DIGITALEDUCATIONREVIEW

Evaluating the evaluators: A comparative study of AI and teacher assessments in Higher Education

Tuğra Karademir Coskun[1,*], Ayfer Alper[2]
[1]University of Sinop, Turkey, tkarademir@sinop.edu.tr, https://orcid.org/0000-0003-4295-2440
[2]University of Ankara, Turkey, aalper@ankara.edu.tr, https://orcid.org/0000-0003-2312-6311

ABSTRACT

This study aims to examine the potential differences between teacher evaluations and artificial intelligence (AI) tool-based assessment systems in university examinations. The research has evaluated a wide spectrum of exams including numerical and verbal course exams, exams with different assessment styles (project, test exam, traditional exam), and both theoretical and practical course exams. These exams were selected using a criterion sampling method and were analyzed using Bland-Altman Analysis and Intraclass Correlation Coefficient (ICC) analyses to assess how AI and teacher evaluations performed across a broad range. The research findings indicate that while there is a high level of proficiency between the total exam scores assessed by artificial intelligence and teacher evaluations; medium consistency was found in the evaluation of visually based exams, low consistency in video exams, high consistency in test exams, and low consistency in traditional exams. This research is crucial as it helps to identify specific areas where artificial intelligence can either complement or needs improvement in educational assessment, guiding the development of more accurate and fair evaluation tools.

KEYWORDS: Artificial intelligence tool-based assessment systems, teacher evaluation, assessments in higher education

1 INTRODUCTION

Artificial Intelligence (AI) significantly contributes to educational processes, possessing the potential to enhance efficiency, productivity, and personalized learning experiences (Dwivedi et al., 2021; Haseski, 2019). AI technologies allow educators to tailor teaching to meet individual student needs, enhance the learning environment, and gain deep insights into student performance (Yin, 2021). Through the automation of tasks, the creation of personalized learning programs, and the optimized use of data, AI facilitates more accurate knowledge acquisition and targeted educational outcomes (Hinojo-Lucena et al., 2019; Hua, 2022). Moreover, AI supports the development of students' cognitive abilities, encourages creativity, maximizes resource utilization, and contributes to the overall advancement of teaching and learning (Jingshan, 2023). AI is also instrumental in the design of intelligent tutoring systems, enhancing learning outcomes across various academic subjects and educational levels (Dermeval et al., 2017). It enables the development of innovative teaching methodologies and curriculum designs, thus creating a more engaging and effective learning environment (Yang & Wang, 2020). Educators can use AI technologies to develop interactive and adaptive learning pathways that accommodate diverse student needs and learning styles (Tapalova & Zhiyenbayeva, 2022). In summary, the use of AI in education is crucial for adapting to rapid technological changes, transforming teaching methodologies, increasing student engagement, and ultimately enhancing teaching standards.

When integrated into measurement and evaluation processes in education, Artificial Intelligence (AI) offers numerous benefits. AI technologies can improve assessment and feedback mechanisms, thereby enhancing students' learning outcomes (Mahligawati, 2023; Choi and McClenen, 2020). These technologies take on functions such as identifying students' needs, providing instant feedback, and facilitating teacher intervention, which in turn boosts teaching effectiveness (Dindar et al., 2022). AI tools enable personalized instruction, efficient assessment, and timely feedback, significantly increasing students' conceptual understanding, engagement, and motivation (Liu et al., 2020). AI-based assessment systems offer accurate and consistent feedback, saving educators time and increasing the effectiveness of the evaluation process (Harry, 2023). The role of AI in education, teaching, and learning processes has been recognized by pre-service teachers, indicating its potential to improve evaluation procedures (Haseski, 2019). The convergence of human and machine learning through AI presents the potential for more effective teaching and learning than could be achieved by humans or AI alone (Luan et al., 2020). However, it is crucial for educators to fulfil their responsibilities in exam evaluations. Human evaluations can vary due to personal preferences, experiences, and individual judgments, leading to inconsistencies in grading and feedback. Particularly in large classes, the time and effort required for manual evaluation can complicate the provision of timely and consistent feedback to students (Cañada et al., 2014). In this context, the role of AI in evaluation processes becomes even more critical as it helps to optimize educational practices, provide personalized feedback, and improve learning outcomes.

Artificial Intelligence (AI) can offer a viable alternative to overcome the limitations associated with self-assessment in examinations. The use of AI in analysing teaching materials and evaluations can enhance the accuracy and objectivity of the process, thereby reducing bias and subjectivity (Amin, 2020; Köse & Arslan, 2017). AI's capability to efficiently and swiftly analyse large datasets enables educators to gain deeper insights into student learning outcomes and tailor educational strategies accordingly (Kazimov et al., 2021). Furthermore, AI helps maintain the integrity of exams and tests by detecting anomalies or irregularities (Amin, 2020;

Popenici & Kerr, 2017). The advancement of the underlying algorithms in artificial intelligence (AI) systems is among the significant factors affecting the ability of AI to make accurate and stable predictions. These algorithms enable the identification of patterns and relationships within datasets that are not apparent through traditional methods (Myszczynska et al., 2020). However, it is pertinent to discuss whether these AI algorithms produce consistent, reliable, and valid results. Numerous studies have emphasized the effectiveness of AI algorithms in enhancing the reliability of evaluations across various medical domains. For example, Keskinbora and Güven (2020) highlighted the use of AI algorithms in ophthalmology, particularly in diabetic retinopathy and age-related macular degeneration, demonstrating the reliability of AI in assisting with disease diagnosis. Additionally, Wang, Zhang, Wu and Wang (2021) illustrated the effectiveness of AI algorithms in multimodal MRI analysis for cervical cancer diagnosis, emphasizing the high diagnostic accuracy achieved through AI assistance. Chen, Chen and Lin (2020) discussed the components of AI education systems, highlighting the role of intelligent algorithms in improving teaching models and learner interactions. This underscores the potential for AI to enhance the validity and reliability of educational assessments by refining the assessment process. Furthermore, Zawacki-Richter, Marin, Bond and Gouverneur (2019) identified assessment and evaluation processes as critical areas for AI applications in higher education. Their systematic review emphasized the use of AI in profiling, prediction, and adaptive systems, showcasing the potential for AI to bolster the validity of evaluations through personalized and intelligent assessment approaches. Additionally, AI has shown potential in improving diagnostic accuracy, managing acute conditions, and even enhancing student presentation skills (Sandhu et al., 2020; Chen et al., 2022). In the context of high-stakes assessments in education and medicine, AI, particularly through neural networks and deep learning, is being utilized to ensure fair, valid, and reliable assessments for various purposes (Richardson & Clesham, 2021; Lee, Wu, Li & Kulasegaram, 2021). AI algorithms have been transforming the analysis, diagnosis, and treatment of medical conditions, emphasizing the need for medical professionals to not only use AI tools but also oversee and evaluate them to ensure safe integration into practice (Wiljer et al., 2021). Moreover, AI literacy among medical students and professionals is crucial for understanding and utilizing AI effectively in healthcare settings (Laupichler et al, 2024). The ethical implications of using AI in scoring assessments, especially in high-stakes situations, need to be carefully considered to ensure fairness and accuracy (Fiske, Henningsen & Buyx, 2019).

While AI shows promise in enhancing assessment processes, the accuracy and reliability of AI tools in grading assessments depend on factors such as data quality, the design of AI models, and the context in which they are applied (Elder et al., 2022). Evaluating the accuracy of AI models in providing clinical insights or grading student performance involves assessing factors such as accessibility, informativeness, and overall effectiveness (Ishaaq & Sohail, 2023). Additionally, it has been suggested that integrating human intelligence with AI algorithms can increase the reliability of predictions in applications (Ed-Driouch, Gourraud, Dumas & Mars, 2022). One key factor in ensuring the reliability of AI algorithms is explainability. Gunning et al. (2019) emphasize the importance of explainable artificial intelligence (XAI) to help users understand, trust, and effectively manage AI applications. Explainability enhances transparency and accountability in the process by allowing users to understand how AI makes decisions (Gunning et al., 2019). This transparency is critical for users to trust the outcomes provided by AI algorithms. Furthermore, the ethical implications of AI algorithms cannot be overlooked. Jobin and Ienca (2019) highlight the global landscape of AI ethical guidelines, emphasizing the importance of transparency and ethical evaluations in the development and implementation of AI systems. Ethical guidelines play a vital role in ensuring the responsible use of AI algorithms and in guaranteeing that the results they provide are reliable and unbiased (Jobin & Ienca, 2019).

Specifically, AI applications in the form of automated assessment and tutoring tools use machine learning and natural language processing to evaluate students' work and provide timely feedback (Bai & Stede, 2022). These tools aim to reduce the workload on teachers and offer immediate feedback to students, thereby enhancing the learning process (Tubino & Adachi, 2022). Additionally, AI models that incorporate generative adversarial networks and attention mechanisms help predict students' attitudes and behaviours in the classroom, assisting teachers in understanding student performance and engagement through data analysis (Zhao & Song, 2022). Compared to self-evaluation by teachers, AI-based evaluation systems provide advantages of objectivity, consistency, efficiency, and scalability.

While the integration of AI into educational assessment processes offers numerous benefits, addressing concerns related to data privacy, security, and ethical issues is crucial. AI tools can analyse large datasets to identify patterns and trends in student performance, which assists educators in making data-driven decisions and interventions (Parapadakis, 2020). However, potential biases in algorithms, a lack of transparency in decision-making processes, and the possibility of errors in automated assessments raise questions about the accuracy and fairness of AI-supported evaluations (Yu & Yu, 2023). Educators can use the power of AI to improve evaluation processes while maintaining ethical standards and ensuring student safety (Collins et al., 2021). While AI can automate aspects of feedback provision, research highlights the importance of integrating human expertise with AI tools for comprehensive feedback, underscoring the potential for AI-human collaboration in education (Nguyen, 2023).

The objectivity and fairness of educational processes, particularly in the evaluation of exam papers, play a critical role in accurately measuring academic success and shaping students' career paths. However, significant inconsistencies exist among current educational evaluation methods. Assessments made by instructors can vary due to subjective biases and personal interpretations, which complicates the objective evaluation of student performance and adversely affects the quality of education (Cañada et al., 2014). Artificial Intelligence (AI) holds great promise for transforming teaching processes by enhancing student achievement, providing personalized learning experiences, and improving educators' efficiency (Dwivedi et al., 2021; Haseski, 2019). While AI-based assessment systems can offer objective and consistent feedback through the analysis of large datasets, they also face inherent issues such as algorithmic biases and a lack of transparency (Yu & Yu, 2023). Studies have shown that AI tools can provide accurate and understandable suggestions (Liu et al., 2023). However, the usefulness, acceptance, and validity of these suggestions may vary. Therefore, the degree of consistency in using generative AI tools for assessment purposes, compared to expert evaluations, may depend on the specific context, the nature of the assessment tasks, and the level of expertise required for evaluation. The reliability of AI in scoring assessments has not yet been fully

validated by experimental data in some contexts (Kawaji et al., 2019).

Therefore, identifying potential discrepancies and alignments between instructor evaluations and AI tools is crucial for understanding how these technologies can be more effectively utilized in education. The compatibility between the scores given by instructors and those generated by AI tools has not been sufficiently explored. Instructor evaluations are prone to variability influenced by subjective factors, such as biases or personal expectations. On the other hand, AI-based evaluation tools promise to provide an objective framework through specific algorithms and datasets. Thus, examining the relationship between these two evaluation methods is essential for ensuring fairness and objectivity in education.

2 RESEARCH OBJECTIVE

Starting from this point; the fundamental goal of this study is to identify the consistencies and potential discrepancies between the scores given by instructors and artificial intelligence tools on exam papers across various formats. This research evaluates exam papers in classical, testing, project, video, and poster formats. The following questions are addressed within this scope:

(1) What is the level of consistency between the scores obtained from artificial intelligence tools and instructor evaluations for university students' exam papers across different types of exams (traditional, test, project – video/poster)?

(2) What is the level of consistency between the scores obtained from artificial intelligence tools and instructor evaluations for university students' project-based assessments?

(3) What is the level of consistency between the scores obtained from artificial intelligence tools and instructor evaluations for university students' tests?

(4) What is the level of consistency between the scores obtained from artificial intelligence tools and instructor evaluations for university students' traditional exams?

3 SIGNIFICANCE OF STUDY

The fundamental goal of this research is critically important as it aims to bridge the gap between traditional educational assessment methods and innovative artificial intelligence (AI) technologies. By examining the consistencies and discrepancies between the scores given by instructors and AI tools across various exam formats—such as classical tests, projects, videos, and posters—this study seeks to validate the reliability and accuracy of AI in educational settings. Investigating these consistencies across different types of assessments (traditional, project-based, tests, and traditional exams) is essential to understand how AI can be effectively integrated to complement and enhance the traditional evaluation processes. This research is pivotal in advancing educational assessment by potentially offering more objective and consistent evaluations compared to the subjective assessments often associated with human graders. Additionally, by exploring how AI evaluations align with human judgments across diverse educational formats, the study will provide valuable insights into customizing AI tools for varied educational purposes, thereby supporting more personalized and effective learning experiences for university students. This alignment check could revolutionize educational practices by ensuring fairer and more accurate assessments, ultimately improving learning outcomes and educational equity.

4 METHOD

4.1 Research Design

The study aims to assess the alignment and potential misalignments between evaluations conducted by instructors and artificial intelligence systems. A cross-sectional research design has been chosen to achieve this objective. This design allows for cross-sectional analyses on data collected from different sample groups within a specific timeframe, facilitating the examination of relationships and differences among variables (Babbie, 2016). This design is particularly effective when the research needs to occur at a single point in time, referred to as a "cross-section."

Several reasons underlie the selection of this design. Firstly, it permits the simultaneous examination of multiple variables, such as the assessment scores from both instructors and artificial intelligence tools. Secondly, the cross-sectional design provides a framework suitable for the collection and analysis of quantitative data, which in turn supports the application of statistical analyses to effectively explore the relationships between variables. Lastly, employing a cross-sectional design allows the research to be conducted in a practical and cost-effective manner, as the data collection process is confined to a single period (Creswell & Creswell, 2018). Additionally, the ability to simultaneously investigate multiple variables makes this design particularly apt for studies examining the relationships and differences between various assessment methods.

4.2 Data Collection Methods

The exam papers analyzed in this research include a variety of formats and originate from different academic departments at the university. This diversity offers a unique opportunity to compare how instructors and artificial intelligence (AI) tools evaluate multidisciplinary student work, particularly in terms of how AI processes and integrates information from different disciplines. Such a comparison is crucial for assessing the support and promotion capacities of AI tools and instructors within multidisciplinary approaches. Therefore, the inclusion of various exam formats was deemed essential in this study, rather than focusing on a single format. The selection criteria for the different types of exams considered are outlined below:

(1) *Numerical and Verbal Course Exams:* The study encompasses exams that assess both numerical and verbal abilities, as these exams evaluate different cognitive and problem-solving skills (Pellegrino & Quellmalz, 2010). Understanding how AI and instructors assess these diverse skills is vital.

(2) *Exams Featuring Open-ended Questions*: Open-ended questions allow students to demonstrate their knowledge and understanding in detail (Brookhart, 2018). Evaluating these types of questions is crucial for observing how instructors and AI analyze unique responses. In this study, classical exams consist of open-ended questions.

(3) *Multiple-choice Tests*: These exams measure students' knowledge quickly and objectively (Haladyna, Downing, & Rodriguez, 2002). This format is where AI excels, making

it suitable for comparing the evaluation processes of instructors and AI.

(4) *Project-based Exams:* These exams assess students' abilities to conduct in-depth research and solve complex problems (Thomas, 2000). They are used to analyze how instructors and AI evaluate creativity and problem-solving strategies.

(5) *Exams Submitted in Visual Formats*: Visual materials, especially in subjects like art and design, assess students' creativity and technical skills (Eisner, 2002). Such exams are crucial for showcasing AI's capacity to analyze visual content and the differences in evaluations between AI and instructors.

(6) *Video-based Exams*: Particularly in communication and performance arts courses, these exams evaluate students' presentation skills and creativity (Dyment & O'Connell, 2011). They provide an opportunity to examine how both instructors and AI assess non-verbal communication and performance.

(7) *Exams from Various Disciplines:* Including exams from different disciplines broadens the scope of the study and shows how instructors and AI evaluate knowledge and skills across various fields (Becher & Trowler, 2001). This diversity is essential for understanding the interdisciplinary differences in evaluation processes.

(8) *Practical and Theoretical Course Exams:* Practical exams evaluate how students apply their practical skills in labs, fieldwork, or studio work, showcasing how they integrate theoretical knowledge into practical applications (Freeman et al., 2014). Theoretical exams, on the other hand, measure how well students understand the concepts, theories, and principles within a specific discipline. Both types of exams provide significant insights into how instructors and AI assess student achievement.

The formats and distributions of the exams evaluated in this study are detailed in Table 1.

Exam Formats	Description of Exam Type	Relevant Course	Number of	Description of Question Types	Number of	Percentage of	Cumulative %
Classical Exams	Theoretical Exams; assess students' knowledge and understanding of a specific course or topic.	Ottoman Modernization Movements	5	Consists of open-ended questions	23	36.4%	36.4%
	Exams Requiring Technical Knowledge; measure expertise in fields like engineering, science, and math.	Technical Engineering Drawings Course	4	Includes open-ended questions and drawings	20		
Test Exams	Verbal and Numerical Course Exams; feature objective types such as multiple-choice and true/false.	Distance and Open Learning Course	40	Composed of multiple-choice test questions	19	16.1%	52.5%
		Robotics with Arduino Course	40	Composed of multiple-choice test questions			
Project Exams (Video)	Requires students to present their projects or research in video format.	Game Development with Machine Learning	1	Involves creating a video of the development and gaming steps	18	15.3%	67.8%
Project Exams (Poster)	Requires students to combine research results or projects with visual and textual content in a poster.	Instructional Technologies	1	Involves developing materials for one question	38	32.2%	100.0%

Table 1: Distribution and Characteristics of Evaluated Exam Papers

Upon reviewing Table 1, it is evident that classical exams are the most frequent, comprising 36.4% of the data analyzed, followed by test exams at 16.1%. Project exams presented in video format involve 18 students and make up 15.3% of the exams, while project exams in poster format are the second most common type, involving 38 students and constituting 32.2% of the total. This table illustrates the diverse assessment methods and their distribution among the student population, providing a comprehensive overview of the examination formats evaluated in this study.

4.3 Data Analysis

This section is divided into two main parts. The first part discusses the scoring criteria used by instructors and the characteristics of the respective exams, while the second part provides examples of

prompts and plugins used in the artificial intelligence assessment of the exam papers.

4.3.1. Instructor Assessment Process

A total of 118 exam papers from various courses were reviewed and scored by six evaluators, whose academic ranks range from Associate Professor to Professor. These evaluators are faculty members responsible for designing and teaching the course content, each possessing extensive experience and expertise in their respective fields. The scores for each exam were assigned on a scale from 0 to 100, with 0 being the lowest possible score and 100 the highest. The process by which instructors evaluate the exams includes specific examples of questions and the criteria for assessment, which are detailed, and examples given below (table 2).

Examination of Classical Question-Based Tests

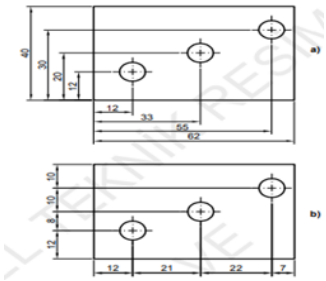
Theoretical Exam: Course- Ottoman Modernization History

***Example Question*:** What term was used to denote political power in the Göktürk Empire?

Assessment Criteria:

The answer must explicitly mention that the term "Kut" was used to describe political power in the Göktürk Empire. Explanation should detail that "Kut" represents the belief, in ancient Turkic culture, that sovereigns were divinely appointed, receiving legitimacy and protection from the gods. Provide examples or explanations of how the "Kut" concept was employed during the Göktürk era, such as during enthronement ceremonies of a Khagan, illustrating the ritualistic aspects of the "Kut" concept. Discuss how the "Kut" influenced political authority, e.g., by enhancing a ruler's legitimacy or by increasing the ruler's authority among the populace and nobility. Analyze how governance and the notion of rulership in Göktürk were intertwined with the "Kut" concept.

Technical Exam: Course-Technical Engineering Drawings



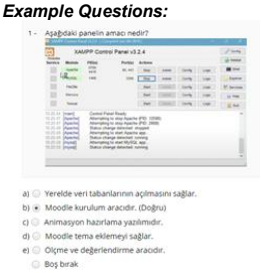
***Example Question*:** Produce parallel and chain dimensioning drawings using the specified software.

Assessment Criteria: The drawings should be performed flawlessly and included in the exam paper with all necessary explanations provided.

Examination of Test Question-Based Exams

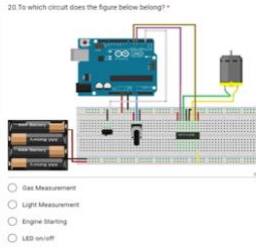
Verbal Course Exam: Course- Open and Distance Learning

Example Questions:



Numerical Course Exam: Course-Arduino and Robotics Applications

Example Questions:



Project Evaluation/ Poster Examination: Course: Instructional Technologies

Assessment Criteria and Guidelines:

Relevance to Selected Days and Weeks (20 Points) - Question: Is the selected day or week relevant to the subject area, and is adequate information provided?

- 0-5 Points: The day or week chosen is unrelated to the topic, or very little information is provided.
- 6-10 Points: The day or week chosen is slightly relevant, but minimal information is offered.
- 11-15 Points: The day or week chosen is moderately appropriate, and sufficient information is provided.
- 16-20 Points: The day or week chosen is highly relevant, and comprehensive information is provided.

Appropriateness of Images Used (20 Points)- Question: Are the images used relevant to the topic and of good visual quality?

- 0-5 Points: Images are irrelevant to the topic and/or of poor quality.
- 6-10 Points: Images are somewhat relevant with moderate quality.
- 11-15 Points: Images are appropriate and of high quality.
- 16-20 Points: Images are very appropriate and of superior quality.

Color Coordination (20 Points)- Question: Do the colors used in the poster harmonize with each other and the topic?

- 0-5 Points: Colors do not harmonize with each other or the topic.
- 6-10 Points: Colors are somewhat harmonious.
- 11-15 Points: Colors generally harmonize well.
- 16-20 Points: Colors are perfectly harmonious and enhance the subject matter.

Typography and Size (20 Points)- Question: Is the font size readable and does it support the visual composition?

- 0-5 Points: The font size is too small or too large and/or hard to read.
- 6-10 Points: The font size is readable but could be optimized.
- 11-15 Points: The font size is appropriate and readable.
- 16-20 Points: The font choice is excellent, both aesthetically and in terms of readability.

Overall Aesthetic and Appearance (20 Points)- Question: Does the overall appearance of the poster look aesthetic and pleasing to the eye?

- 0-5 Points: The poster is not visually appealing and is disorganized.
- 6-10 Points: The poster is slightly appealing with some disorganization.
- 11-15 Points: The poster is aesthetically pleasing and well organized.
- 16-20 Points: The poster is highly aesthetic and all elements are perfectly integrated.

Project Evaluation/ Video Examination: Course- Game Development using Machine Learning

Example: Creating a video that outlines the development steps in a game development project using machine learning.

Assessment Criteria and Guidelines:

Alignment with Learning Outcomes (25 Points): Evaluate how well the project or study meets the defined learning outcomes or objectives.

- 0-6 Points: Little to no relevance to the learning outcomes.
- 7-13 Points: Partially meets the learning outcomes, but with several deficiencies.
- 14-19 Points: Covers most of the learning outcomes, good level of alignment.
- 20-25 Points: Fully meets all the learning outcomes, excellent coverage.

Usage and Explanation of Code (25 Points): Assess the use of programming codes within the project and how these codes are explained.

- 0-6 Points: No use of code or no information provided about the codes.
- 7-13 Points: Codes are used but not enough information is provided about their usage.
- 14-19 Points: Good use of code; detailed information about the codes is partially provided.
- 20-25 Points: Excellent use of code; detailed and clear explanations about how the codes are used.

Appropriateness of Visual Design (25 Points): Evaluate how suitable the visual elements (colors, shapes, fonts, etc.) are for the purpose and content of the project.

- 0-6 Points: The visual design is not suitable for the purpose of the project.
- 7-13 Points: The design is somewhat appropriate but has significant disorganization.
- 14-19 Points: The visual design is mostly appropriate and organized.
- 20-25 Points: The visual design is completely appropriate and aesthetically flawless.

Aesthetics and Playability (25 Points): Assess the aesthetic appeal and the ease of use for players or users.

- 0-6 Points: Neither aesthetic nor playable.
- 7-13 Points: Somewhat aesthetic and/or has playability issues.
- 14-19 Points: Good level of aesthetics and playability.
- 20-25 Points: Very aesthetic and has excellent playability.

4.3.2. Evaluation by Artificial Intelligence Tools

Artificial intelligence tools, specifically the latest language model developed by OpenAI, ChatGPT-4, along with various plugins, were employed in the evaluation process of the exams. This model is considered a significant advancement in the fields of artificial intelligence and natural language processing (NLP), reputed for its superior comprehension and language production capabilities compared to its predecessors. Studies have also benchmarked ChatGPT-4 against the earlier version, ChatGPT-3.5, in evaluations such as the Ophthalmology Knowledge Assessment Program (OKAP) examination, where it demonstrated enhanced performance (Teebagy et al., 2023). Trained on large datasets, ChatGPT-4 can generate natural and human-like texts across a broad range of topics. It holds the potential to transform academic and library sciences in various ways, offering novel opportunities for educational and research settings (Lund & Wang, 2023). In this context, ChatGPT-4 was directly utilized for grading classical exams and tests, while its "Web Search" plugin was used for poster evaluations. The "Web Search" tool can gather information from internet links and perform visual analysis using the infrastructure of ChatGPT-4. For video content evaluations, the "Video Summarize" plugin of ChatGPT-4 was chosen. This plugin allows for the transcription and subsequent analysis of video content from platforms such as YouTube. The assessment process primarily leveraged the natural language processing capabilities of ChatGPT-4.

During the evaluation processes, the use of artificial intelligence (AI)-based tools played a crucial role in examining the exam papers. Initially, students' exam papers were introduced to the AI system, which then conducted assessments based on provided instructions. The prompts used in the AI-supported evaluation process were reviewed by experts in education to ensure they accurately measured the students' knowledge and skills in alignment with curriculum goals and learning outcomes. These instructions were designed to establish a general framework for assessment rather than specific details, allowing the AI tools to develop and apply their own assessment criteria. Based on these criteria, the AI assigned scores on a 100-point scale for each student's exam paper. The resulting data were subsequently categorized and analyzed in Microsoft Excel for further analysis and comparison. This categorization considered the varying assessment criteria and AI plugins used, depending on the type of exam. For example, specific AI plugins that best met the assessment criteria for exams in different disciplines, such as mathematics and literature, were selected and applied. These plugins were chosen and implemented to meet the unique requirements of each exam type. Details of the evaluation process and the analysis of the collected data are presented in the appendices of the study. Below are the prompts used in the evaluation process and examples of a few assessment outcomes:

Evaluation of Classical Exams Comprising Theoretical and Technical Questions: The ChatGPT-4 was utilized to assess exams containing classical questions. An example question, the corresponding prompt, and the evaluation by ChatGPT-4 are provided below.

Course	Ottoman Modernization History
Software Used:	Chat GPT 4
Prompt:	Before starting the evaluation process, please assess each student's answers individually based on the following criteria. Each question can be awarded a maximum of 20 points, with the total score for the entire exam not exceeding 100 points. The assessment criteria are as follows: **Content of the Answer (10 Points)** Comprehensive and Complete Answer (8-10 Points): The question is fully and correctly answered, covering all relevant key points. Adequate but Incomplete Answer (5-7 Points): The question is partially answered, missing some key information. Insufficient, Incorrect, or Irrelevant Answer (0-4 Points): The answer is inadequate, contains little information, or includes incorrect details. **Depth and Detail of the Answer (5 Points)** Rich Content and Deep Analysis (4-5 Points): The answer provides a detailed discussion of the topic, supported by examples and explanations. Moderate Detail and Analysis (2-3 Points): There are some explanations and examples, but more detail could be added. Superficial and Minimal Detail (0-1 Point): The answer is superficial, lacking in detail and examples. **Use of Language and Expression (5 Points)** Flawless Language Use (4-5 Points): There are no spelling, grammar, or punctuation errors; the response is fluid and comprehensible. Understandable but Grammatically Incorrect Language (2-3 Points): There are some errors in language use, but the answer is understandable. Numerous Grammatical Errors and Difficult to Understand (0-1 Point): There are extensive spelling and grammar mistakes, making the answer hard to understand. Implementation Instructions: Upload each student's answers separately. Score each answer according to the criteria outlined above. Record the scores and use them to provide feedback to students.
Evaluation Results	**Question 1.** The Role of Proverbs in Uyghur Education History: (17/20) Mastery of Subject: 4/4; Clarity and Detailing: 3/4; Creativity and Originality: 3/4,Organization and Presentation: 4/4; Grammar and Spelling: ¾; **Question 2.** Rules of the Manichaean Religion: (16/20) Mastery of Subject: 4/4; Clarity and Detailing: ¾ ; Creativity and Originality: ¾ ; Organization and Presentation: ¾ ; Grammar and Spelling: 3/4 **Question 3.** Characteristics and Duties of Göktürk Khagans: (18/20) Mastery of Subject: 4/4 ; Clarity and Detailing: 4/4 ; Creativity and Originality: 4/4 ; Organization and Presentation: ¾; Grammar and Spelling: 3/4 **Question 4.** Political Power in the Göktürk Empire: (15/20) Mastery of Subject: 4/4; Clarity and Detailing: ¾ ; Creativity and Originality: 2/4; Organization and Presentation: ¾; Grammar and Spelling: 3/4 **Question 5**. Dietary Practices in Hun Society: (17/20) Mastery of Subject: 4/4; Clarity and Detailing: 4/4; Creativity and Originality: ¾; Organization and Presentation: 3/4; Grammar and Spelling: 3/4 **Total Score for Student 1: 83/100**

Evaluation of Exams Comprising Test Questions Using ChatGPT-4. For the evaluation of verbal and numerical course exams, ChatGPT-4 has been utilized. Below is an example question, the associated prompt used for AI evaluation, and the outcomes based on ChatGPT-4's assessment:

Course	Open and Distance Learning
Software Used:	Chat GPT 4
Prompt:	*Control* Panel Data Reporting Administrator I-II-III Quality ... (other questions and responses) *Evaluation of Student Answers*: Each student's answer sheet will be uploaded. ChatGPT will compare each given answer to the correct answer. Each correct answer will earn the student 5 points. After all questions have been evaluated, the total score for each student will be calculated.

Evaluation Results	# Student answers from the previous input before the technical issue
	student_answers_from_previous_input = {
	"id1": ["d", "b", "d", "d", "c", "b", "a", "e", "e", "c", "c", "a", "a", "e", "c", "c", "e", "c", "c", "c"],
	"id2": ["d", "b", "d", "d", "c", "b", "d", "e", "e", "c", "e", "a", "a", "e", "a", "a", "e", "b", "c", "c"],
	"id3": ["c", "a", "d", "d", "c", "e", "a", "e", "e", "c", "d", "c", "a", "a", "a", "a", "e", "c", "c", "c"],
	"id4": ["d", "a", "d", "a", "c", "b", "a", "e", "e", "c", "a", "a", "a", "e", "c", "d", "e", "b", "c", "c"]
	} # Correct answers provided just now by the user
	correct_answers_updated = [
	"d", "a", "d", "d", "c", "b", "a", "a", "e", "c", "a", "a", "a", "d", "c", "d", "e", "b", "c", "c"
	]# Now let's recalculate the scores for each student with the updated correct answers

Evaluation of Project Exams. Project exams have been evaluated in two different formats. For the evaluation of visuals (posters), the "Web Search" plugin of GPT-4 was utilized. GPT-4's "Web Search" plugin combines AI's natural language processing capabilities with the ability to search information on the internet, allowing it to instantly generate answers to user queries from the web. In the context of project exams, the information on the posters created by students is verified online using the "Web Search" plugin. This process involves comparing historical data, statistical figures, and scientific claims found on the posters with reliable sources to ensure the information is current and accurate, thereby validating the sources' authenticity.

Below, the prompt used in the poster evaluation process and the assessment by Chat GPT-4 are provided.

Course	Instructional Technologies
Software Used:	Chat GPT 4 + Web Search
Prompt:	Students have prepared posters about specific days and weeks. You are required to evaluate these posters according to the criteria listed below. A maximum of 20 points can be awarded for each criterion, with a total possible score of 100 points. **Evaluation Criteria and Questions:** 1.Selection and Information Content of Specific Days and Weeks (20 Points) Does the poster contain sufficient information about the chosen day or week? Are the presented details accurate and comprehensive? 2.Visual Appropriateness (20 Points) Are the images used relevant to the topic and of standard visual quality? 3.Color Harmony (20 Points) Is there harmony among the colors used in the poster? Do the colors create a visual unity? 4.Font Type and Size (20 Points) Is the font type and size legible and aesthetically pleasing? Do the texts support the visual design? 5.Overall Aesthetic and Appearance (20 Points) Is the poster aesthetically pleasing and visually appealing overall? Are the design elements well integrated? **Evaluation Method:** Evaluate each poster based on the above criteria. Assign points ranging from 0 to 20 for each criterion. Sum the points to calculate the overall score for the poster.
Evaluation Results	**Special Day/Week Selection and Information (20 points):** The poster clearly marks "World Animal Protection Day" on October 4th, and includes activities, a precise location, and time for the event. This excellent use of space provides relevant information. Score: 20/20 **Relevance of Images (20 points):** The background features a variety of animal illustrations, which are highly pertinent for Animal Protection Day. The diversity of animals highlighted underlines the significance of all species. Score: 20/20 **Color Harmony (20 points):** The colors are vibrant and appealing, potentially attracting a broad audience, especially for an animal-related event. However, the busy background may slightly obscure the text. Score: 15/20 **Font Type and Size (20 points):** The main title uses a clear and large font, although the event details are slightly overshadowed by the vibrant background, impacting readability. Score: 14/20 **Overall Aesthetic and Look of the Poster (20 points):** The poster is visually striking and thematic. The circular text area effectively contains the information, although the design would benefit from more whitespace to improve readability and visual flow. Score: 16/20 **Total Score for the Poster: 85/100**. The poster is thematic and engaging, with excellent imagery for World Animal Protection Day. For improvement, balancing the busy background with clearer text areas would enhance readability. Well done! 🐾

Another evaluation was conducted through video submissions, utilizing the "Video Summarize" plugin. This tool is designed to efficiently analyse student video assignments and assign a score out of 100 based on predefined criteria. By providing concise summaries of the main topics, key points, and instructional content, the plugin enables educators and evaluators to objectively assess the video content. This ensures that each video assignment is graded according to the established criteria. Below are the details of the video evaluation process, including the prompt used and the assessment by Chat GPT 4.

Course	Game Development with Machine Learning
Software Used:	Video Summarize + Chat GPT 4
Prompt:	You are required to review the game videos designed by the students and score them based on specific criteria up to a total of 100 points. Each criterion corresponds to a set portion of the total score. **Evaluation Criteria and Scoring Method:**

	1.Alignment with Educational Objectives (25 Points): Assess how well the game meets the educational objectives. How integrated is the game content with the teaching goals and outcomes? 2.Code Usage and Information Provision (25 Points): Check whether the game development involved coding and programming techniques and if these were explained. Does the video provide sufficient explanation about the use of code? 3.Visual Design (25 Points): Evaluate the game's visual design considering the professionalism, quality of graphics, and visual coherence. Are the visuals consistent with the overall theme and purpose of the game? 4.Aesthetics and Playability (25 Points): Consider the aesthetic appeal and user experience of the game. Is the game easy to understand and does it have high playability? **Video Review and Scoring:** Watch the game video using the provided link: [Game Video Link] Evaluate each criterion within the specified scoring range. Assign points from 0 to 25 for each criterion based on performance. Calculate the total score for the video by summing the points.
Evaluation Results	**Alignment with Educational Objectives (25 Points):** The video showcases an interactive game that teaches English language structures concerning animals and likes/dislikes. The game clearly targets specific educational outcomes, justifying a full score: 25/25 **Code Usage and Information Provision (25 Points):** The narrator explains how student responses to in-game questions are evaluated and how feedback (sounds) is adjusted based on correct or incorrect answers. The coding aspect of the game is thoroughly detailed, meriting a full score: 25/25 **Visual Design (25 Points):** The visual design is structured to maintain student engagement and is appropriate for educational purposes. However, detailed information on graphical aspects and color harmony is lacking, making it difficult to award full points: 20/25 **Aesthetics and Playability (25 Points):** The video demonstrates that the game is aesthetically simple and understandable. It is designed to be highly playable and to encourage user interaction, offering a functional and user-friendly design: 25/25 **Total Score: 95/100** This game provides an interactive learning experience targeting specific English language outcomes. The coding aspects are comprehensively covered, and auditory feedback mechanisms reinforce learning with correct responses. The visual design and playability are appealing to students. The only minor drawback is the lack of detailed information on graphical elements and color harmony, which slightly lowers the overall score.

4.4 Data Analysis

During the data analysis phase, two distinct statistical methods were utilized to meticulously examine the students' exam results and to measure the consistency of the scores provided by both artificial intelligence and human evaluators. These methods are the Bland-Altman analysis (Bland & Altman, 1986) and the Interclass Correlation Coefficient (ICC) (Shrout & Fleiss, 1979). Selected for their relevance to the research questions and their inherent advantages, these analytical approaches provided a comprehensive assessment of the exam results. The Bland-Altman analysis is used to evaluate the agreement between two different measurement methods. It graphically represents the consistency between measurements by calculating the mean difference between methods and the limits of agreement (mean ± 1.96 SD) (Bland & Altman, 1986). In this study, the Bland-Altman analysis was employed to graphically assess the concordance between the scores from AI and human evaluators. This method was chosen specifically for its ability to visualize the magnitude of differences and their distribution relative to the measurement range.

The Interclass Correlation Coefficient (ICC) analysis provides a quantitative assessment of the agreement or consistency between two or more measurements. An ICC value closer to 1 indicates greater agreement (Shrout & Fleiss, 1979). In this research, ICC analysis was used to quantitatively measure the consistency of scoring between AI and human evaluators. The rationale for selecting ICC analysis lies in its capacity to provide a definitive measure of the reliability and consistency of different evaluation methods.

Both analytical methods were applied across overall scores and for individual types of exams, facilitating the evaluation of both general agreement and specific agreement relative to different exam types. The application of these analyses highlighted potential discrepancies and consistencies between AI and human scores, thereby providing valuable insights into the objectivity and reliability of the evaluation processes.

Validity and Reliability: Validity in research refers to whether the measurement instruments accurately measure the concepts they are intended to measure. In this study, to enhance validity, two critical strategies-content validity and construct validity-were implemented. Regarding content validity, the prompts used in the AI-supported evaluation process were meticulously reviewed by experienced educational experts. This review ensured that the prompts accurately measured students' knowledge and skills in alignment with curriculum objectives and learning outcomes. Construct validity was supported by the breadth of analyses across various exam types and the comprehensive scope of the evaluation process. Hence, instead of focusing on a single type of exam, different formats were analysed both individually and collectively to ensure a broad-based assessment.

Reliability refers to the consistency of a study's results and its ability to produce similar outcomes upon repetition. To ensure reliability in this study, three main approaches were followed: internal consistency, test-retest reliability, and inter-rater reliability. Internal consistency was assessed through multiple evaluations of the same set of exams using AI, comparing the scores to evaluate the consistency of the AI evaluations. Moreover, the reliability of the results was secured by employing two different analytical techniques. Inter-rater reliability was examined by comparing the scores assigned by human evaluators and AI, measuring the level of agreement and objectivity of the evaluation process.

The implementation of these methodological approaches was critical to ensuring that the research produced reliable and valid results, thereby confidently addressing the research questions with consistent and trustworthy findings.

5 FINDINGS

(1) What is the level of consistency between the scores obtained from artificial intelligence tools and instructor evaluations for university students' exam papers across different types of exams (traditional, test, project – video/poster)?

In the initial phase of the research, a detailed examination of the differences between scores given by artificial intelligence tools and human evaluators on all exam papers was conducted. The Bland-Altman analysis was employed to analyse these differences, and

the results are presented in Table 3, with the corresponding graph displayed in Figure 1.

	Estimate	Lower	Upper
Bias (n = 118)	5.415	3.072	7.758
Lower limit of agreement	-19.772	-23.787	-15.756
Upper limit of agreement	30.602	26.587	34.617

Tablo 3. Bland-Altman analysis for all exam papers

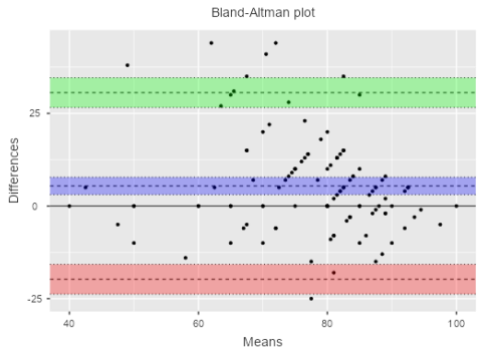


Figure 1: The Bland-Altman Plot

Examination of Table 3 reveals that the estimated average difference is 5.415, indicating that scores from AI tools are, on average, 5.415 units higher than those given by human evaluators. The calculated 95% confidence interval for this difference ranges from 3.072 to 7.758, suggesting a statistically significant bias between the two measurement methods as this interval does not include zero (Bland & Altman, 1986). The lower agreement limit is -19.772 with a 95% confidence interval between -23.787 and -15.756, and the upper agreement limit is 30.602, with its confidence interval ranging from 26.587 to 34.617. These limits indicate the potential range of differences in the worst and best-case scenarios, respectively. The breadth of these bias and agreement limits is an important indicator of the substantial individual variation that might exist between measurements. The graph analysis shows that most differences lie within the blue band, suggesting a general agreement. The presence of several data points just above and below the central band where differences cluster indicates that the limits of consistency are being tested in some measurements. Overall, the graph suggests that while the measurement methods generally produce consistent results, there are points of inconsistency that may require more careful analysis.

To further test the intensity and significance of the observed consistency, the Intraclass Correlation Coefficient (ICC) analysis was performed, and the results are documented in Table 4.

Intraclass Correlation		95% Confidence Interval		F Test with True Value 0				Cronbach's Alpha
		Lower Bound	Upper Bound	Value	df1	df2	Sig	
Single Measures	,527[a]	,383	,646	3,225	117	117x	,000	,690
Average Measures	,690[c]	,554	,785	3,225	117	117	,000	

Table 4: Intraclass Correlation Coefficient results for all exam papers

According to Table 4, the reported ICC value for single measures is 0.527, which indicates moderate consistency, typically considered within the range of 0.41 to 0.60. The ICC for average measures is 0.690, suggesting high consistency, usually denoted by values between 0.61 and 0.80. The 95% confidence intervals for these measurements range from 0.383 to 0.646 for Single Measures and from 0.554 to 0.785 for Average Measures, further validating the reliability of these ICC values. The F-test produced a value of 3.225 with a p-value <0.001, demonstrating that the consistency between the measurements is statistically significantly different from zero. The Cronbach's Alpha value found was 0.690, indicating a high level of internal consistency.

These findings suggest that the scores from AI tools and human evaluators are moderately consistent, highlighting the effectiveness of integrating artificial intelligence in evaluation processes while also pointing to areas where human oversight remains crucial.

(2) What is the level of consistency between the scores obtained from artificial intelligence tools and instructor evaluations for university students' project-based assessments?

In the research process, the consistency of scores within each type of exam was also compared. As part of these analyses, the distribution and consistency of scores from poster evaluations were examined. The Bland-Altman analysis was applied for these assessments, and the results are displayed in Table 5, with the corresponding graph shown in Figure 2.

	Estimate	Lower	Upper
Bias (n = 38)	3.684	1.048	6.321
Lower limit of agreement	-12.036	-16.582	-7.491
Upper limit of agreement	19.405	14.859	23.950

Table 5. Bland-Altman analysis for poster exams

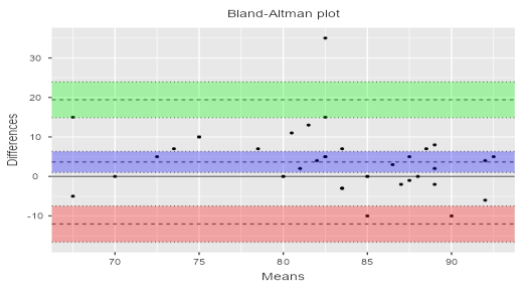


Figure 2. The Bland-Altman Plot

The analysis in Table 5 indicates that the average difference (bias) between the two measurement methods is 3.684, suggesting that one method typically measures an average of 3.684 units higher or lower than the other. The 95% confidence interval for this average difference is set between 1.048 and 6.321, which indicates a high level of confidence that the average difference lies within this range. The graph shows that most differences between the two measurement methods fall within the blue zone, indicating good agreement between them. Data points above the green line represent larger than expected positive differences, while those below the red line indicate larger than expected negative differences. This graphical representation demonstrates the overall

alignment and areas of significant deviation between the measurement methods.

To further test the significance and strength of the observed agreement, results from the Intraclass Correlation Coefficient (ICC) analysis are provided in Table 6.

	Intraclass Correlation[b]	95% Confidence Interval	
		Lower Bound	Upper Bound
Single Measures	,451[a]	,158	,671
Average Measures	,622[c]	,272	,803

	F Test with True Value 0				Cronbach's Alpha
	Value	df1	df2	Sig	
Single Measures	2,644	37	37	,002	,622
Average Measures	2,644	37	37	,002	

Table 6: Intraclass Correlation Coefficient results for poster exams

Reviewing Table 6, the single measurement ICC value calculated using ICC is 0.451, which points to moderate consistency between the measurements (Koo & Li, 2016). From a broader perspective, the ICC for average measurements is found to be 0.622, indicating a good level of consistency. The 95% confidence intervals for these measurements range from 0.158 to 0.671 for single measurements and from 0.272 to 0.803 for average measurements; these intervals support the reliability of the ICC values and their statistical significance from zero. The p-values for both single and average measurements are calculated as 0.002, further substantiating that the findings are statistically significant. The Cronbach's Alpha value is determined to be 0.622, reflecting a moderate level of internal consistency and general correlation among the items within the scale (Cronbach, 1951).

These results indicate that there is moderate consistency between AI-generated scores and human evaluator scores in poster assessments, although not perfect. The findings underscore the potential for AI tools in assessment processes while also highlighting the importance of careful statistical analysis to understand the extents and limits of this technology in educational settings.

The evaluation of video exams used the Bland-Altman analysis to study the distribution and consistency of scores obtained. The results of this analysis are presented in Table 7, and the corresponding graphical representation is shown in Figure 3.

	Estimate	Lower	Upper
Bias (n = 18)	-2.500	-9.027	4.027
Lower limit of agreement	-28.225	-39.593	-16.857
Upper limit of agreement	23.225	11.857	34.593

Table 7: Bland-Altman analysis for video exams

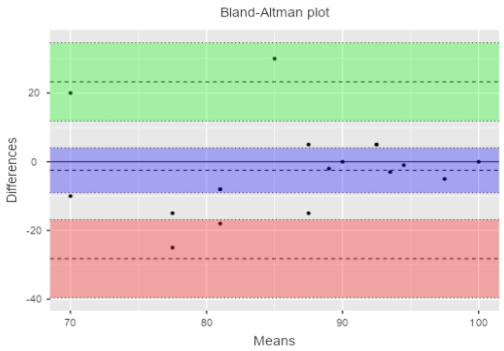


Figure 3: The Bland-Altman Plot

Table 7 reveals that the average difference (bias) between the two measurement methods is -2.500, indicating that, on average, instructor evaluations are 2.500 units lower than those made by artificial intelligence tools. The analysis of agreement limits shows that the lowest difference is -28.225, with a 95% confidence interval ranging from -39.593 to -16.857; the highest difference is 23.225, with a 95% confidence interval from 11.857 to 34.593. These values delineate the potential range of differences under the worst and best-case scenarios respectively, illustrating the extent to which the two measurement methods might diverge. The fact that the confidence interval of the average difference includes zero suggests there may not be significant bias between the two methods, although the broad range of agreement limits indicates significant individual variability in measurements. These findings from the Bland-Altman analysis (Bland & Altman, 1986) suggest that significant discrepancies can exist under certain conditions, necessitating careful consideration of the suitability of these methods. The Bland-Altman plot indicates that most differences cluster within the blue band, signifying generally good agreement between measurements. However, several data points in the upper green and lower red bands highlight the presence of notable outliers, suggesting some measurements exceed the limits of agreement and could be considered potential outliers, which might require re-evaluation of the measurement methods used.

To further test the significance and extent of consistency identified in the Bland-Altman analysis, results from the Intraclass Correlation Coefficient (ICC) analysis are presented in Table 8.

	Intraclass Correlation[b]	95% Confidence Interval		F Test with True Value 0				Cronbach's Alpha
		Lower Bound	Upper Bound	Value	df1	df2	Sig	
Single Measures	,281[a]	-,201	,653	1,780	17	17	,122	,438
Average Measures	,438[c]	-,502	,790	1,780	17	17	,122	

Table 8: Intraclass Correlation Coefficient results for video exams

Analysis of Table 8 shows that the ICC value for single measurements is 0.281, indicating low consistency as values generally between 0.00 and 0.40 are considered low (Koo & Li, 2016). The ICC for average measurements is 0.438, showing some

improvement in consistency but still not reaching the ideal level. The wide 95% confidence intervals for both measurements include zero, suggesting the ICC values do not achieve statistical significance. The F-test value is 1.780 with a p-value of 0.122, indicating that the consistency measurements are not statistically significant from zero. The Cronbach's Alpha value of 0.438 indicates low internal consistency within the scale, typically values above 0.7 are deemed acceptable (Cronbach, 1951). These results demonstrate very low consistency between AI tools and instructor scores in video evaluations, highlighting significant challenges in achieving reliable assessment outcomes in this context.

(3) What is the level of consistency between the scores obtained from artificial intelligence tools and instructor evaluations for university students' tests?

In the assessment of standardized tests, the Bland-Altman methodology was employed to analyse the distribution and consistency of the scores obtained. The results of this study are detailed in Table 9 and the corresponding graph is shown in Figure 4.

	Estimate	Lower	Upper
Bias (n = 19)	-0.789	-2.632	1.053
Lower limit of agreement	-8.284	-11.490	-5.077
Upper limit of agreement	6.705	3.498	9.911

Table 9. Bland-Altman analysis for test exams

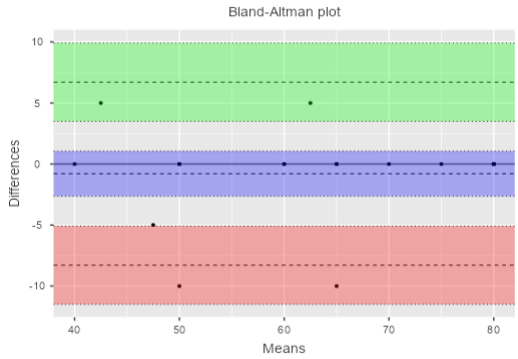


Figure 4. The Bland-Altman Plot

According to Table 9, the average difference (bias) between the two measurement methods is -0.789, indicating that instructor scoring is, on average, approximately 0.789 units lower than AI scoring. This negative value suggests that the first measurement method generally yields lower results compared to the second. The analysis of agreement limits reveals that the lowest difference is -8.284, with a 95% confidence interval between -11.490 and -5.077, and the highest difference is 6.705, with a 95% confidence interval between 3.498 and 9.911. These figures indicate that differences between the two measurement methods can range from as low as 8.284 units less to as high as 6.705 units more in the worst and best-case scenarios, respectively. The analysis of confidence intervals and agreement limits, while showing potential significant differences between methods, also suggests that these differences are not statistically significant as the confidence interval for the average difference includes zero. The Bland-Altman plot illustrates that the majority of data points fall within the blue band, indicating good overall agreement between the two measurement methods. However, the presence of a few data points in the upper green and lower red bands indicates that some individual measurements fall outside the acceptable limits of agreement, highlighting potential outliers. This suggests that the measurement methods might need to be reviewed in some cases. The overall trend of the graph indicates that although the measurement methods are largely consistent, caution is needed for potential extreme values.

	Estimate	Lower	Upper
Bias (n = 43)	13.000	8.420	17.580
Lower limit of agreement	-16.169	-24.057	-8.280
Upper limit of agreement	42.169	34.280	50.057

Furthermore, to test the degree of consistency and significance of the data, results from the Intraclass Correlation Coefficient (ICC) analysis are presented in Table 10.

Intraclass Correlation[b]		95% Confidence Interval		F Test with True Value 0				Cronbach's Alpha
		Lower Bound	Upper Bound	Value	df1	df2	Sig	
Single Measures	,962[a]	,903	,985	51,120	18	18	,000	,980
Average Measures	,980[c]	,949	,992	51,120	18	18	,000	

Table 10. Intraclass Correlation Coefficient results for test exams

The ICC value for single measurements, as shown in Table 10, is 0.962, indicating an extremely high level of consistency, which is considered excellent by standard criteria (Koo & Li, 2016). The ICC for average measurements is even higher at 0.980, suggesting nearly perfect agreement among multiple measurements. The F-test results, with an F value of 51.120 and a p-value of 0.000, confirm that the consistency between measurements is not by chance and is statistically highly significant (Shrout & Fleiss, 1979). The Cronbach's Alpha value calculated at 0.980 indicates almost perfect internal consistency within the scale, supporting that the items within the measurement tool are highly consistent and well-correlated; typically, values above 0.7 are accepted as an adequate level of internal consistency (Cronbach, 1951).

These findings demonstrate a high degree of consistency between the scores from AI tools and instructor evaluations, underscoring the reliability of integrating AI in educational assessment practices.

(4) 4. What is the level of consistency between the scores obtained from artificial intelligence tools and instructor evaluations for university students' traditional exams?

The Bland-Altman analysis was utilized to investigate the distribution and consistency of the scores derived from classical exams. The findings from this analysis are detailed in Table 11, and a visual summary of the evaluation results is provided in Figure 5

Table 11. Bland-Altman analysis for classic exams

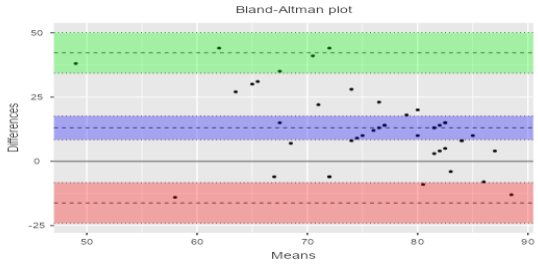


Figure 5. The Bland- Altman Plot

Upon reviewing Table 11, it is noted that the average difference between the two measurement methods is 13 units, indicating that AI tools typically award higher scores than human evaluators by an average of 13 units. This significant positive mean difference suggests a noticeable systemic bias in the measurements. he analysis of agreement limits shows that the range of differences between the two methods can be as low as -16.169 (with a 95% confidence interval between -24.057 and -8.280) and as high as 42.169 (with a 95% confidence interval between 34.280 and 50.057). These wide limits indicate substantial variance in the differences between measurements (Bland & Altman, 1986). Furthermore, most data points in the Bland-Altman plot fall within the blue band, suggesting that the differences between measurements generally remain within acceptable limits. However, some data points that lie within the upper green and lower red bands indicate that certain measurements exceed these limits, potentially signaling problematic evaluations.

Following this analysis, the Intraclass Correlation Coefficient (ICC) analysis was conducted to test the degree of consistency and significance of the data, with the results presented in Table 12.

Intraclass Correlation[b]		95% Confidence Interval		F Test with True Value 0				Cronbach's Alpha
		Lower Bound	Upper Bound	Value	df1	df2	Sig	
Single Measures	,128[a]	-,176	,410	1,294	42	42	,204	,227
Average Measures	,227[c]	-,427	,581	1,294	42	42	,204	

Table 12: Intraclass Correlation Coefficient results for classic exams

The ICC value for single measurements stands at 0.128, indicating low consistency; this suggests a statistically significant lack of harmony between the measurements. According to the classification by Koo and Li (2016), values between 0.00 and 0.40 are considered to show low consistency, and our findings fall within this range. The ICC for average measurements, though slightly higher, is still low at 0.227. The confidence intervals for both measurements-0.176 to 0.410 for Single Measures and -0.427 to 0.581 for Average Measures-include zero, meaning the ICC values do not achieve statistical significance. Additionally, the F value of 1.294 and a p-value of 0.204 for both measurements indicate that the results are not statistically significant. The Cronbach's Alpha value of 0.227 also points to low internal consistency within the measurement tool, considerably lower than the generally accepted threshold of 0.7 and above (Cronbach, 1951).

In summary, there is low consistency between the scores from AI tools and human evaluators in the evaluation of classical exams, highlighting significant discrepancies that warrant further investigation.

6 CONCLUSION AND DISCUSSION

This study aimed to test the consistency of scores derived from various exam formats (classical, test, and project - video, poster) assessed by both artificial intelligence tools and human evaluators among university students. Initially, all exam types were evaluated collectively, and both Bland-Altman and ICC tests indicated a moderate to high level of consistency between the two sets of scores. In an era where AI-based assessment systems are increasingly utilized in education, demonstrating the alignment of these systems with human evaluations is of critical importance. This study has shown a general concordance between AI and human evaluations, though some individual differences exist. According to the literature, such differences could stem from the inherent nature of the assessment methods (Giavarina, 2015). The consistency between AI and human evaluations suggests that AI-based assessment systems can be effectively used in education. AI technologies offer significant opportunities to enhance the assessment process for educators by providing automated grading, personalized feedback, and data-driven insights (González-Calatayud, Prendes-Espinosa, & Roig-Vila, 2021; Zhai & Nehm, 2023). These tools can simplify the evaluation process and allow teachers to focus more on teaching strategies and student support (Celik, Dindar, Muukkonen, & Järvelä, 2022). AI-based assessment systems could assist teachers in identifying individual student needs and customizing instruction accordingly (Celik et al., 2022). However, the individual differences highlighted by the research indicate that AI evaluations might not always replace teacher assessments. These systems could be further developed or used to support educators. By leveraging AI technologies, teachers can enhance the effectiveness of their assessment strategies and promote student learning outcomes (Mahligawati, Allanas, Butarbutar & Nordin, 2023). For AI to be effectively used in educational assessments, it is necessary to ensure consistency between AI and human evaluations, as well as to diversify assessment criteria and comprehensively examine individual student performance. AI systems have high potential to ensure objectivity and consistency; however, limitations exist, particularly in reflecting subjective assessment elements (Zhou & Shen, 2018). The benefits and potential limitations of using AI in education should be considered, along with the importance of human intervention when necessary.

During the research, separate evaluations for different exam types were conducted. It was observed that AI generally awarded higher scores than human evaluations in image-based exams, but both displayed moderate to high consistency. AI software has made significant advancements in image analysis across various fields. For instance, in detecting diabetic retinopathy, an AI-based diagnostic system achieved 87.2% sensitivity and 90.7% specificity, surpassing established superiority thresholds (Abràmoff et al., 2018). Similarly, deep learning models have shown impressive area under the curve (AUC) values ranging from 0.864 to 0.937 in diagnosing lung nodules or lung cancer through imaging techniques such as chest X-rays or CT scans (Aggarwal et al., 2021). In the field of phenomics, the integration of machine and deep learning has significantly enhanced data collection and analysis efficiency, improving image analysis processes (Nabwire et al., 2021). The moderate to high consistency between teacher scores and AI evaluation scores may be attributed to the success

in image processing. These successful outcomes suggest that AI could be an effective tool in exam evaluations, particularly in assessments involving visual data, where AI tools can ensure objective and consistent analysis of exam papers.

This study has observed low consistency between instructor scores and AI scores in video format exam evaluations. These findings suggest that Artificial Intelligence may have certain limitations in assessing student performance in exams conducted in video format. In particular, the low consistency and negative bias between AI and instructor evaluations reflect the challenges faced by AI systems in evaluating such complex and visual materials. Despite advancements in educational technologies that have improved AI's capability to analyse and understand visual content (Abràmoff, Lavin, Birch, Shah, & Folk, 2018; Aggarwal et al., 2021; Hung, Montalvao, Tanaka, Kawai, & Bornstein, 2020), this study indicates that AI may not fully replace the insight and assessment capabilities of human evaluators, especially in complex tasks like student performance evaluations in video formats. AI technologies could help educators embrace AI-supported assessment tools in e-learning, providing insights into the effectiveness and usability of these educational innovations (Sánchez-Prieto, Cruz-Benito, Therón, & García-PeñalvoPrieto, 2019). However, the results show that for AI to be more effective in these evaluations, there needs to be further development in the complexity of algorithms and the assessment of visual materials.

In contrast, an excellent level of agreement has been observed between instructor scores and AI scores in test exams. These findings demonstrate the effectiveness and reliability of using Artificial Intelligence in the educational sector, particularly in the format of test exams. AI evaluations are known to ensure objectivity and consistency, particularly in standardized tests. However, this technology also has limitations in assessing subjective elements. Research aligned with these findings suggests that AI-based systems can offer immediate feedback to students and reduce educators' workload, providing effective and accurate grading mechanisms (Qu, Zhao & Xie, 2022; Lübke, 2023). Additionally, AI can facilitate the automation of assessment processes, thus enhancing consistency and objectivity in evaluations (Lübke, 2023). The high level of consistency between AI and instructor evaluations suggests that AI can be a reliable tool in test exam formats.

Lastly, the findings reveal a significant mismatch in the classic exam format between AI and instructor scores. Despite AI's strength in evaluating objective responses, it may have limitations in more nuanced and subjective assessment scenarios typical of classical exams. This issue persists despite advancements in fields like Natural Language Processing (NLP), where AI has made significant strides. AI's success in NLP has been notable in areas such as computer vision, problem-solving, and language understanding (Blankenship, 2023; Hamal & Faddouli, 2022). High-performance AI question-answer systems have surpassed human abilities in automated language processing (Hamal & Faddouli, 2022). Yet, the complexity of natural languages, syntactic intricacies, and semantic ambiguities continue to pose challenges (Choi, 2014). The mismatch in classical exams might stem from AI systems' limited ability to fully comprehend the nuances and subjectivities across different languages.

6.1 Limitations and Recommendations

This study has examined the relationship between AI and instructor scores in exam paper evaluations, highlighting the potential and limitations of AI use in education. The findings provide valuable insights into how AI systems can be effectively used alongside instructors in educational assessment processes and may serve as a basis for future research to further develop these systems. The research was conducted using a single AI software; using various AI tools could diversify assessment outcomes and strengthen the validity of the findings. Therefore, future studies are recommended to use multiple AI tools for a more comprehensive and multifaceted evaluation. Keeping evaluation prompts constant could lead to AI performing better or worse on certain types of questions. Future research should increase the diversity and adaptation of prompts to more thoroughly test AI's evaluation capabilities.

REFERENCES

Amin, A. (2020). A face recognition system based on deep learning (frdls) to support the entry and supervision procedures on electronic exams. International Journal of Intelligent Computing and Information Sciences, 20(1), 40-50. https://doi.org/10.21608/ijicis.2020.23149.1015

Babbie, E. R. (2016). The Practice of Social research. Nelson Education.

Bai, X. and Stede, M. (2022). A survey of current machine learning approaches to student free-text evaluation for intelligent tutoring. International Journal of Artificial Intelligence in Education, 33(4), 992-1030. https://doi.org/10.1007/s40593-022-00323-0

Becher, T., & Trowler, P. R. (2001). *Academic tribes and territories: Intellectual enquiry and the culture of disciplines.* Open University Press.

Bland, J. M., & Altman, D. G. (1986). Statistical methods for assessing agreement between two methods of clinical measurement. *The Lancet, 327*(8476), 307-310.

Brookhart, S. M. (2008). *How to give effective feedback to your students*. Association for Supervision and Curriculum Development.

Cañada, J., Sanguino, T., Merelo, J., & Santos, V. (2014). Open classroom: enhancing student achievement on artificial intelligence through an international online competition. Journal of Computer Assisted Learning, 31(1), 14-31. https://doi.org/10.1111/jcal.12075

Celik, I., Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: A systematic review of research. *TechTrends, 66*(4), 616–630. https://doi.org/10.1007/s11528-022-00715-y

Chen, J., Lai, P. P. Y., Chan, A., Man, V., & Chan, C. (2022). Ai-assisted enhancement of student presentation skills: challenges and opportunities. *Sustainability, 15*(1), 196. https://doi.org/10.3390/su15010196

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

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

Choi, Y. and McClenen, C. (2020). Development of adaptive formative assessment system using computerized adaptive testing and dynamic bayesian networks. Applied Sciences, 10(22), 8196. https://doi.org/10.3390/app10228196

Collins, G., Dhiman, P., Navarro, C., Ma, J., Hooft, L., Reitsma, J., … & Moons, K. (2021). Protocol for development of a reporting guideline (tripod-ai) and risk of bias tool (probast-ai) for diagnostic and prognostic prediction model studies based on artificial intelligence. BMJ Open, 11(7), e048008. https://doi.org/10.1136/bmjopen-2020-048008

Creswell, J. W., & Creswell, J. D. (2018). *Research design: Qualitative, quantitative, and mixed methods approaches*. Sage Publications.

Cronbach, L. J. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika, 16*(3), 297-334.

Dermeval, D., Paiva, R., Bittencourt, I., Vassileva, J., & Borges, D. (2017). Authoring tools for designing intelligent tutoring systems: a systematic review of the literature. International Journal of Artificial Intelligence in Education, 28(3), 336-384. https://doi.org/10.1007/s40593-017-0157-9

Dindar, M., Muukkonen, H., & Järvelä, S. (2022). The promises and challenges of artificial intelligence for teachers: a systematic review of research. Techtrends, 66(4), 616-630. https://doi.org/10.1007/s11528-022-00715-y

Dwivedi, Y., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., … & Williams, M. (2021). Artificial intelligence (ai): multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management, 57, 101994. https://doi.org/10.1016/j.ijinfomgt.2019.08.002

Dyment, J. E., & O'Connell, T. S. (2011). Assessing the quality of reflection in student journals: A review of the research. *Teaching in Higher Education, 16*(1), 81-97.

Ed-Driouch, C., Gourraud, P., Dumas, C., & Mars, F. (2022). The integration of human intelligence into artificial intelligence to provide medical practice-based predictions. *HHAI2022: Augmenting Human Intellect.* https://doi.org/10.3233/faia220221

Eisner, E. W. (2002). *The arts and the creation of mind*. Yale University Press.

Elder, H., Rieger, T., Canfield, C., Shank, D. B., & Hines, C. (2022). Knowing when to pass: the effect of ai reliability in risky decision contexts. *Human Factors: The Journal of the Human Factors and Ergonomics Society, 66*(2), 348-362. https://doi.org/10.1177/00187208221100691

Fiske, A., Henningsen, P., & Buyx, A. (2019). Your robot therapist will see you now: ethical implications of embodied artificial intelligence in psychiatry, psychology, and psychotherapy. *Journal of Medical Internet Research, 21*(5), e13216. https://doi.org/10.2196/13216

Freeman, S., Eddy, S. L., McDonough, M., Smith, M. K., Okoroafor, N., Jordt, H., & Wenderoth, M. P. (2014). Active learning increases student performance in science, engineering, and mathematics. *Proceedings of the National Academy of Sciences, 111*(23), 8410-8415.

Giavarina, D. (2015). Understanding Bland Altman analysis. *Biochemia Medica, 25*(2), 141-151.

González-Calatayud, V., Prendes-Espinosa, P., & Roig-Vila, R. (2021). Artificial intelligence for student assessment: A systematic review. *Applied Sciences, 11*(12), 5467. https://doi.org/10.3390/app11125467

Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G. (2019). Xai—explainable artificial intelligence. *Science Robotics, 4*(37). https://doi.org/10.1126/scirobotics.aay712,

Güven, H. and Güven, E. (2023). Use of artificial intelligence applications in e-commerce. International Journal of Management and Administration, 7(13), 69-94. https://doi.org/10.29064/ijma.1194949

Haladyna, T. M., Downing, S. M., & Rodriguez, M. C. (2002). A review of multiple-choice item-writing guidelines for classroom assessment. *Applied Measurement in Education, 15*(3), 309-334.

Harry, A. (2023). Role of ai in education. Interdiciplinary Journal and Hummanity (Injurity), 2(3), 260-268. https://doi.org/10.58631/injurity.v2i3.52

Haseski, H. (2019). What do turkish pre-service teachers think about artificial intelligence?. International Journal of Computer Science Education in Schools, 3(2), 3-23. https://doi.org/10.21585/ijcses.v3i2.55

Hinojo-Lucena, F., Díaz, I., Reche, M., & Rodríguez, J. (2019). Artificial intelligence in higher education: a bibliometric study on its impact in the scientific literature. Education Sciences, 9(1), 51. https://doi.org/10.3390/educsci9010051

Hua, Y. (2022). Design of online music education system based on artificial intelligence and multiuser detection algorithm. Computational Intelligence and Neuroscience, 2022, 1-11. https://doi.org/10.1155/2022/9083436

Ishaaq, N. & Sohail, S. S. (2023). Re: investigating the impact of innovative ai chatbot on post-pandemic medical education and clinical assistance: a comprehensive analysis. *ANZ Journal of Surgery, 94*(3), 494-494. https://doi.org/10.1111/ans.18721

Jingshan, H. (2023). Analysis of the Application of Artificial Intelligence in Education and Teaching. Advances in Educational Technology and Psychology, doi: 10.23977/aetp.2023.070210

Jobin, A. & Ienca, M. (2019). The global landscape of ai ethics guidelines. Nature Machine Intelligence, 1(9), 389-399. https://doi.org/10.1038/s42256-019-0088-

Kawaji, T., Hojo, S., Kushiyama, A., Nakatsuma, K., Kaneda, K., Kato, M., … & Sato, M. (2019). Limitations of lesion quality estimated by ablation index: an in vitro study. *Journal of Cardiovascular Electrophysiology, 30*(6), 926-933. https://doi.org/10.1111/jce.13928

Kazimov, T., Bayramova, T., & Malikova, N. (2021). Research of intelligent methods of software testing. System Research and Information Technologies, (4), 42-52. https://doi.org/10.20535/srit.2308-8893.2021.4.03

Keskinbora, K. & Güven, F. (2020). Artificial intelligence and ophthalmology. *Turkish Journal of Ophthalmology, 50*(1), 37-43. https://doi.org/10.4274/tjo.galenos.2020.78989

Koo, T. K., & Li, M. Y. (2016). A guideline of selecting and reporting intraclass correlation coefficients for reliability research. *Journal of Chiropractic Medicine, 15*(2), 155-163.

Köse, U. and Arslan, A. (2017). Optimization of self-learning in computer engineering courses: an intelligent software system supported by artificial neural network and vortex optimization algorithm. Computer Applications in Engineering Education, 25(1), 142-156. https://doi.org/10.1002/cae.21787

Laupichler, M. C., Aster, A., Meyerheim, M., Raupach, T., & Mergen, M. (2024). Medical students' ai literacy and attitudes towards ai: a cross-sectional two-center study using pre-validated assessment instruments. *BMC Medical Education, 24*(1). https://doi.org/10.1186/s12909-024-05400-

Liu, H., Liu, Z., Wu, Z., & Tang, J. (2020). Personalized multimodal feedback generation in education.. https://doi.org/10.18653/v1/2020.coling-main.166

Liu, S., Wright, A., Patterson, B., Wanderer, J., Turer, R., Nelson, S., … & Wright, A. (2023). Using ai-generated suggestions from chatgpt to optimize clinical decision support. *Journal of the American Medical Informatics Association, 30*(7), 1237-1245. https://doi.org/10.1093/jamia/ocad072

Luan, H., Géczy, P., Lai, H., Gobert, J., Yang, S., Ogata, H., … & Tsai, C. (2020). Challenges and future directions of big data and artificial intelligence in education. Frontiers in Psychology, 11. https://doi.org/10.3389/fpsyg.2020.580820

Lund, B. D., & Wang, T. (2023). Chatting about ChatGPT: How may AI and GPT impact academia and libraries? *Library Hi Tech News, 40*(3), 26–29. https://doi.org/10.1108/lhtn-01-2023-0009

Mahligawati, F. (2023). Artificial intelligence in physics education: a comprehensive literature review. Journal of Physics Conference Series, 2596(1), 012080. https://doi.org/10.1088/1742-6596/2596/1/012080

Mahligawati, F., Allanas, E., Butarbutar, M. H., & Nordin, N. A. N. (2023). Artificial intelligence in Physics Education: A comprehensive literature review. *Journal of Physics: Conference Series, 2596*(1), 012080. https://doi.org/10.1088/1742-6596/2596/1/012080

Myszczynska, M. A., Ojamies, P. N., Lacoste, A. M. B., Neil, D., Saffari, A., Mead, R., … & Ferraiuolo, L. (2020). Applications of machine learning to diagnosis and treatment of neurodegenerative diseases. *Nature Reviews Neurology, 16*(8), 440-456. https://doi.org/10.1038/s41582-020-0377-8

Nguyen, T. (2023). Exploring the efficacy of chatgpt in language teaching. Asiacall Online Journal, 14(2), 156-167. https://doi.org/10.54855/acoj.2314210

Parapadakis, D. (2020). Can artificial intelligence help predict a learner's needs? lessons from predicting student satisfaction. London Review of Education, 18(2). https://doi.org/10.14324/lre.18.2.03

Pellegrino, J. W., & Quellmalz, E. S. (2010). Perspectives on the integration of technology and assessment. *Journal of Research on Technology in Education, 43*(2), 119-134.

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

Reiß, M. (2021). The use of ai in education: practicalities and ethical considerations. London Review of Education, 19(1). https://doi.org/10.14324/lre.19.1.05

Richardson, M. & Clesham, R. (2021). Rise of the machines? the evolving role of ai technologies in high-stakes assessment. *London Review of Education, 19*(1). https://doi.org/10.14324/lre.19.1.09

Sandhu, S., Lin, A., Brajer, N., Sperling, J., Ratliff, W., Bedoya, A., … & Sendak, M. (2020). Integrating a machine learning system into clinical workflows: qualitative study. *Journal of Medical Internet Research, 22*(11), e22421. https://doi.org/10.2196/2242

Sapci, A. and Sapci, H. (2020). Artificial intelligence education and tools for medical and health informatics students: systematic review. Jmir Medical Education, 6(1), e19285. https://doi.org/10.2196/19285

Shrout, P. E., & Fleiss, J. L. (1979). Intraclass correlations: Uses in assessing rater reliability. *Psychological Bulletin, 86*(2), 420-428.

Tapalova, O. and Zhiyenbayeva, N. (2022). Artificial intelligence in education: aied for personalised learning pathways. The Electronic Journal of E-Learning, 20(5), 639-653. https://doi.org/10.34190/ejel.20.5.2597

Teebagy, S., Colwell, L., Wood, E., Yaghy, A., & Faustina, M. (2023). Improved performance of ChatGPT-4 on the OKAP examination: A comparative study with ChatGPT-3.5. *Journal of Academic Ophthalmology, 15*(02), e184–e187. https://doi.org/10.1055/s-0043-1774399

Thomas, J. W. (2000). *A review of research on project-based learning*. San Rafael, CA: Autodesk Foundation

Tubino, L. and Adachi, C. (2022). Developing feedback literacy capabilities through an ai automated feedback tool. Ascilite Publications, e22039. https://doi.org/10.14742/apubs.2022.39

Wang, B., Zhang, Y., Wu, C., & Wang, F. (2021). Multimodal mri analysis of cervical cancer on the basis of artificial intelligence algorithm. *Contrast Media & Amp; Molecular Imaging, 2021*, 1-11. https://doi.org/10.1155/2021/1673490

Wiljer, D., Salhia, M., Dolatabadi, E., Dhalla, A., Gillan, C., Al-Mouaswas, D., … & Tavares, W. (2021). Accelerating the appropriate adoption of artificial intelligence in health care: protocol for a multistepped approach. *JMIR Research Protocols, 10*(10), e30940. https://doi.org/10.2196/30940

Yang, D. and Wang, Y. (2020). Hybrid physical education teaching and curriculum design based on a voice interactive artificial intelligence educational robot. Sustainability, 12(19), 8000. https://doi.org/10.3390/su12198000

Yin, W. (2021). Modeling method and application of college comprehensive teaching mode based on artificial intelligence. Converter, 566-573. https://doi.org/10.17762/converter.231

Yu, L. and Yu, Z. (2023). Qualitative and quantitative analyses of artificial intelligence ethics in education using vosviewer and citnetexplorer. Frontiers in Psychology, 14. https://doi.org/10.3389/fpsyg.2023.1061778

Zawacki-Richter, O., Marín, V. I., Bond, M., & Gouverneur, F. (2019). Systematic review of research on artificial intelligence applications in higher education – where are the educators*?. International Journal of Educational Technology in Higher Education, 16*(1). https://doi.org/10.1186/s41239-019-0171-0

Zhai, X., & Nehm, R. H. (2023). AI and formative assessment: The train has left the station. *Journal of Research in Science Teaching, 60*(6), 1390–1398. https://doi.org/10.1002/tea.21885

Zhao, T. and Song, T. (2022). Establishing a fusion model of attention mechanism and generative adversarial network to estimate students' attitudes in english classes. Tehnicki Vjesnik - Technical Gazette, 29(5). https://doi.org/10.17559/tv-20210922053009

Zhou, J., & Shen, M. (2018). When human intelligence meets artificial intelligence. *PsyCh Journal, 7*(3), 156–157. https://doi.org/10.1002/pchj.216

AVALUANT ELS AVALUADORS: UN ESTUDI COMPARATIU ENTRE LA IA I LES AVALUACIONS DOCENTS A L'EDUCACIÓ SUPERIOR

Aquest estudi pretén examinar les diferències potencials entre les avaluacions dels professors i els sistemes d'avaluació basats en eines d'intel·ligència artificial (IA) en els exàmens universitaris. La investigació ha avaluat un ampli espectre d'exàmens que inclouen exàmens de cursos numèrics i verbals, exàmens amb diferents estils d'avaluació (projecte, examen de prova, examen tradicional) i exàmens de cursos tant teòrics com pràctics. Aquests exàmens es van seleccionar mitjançant un mètode de mostreig de criteris i es van analitzar mitjançant l'anàlisi de Bland-Altman i les anàlisis del coeficient de correlació intraclasse (ICC) per avaluar el rendiment de l'IA i les avaluacions dels professors en una àmplia gamma. Els resultats de la investigació indiquen que si bé hi ha un alt nivell de competència entre les puntuacions totals dels exàmens avaluades per intel·ligència artificial i les avaluacions dels professors; Es va trobar una consistència mitjana en l'avaluació dels exàmens basats en la visualitat, una consistència baixa en els exàmens de vídeo, una consistència alta en els exàmens de prova i una consistència baixa en els exàmens tradicionals. Aquesta investigació és crucial ja que ajuda a identificar àrees específiques on la intel·ligència artificial pot complementar o necessitar millora en l'avaluació educativa, orientant el desenvolupament d'eines d'avaluació més precises i justes. (Word Style "Article Text").

PARAULES CLAU: Sistemes d'avaluació basats en eines d'intel·ligència artificial, avaluació docent, avaluacions en educació superior

EVALUANDO A LOS EVALUADORES: UN ESTUDIO COMPARATIVO ENTRE LA IA Y LAS EVALUACIONES DOCENTES EN LA EDUCACIÓN SUPERIOR

Este estudio pretende examinar las potenciales diferencias entre las evaluaciones de los profesores y los sistemas de evaluación basados en herramientas de inteligencia artificial (IA) en los exámenes universitarios. La investigación ha evaluado un amplio espectro de exámenes que incluyen exámenes de cursos numéricos y verbales, exámenes con distintos estilos de evaluación (proyecto, examen de prueba, examen tradicional) y exámenes de cursos tanto teóricos como prácticos. Estos exámenes se seleccionaron mediante un método de muestreo de criterios y se analizaron mediante el análisis de Bland-Altman y los análisis del coeficiente de correlación intraclase (ICC) para evaluar el rendimiento del IA y las evaluaciones de los profesores en una amplia gama. Los resultados de la investigación indican que si bien existe un alto nivel de competencia entre las puntuaciones totales de los exámenes evaluadas por inteligencia artificial y las evaluaciones de los profesores; Se halló una consistencia media en la evaluación de los exámenes basados en la visualidad, una consistencia baja en los exámenes de vídeo, una consistencia alta en los exámenes de prueba y una consistencia baja en los exámenes tradicionales. Esta investigación es crucial puesto que ayuda a identificar áreas específicas donde la inteligencia artificial puede complementar o necesitar mejora en la evaluación educativa, orientando el desarrollo de herramientas de evaluación más precisas y justas.

PALABRAS CLAVE: Sistemas de evaluación basados en herramientas de inteligencia artificial