupil, and improve performance and develop skills. The importance of

assessment is highlighted in providing feedback to students that helps them know their progress and guides them towards determining their own learning strategies (Wahyudin et al., 2021). In addition, assessment contributes to enhancing students' self-confidence and motivation to learn by encouraging them to achieve their academic goals. On a broader level, assessment provides important data for schools and educational institutions to improve curricula and teaching methods to meet students' needs and enhance the quality of education (Palm, 2019).

The manifestations of technological development have affected assessment by shifting from relying on the traditional teacher-based approach, which is often no more than tests whose ability to capture the true and complete experiences of students is limited, to a modern approach that relies on technology and innovation. In the past, assessment was largely limited to traditional paper tests and final exams that do not claim to be information retrieval to measure the extent of student memorization (Rudolph et al., 2023) .These methods focused primarily on the final results and did not take into account the entire educational process (Coskun & Alper, 2024).

With the emergence of modern educational theories, this shift began towards authentic and qualitative assessment methods that are different from the prevailing ones, as they focus on measuring and enhancing deep understanding of knowledge and developing skills instead of relying only on memorization or lower-level thinking skills (Norova, 2020). Among these methods, self-assessment stands out as an important means that helps students identify their strengths and weaknesses, which encourages them to take responsibility for their learning and promotes reflective and critical thinking (Yan, Panadero, et al., 2023), Self-assessment is also one of the assessment methods that plays a pivotal role in promoting self-learning and personal awareness for students (London et al., 2023) .This type of assessment allows students to evaluate their own performance in a way that encourages them to take responsibility for their learning and develops their skills of analysis and objective assessment (Kadri & Amziane, 2021) .Self-assessment supports learner's confidence by encouraging them to utilize positive dispositions and experience significant success. As a result, they become more aware of their areas of weakness and are better equipped to address and improve them over time (Yan et al., 2020). Self-assessment is considered part of the continuous learning process that enables students to develop their skills more independently, which leads to enhancing academic achievement and developing basic life skills (Butler, 2024).

Similarly, peer assessment is an effective tool to enhance collaboration and exchange of knowledge, experiences, and opinions among students, as it provides them with an opportunity to provide feedback and learn from the perspective of their peers who are in some situations more capable of evaluating their peers than the teacher (Iglesias Pérez et al., 2022). Peer assessment gives students the opportunity to exchange roles as evaluators, as students evaluate their peers' work based on pre-determined criteria. Hence, it is considered a good way to increase and enhance students' critical thinking and raise their awareness of the quality standards of their academic performance.(Ismail & Heydarnejad, 2023). This assessment also increases students' interaction with their peers and the environment around them, which fosters effective communication within the classroom, allowing for the exchange of ideas, opinions, and constructive feedback with high objectivity (Ayyoub et al., 2017). The development of technology and the continuous increase in interest in this field, as well as artificial intelligence, as one of the components of the educational process and part of the tools that can be relied upon in evaluating students, as they allow for the provision of direct, immediate and individual evaluation as well when needed for each student to receive special feedback appropriate to his capabilities. Therefore, this evaluation will support individual and group learning.(Hooda et al., 2022), Thus, attention has shifted to a new level of assessment methods, utilizing digital technologies that allow for accurate and honest analysis of student performance (Swiecki et al., 2022).

AI-based assessment is not limited to student cognitive assessment, but extends to skill, performance, and attitude assessment as well. It can also be used to analyze video recordings of student skills, thus improving not only the final outcome, but also the speed, effectiveness, and smoothness of the assessment process (Yan, Wang, et al., 2023), this distinguishes AI-based assessment from other assessment methods in terms of objectivity, accuracy, speed, and impartiality (Dimitriadou & Lanitis, 2023) .In general, modern

assessment methods are drawing attention to a deeper understanding of assessment as a driver of educational improvement, rather than simply a method of measuring student knowledge (Mao et al., 2024). These major shifts, driven by interest in technology and its integration into education and curricula, have been reflected in measurement and assessment systems and methods (Planas-Lladó et al., 2021), however, there are some doubts about the reliability of the assessment compared to the teachers' own assessment, and hence the main research question of this study was about the reliability of self-assessment, peer assessment, and artificial intelligence assessment compared to the teachers' assessment.

Despite the ongoing developments in assessment and its methods, the teacher's role remains essential. The teacher possesses the experience and ability to provide a comprehensive and in-depth view of students and their abilities as individuals with their own circumstances and skills driven by their surrounding influences. Teacher assessment guides the educational process as a whole through constructive feedback (Dhara et al., 2022) .The teacher is also the one who can employ the results extracted from other assessments and enrich them with knowledge and personal experience to ensure that educational goals are achieved in a balanced manner. No matter how much there is a shift away from teacher assessment, the teacher's human touch remains an indispensable element for achieving comprehensive assessment that supports students' academic and personal growth (Yu, 2024). However, the diversity and rapid development of assessment methods lead us not to limit the assessment process to teacher-based assessment. Teacher assessment is useful in describing students' progress, while it is unable to play the role of self-assessment, which enhances metacognitive thinking skills and self-awareness, the role of peer assessment in exchanging ideas, increasing familiarity and cooperation, and the role of artificial intelligence assessment in providing insights tailored to students' individual needs for improvement, and integrating them to enhance the balance and efficiency of the learning experience and the accuracy of measurement. (Yan, Wang, et al., 2023)

With this qualitative shift, the assessment process has become more complex and accurate by focusing on the practical results of the educational process. Hence, authentic assessment has become an urgent necessity to meet this need to measure what the student can do in real, realistic contexts based on their skills and knowledge, which enhances learning based on experience and application in a way that opens new horizons for students by providing the opportunity for deep and sustainable learning that does not stop at the boundaries of classrooms (Hu, 2021). Authentic assessment depends on designing activities that contribute to building students' confidence in their ability to confront authentic problems in their work environments, which enhances their sense of responsibility and helps them uncover their skills gap, which in our current era has emerged research, cognitive and digital skills, by identifying strengths and weaknesses in a way that provides additional support to improve them, and stimulates student involvement in developing the educational environment (Al Maktoum & Al Kaabi, 2024). This development in assessment processes has also changed the view of educators in the twenty-first century towards the assessment tools used by changing the objectives of the educational process. Tests are no longer a sufficient means to measure all the outcomes of the learning process, whose primary goal has become to prepare students for their practical life and enhance their mental and research abilities in dealing with problems, solving them and adapting to circumstances flexibly (Tomczyk et al., 2024), Hence, educational research has become directed towards using rubrics primarily in authentic assessment to evaluate students' performance on research tasks, especially at the university level, in which the student is supposed to acquire performance skills in which he applies his knowledge in producing new knowledge and performances. Or use it appropriately with a minimum that reflects on his professional future (Castillo-Martínez et al., 2024).

The rubric came as a tool that is compatible with the transition to authentic assessment as a strong and organized tool, formed within a measurement framework that includes standards that include the aspects that will be evaluated and an accurate and comprehensive description that specifies the levels of performance for each standard, to provide the student with a clear vision of what is expected of him to achieve in a way that enhances him to work to exert his maximum energy and direct his efforts and focus them to achieve these standards, which increases the fairness and objectivity of the assessment process (Taylor et al., 2024). For students, graduation projects showcase their mastery of certain skills and the culmination of certain milestones, which is why self-directed learning is

vital for students to accomplish these projects on time. Hence, it makes sense to find assessment techniques that are fair, and reliable. Within Palestinian universities, where the majority of assessment techniques are still focused on teaching, there are increasing calls for more variety in assessment techniques. This could lead to fostering critical fairness, and increasing levels of active participation within students.

New studies show how the use of exams and grades, in particular, can lead to evaluative discrepancies and gaps in the students' ability to assess their own work (BioMed Central, 2024). Exploring the relations of expert, self, peer, and AI assessments is vital to find consistency and objectivity within assessment strategies as proven in reliance within studies. Global studies have shown the inclusion of peer and self-assessments with expert ones tend to raise motivation, promote reflective thinking, and enhance the writing skills of students (SpringerLink, 2025).As more AI tools become available and more popular in automatic assessments, understanding the degree to which AI feedback aligns with professional feedback and reviews becomes more and more needed. AI feedback seems more positive or glowing than human feedback which leads to concerns on how much positive feedback results in grade inflation or higher grades and concerns on the standards distributed during assessments (Taylor & Francis Online, 2025).

Assessment practices and methods have received considerable attention in academic research and writings across the globe, yet the systematic research that studies the application and comparison of AI, peer, self, and expert assessment in Palestinian universities, particularly on graduation projects, is still lacking. Most existing studies focus on theoretical models, small classroom activities, or traditional teacher-led evaluations, leaving a need for deeper understanding of how these methods function in Palestine’s unique cultural and institutional environment. The positive aspects reflected as a result of the development of assessment methods and tools have created a controversial situation about the effectiveness and efficiency of these methods and approaches in authentic assessment. Here, the question arose about the consistency of the results of authentic assessment using artificial intelligence, peer assessment, self-assessment, and expert assessment. With the multiplicity of assessment methods, challenges have emerged related to the effectiveness of each type in improving student performance and helping them achieve their educational goals and evaluate them from a different perspective. Hence, the need for such a study to highlight the possibility of using these assessment methods (Hodges & Kirschner, 2024), especially with many studies indicating the necessity of integrating assessment methods and types and not relying on one tool or aspect to ensure consistency and fairness of assessment and deviating from the stereotype in a way that ensures strengthening students, even if implicitly (Bower et al., 2024), Many studies have indicated the role of self-assessment as a tool to strengthen the student and evaluate his work and the effect of overlapping assessment methods on performance. This does not eliminate the role of the teacher in the assessment process, but on the contrary, it may assign him new tasks that highlight his leadership role as a facilitator and guide to the educational process (Yan et al., 2022).

This research aimed to investigate the consistency of the mentioned assessment methods in the actual assessment of university students in the Palestinian context, considering that it has become an urgent need to help reveal the strengths and weaknesses of each assessment method and how these methods can be alternatives or complements to teacher assessment. These types may also affect students' motivation, and develop and enhance their skills differently. Hence, the researchers built an assessment tool for students' graduation projects from colleges of education according to the appropriate validity and reliability standards to evaluate students' graduation projects using them and relying on expert assessment, artificial intelligence assessment, peer assessment, and self-assessment. Hence, the research questions came as follows:

- To what extent are the expert assessment results (as a reference standard) consistent with the results of AI-based assessment, peer assessment, and self-assessment?
- Which assessment methods (AI-based, peer, or self-assessment) correlates most strongly with expert evaluations?
- Are there statistically significant differences in the results of evaluating students’ graduation projects when using the four assessment methods (experts, AI-based, peer, and self-assessment)?

Based on the above research questions, this study aims to examine the following hypotheses to clarify the relationship between different assessment methods and their consistency with expert assessment, as well as to compare the assessment results using artificial intelligence, self-assessment, and peer assessment in undergraduate graduation projects.

- **H1:** The results of AI-based assessment, peer assessment, and self-assessment are consistent with expert assessment results when evaluating students' graduation projects.
- **H2:** Among AI-based, peer, and self-assessment methods, AI-based assessment shows the strongest correlation with expert evaluations.
- **H3:** There are statistically significant differences between the four assessment methods (experts, AI-based, peer, and self-assessment) in evaluating students' graduation projects.

With the rapid development in the field of education and the adoption of various assessment methods, the importance of revealing the efficiency of these methods in evaluating university students' graduation projects has emerged. This research aims to compare the studied assessment methods by examining the consistency of the assessment results based on them and comparing them in the assessment of experts as a standard, which gives an idea of the reliability of these methods in the authentic assessment and the extent of the possibility of relying on them in a way that helps achieve and examine the quality of the outcomes of the educational process. This research also aims to provide a standardized tool for evaluating graduation projects, investigate the reliability and validity of assessments using artificial intelligence, peer assessment, and self-assessment compared to expert assessment, compare AI, peer, and self-assessment methods to determine the degree of consistency of their results in evaluating graduation projects.

The importance of this research is that it integrates several topics in the field of authentic assessment in the educational process in different ways, as it tries, through examining assessment methods, to provide good suggestions for assessment methods that help teachers overcome time and effort as a basic problem they have when adopting authentic assessment. This research also deals with assessment with artificial intelligence in a way that highlights the extent to which its tools can be employed in the assessment process, in addition to drawing the attention of educators to the possibility of relying on other methods in assessment that do not depend on the teacher only in an attempt to reveal the degree of reliability of these methods, which contributes to supporting and guiding decision-makers in the educational field and teachers to work to build assessment strategies directed to achieve the quality of the outcomes of the educational process by building appropriate assessment strategies based on the results of this research.

2. Research Literature

Academic assessment methods have witnessed significant developments in light of the shift towards authentic assessment. This has highlighted the need to examine the reliability of modern methods such as self-assessment, peer assessment, and AI-based assessment compared to expert assessment as the benchmark.

Different scholars have examined assessment methods in several ways. Coskun & Alper (2024) pointed out that ChatGPT-4's evaluations were very consistent with teacher assessments across different task types, particularly during written assessments, but had difficulty with more complex visual content. In the same manner, Shabara et al. (2024) found that ChatGPT-3.5 was inaccurate in assessing second-language writing, culminating with teacher evaluations misalignment, and thus, failure in understanding the assessment criteria. Awidi (2024) noted that while the AI feedback was consistent, it poorly lacked personalization and the qualitative assessment needed. Chunping et al. (2024) pointed out that AI accuracy was greater than that of peer assessment, but as with many other scholars, needed additional refinement for complex educational settings. Dimitriadou & Lanitis (2023) pointed out that having AI as a tool in the assessment process is a bright prospect, but it is far from being able to replace the human component required for more complex task evaluations.

Other work has analyzed self-assessment in the context of improving student performance, encouraging reflective learning, and increasing self-awareness. Milic & Simeunovic (2022) noted that self-assessment only ranked behind the consistency of teacher evaluation. Ayoub et al. (2021) argued that self-assessment helps performance improve via repetition and consistency. Research conducted by Butler (2024) and Yan et al. (2020, 2022, 2023, 2023) focused on the precision of self-assessment being dependent on the clarity of the self-assessment criteria, as well as students' prior training on the criteria, and argued that this training results in the formation of self-calibration in metacognitive reasoning. London et al. (2023) stated that self-assessment increases self-awareness, thus confirming it as a mechanism for bolstering independent and self-directed learning. Ismail & Heidarnejad (2023) stated self-assessment increased students' perception of self-efficacy, which has a direct effect on their psychological well-being and academic performance.

There is also attention given to peer assessments. Power & Tanner (2023) noted that at the intermediate levels, students tend to rate their peers more favorably. This raises questions about their objectivity. Ayyoub et al. (2017) recognized the importance of peer assessment for skill development, but for developing accuracy, they emphasized the need to provide students with basic evaluative tools. Iglesias Pérez et al. (2022) noted that while peer assessment improves interaction among students, they need training to apply peer assessments consistently. Consistency, as noted by Chunping et al. (2024) while presenting peer assessments, is critical to the development of the presentation and delivery skills. Personal biases, however, are a problem. According to Milic & Simeunovic (2022), peer assessment comes third in accuracy after self-assessment and assessment by peers.

Other studies, like those by BioMed Central (2024), SpringerLink (2025), and Taylor & Francis Online (2025), have underscored the need to assess other evaluation alternatives in comparison to expert evaluation as the standard. BioMed Central pointed out that depending solely on standard assessments without alternatives may result in unreliable assessments on the part of the evaluator. SpringerLink established that self, peer, and expert assessments together can foster student motivation and reflective thinking. Regarding grade standards, Taylor & Francis Online raised the concern that AI gives higher scores than teachers and will have to be monitored closely to avoid grade inflation. Chunping et al. (2024) concluded that AI assessments are consistent and reliable compared to peer assessments, though remaining criteria need to be better articulated for AI-based evaluations.

Overall, the literature demonstrates sharp disagreement in the consistency and validity of the various methods of assessment. This demonstrates the need to incorporate other methods in order to achieve equitable, precise, and dependable assessments especially in higher education, which has the greatest range of tools designed to measure student growth and development. One should study which assessment techniques are best suited to each educational situation.

3. Research Methodology

This study evaluates how self, peer, and AI assessments compare to expert evaluation in undergraduate graduation projects. A cross-sectional, descriptive quantitative approach was apapted for this study, where all evaluations were conducted at the same time to maintain consistency and mitigate bias (Vogt et al., 2012). The AI ChatGPT-3.5 model was used to assess the projects which were being evaluated in one session. Human evaluators also used the same rubric, and this approach minimized differences in evaluation and ensured consistent application of the defined evaluation criteria. AI was aligned with the evaluation rubric through carefully crafted prompts to ensure reliability in the evaluation. The study sample encompassed 18 graduation projects within the random selection from the students of the Procedural Research course during the first semester of the 2024/2025 academic year at the Faculty of Education, An-Najah National University. In all three assessments, students reported: self-assessment, where the project evaluator was assessed anonymously using the same rubric (peer assessment) and then, the project was assessed by ChatGPT-3.5 (AI evaluation).

Determining sample size needs to take practical aspects and statistics into account. As Bujang and Baharum (2016) explain, a sample can be considered adequate in Pearson correlation analysis when it can detect a moderate effect size ($r = 0.5$) with 80 percent power

and 0.05 level of significance. This shows that the size of the sample was reliable in this study. Participants were advised in writing and in a virtual meeting about their information being kept confidential and meeting contours, and consent was properly collected. In the meeting, the researcher explained the assessment procedures, clarified the rubric, and asked students to assess a sample to pilot their evaluations. This confirmed their understanding, and the researcher discussed the results with them and ensured the assessment was valid. For standardizing evaluations, the researchers opted for a specifically designed rubric. This rubric explained criteria better, which helped the researchers in upholding fairness, assessment accuracy, and reliability. It also gave the researchers the ability to comment on the students' work in a constructive manner, thereby aiding their growth and improvement (Stevens&Levi, 2023). The following are selected examples of rubric criteria and their descriptions to illustrate how student performance was evaluated:

Clarity of Research Title and Components
- Complete elements
- Clear terminology and definitions
- Title does not exceed 15 words
- Originality and novelty of the topic
- Clear description of the study population and sample
- Provides an overview of other research components

Relevance to Cultural and Social Context
- Addresses community needs
- Respects local beliefs and traditions
- Benefits the community
- Considers cultural diversity
- Acknowledges the uniqueness of the community
- Balances global and local perspectives

Clarity of Research Objectives
- Clear and precise objectives
- Measurable and assessable
- Aligned with methodology
- Derived from the main research question
- Linked to research questions
- Clearly expresses key aspects of the objective

Presentation and Organization
- Logical sequence of content without gaps
- Cohesive paragraphs
- Text is understandable and clear
- Starts with general topic, ends with specific focus
- Objective writing style
- Summarizes the main outlines of the topic and research

It was built based on the following steps: First: Collecting criteria for evaluating university graduation projects: The arbitration criteria for graduation projects were collected from the suggestions of 21 specialists with higher degrees by asking the following questions:

- What are the graduation project judging criteria that you have noticed through your experience that the examiners care about?
- From your own experience, what are the questions that were directed to you and that you felt were among the graduation project judging criteria?
- From their answers, the researchers collected 40 criteria attached in Appendix.(1)

Second: Calculating the validity and reliability of the assessment criteria: After collecting the assessment criteria, they were presented to 10 arbitrators to calculate the logical validity using CVR: Coefficient of Variation Ratio to calculate the relative coefficient of variation between the arbitrators' responses to all assessment criteria: (De La Rosa Gómez et al., 2019).

$$CVR = \frac{n_e - \frac{N}{2}}{\frac{N}{2}}$$

Where N represents the number of referees and n_e represents the number of referees who agreed that the paragraph is important and necessary as a criterion in the rubric. CVI: Certainty of Validity Index was also calculated to ensure the validity of the paragraphs, according to the following equation: (De La Rosa Gómez et al., 2019)

$$CVI = \frac{\text{Number of items agreed upon by experts}}{\text{Total number of it Number of items}}$$

The paragraphs with a CVR or CVI higher than (0.6) were considered acceptable and those lower were considered rejected. Appendix (2) shows the results, where 24 criteria remained after deleting 16 of them. To ensure the reliability of the rubric, one of the graduation projects was used as a survey sample, by presenting it to two evaluators to evaluate it according to the rubric prepared in Appendix (3). The reliability between evaluators (Interrater reliability) was calculated in two ways after obtaining their assessment according to the criteria as attached in Appendix (4).

First method (Hollisty equation): R = (2M) / (N1 + N2) (Dasgupta et al., 2014)

- R: coefficient of agreement between the two analysts
- M: number of analysis units agreed upon by two analysts
- N1: number of analysis units evaluated by the first analyst
- N2: number of analysis units evaluated by the second analyst

$$R = \frac{2 \times 20}{24 + 24} = \frac{40}{48} = 0.83$$

The second method: Calculating the correlation (Pearson correlation)

Table 1: Pearson Correlation Coefficient for Inter-Rater Reliability Between the Two Evaluators (CVR method)

		First Rater	Second Rater
First Rater	Pearson Correlation	1	.731**
	Sig. (2-tailed)		.000
	N	24	24
Second Rater	Pearson Correlation	.731**	1
	Sig. (2-tailed)	.000	
	N	24	24

In both methods, the results showed an acceptable level of stability. Intrarater reliability was also calculated with a two-week time difference for the first evaluator. The results were as in Appendix (5), which shows that the second assessment differed from the first in 3 paragraphs, while it agreed in 21 paragraphs. Reliability was calculated in two ways:

The first method (Hollisty equation): R = (2M) / (N1 + N2)

$$R = \frac{2 \times 21}{24 + 24} = \frac{42}{48} = 0.875$$

The second method: Calculating the correlation (Pearson correlation)

Table 2: Pearson Correlation Coefficient for Intra-Rater Reliability of the First Evaluator (CVI Method)

		First Rater	Second Rater
First Rater	Pearson Correlation	1	.760**
	Sig. (2-tailed)		.000
	N	24	24
Second Rater	Pearson Correlation	.760**	1
	Sig. (2-tailed)	.000	
	N	24	24

In both methods, the results showed high and acceptable reliability between the first and second rater's assessment.

4. Results and Discussion

The researchers used SPSS version 23 to analyze the data, where the arithmetic means were calculated for the assessment of experts, peers, and artificial intelligence (4 experts and 18 students for peer assessment and two repeated assessments of the same research from artificial intelligence), and at first the descriptive statistics were calculated for each set of data according to the assessment method used, and they were as follows:

Table 3: Descriptive Statistics

	N	Mean	Std. Deviation
Self-Assessment	18	61.7778	5.30877
AI Assessment Chat-GPT	18	55.2222	7.71214
Peer Assessment	18	53.5556	10.52852
Experts Assessment	18	57.2222	7.82572

Table 3 shows the existence of differences between the averages of students' marks and standard deviations in all assessments compared to the experts' assessment. To measure the degree and strength of the linear relationship between the marks classified as continuous variables, the correlation coefficient was calculated between them after ensuring the normal distribution of the results of the four assessments based on the Shapiro-Wilk test. At first, to verify the assumption of normal distribution of the data, the Shapiro-Wilk test was used. This test is one of the most powerful and reliable tests, especially for small to medium-sized samples ($n < 50$), and is widely recommended in studies relying on parametric analysis (Murray, 2025).

Table 4: Tests of Normality

	Shapiro-Wilk		
	Statistic	Df	Sig.
Self-Assessment	.967	18	.734
AI Assessment Chat-GPT	.980	18	.955
Peer Assessment	.919	18	.124
Experts Assessment	.955	18	.515

*. This is a lower bound of the true significance.
a. Lilliefors Significance Correction

The table shows a normal distribution of the results of the four assessments, where the results of the four assessments: self, artificial intelligence, peers, and experts were as follows, in order: ($W= 0.967$, $p= 0.734$; $W= 0.980$, $p= 0.955$; $W = 0.919$, $p=0.124$; $W=0.955$, $p= 0.515$). The test results showed that all variables followed a normal distribution ($p > 0.05$), justifying the use of parametric tests such as Pearson's correlation coefficient and repeated measures ANOVA. Below are the results of the Pearson correlation coefficient test:

Table 5: Correlations

		Self-Assessment	AI Assessment Chat-GPT	Peer Assessment
Experts Assessment	Pearson Correlation	.484*	.897**	.380
	Sig. (2-tailed)	.042	.000	.120
	N	18	18	18
	Cohen's d	-0.682	0.257	0.394

*. Correlation is significant at the 0.05 level (2-tailed).
**. Correlation is significant at the 0.01 level (2-tailed).

The table shows the correlation coefficient between expert assessment and three types of assessments: AI assessment using Chat-GPT, peer assessment, and self-assessment. For AI assessment, the results showed that the correlation coefficient was 0.897R= 0.00, which reflects a very strong and positive relationship between AI assessment and expert assessment as a criterion. AI can accurately replicate expert evaluation, proving to be a dependable tool

for assessing graduation projects. There are a few reasons why the assessments made my AI matched so well with the expert evaluator assessments. First, AI evaluation was completely unbiased because the AI calculations were made using purely objective algorithms regarding the criteria set in the rubric. In addition, the AI evaluations were made in a single session which means the same prompts were used, thereby reducing variability in rubric interpretations. This consistency explains the strong equilibrium between the AI evaluations and the expert assessments.

This level of accuracy AI achieves, is because of superior algorithm design that employs deep learning and extensive datasets. Research like Coskun & Alper (2024) and Milic & Simeunovic (2022) highlight tools that AI uses to objectively assess data inputs and sophisticated metrics, thereby eliminating the biases related to peer and self-assessments. With the ability to quickly and accurately analyze huge amounts of data, the assessments made and the results produced can be reconciled reliably. According to Swiecki et al. (2022), the implementation of AI technology within the evaluative framework of education encourages the uniformity and trustworthiness of assessment within the AI spans precision control. However, within the scope of peer assessment, the outcomes showed the least correlation with the expert assessments. The correlation score of $r = 0.380$ with a p value of 0.120 indicates correlation absence. The outcomes of peer assessments present a considerable range, which negatively affects the reliability and objectivity of the assessments. Different explanations can account for this lack of consistency. The inability to assess complex evaluative criteria of a project and the implicit bias to score within a rubric could stem from a lack of experience and knowledge. Social bias, through peer and friend structures, can also be highly influential. Lastly, the rubric allows for subjective criteria, which makes it even harder to score objectively. These issues account for a greater lack of reliability within peer assessments than within assessments made by experts.

Peer assessments show the impact of biases mentioned in Power and Tanner (2023). Students' personal biases and lack of experiences in assessing work may lead to evaluations missing important components of the academic work. This is the reason why peer assessment is probably the least dependable as a tool in measuring the academic outcomes. This is not to say that peer assessment does not provide any value. As Chunping et al. (2024) notes, peer assessments can provide skill development and collaborative learning. Self-assessment, on the other hand, did show a moderate, positive, and statistically significant correlation with expert evaluation, with $r = 0.484$ and $p = 0.042$. This indicates that while students may not match the expert evaluation exactly, they do have a sense of their performance levels. In connection to these results, It can be said that even if students self-assess, they still lack the precision of expert work.

The moderate alignment serves as a positive indicator, particularly after students have been taught to criterion-based outline and constructive reflection elements. The findings from Milic & Simeunovic (2022) show that self-assessment can help evaluate a student's performance as a self-assessment tool; however, they indicate that in more challenging situations, expert evaluations remain more accurate. Moreover, (Ayyoub et al., 2021) noted that past training, established rubrics, and self-assessment process iterations are more reliable and accurate in self-assessment outcomes. Altogether, the data indicates that AI evaluation is the most consistent and dependable assessment compared to expert evaluation, with self-evaluation being slightly dependable, while peer evaluation was the most inconsistent with expert evaluation. This shows the prospect of AI being a dependable resource in assessing project work in a fair and precise manner.

In the described results, there is a statistically significant relationship ($\alpha=0.001$) that shows there is a relationship between the expert evaluation criterion to AI evaluation alone, which leads to the conclusion that there is a close alignment between the evaluations, and that AI can closely mimic expert evaluation, thus affirming its use in authentic assessment. In contrast to this, other evaluation methods do not show a significant relationship. In general, the results show that AI assessment is the most consistent and robust with expert assessment, followed by self-assessment to a lesser extent, while peer assessment was the least consistent with expert assessment, highlighting the limitations of this method due to the disparity in experience and objectivity among students.

In general, this was confirmed by (Dimitriadou & Lanitis, 2023) who indicated that AI excels as an independent assessment tool and is a promising option in the near future to replace or support traditional assessment methods. To see how expert assessments compare with other methods (AI, self-assessment, and peer assessment), Cohen's d was calculated for effect size. The small effect size from expert assessments and AI assessments show there is some alignment between the two (d = 0.257).

On the other hand, self-assessment showed a medium negative effect size (d = -0.682) which indicates that, on average, students rated their work higher than the experts did, showing a significant overestimation. Peer assessment showed a small to medium effect size (d = 0.394) which indicates some variability in the accuracy of students' assessments of each other's work .These differences highlight the challenges of maintaining consistency and objectivity in peer evaluations. These values enhance our understanding of the practical differences between the different assessment methods and demonstrate that AI provides more consistent results with expert assessment than self-assessment or peer assessment, supporting its potential use as a reliable assessment tool in higher education contexts.

Repeated Measures ANOVA test was also applied to study the differences between the results of the actual assessment of students' graduation projects using assessment methods (AI, peer and self-assessment). At first, the assumptions of the Repeated test were examined, and the results showed that the normal distribution was achieved, as shown in Table (1). The following was an examination of Sphericity, and the results were as shown in Table.(4)

Table 6: Mauchly's Test of Sphericity[a]

Measure: MEASURE_1

Within Subjects Effect	Mauchly's W	Approx. Chi-Square	Df	Sig.
assessment	.916	1.397	2	.497

Tests the null hypothesis that the error covariance matrix of the orthonormalized transformed dependent variables is proportional to an identity matrix.

a. Design: Intercept Within Subjects Design: Assessment Method

b. May be used to adjust the degrees of freedom for the averaged tests of significance. Corrected tests are displayed in the Tests of Within-Subjects Effects table.

The results showed that the P value for Mauchly's Test of Sphericity was 0.497, indicating that sphericity was achieved, and therefore there was no need to apply any corrections to the degrees of freedom in the repeated measures analysis of variance.

Table 7: Within Subjects Effects

Cases	Sum of Squares	Df	Mean Square	F	Sig.	η^2
RM Factor 1	680.148	2	340.074	6.768	0.003	0.285
Residuals	1708.519	34	50.251			

Note. Type III Sum of Squares

Analysis of variance to examine the effect of different assessment methods (self-assessment, peer assessment, and AI-based assessment) on students' grades in graduation projects showed statistically significant differences between the three assessment methods ($F(2, 34) = 6.77$, $p = 0.003$, Partial Eta Squared (η^2) = 0.285). This indicates that the assessment method significantly affects grades, with the assessment type explaining 28.5% of the grade variance, which confirms the practical importance of choosing the appropriate assessment method in the context of evaluating graduation projects. This makes the researcher want to dig deeper to find out what caused the differences in the three assessment methods that were analyzed. It becomes necessary to find out if they could unify assessment methods, and to see if they could combine the good things from the different methods to build an integrated assessment system that is fair and objective in measuring graduation projects. It also highlights the need to better understand the different factors that may lead to varying assessments as well as the different facets of student performance.

5. Conclusion

Undeniably, integrating AI with traditional methods and improving the assessment system will be indispensable in future assessments in education. They should also teach students the methods of assessing peers so that students understand the importance of objectivity before students' evaluations are used. In self-assessment, students need to explain the criteria they used, and instructors should provide constructive feedback to help deepen students' understandings of their performances and assist them in self-evaluating accurately. To integrate AI into assessment systems, the study advises policymakers and universities to devise focused and practical approaches followed by appropriate stakeholder training on AI tool usage. Assist in a seamless shift, AI integration should occur steadily, beginning with pilot initiatives to scale focused academic programs, faculties, and schools. In developing countries, free and open-source AI solutions may minimize the expenses associated with integration into academic programs and systems. Complementing the AI systems with human judgment, incorporating traditional techniques like peer and self-assessments, are also required to make assessments' systems more effective . In addition , they should incorporate the principles of incremental change, systems refinement, and feedback in relation to the prevailing educational context.

The assessment strategies should include innovative options within the scope of new assessment strategies. planned research should also focus on AI-based assessments as a factor on student motivation and engagement in various subjects and educational levels. In the end, the study also points out the need to resolve important issues and practical issues like data privacy, bias in the algorithms, and the transparency and fairness of how the assessments are carried out. It also recommends longitudinal studies to look at the long-term impacts of using different assessment techniques on education systems.

References

Al Maktoum, S. B., & Al Kaabi, A. M. (2024). Exploring teachers' experiences within the teacher assessment process: A qualitative multi-case study. *Cogent Education*, *11*(1), 2287931.

Awidi, I. T. (2024). Comparing expert tutor assessment of reflective essays with marking by generative artificial intelligence (AI) tool. *Computers and Education: Artificial Intelligence*, *6*, 100226.

Ayyoub, A. A., Assali, A., Eideh, B. A., & Suleiman, M. (2017). An Action Research Approach for Using Self/Peer Assessment to Enhance Learning and Teaching Outcomes. *Journal of Teaching and Teacher Education*, *5*(01), 33–42.

Ayyoub, A. A., Bsharat, A., & Suleiman, M. (2021). The impact of alternative authentic assessment outcomes in Palestinian fourth grade math classrooms. *Studies in Educational Assessment*, *70*(May), 101056. https://doi.org/10.1016/j.stueduc.2021.101056

Bower, M., Torrington, J., Lai, J. W. M., Petocz, P., & Alfano, M. (2024). How should we change teaching and assessment in response to increasingly powerful generative Artificial Intelligence? Outcomes of the ChatGPT teacher survey. *Education and Information Technologies*, 1–37.

Butler, Y. G. (2024). Self-assessment in second language learning. *Language Teaching*, *57*(1), 42–56.

Bujang, M. A., & Baharum, N. (2016). Sample size guideline for correlation analysis. World, 3(1), 37-46. doi:10.22158/wjssr.v3n1p37

Castillo-Martínez, I. M., Ramírez-Montoya, M. S., Glasserman-Morales, L. D., & Millán-Arellano, J. A. (2024). eComplex: validity and reliability of rubric for assessing reasoning for complexity competency. *Quality & Quantity*, *58*(2), 1545–1563.

Chunping, Z., Xu, C., Huayang, Z., & Ching Sing, C. (2024). Automated versus Peer Assessment: Effects on Learners' English Public Speaking. *Language Learning & Technology*, *28*(2), 210–228. https://search.ebscohost.com/login.aspx?direct=true&db=eric&AN=EJ1427980&site=ehost-live

Coskun, T. K., & Alper, A. (2024). Evaluating the evaluators: A comparative study of AI and teacher assessments in Higher Education. *Digital Education Review*, *45*, 124–140. https://doi.org/10.1344/der.2024.45.124-140

Dasgupta, A. P., Anderson, T. R., & Pelaez, N. (2014). Development and validation of a rubric for diagnosing students' experimental design knowledge and difficulties. *CBE Life*

Sciences Education, *13*(2), 265–284. https://doi.org/10.1187/cbe.13-09-0192

De La Rosa Gómez, A., Cano, J. M. M., & Díaz, G. A. M. (2019). Validation of a rubric to evaluate open educational resources for learning. *Behavioral Sciences*, *9*(12). https://doi.org/10.3390/bs9120126

Dhara, S., Chatterjee, S., Chaudhuri, R., Goswami, A., & Ghosh, S. K. (2022). Artificial Intelligence in Assessment of Students' Performance. In *Artificial Intelligence in Higher Education* (pp. 153–167). CRC Press.

Dimitriadou, E., & Lanitis, A. (2023). A critical assessment, challenges, and future perspectives of using artificial intelligence and emerging technologies in smart classrooms. *Smart Learning Environments*, *10*(1), 12.

Hodges, C. B., & Kirschner, P. A. (2024). Innovation of instructional design and assessment in the age of generative artificial intelligence. *TechTrends*, *68*(1), 195–199.

Hooda, M., Rana, C., Dahiya, O., Rizwan, A., & Hossain, M. S. (2022). Artificial intelligence for assessment and feedback to enhance student success in higher education. *Mathematical Problems in Engineering*, *2022*(1), 5215722.

Hu, J. (2021). Teaching Assessment System by use of Machine Learning and Artificial Intelligence Methods. *International Journal of Emerging Technologies in Learning (IJET)*, *16*(5), 87–101. https://www.learntechlib.org/p/220079

Iglesias Pérez, M. C., Vidal-Puga, J., & Pino Juste, M. R. (2022). The role of self and peer assessment in Higher Education. *Studies in Higher Education*, *47*(3), 683–692.

Ismail, S. M., & Heydarnejad, T. (2023). Probing into the influence of EFL learners' self-assessment and assessment apprehension in predicting their personal best goals and self-efficacy skills: a structural equation modeling. *Language Testing in Asia*, *13*(1), 8.

Kadri, N., & Amziane, H. (2021). Students' Attitudes about Self-Assessment: A Neglected Aspect in the Algerian EFL Classrooms. *The Educational Review, USA*, *5*(8), 275–286.

London, M., Sessa, V. I., & Shelley, L. A. (2023). Developing self-awareness: Learning processes for self-and interpersonal growth. *Annual Review of Organizational Psychology and Organizational Behavior*, *10*(1), 261–288.

Mao, J., Chen, B., & Liu, J. C. (2024). Generative Artificial Intelligence in Education and Its Implications for Assessment. *TechTrends*, *68*(1), 58–66. https://doi.org/10.1007/s11528-023-00911-4

Milic, S., & Simeunovic, V. (2022). Concordance between Giftedness Assessments by Teachers, Parents, Peers and the Self-Assessment Using Multiple Intelligences. In *High Ability Studies* (Vol. 33, Issue 1, pp. 1–19).

Murray, S. A. (2025). Exploring Quantum Machine Learning-Enhanced Models for EEG Data Classification (Doctoral dissertation).URL: https://hdl.handle.net/1773/53503

Norova, M. (2020). Advantages And Disadvantages of Traditional and Alternative Ways of Assessment. *ЦЕНТР НАУЧНЫХ ПУБЛИКАЦИЙ (Buxdu. Uz)*, *2*(2).

Palm, T. (2019). Performance assessment and authentic assessment: A conceptual analysis of the literature. *Practical Assessment, Research, and Assessment*, *13*(1), 4.

Planas-Lladó, A., Feliu, L., Arbat, G., Pujol, J., Suñol, J. J., Castro, F., & Mart\'\i, C. (2021). An analysis of teamwork based on self and peer assessment in higher education. *Assessment \& Assessment in Higher Education*, *46*(2), 191–207.

Power, J. R., & Tanner, D. (2023). Peer Assessment, Self-Assessment, and Resultant Feedback: An Examination of Feasibility and Reliability. In *European Journal of Engineering Education* (Vol. 48, Issue 4, pp. 615–628).

Rudolph, J., Tan, S., & Tan, S. (2023). War of the chatbots: Bard, Bing Chat, ChatGPT, Ernie and beyond. The new AI gold rush and its impact on higher education. *Journal of Applied Learning and Teaching*, *6*(1), 364–389.

Shabara, R., ElEbyary, K., & Boraie, D. (2024). TEACHERS or CHATGPT: The ISSUE of ACCURACY and CONSISTENCY in L2 ASSESSMENT. *Teaching English with Technology*, *24*(2), 71–92. https://doi.org/10.56297/vaca6841/LRDX3699/XSEZ5215

Stevens, D. D., & Levi, A. J. (2023). *Introduction to rubrics: An assessment tool to save grading time, convey effective feedback, and promote student learning*. Routledge.

Swiecki, Z., Khosravi, H., Chen, G., Martinez-Maldonado, R., Lodge, J. M., Milligan, S., Selwyn, N., & Gašević, D. (2022). Assessment in the age of artificial intelligence. *Computers and Education: Artificial Intelligence*, *3*, 100075.

Taylor, B., Kisby, F., & Reedy, A. (2024). Rubrics in higher education: an exploration of undergraduate students' understanding and perspectives. Assessment & Assessment in Higher Education, 49(6), 799–809. URL:

https://doi.org/10.1080/02602938.2023.2299330
Tomczyk, Ł., Limone, P., & Guarini, P. (2024). Assessment of modern educational software and basic digital competences among teachers in Italy. *Innovations in Education and Teaching International*, *61*(2), 355–369.
Vogt, W. P., Gardner, D. C., & Haeffele, L. M. (2012). *When to use what research design*. Guilford Press.
Wahyudin, A. Y., Pustika, R., & Simamora, M. W. B. (2021). Vocabulary learning strategies of EFL students at tertiary level. *The Journal of English Literacy Education: The Teaching and Learning of English as a Foreign Language*, *8*(2), 101–112.
Yan, Z., Brown, G. T. L., Lee, J. C.-K., & Qiu, X.-L. (2020). Student self-assessment: Why do they do it? *Educational Psychology*, *40*(4), 509–532.
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*, 100484.
Yan, Z., Panadero, E., Wang, X., & Zhan, Y. (2023). A systematic review on students' perceptions of self-assessment: usefulness and factors influencing implementation. *Educational Psychology Review*, *35*(3), 81.
Yan, Z., Wang, X., Boud, D., & Lao, H. (2023). The effect of self-assessment on academic performance and the role of explicitness: a meta-analysis. Assessment \& Assessment in Higher Education, 48(1), 1–15. URL: https://doi.org/10.1080/02602938.2021.2012644
Yu, H. (2024). The application and challenges of ChatGPT in educational transformation: New demands for teachers' roles. *Heliyon*.
BioMed Central. (2024). Assessment practices in Palestinian higher education: Consistency and challenges. https://edintegrity.biomedcentral.com/articles/10.1007/s40979-024-00160-9
SpringerLink. (2025). Peer and self-assessment in higher education: Motivational and reflective impacts. https://link.springer.com/article/10.1186/s41239-025-00522-4
Taylor & Francis Online. (2025). Generative AI vs. instructor vs. peer assessments: A comparison of grading and feedback in higher education. https://www.tandfonline.com/doi/full/10.1080/02602938.2025.2487495