Ed Process Int J

www.edupij.com

ISSN 2147-0901 (Print), 2564-8020 (Online)

EDUCATIONAL PROCESS
INTERNATIONAL JOURNAL

Research Article

Cite this article: Zacharis, G., & Papadakis, S. (2025). Can AI Grade Like a Human? Validity, Reliability, and Fairness in University Coursework Assessment. *Educational Process: International Journal, 19*, e2025591. https://doi.org/10.22521/edupij.2025.19.591

Can AI Grade Like a Human? Validity, Reliability, and Fairness in University Coursework Assessment

Georgios Zacharis, Stamatios Papadakis

Abstract

Background/purpose. Generative artificial intelligence (GenAI) is often promoted as a transformative tool for assessment, yet evidence of its validity compared to human raters remains limited. This study examined whether an AI-based rater could be used interchangeably with trained faculty in scoring complex coursework.

Materials/methods. Ninety-one essays from teacher education courses at two Greek universities were independently evaluated by two human raters and an AI system, using a common rubric.

Results. Human inter-rater reliability was excellent (ICC(2,1) = .884; ICC(2,k) k=2 = .938). In contrast, AI–human agreement was substantially weaker (AI vs Human-Z: ICC(2,1) = .406; ICC(2,k) = .578; AI vs Human-S: ICC(2,1) = .279; ICC(2,k) = .436). The AI consistently inflated scores by 2.71–3.32 points and compressed distributions, limiting its ability to discriminate across performance levels. Bland–Altman analyses confirmed systematic proportional bias, with over-scoring of weaker work and under-scoring of stronger work. Results revealed significant inconsistency in AI performance: while the model failed to align with Human-S (κ = .017), it demonstrated statistically significant, moderate agreement with Human-Z (κ = .367). This discrepancy highlights the lack of standardization in human grading and the sensitivity of algorithms to divergent interpretive frameworks. A principal component analysis suggested that AI captured a narrower construct of quality than human raters.

Conclusion. These findings indicate that current GenAI tools are not suitable for high-stakes assessment in higher education, where fairness and construct validity are essential. They may, however, offer value in formative feedback or administrative support if used transparently and under human oversight.

Received August 10, 2025
Accepted November 19, 2025

Keywords: Higher education assessment, human–AI agreement, AI grade, generative AI in education

Author for correspondence:
Georgios Zacharis
gzacharis@nured.auth.gr
Department of Early Childhood Education, Aristotle University of Thessaloniki, Greece

1. Introduction

Generative AI (GenAI) is rapidly moving from curiosity to infrastructure across education systems. In primary and secondary schooling, national actors are already piloting or scaling policy, tooling, and professional learning: New Zealand's NZCER documents widespread teacher uptake alongside uneven policy support and persistent concerns about accuracy, equity, and cultural responsiveness (e.g., limited te reo/mātauranga Māori coverage) (Coblentz, Dong, & Gibbs, 2025). Australia is transitioning from blanket bans toward risk-managed roll-outs, guided by a national framework on the responsible use of generative AI in schools and pilot tools such as South Australia's EdChat (The Guardian, 2024; The Australian, 2024). In Canada, federal guidance has been issued under the "FASTER" principles, while provincial activity remains uneven: some jurisdictions, such as Ontario's Peel District School Board, have begun publishing classroom-focused guidance, though many provinces still lack formal frameworks (Government of Canada, 2024; Canadian Teachers' Federation, 2024). This is not theoretical anymore; it is governance, procurement, and classroom planning in motion.

Universities feel the same gravitational pull. A fast-growing body of higher-education research chronicles two tensions: efficiency and feedback gains versus integrity, bias, and over-reliance. Systematic reviews and sector syntheses describe opportunities in formative feedback and learner support, but warn about assessment validity, equity, and shifting academic literacies (Crompton & Burke, 2023). At the same time, empirical work shows that teacher adoption is not simply a function of tool availability: attitudes, trust, perceived risk, and local policy signals shape whether GenAI enters authentic pedagogy or remains a background convenience (Gamlem et al., 2025).

Technically, GenAI systems can produce fluent teaching artefacts; pedagogically, fluency is not enough. Large-scale language models hallucinate, confabulate, and "lie" in ways that are predictable from the models' training and inference mechanics; the point is not moral condemnation but understanding what kinds of statements you should never grade at face value (Mitchell, 2025). These reliability limits ripple straight into assessment: if we use AI to generate, critique, or score student work, the measurement argument must survive psychometric scrutiny (construct coverage, reliability, invariance, fairness).

Meanwhile, the surrounding ecosystem is changing. Editorial processes are testing AI's role both as a tool and as an actor: a medical journal study proposes practical methods to detect AI co-authorship in manuscripts, reflecting a field trying to keep provenance visible without defaulting to prohibition (Öztürk et al., 2025). Peer-review experiments comparing AI-generated to human comments suggest AI can be serviceable for structure and tone but inconsistent on nuanced judgment, a pattern with obvious analogues to grading (Flodén, 2025; Wetzler et al., 2024). School systems are simultaneously exploring student-facing creation: "babyGPTs" projects show secondary students can curate datasets and prompts to build task-specific agents—energising, but also a reminder that data control and scope conditions matter for any claim about learning gains or fairness (Morales-Navarro et al., 2025).

Closer to practice, public reporting keeps surfacing inequities that automated systems can amplify. Analyses of teacher-facing AI recommenders show racially biased suggestions for "struggling students," a stark warning that convenience features can encode historical patterns unless audited (Chai et al., 2024; Pham et al., 2025). In parallel, nation-scale experiments like Estonia's AI Leap 2025 are moving to embed AI throughout schooling, positioning access, privacy, and professional development as policy levers, a macro-signal that evaluation and assurance will need to keep pace with deployment (Chai et al., 2024; Pham et al., 2025; Wetzler et al., 2024).

Recent research in artificial intelligence in education points to both opportunities and limitations of using large language models (LLMs) in real learning contexts. Editorial perspectives stress their

https://doi.org/10.22521/edupij.2025.19.591 Published online by Universitepark Press

potential for content generation, feedback, and inclusion, while highlighting challenges such as domain adaptation, accuracy, and reliability (Xing et al., 2025). In programming and data science education, automated systems have been deployed to provide formative assessment and feedback, aligning with instructor judgments and improving efficiency (Vittorini et al., 2023; Vittorini et al., 2021). When extended with LLM-generated feedback, these systems fostered high student engagement and improved outcomes, although verbosity sometimes hindered focus (Letteri & Vittorini, 2025). These insights emphasize that while AI systems can influence learning experiences, the structure of programming activities themselves may be even more decisive in shaping students' emotional engagement.

Within this shifting landscape, higher education institutions are asking a targeted question: Can AI-assisted scoring be used interchangeably with human ratings for complex coursework? The answer cannot rely on correlation snapshots or single-criterion pilots. It calls for (i) rigorous inter-rater reliability among humans, (ii) agreement analyses that go beyond simple correlation, and (iii) exploration of construct coverage. That is the measurement argument this paper takes up, using a uniquely clean setting: identical essays, a detailed rubric, two human raters, and an AI scorer, drawn from two Greek universities.

We position our study at the intersection of assessment validity and responsible GenAI integration. Specifically, we (a) quantify human–human and human–AI agreement using intraclass correlation coefficients (ICC) and Bland–Altman analyses, and (b) explore the internal structure of criterion scores with principal component analysis (PCA). We do not test equivalence margins or cross-institution fairness. Our claim is therefore not that AI grading is proven interchangeable, but that we provide evidence about the scope and limits of its agreement with human raters in a tertiary education setting.

2. Literature Review

2.1. Global policy and governance context

The rapid rise of generative artificial intelligence (GenAI) in education has spurred international organizations and national governments to produce guidance and regulation at an unprecedented pace. UNESCO's Guidance for Generative AI in Education and Research (2023) frames AI adoption within a "human-centered" vision, emphasizing equity, agency, and cultural responsiveness. However, Karakose et al. (2025) emphasize the fact that widening digital divides may lead to significant threats in the future and recommend that professional learning of teachers should be seen as a prerequisite for meaningful integration. Similarly, the OECD's Digital Education Outlook 2023 stresses the need for "responsible adoption" and recommends that governments co-design governance frameworks with educators rather than imposing top-down mandates (OECD, 2023).

National responses are uneven (Department for Education, 2025; OECD, 2023; UNESCO, 2023). The UK Department for Education (2025) released an updated policy document encouraging schools to integrate GenAI cautiously but actively, positioning it as a support tool rather than a replacement for teaching. In the United States, at least 20 state education departments have issued K–12 AI frameworks, moving away from blanket bans toward more nuanced approaches (Gu & Ericson, 2025). Estonia represents perhaps the most aggressive national deployment: its AI Leap 2025 initiative seeks to position the country as a global model for AI in education by embedding AI systems across primary, secondary, and tertiary sectors (Petrone, 2025).

What unites these efforts is the recognition that AI in education is no longer a speculative issue. Instead, it is a matter of system design, procurement, accountability, and teacher capacity. Yet gaps persist. The New Zealand Council for Educational Research's survey of teachers and students (Coblentz, Dong, & Gibbs, 2025) highlights significant disparities in access: while teachers experiment widely with free tools like ChatGPT and Gemini, three-quarters report lacking institutional support for premium AI systems, leaving them with lower-performing models prone to errors and bias. Policy frameworks exist, but capacity building and equitable resourcing remain incomplete.

2.2. Teacher adoption, attitudes, and professional practice

Teachers are simultaneously intrigued and cautious. Empirical work shows adoption clustered around low-stakes, efficiency-oriented tasks: lesson planning, rubric drafting, and generation of teaching materials. Gamlem et al. (2025) conducted a mixed-methods study of pre-service teachers in Norway, finding that while most had tried GenAI for planning and idea generation, many expressed skepticisms about its pedagogical appropriateness. Importantly, their reservations were not rooted in fear of technology but in professional judgment: concerns about accuracy, bias, and a loss of human connection in teaching.

Similar results emerge from primary school teachers in Aotearoa, New Zealand. Coblentz et al. (2025) report that nearly 70% of surveyed teachers used AI tools weekly, mostly for administrative or planning support, yet fewer than half felt confident teaching students how to use AI responsibly. Most desired additional training, suggesting adoption is pragmatic but fragile.

Teachers' experiences also reveal what Neil Selwyn and colleagues call the "frailties" of GenAI. Their qualitative study of Swedish and Australian teachers (Selwyn, Ljungqvist, & Sonesson, 2025) shows that far from being automated, AI-enabled teaching often generates more work: teachers must review, repair, and contextualize AI outputs to ensure pedagogical fit. This "hidden labor" challenges industry claims of workload reduction and underscores that teachers remain the central arbiters of educational quality.

At the same time, resistance is not trivial. An open letter published in mid-2025, signed by educators worldwide, explicitly rejected the normalization of GenAI in classrooms, particularly for grading and assessment (Dusseau et al., 2025). The letter frames AI adoption as a political choice rather than an inevitability, warning against a narrative where commercial edtech firms set the agenda for schools.

2.3. Student perspectives, trust, and reliance

Student adoption is equally heterogeneous. Wang, Li, Cheung, and Wong (2025) examined university students in Hong Kong, proposing a trust–reliance–resistance model of GenAI use in language learning. Their findings reveal that trust strongly predicts intention to use AI, while perceived risk increases resistance. Most crucially, "appropriate reliance" mediates these relationships: students are not simply for or against AI; they calibrate their use depending on perceived accuracy, task stakes, and institutional signals.

At the primary school level, Coblentz et al. (2025) show that while 90% of surveyed students in New Zealand had heard of GenAI, fewer than half used it regularly, and most reported confusion about school policies. Students described AI as both "helpful" and "a bit like cheating," echoing similar ambivalence found in surveys of U.S. high school students (Morales-Navarro et al., 2025). This uncertainty reflects the absence of explicit guidance in many schools, leaving young learners to improvise their own norms.

Critically, not all students experience GenAI equally. Empirical studies show fairness concerns are salient to students and vary by context (Chai et al., 2024; Pham et al., 2025). Cultural responsiveness is another gap: mainstream models perform poorly with Indigenous languages and contexts, leading New Zealand teachers to worry that AI tools might undermine Māori data sovereignty (Coblentz et al., 2025). These findings underscore that trust is not simply about accuracy; it is also about fairness and cultural legitimacy.

2.4. Generative AI in assessment and feedback

Assessment is arguably the most contested terrain. Proponents emphasize efficiency: AI can grade papers in seconds, offer instant formative feedback, and help identify struggling learners early. Critics counter that such efficiency risks hollowing out the very qualities that make assessment meaningful: empathy, nuance, and contextual judgment. Researchers argue that grading student writing with AI reduces writing to mechanical compliance with rubrics, stripping away the relational aspects of feedback. Recent university-level comparisons show that GenAI grading disagrees with instructors on both mean scores and categorical classifications, especially for borderline work (Flodén, 2025; Karakose & Tulubas, 2025; Papadakis & Karakose, 2025; Wetzler et al., 2025).

Technical experiments lend mixed evidence. Ghazawi and Simpson (2025) tested AI grading of Arabic essays, demonstrating that while LLMs can approximate human scoring under constrained conditions, issues of validity and construct coverage remain unresolved. Öztürk et al. (2025) proposed a methodology for identifying AI co-authorship in academic papers, indirectly raising questions about provenance and authorship in educational assessment contexts. Similarly, Marrella et al. (2025) compared AI-generated peer review to human reviewers, finding that while AI could mimic structural feedback, it often failed to capture nuanced evaluation of methods or originality, a pattern with clear implications for student grading.

2.5. Ethics, bias, and reliability

Reliability remains a central obstacle. Large language models are prone to "hallucinations," fabricating sources or producing confidently incorrect statements. Mitchell (2025) in Science characterizes this behavior not as a rare malfunction but as systemic, cautioning that AI can "lie" in ways indistinguishable from truth to lay users. This unreliability carries severe consequences in educational contexts where accuracy and trust are paramount.

Bias compounds these risks. Chai et al. (2024) and Pham et al. (2025) demonstrate racially skewed recommendations in AI-driven teacher assistants. In New Zealand, teachers reported that AI outputs often overlooked culturally responsive teaching practices (Coblentz et al., 2025). Scholars warn that without intentional auditing and adaptation, GenAI may reproduce historical inequities rather than correct them (Selwyn et al., 2025).

Beyond accuracy and bias lies the question of labor and agency. Selwyn et al. (2025) highlight the "hidden work" teachers must perform to adapt AI outputs, undermining claims of efficiency. The open letter by Dusseau et al. (2025) frames AI adoption as a political economy issue: who benefits financially, who carries the risk, and who retains control over educational decision-making.

2.6. Future-oriented models and research gaps

Not all developments are cautionary. Jurenka et al. (2024), in Google DeepMind's LearnLM report, describe a participatory, evaluation-driven approach to building an AI tutor optimized for pedagogy rather than generic helpfulness. Their framework emphasizes co-design with educators, iterative evaluation, and transparency—principles that could guide responsible AI adoption in assessment. Similarly, research on AI in creative education suggests that, with scaffolding, AI tools can enhance exploration and ideation without replacing originality. Yet even in promising contexts, peer-reviewed evaluations show mixed reliability when LLMs are applied to grading student writing (Pack et al., 2024; Tate et al., 2024).

Empirical studies across disciplines reinforce these concerns. In computing, LLMs correctly flagged only about half of concurrency errors compared to expert raters (Estévez-Ayres et al., 2024). In mathematics, GPT-4 achieved moderate agreement with human grading but produced systematic discrepancies across problem types (Lee et al., 2024). In medical education, ChatGPT-generated multiple-choice items displayed weaker psychometric quality than faculty-authored ones (Chauhan et al., 2025). Collectively, these studies suggest that while LLMs can scale feedback and assessment, their judgments often diverge from expert standards in ways that matter for fairness and validity.

Other research underscores how the quality of feedback shapes learner responses. NLP pipelines for essay assessment increased perceived helpfulness by generating more detailed formative comments (Gombert et al., 2024), and intelligent textbooks using encoder-based models explained substantial variance in human ratings of student summaries (Morris et al., 2024). These systems show potential for supporting competence and autonomy, but also reveal risks when feedback becomes overwhelming or poorly targeted (Letteri & Vittorini, 2025). Broader perspectives in artificial intelligence in education further emphasize the need for asset-based design (Ocumpaugh et al., 2024), privacy-aware open learner models (Zapata-Rivera, 2021), and blended approaches where AI complements but does not replace expert mentoring (Nygren et al., 2025).

Despite these advances, three gaps remain. First, most studies report correlations between AI and human scores rather than testing equivalence, leaving unanswered whether AI and human raters can be considered interchangeable within pedagogically meaningful bounds. Second, construct validity is under-explored: few analyses demonstrate whether AI scoring respects rubric criteria or privileges surface features such as fluency. Third, fairness across contexts is rarely tested, with limited evidence on invariance across institutions, languages, or demographic groups. Addressing these gaps is essential for moving beyond efficiency narratives toward evidence-based integration of GenAI in assessment.

3. Methodology

3.1. Participants and data

The dataset comprised student essays from teacher education courses in two universities in Greece. Specifically, essays were developed within the courses 'Teaching Programming and Algorithmic Thinking in Early Childhood Education' at the University of Crete (UoC) and 'Interdisciplinary Approaches to Science and Information and Communication Technologies (ICT) in Education' at the Aristotle University of Thessaloniki (AUTh), both in Greece. Each assignment required students to design an educational activity in ScratchJr to teach science concepts to preschool children (ages 4–6). A total of N = 91 essays were collected across the two universities (Uni1 = 48, Uni2 = 43). All essays were anonymized prior to evaluation, with identifiers removed to ensure confidentiality.

3.2. Assignment and task design

The assignment brief instructed students, at the end of the semester, to design a ScratchJr activity aligned with the national preschool science curriculum, focusing on phenomena such as living organisms, physical changes, or planetary systems. Each student was required to submit (a) a written report of 1,500–2,000 words, (b) screenshots of their ScratchJr implementation, and (c) a justification of their pedagogical design choices, referencing both curriculum standards and their ScratchJr project. ScratchJr is a block-based programming environment designed for preschool and early primary students (https://www.scratchjr.org/). The task emphasized alignment with learning objectives, creativity, technical execution, classroom feasibility, and assessment planning.

3.3. Instruments: Rubric and evaluation guidelines

Evaluation was guided by a nine-criterion rubric developed and validated within the course context. Each criterion was scored on a 1–10 scale and addressed:

1. Pedagogical alignment with curriculum goals
2. Originality and creativity of the activity

3. Technical implementation in ScratchJr
4. Accuracy and integration of science content
5. Use of ScratchJr features to support learning
6. Coherence and organization of the activity
7. Feasibility of classroom implementation
8. Evaluation methods for children's learning
9. Quality of the written report

Scores were aggregated to produce totals up to 90. The rubric was accompanied by evaluation guidelines, which the researchers used to interpret the assignment brief, review each student submission, and apply the rubric criterion by criterion while providing detailed, pedagogically constructive feedback.

3.4. Human raters

The two authors of this study, both faculty members with expertise in educational technology and preschool science education, served as human raters. They first completed a calibration session using three pilot essays not included in the final dataset. In this session, they applied the rubric independently and then discussed discrepancies to establish common scoring interpretations. During the main scoring phase, the raters worked independently, blind to each other's evaluations and to the AI results.

3.5. AI-based evaluation

Essays were also evaluated by a large language model (LLM). The AI received the anonymized student essays together with the exercise description, the full rubric, and the evaluation guidelines (see Supplementary Material: Exercise Description, Rubric, and Evaluation Guidelines). A structured prompt instructed the model to (a) assign a score from 1 to 10 for each rubric criterion, (b) provide a brief justification for each score using rubric language, and (c) present the results in a structured table for aggregation.

AI system. For AI-based scoring, we used ChatGPT-4o via the ChatGPT Plus subscription (OpenAI, June–July 2025). At the time of data collection, ChatGPT was the most commonly used generative AI tool among university educators, with surveys reporting that a large majority of instructors relied on it for both content creation and assessment tasks (Jaschik, 2024); GPT-5 had not yet been released. This prevalence, combined with its accessibility through the ChatGPT Plus interface, informed our decision to use ChatGPT-4o for automated scoring. We considered alternative services such as MagicSchool and Brisk Teaching. MagicSchool is known to support GPT-4o among other models, while Brisk Teaching offers a suite of generative-AI tools, though its underlying LLM isn't publicly disclosed. To maintain transparency and reproducibility, we used the model directly via the ChatGPT interface. Competing systems (e.g., Claude, Copilot, or Chinese LLMs) were not used, ensuring the study remained transparent and reproducible by focusing on one accessible platform.

Prompting procedure. For each essay, we provided the complete Greek text together with the analytic rubric, instructing the model to produce criterion-level scores and brief rationales. To minimize variability, the same prompt template was used for all essays, and responses were requested in a structured format (scores plus rationale; see Supplementary Material – Prompt template). This approach avoided potential meaning distortion or stylistic bias that can occur when texts are translated into English prior to scoring.

https://doi.org/10.22521/edupij.2025.19.591 Published online by Universitepark Press

3.6. Procedure

Each essay was evaluated three times: by Rater A, Rater B, and the AI tool. Scores were first exported into spreadsheets for preliminary checks and then imported into IBM SPSS Statistics (Version 29). All statistical analyses were conducted in SPSS, while figures were generated in Python using the matplotlib and seaborn libraries. Anonymized data outputs are available on OSF to support reproducibility. Human raters completed their evaluations within two weeks, and AI scoring was conducted afterwards to prevent any contamination of human ratings.

3.7. Data analysis

Analyses followed three steps:

1. Reliability and agreement. Inter-rater reliability between human graders was assessed using intraclass correlation coefficients (ICCs). Agreement between human and AI scores was further examined with Bland–Altman plots.

2. Construct coverage. A principal component analysis (PCA) was conducted as a descriptive check of the dimensionality of rubric scores, comparing the variance structure produced by human and AI ratings.

3. Group comparisons. Differences in AI–human agreement across institutions were explored descriptively; no confirmatory invariance testing was conducted.

4. Results

4.1. Descriptive statistics and distributions

In comparison to the human experts, the AI demonstrated a marked tendency toward score inflation and variance compression. As detailed in Table 11, the AI effectively established a high 'floor' for performance (min = 55), whereas the human raters utilized the full breadth of the scale to distinguish between failing and excellent work (Human-Z min = 31; Human-S min = 26). Consequently, the AI produced a highly compressed, uniform distribution (SD = 8.41) that failed to discriminate between proficiency levels as effectively as the faculty raters (Human SDs > 10.7).

Distributional analyses confirmed these patterns. The AI’s distribution was nearly symmetric, with slight negative skew (–0.873) and kurtosis near zero (0.156). As shown in Figure 2, both human distributions were more heavily skewed to the left (Z = –1.628, S = –1.549) and leptokurtic (kurtosis ≈ 4.0), reflecting a concentration of higher scores with a minority of very low scores.

Table 1. Descriptive statistics for AI, Human-Z, and Human-S scores

Rater	M	SD	Min	Max	Skewness	Kurtosis
AI	78.37	8.41	55	90	−0.873	0.156
Human-Z	75.49	10.72	31	90	−1.628	4.042
Human-S	74.89	11.49	26	90	−1.549	4.118

Note. AI = automated rater; Human-Z and Human-S = human raters.

https://doi.org/10.22521/edupij.2025.19.591 Published online by Universitepark Press

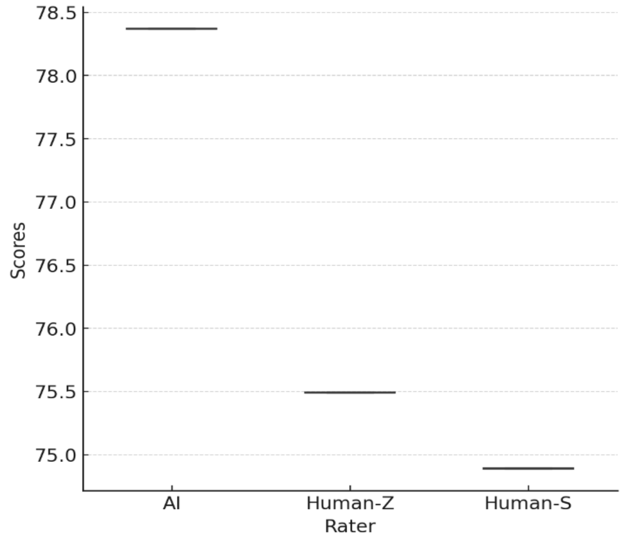

Figure 1. Boxplots comparing distributions of scores assigned by AI, Human-Z, and Human-S. The AI scores show a narrower range and a higher floor, while human raters made fuller use of the rubric scale.

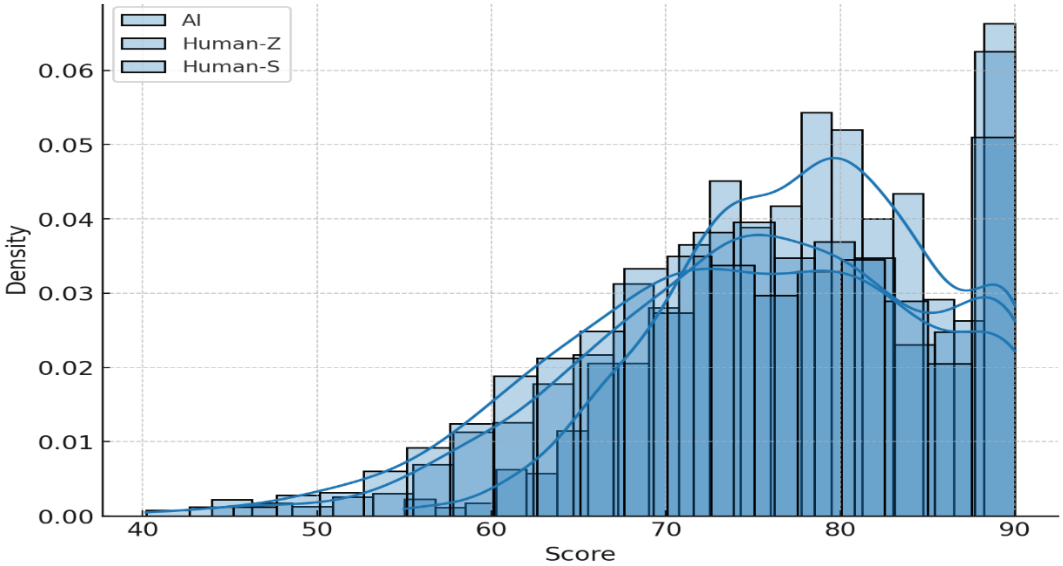

Figure 2. Histograms with kernel density estimates (KDEs) for each rater. AI scores cluster more tightly around the mean, whereas human raters' distributions are broader and more skewed, reflecting greater differentiation across the performance spectrum

4.2. Internal consistency of rubric criteria

Cronbach's α values were high for all raters: AI (α = .931), Human-Z (α = .954), and Human-S (α = .962). This confirms internal coherence of the nine rubric criteria, though the high magnitude also reflects some redundancy, consistent with the dominance of the first principal component in later analyses.

4.3. Human–human agreement

Human raters exhibited excellent agreement. Table 2 illustrates that Pearson's correlation between Human-Z and Human-S totals was $r = .886$ ($p < .001$), and Spearman's $\rho = .803$ ($p < .001$). Intraclass correlations confirmed this, with ICC = .884 (single measures) and ICC = .938 (average measures), both $p < .001$. Criterion-level ICCs ranged from .76 to .87 (single) and .86 to .93 (average), indicating consistent alignment across all dimensions. The PCA indicated that most variance was captured by the first component, suggesting that the AI applied a narrower construct of quality than human raters. No confirmatory factor analysis was conducted.

Table 2. Descriptive statistics for AI, Human-Z, and Human-S scores

Panel A. Total-score correlations

Measure	Value (p)
Pearson's r	.886 ($p < .001$)
Spearman's ρ	.803 ($p < .001$)

Panel B. ICC (2,1) absolute agreement (two-way random)

Criterion	ICC Single	ICC Average
1	.857	.923
2	.803	.891
3	.782	.878
4	.760	.864
5	.806	.892
6	.809	.894
7	.853	.921
8	.754	.860
9	.865	.927
Total	.884	.938

4.4. AI–human correlations

Table 3 illustrates that AI's agreement with humans was weaker. With Human-Z, Pearson's $r = .431$ ($p < .001$) and Spearman's $\rho = .468$ ($p < .001$). With Human-S, Pearson's $r = .305$ ($p = .003$) and Spearman's $\rho = .301$ ($p = .004$). Criterion-level associations ranged from .27 to .45, consistent with only moderate alignment.

Table 3. AI–human correlations

Totals

Pair	Pearson r (p)	Spearman ρ (p)
AI – Human-Z	.431 ($p < .001$)	.468 ($p < .001$)
AI – Human-S	.305 ($p = .003$)	.301 ($p = .004$)

Per-criterion (Pearson, AI – Human-Z only)

Crit.	r (p)
1	.374 ($p < .001$)
2	.289 ($p = .005$)
3	.273 ($p = .009$)
4	.357 ($p < .001$)
5	.322 ($p = .002$)
6	.379 ($p < .001$)
7	.280 ($p = .007$)

https://doi.org/10.22521/edupij.2025.19.591 Published online by Universitepark Press

8	.385 ($p < .001$)
9	.451 ($p < .001$)

4.5. Mean comparisons

Paired-samples t-tests demonstrated systematic inflation by the AI. As shown in Table 4, AI scores were significantly higher than Human-Z (+2.71 points, $t(90) = 2.48$, $p = .015$, $d = .26$) and Human-S (+3.32 points, $t(90) = 2.63$, $p = .010$, $d = .28$). Both differences remained significant under Bonferroni correction. The difference between Human-Z and Human-S (+0.60) was nonsignificant. At the criterion level, AI assigned consistently higher marks, with the largest gaps in criteria 6 and 7 (effect sizes up to .71).

Table 4. Paired-samples comparisons (AI vs Human-Z, AI vs Human-S, Human-Z vs Human-S) with Bonferroni adjustment

Comparison	Mean difference	t(90)	p (two-tailed)	Bonferroni-adj p	Cohen's d
AI – Human-Z	+2.714	2.484	.015	.045 *	.260
AI – Human-S	+3.319	2.632	.010	.030 *	.276
Human-Z – Human-S	+0.604	1.078	.284	.852 (n.s.)	.113

*Note. N = 91 essays. Bonferroni adjustment over three pairwise tests. Cohen's d values are from the paired-samples output. * Indicates significant after Bonferroni correction.*

4.6. Variance compression and distributional differences

Levene's tests indicated significant variance differences between AI and human scores, reflecting AI's narrower spread (Figure 3). Kernel density plots highlight this compression, showing the AI's reduced ability to discriminate at the lower and upper ends of performance.

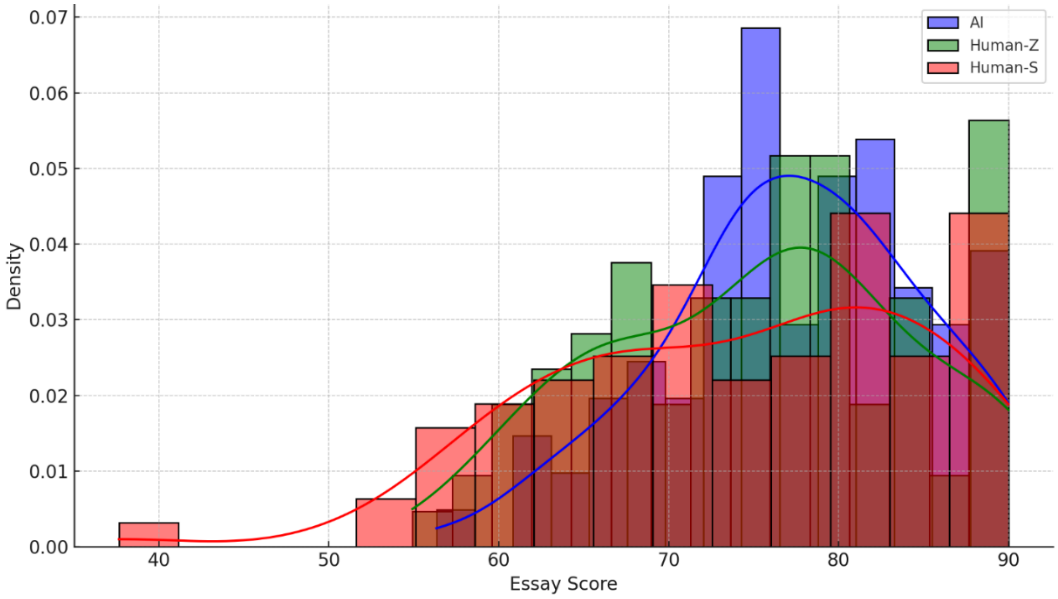

Figure 3. *Histograms with kernel density overlays for AI, Human-Z, and Human-S scores. AI scores show a compressed distribution with a higher floor and narrower spread, whereas human raters display broader, left-skewed distributions, reflecting greater discrimination across the performance spectrum.*

4.7. Bland–Altman analyses and proportional bias

According to Figure 4, Bland–Altman plots (AI–Z, AI–S) revealed systematic proportional bias. Regression slopes were negative and significant (AI–Z: $b = –0.253$, $p = .016$; AI–S: $b = –0.160$, $p =$

https://doi.org/10.22521/edupij.2025.19.591 Published online by Universitepark Press

.003), showing that AI over-scored weak essays and under-scored strong ones. This regression-to-the-mean effect explained 6–9% of variance in score differences.

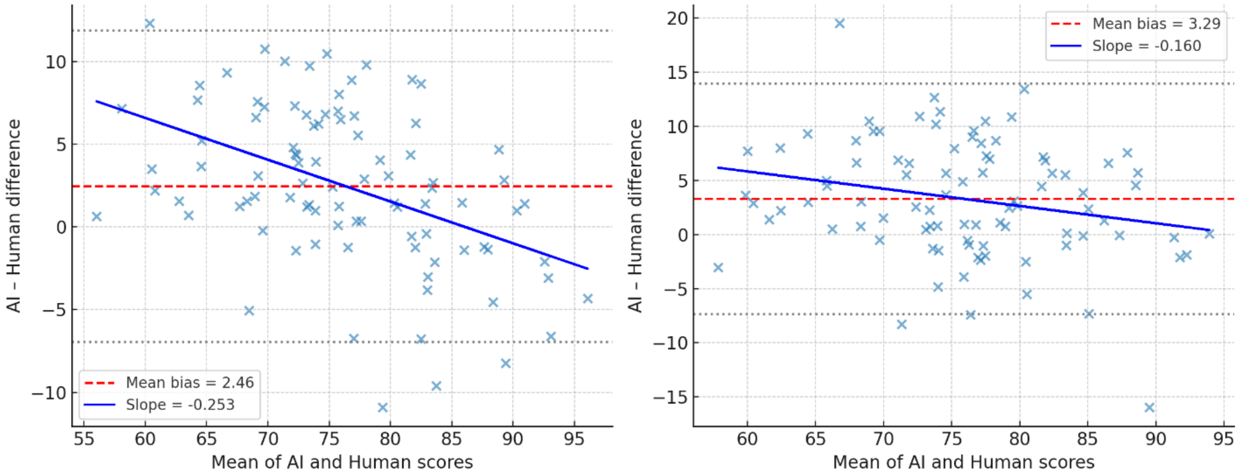

Figure 4. *Bland–Altman plots comparing AI scores with Human-Z (a) and Human-S (b). The dashed red line indicates mean bias, dotted grey lines mark ±1.96 SD limits of agreement, and the solid blue line shows regression fit. Negative slopes indicate systematic proportional bias: AI inflated weaker essays and under-scored stronger ones.*

4.8. Criterion-level intraclass agreement

AI–human ICC were consistently weak. As shown in Table 5, for AI–Z, single-measure ICCs across criteria ranged from .18 to .44, with overall ICC = .406 (average = .578). For AI–S, ICCs were lower still, averaging around .28 (single) and .44 (average). By contrast, human–human ICCs were consistently above .80, confirming stark divergence in reliability.

Table 5. Intraclass correlation coefficients (ICC[2,1], absolute agreement) per criterion and overall

Criterion	AI–Human Z (single)	AI–Human Z (average)	AI–Human S (single)	AI–Human S (average)	Human–Human (single)	Human–Human (average)
1	.367	.537	.259	.411	.857	.923
2	.268	.423	.118	.210	.803	.891
3	.270	.425	.241	.388	.782	.878
4	.352	.521	.219	.359	.760	.864
5	.321	.486	.289	.448	.806	.892
6	.315	.479	.259	.412	.809	.894
7	.181	.307	.152	.264	.853	.921
8	.384	.555	.160	.275	.754	.860
9	.443	.614	.365	.535	.865	.927
Total	.406	.578	.279	.436	.884	.938

Note. N = 91 essays (criterion 6: *N* = 90). ICCs estimated from a two-way random-effects model with absolute agreement. "Single" = ICC(2,1); "Average" = ICC(2,2). (Note: One case was excluded from this analysis due to missing data.)

4.9. Grade-band agreement

When scores were collapsed into grade bands (Fail <50, Pass 50–69, Very Good 70–84, Excellent ≥85), the categorical agreement between the AI and human raters was notably poor. As shown in

Table 6, the AI and Human-S agreed exactly in only 40.7% of cases. More concerningly, the AI misclassified students by two or more grade bands in 5.5% of instances. Cohen's κ indicated no meaningful agreement with Human-S ($\kappa = .017$, $p = .821$). In contrast, agreement with Human-Z was statistically significant ($\kappa = .367$, $p < .001$), with exact agreement occurring in 50.5% of cases. However, 4.4% of cases still exhibited discrepancies of two or more grade bands. This suggests that while the AI aligned better with Human-Z than with Human-S, its accuracy remains insufficient for the high stakes of summative assessment.

Table 6. Grade-band agreement between AI and human raters

Comparison	Exact agreement	±1 category	≥2 categories	Weighted Cohen's κ
AI vs Human-S	40.7%	53.8%	5.5%	0.017 (ns)
AI vs Human-Z	50.5%	45.1%	4.4%	0.367 (p<.001)

Note. N = 91 essays. Grade bands: Fail (<50), Pass (50–69), Very Good (70–84), Excellent (≥85). Percentages calculated from crosstabs. Agreement with Human-Z could not be reliably estimated because the grade-band distribution was constant; SPSS returned a κ of .000 with no standard error.

4.10. Principal component analysis

PCA confirmed unidimensionality across raters, with PC1 explaining 67.8% of variance for AI, 74.2% for Human-Z, and 77.8% for Human-S (Figure 4). Although the AI's scores were internally consistent, the lower proportion of variance explained suggests a narrower capture of essay quality than humans do.

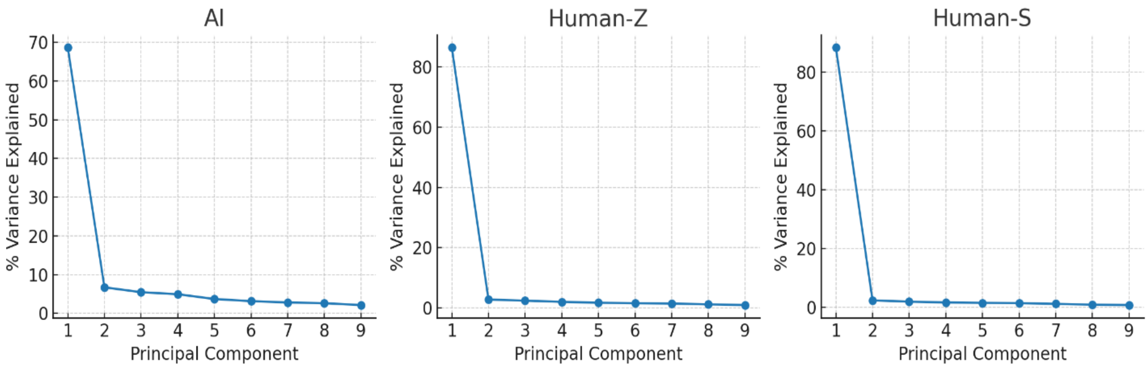

Figure 5. *Scree plots from principal component analyses of rubric criteria for AI, Human-Z, and Human-S. The first component dominated variance for all raters, though AI explained a smaller proportion (~68%) compared to Human-Z (~74%) and Human-S (~78%), indicating narrower construct capture by the AI.*

5. Discussion

The results of this study demonstrate a consistent pattern of distributional compression in the AI's scoring. Compared to human raters, the AI produced higher mean scores but with a markedly narrower range. While humans made fuller use of the rubric scale to distinguish weak from excellent essays, the AI flattened performance differences. This compression undermines sensitivity, as weaker essays were rewarded beyond merit while the strongest essays risked being under-recognized. Although internal consistency was high across all raters, the reliability indices revealed a critical divergence: human–human agreement was excellent, with ICCs approaching .90, whereas AI–human ICCs rarely exceeded .45. This indicates that while the AI applied its own scoring consistently, it failed to converge with the evaluative standards applied by human experts.

Correlational analyses further reinforce this point. The AI's associations with human scores ($r \approx .30–.43$) are insufficient for high-stakes assessment, and the consistent inflation of 2.7–3.3 points indicates a systematic upward bias. Such inflation echoes patterns reported in previous research on

https://doi.org/10.22521/edupij.2025.19.591 Published online by Universitepark Press

automated scoring, where algorithms tend toward leniency. The narrower standard deviation of AI scores also points to reduced discriminatory power, limiting construct validity by blurring the distinctions between weak, average, and outstanding performance. Bland–Altman analyses confirmed that this was not random error but systematic proportional bias: the AI inflated low scores and depressed high ones, effectively regressing all performances toward the mean. Such distortions are pedagogically damaging, as they obscure genuine excellence and risk undermining student motivation.

At the criterion level, intraclass correlations revealed that the AI was unable to replicate the nuanced judgments that human raters applied when interpreting the rubric. Whereas humans consistently differentiated dimensions of writing quality, the AI's pattern suggested a more superficial reliance on surface regularities rather than deep engagement with rubric constructs. The categorical analysis of grade bands illustrates the practical consequences of these weaknesses. Agreement between the AI and humans was inconsistent. While the AI achieved fair alignment with Human-Z ($\kappa = .367$), agreement with Human-S was negligible ($\kappa = .017$). Crucially, exact categorical agreement remained low (40–51%), meaning the AI assigned the wrong grade band in nearly half of all cases. Misclassifications clustered at key thresholds, particularly between "Very Good" and "Excellent," raising concerns about fairness, appeals, and trust in automated systems.

The fact that the AI reached moderate agreement with Human-Z it demonstrates that the model wasn't just guessing. It successfully identified underlying grading patterns. This makes its failure to align with Human-S even more significant. Essentially, the AI acted as a mirror, revealing a lack of standardization between the two human raters. Even though they both used the same rubric, they clearly interpreted it through different lenses. Rubrics rely on subjective terms like 'critical depth' or 'coherent structure,' which leaves room for interpretation. It seems Human-Z focused on criteria the AI could easily spot—like sentence structure and vocabulary—while Human-S likely valued more subtle qualities, such as originality or rhetorical flair, that the algorithm simply couldn't measure.

A principal component analysis (PCA) provided only an exploratory perspective on construct coverage. It suggested that the AI's scoring was largely unidimensional and captured less variance than the human raters, implying a shallower and less discriminative account of writing quality. Taken together, these findings indicate that AI is not yet suitable as a stand-alone grader in high-stakes contexts. Its internal coherence and moderate alignment with human ratings suggest potential utility in formative assessment or as a support tool for managing large cohorts, where it could provide preliminary feedback, triage submissions, or flag problematic cases for human review.

The implications extend beyond psychometric performance to institutional policy. In higher education, where assessment outcomes underpin progression, accreditation, and degree quality, interchangeability with human judgment is a minimum requirement (Ozdogru et al., 2025). Without this, deploying AI in summative grading risks breaching standards of validity and fairness that universities are legally and ethically bound to uphold. Ethical considerations further underline this point: systematic bias toward mid-range scores undermines equity, black-box outputs reduce transparency in appeals, and visible inconsistencies may erode student trust. Until models can demonstrate reproducible equivalence with expert raters, AI's role should remain limited to low-stakes, human-supervised functions.

6. Conclusion

This study tested whether a state-of-the-art large language model could serve as an interchangeable grader for complex coursework. While human raters demonstrated excellent reliability, agreement between AI and humans was weak, with systematic bias and variance compression evident in Bland–Altman analyses. Grade-band agreement was poor, underscoring that current GenAI tools cannot yet replace trained faculty in high-stakes assessment.

https://doi.org/10.22521/edupij.2025.19.591 Published online by Universitepark Press

At the same time, the AI's ability to generate consistent outputs suggests potential for low-stakes applications such as formative feedback, triage, or administrative support—roles where transparency and human oversight remain essential. Broader adoption must also reckon with fairness across languages and contexts, since model performance varies outside English and across academic disciplines.

Future work should extend this analysis to other subject areas, assignment types, and LLM architectures, and evaluate how emerging multilingual models handle authentic student writing. Such evidence will be critical for institutions weighing the integration of AI into assessment policy and practice.

7. Implications for higher education assessment

These findings carry clear implications for higher education. Current GenAI graders should not be treated as interchangeable with human raters in summative assessment, as their systematic bias and limited construct coverage risk unfair outcomes and erode trust. Institutions considering adoption must ensure any AI use remains under human oversight, with transparency about how scores are generated and safeguards for student appeals. At the policy level, quality assurance frameworks and accreditation bodies should require robust evidence of agreement and validity before allowing AI to influence high-stakes grading. In the meantime, universities may explore AI's role in formative feedback, workload reduction, and administrative triage, provided its limitations are openly communicated and qualified educators review outputs.

For higher education institutions, the findings highlight four contributions to practice and policy:

1. Evidence-based boundary-setting – The study clarifies that GenAI is not yet interchangeable with human raters in summative assessment.

2. Productive formative roles – AI may add value as a tool for formative feedback, triaging large cohorts, or administrative streamlining, under transparent human oversight.

3. Methodological guidance – Agreement analyses (ICC, Bland–Altman) and an exploratory PCA offer a practical way for institutions to pre-assess AI graders before adoption.

4. Policy leverage – The results reinforce the need for AI literacy programs for staff and students, and for institutional frameworks that balance efficiency with fairness and equity.

8. Limitations

This study has several limitations. The dataset was drawn from two universities within a single national context, limiting the generalizability of findings to other disciplines or cultural settings. The sample size, while sufficient for the analyses conducted, remains modest compared to large-scale assessment studies. The study also focused on one assignment type—designing a ScratchJr activity for preschool science—so results may not transfer directly to other forms of student work. Furthermore, only one version of a large language model was tested; different models or prompting strategies may yield different outcomes. Unlike many studies that translate non-English student work into English before using AI scoring, our study preserved the original Greek essays. This strengthened construct validity by ensuring that scores reflected the students' authentic language use. At the same time, current large language models vary in performance across languages, and further research should examine whether results replicate with essays written in English or in other linguistic contexts. As multilingual LLMs continue to improve, future work should evaluate whether advances in cross-lingual performance can reduce these constraints and support more equitable use of AI scoring across diverse linguistic settings.

9. Future research directions

Future studies should address these limitations by testing AI–human agreement across larger, more diverse datasets, including assignments from other subject domains and at different levels of education. Comparative studies involving multiple AI models and prompt designs would help clarify whether observed biases are model-specific or systematic. Cross-linguistic research is also necessary to understand how AI scoring behaves in different languages, particularly those underrepresented in training data. Longitudinal work could explore how repeated exposure to AI feedback influences student learning and teacher practices over time. Finally, further inquiry should examine institutional and policy frameworks that support the responsible integration of AI in assessment while maintaining fairness, transparency, and academic standards.

Declarations

Author Contributions. The article was written by two authors who read and approved the final version of the article.

Conflicts of Interest. The authors declare no conflict of interest.

Funding. The authors received no financial support for the research, authorship, and/or publication of this article.

Ethical Approval. In accordance with both universities' ethics policies, analysis of anonymised coursework produced in the normal course of instruction did not require additional REC approval (Research Ethics Committee). All materials were anonymised before analysis, and no identifying information was collected or retained. The study complied with GDPR and institutional guidelines for educational research.

Data Availability Statement. Anonymized rubric scores, AI outputs, and analysis scripts will be available at OSF (https://osf.io/7ay9b/?view_only=dbbceae688584cd3ad2d93ec3b6a7405). Full student essays cannot be shared due to privacy restrictions; summary data and code are provided for replication.

Generative AI Usage Declaration. During the preparation of this study, the authors used Generative AI tools to assist with language editing and clarity of expression. Specifically, ChatGPT and Grammarly were used to improve the language and enhance the text's overall readability. These AI assistants were employed to enhance the linguistic quality of the manuscript, including grammar, style, and clarity of expression. After using these tools, the authors reviewed and revised the content as necessary.

References

Canadian Teachers' Federation. (2024). *Policy brief: Education and AI in Canada*. Canadian Teachers' Federation. https://www.otffeo.on.ca/en/wp-content/uploads/sites/2/2025/05/CTF-ENAI-policy-brief-AGM-2024.pdf

Chai, F., Ma, J., Wang, Y., Zhu, J., & Han, T. (2024). Grading by AI makes me feel fairer? How different evaluators affect college students' perception of fairness. *Frontiers in Psychology, 15*, Article 1221177. https://doi.org/10.3389/fpsyg.2024.1221177

Chauhan, A., Khaliq, F., & Nayak, K. R. (2025). Assessing quality of scenario-based multiple-choice questions in physiology: Faculty-generated vs. ChatGPT-generated questions among phase I medical students. *International Journal of Artificial Intelligence in Education*. Advance online publication. https://doi.org/10.1007/s40593-025-00471-z

Coblentz, D., Dong, J., & Gibbs, B. (2025). Generative artificial intelligence in Aotearoa New Zealand primary schools: Teacher and student survey findings. *New Zealand Council for Educational Research.* https://doi.org/10.18296/rep.0077

Crompton, H., & Burke, D. (2023). Artificial intelligence in higher education: The state of the field. International Journal of Educational Technology in Higher Education, 20(1), 22. https://doi.org/10.1186/s41239-023-00392-8

Department for Education. (2025). *Generative AI in education: Updated guidance for schools*. UK Government. https://www.gov.uk/government/publications/generative-artificial-intelligence-in-education

Dusseau, M., Reynoldson, M., Bone, E., Miciek, C., Kurowski, M., Osmond, M., & Lovelace, G. (2025, July 6). An open letter from educators who refuse the call to adopt GenAI in education. [*Blog post*]. https://openletter.earth/an-open-letter-from-educators-who-refuse-the-call-to-adopt-genai-in-education-cb4aee75

Estévez-Ayres, I., Callejo, P., Hombrados-Herrera, M. Á., Alario-Hoyos, C., & Delgado Kloos, C. (2024). Evaluation of LLM tools for feedback generation in a course on concurrent programming. *International Journal of Artificial Intelligence in Education, 34*(3), 774–790. https://doi.org/10.1007/s40593-024-00406-0

Flodén, J. (2025). Grading exams using large language models: A comparison between human and AI grading of exams in higher education using ChatGPT. *British Educational Research Journal, 51*(1), 201–224. https://doi.org/10.1002/berj.4069

Gamlem, S. M., McGrane, J., Brandmo, C., Moltudal, S., Sun, S. Z., & Hopfenbeck, T. N. (2025). Exploring pre-service teachers' attitudes and experiences with generative AI: A mixed methods study in Norwegian teacher education. *Educational Psychology*. Advance online publication. https://doi.org/10.1080/01443410.2025.2528663

Ghazawi, R., & Simpson, E. (2025). How well can large language models grade essays in Arabic? *Computers & Education: Artificial Intelligence, 6*, 100449. https://doi.org/10.1016/j.caeai.2025.100449

Gombert, S., Fink, A., Giorgashvili, T., & colleagues. (2024). From the automated assessment of student essay content to highly informative feedback: A case study. International Journal of *Artificial Intelligence in Education, 34*(4), 1378–1416. https://doi.org/10.1007/s40593-023-00387-6

Government of Canada. (2024). *Guidance on the use of generative AI. Government of Canada*. https://www.canada.ca/en/government/system/digital-government/digital-government-innovations/responsible-use-ai/guide-use-generative-ai.html

Gu, X., & Ericson, B. J. (2025). AI literacy in K-12 and higher education in the wake of generative AI: An integrative review. In *Proceedings of the 2025 ACM Conference on International Computing Education Research V. 1* (pp. 125–140). ACM.

Jaschik, S. (2024, June 28). One-third of college instructors are using generative AI. Here's how*. Inside Higher Ed.* https://www.insidehighered.com/news/student-success/academic-life/2024/06/28/one-third-college-instructors-are-using-genai-heres

Jurenka, I., Kunesch, M., McKee, K. R., Gillick, D., Zhu, S., Wiltberger, S., & Ibrahim, L. (2024). *Towards responsible development of generative AI for education: An evaluation-driven approach (LearnLM Technical Report)*. Google DeepMind. https://goo.gle/LearnLM

Karakose, T. & Tulubas, T. (2025). The Role of Educational Leaders in the Age of Artificial Intelligence (AI). *Educational Process: International Journal, 16*, e2025267. https://doi.org/10.22521/edupij.2025.16.267

Karakose, T., Tülübaş, T., Kanadli, S., & Gurr, D. (2025). What factors mediate the relationship between principal leadership and teacher professional learning? Evidence from meta-analytic structural equation modelling (MASEM). *Journal of Educational Administration*, *63*(1), 6376. https://doi.org/10.1108/JEA-05-2024-0160

Lee, U., Kim, Y., Lee, S., Park, J., Mun, J., Lee, E., Kim, H., Lim, C., & Yoo, Y. J. (2024). Can we use GPT-4 as a mathematics evaluator in education? Exploring the efficacy and limitations of LLM-based assessment for open-ended mathematics questions. *International Journal of Artificial Intelligence in Education, 34*(4), 1123–1145. https://doi.org/10.1007/s40593-024-00448-4

Letteri, I., & Vittorini, P. (2025). Enhancing student feedback in data science education: Harnessing the power of AI-generated approaches. *International Journal of Artificial Intelligence in Education, 35*(5), 921–940. https://doi.org/10.1007/s40593-025-00492-8

Marrella, D., Jiang, S., Ipaktchi, K., & Liverneaux, P. (2025). Comparing AI-generated and human peer reviews: A study on 11 articles. *Hand Surgery and Rehabilitation, 14*, 102225. https://doi.org/10.1016/j.hansur.2025.102225

Mitchell, M. (2025). Why AI chatbots lie to us. *Science, 389*(6705), 842–844. https://doi.org/10.1126/science.aea3922

Morales-Navarro, S., Morales, A., Nápoles, R., Ceriani, L., & Hernández, A. (2025). High school students building babyGPTs: Engaging in data and prompt engineering practices. *International Journal of Child-Computer Interaction, 37*, 100769. https://doi.org/10.1016/j.ijcci.2025.100769

Morris, W., Crossley, S., Holmes, L., Ou, C., Dascalu, M., & McNamara, D. (2024). Formative feedback on student-authored summaries in intelligent textbooks using large language models. *International Journal of Artificial Intelligence in Education, 34*(2), 321–345. https://doi.org/10.1007/s40593-024-00395-0

Nygren, T., Samuelsson, M., Hansson, P.-O., Efimova, E., & Bachelder, S. (2025). AI versus human feedback in mixed reality simulations: Comparing LLM and expert mentoring in preservice teacher education. *International Journal of Artificial Intelligence in Education, 35*(6), 1001–1022. https://doi.org/10.1007/s40593-025-00484-8

Ocumpaugh, J., Roscoe, R. D., Baker, R. S., & colleagues. (2024). Toward asset-based instruction and assessment in artificial intelligence in education. *International Journal of Artificial Intelligence in Education, 34*(4), 1559–1598. https://doi.org/10.1007/s40593-023-00382-x

Organisation for Economic Co-operation and Development. (2023). *OECD digital education outlook 2023*. OECD Publishing. https://doi.org/10.1787/20769679

Ozdogru, M., Tulubas, T., Karakose, T., Kanadlı, S., Kardas, A., & Papadakis, S. (2025). How does teacher self-efficacy mediate the relationship between student outcomes and principal leadership for learning? Results from meta-analytic structural equation modelling (MASEM). *Acta Psychologica, 258*, 105144. https://doi.org/10.1016/j.actpsy.2025.105144

Öztürk, A., Karahan, A. T., Günay, S., Erdal, A. S., Komut, S., Komut, E., & Yiğit, Y. (2025). A methodology to identify generative AI tools used as co-authors in writing scientific articles. *The American Journal of Emergency Medicine. Advance online publication*. https://doi.org/10.1016/j.ajem.2025.07.034

Pack, A., Barrett, A., & Escalante, J. (2024). Large language models and automated essay scoring of English language learner writing: Insights into validity and reliability. *Computers and Education: Artificial Intelligence, 6*, Article 100234. https://doi.org/10.1016/j.caeai.2024.100234

Papadakis, S., & Karakose, T. (2025). Gamification and Student Achievement: Potential Benefits, Limitations, and Effective Use in Educational Environments. *Educational Process: International Journal, 19,* e2025529. https://doi.org/10.22521/edupij.2025.19.529

Petrone, J. (2025). AI Leap 2025: Estonia sets the standard for AI in education. *e-Estonia*. https://e-estonia.com/

Pham, N., Pham, N. H., & Nguyen-Duc, A. (2025). Fairness for machine learning software in education: A systematic mapping study. *Journal of Systems and Software, 219*, Article 112244. https://doi.org/10.1016/j.jss.2024.112244

Selwyn, N., Ljungqvist, M., & Sonesson, A. (2025). When the prompting stops: Exploring teachers' work around the educational frailties of generative AI tools. *Learning, Media and Technology*. Advance online publication. https://doi.org/10.1080/17439884.2025.2537959

Tate, T. P., Steiss, J., Bailey, D., Graham, S., Moon, Y., Ritchie, D., Tseng, W., & Warschauer, M. (2024). Can AI provide useful holistic essay scoring? *Computers and Education: Artificial Intelligence, 7*, Article 100255. https://doi.org/10.1016/j.caeai.2024.100255

The Australian. (2024, March 2). Shared networks explore AI resources. *The Australian*. https://www.theaustralian.com.au/special-reports/shared-networks-explore-ai-resources/news-story/e8d377b21401b34f8b1dc55a3a4360ed

The Guardian. (2024, January 23). ChatGPT in Australian schools: What you need to know about law changes. *The Guardian*. https://www.theguardian.com/australia-news/2024/jan/23/chatgpt-in-australian-schools-what-you-need-to-know-law-changes

United Nations Educational, Scientific and Cultural Organization. (2023*). Guidance for generative AI in education and research.* UNESCO Publishing. https://www.unesco.org/en/articles/guidance-generative-ai-education-and-research

Vittorini, P., Menini, S., & Tonelli, S. (2021). An AI-based system for formative and summative assessment in data science courses. *International Journal of Artificial Intelligence in Education, 31*(1), 159–185. https://doi.org/10.1007/s40593-020-00230-2

Wang, F., Li, N., Cheung, A. C. K., & Wong, G. K. W. (2025). In GenAI we trust: An investigation of university students' reliance on and resistance to generative AI in language learning. *International Journal of Educational Technology in Higher Education, 22*, 59. https://doi.org/10.1186/s41239-025-00547-9

Wetzler, E. L., Cassidy, K. S., Jones, M. J., Frazier, C. R., Korbut, N. A., Sims, C. M., Bowen, S. S., & Wood, M. (2024). Grading the graders: Comparing generative AI and human assessment in essay evaluation. *Teaching of Psychology, 52*(3), 298–304. https://doi.org/10.1177/00986283241282696

Xing, W., Nixon, N., Crossley, S., Denny, P., Lan, A., Stamper, J., & Yu, Z. (2025). The use of large language models in education. *International Journal of Artificial Intelligence in Education, 35*(1), 439–443. https://doi.org/10.1007/s40593-025-00457-x

Zapata-Rivera, D. (2021). Open student modeling research and its connections to educational assessment. *International Journal of Artificial Intelligence in Education, 31*(2), 380–396. https://doi.org/10.1007/s40593-020-00206-2

About the Contributor(s)

Stamatios Papadakis is an Assistant Professor at the Department of Preschool Education at the University of Crete, Greece. His scientific and research interests include mobile learning, especially the use of smart mobile devices and their accompanying mobile applications (apps) in preschool and primary education, with a focus on developing computational thinking and students' understanding of numbers.
Email: stpapadakis@uoc.gr
ORCID: http://orcid.org/0000-0003-3184-1147

Georgios Zacharis is an Assistant Professor in the Department of Early Childhood Education at Aristotle University of Thessaloniki. His scientific and research interests focus on ICT in Early Childhood Education, with emphasis on pre-scientific learning, computational thinking, and the development of educational tools and interventions that promote scientific and technological literacy in young children.
Email: gzacharis@nured.auth.gr
ORCID: http://orcid.org/0000-0003-1158-9175

Appendix I.

Exercise Description

Assignment Topic

Designing an Educational Activity in ScratchJr for Teaching Science in Early Childhood Education

The aim of this assignment is to develop an interactive activity in the ScratchJr programming environment that supports the teaching of science concepts to preschool children (ages 4–6). You are asked to create an educational scenario aligned with the national preschool curriculum's learning objectives. The scenario should help children engage with basic concepts and phenomena from the natural world through programming and play.

Structure of the Assignment

1. Introduction
    - Briefly describe ScratchJr and its relevance in early childhood education.
    - Present pedagogical theories that support the use of ICT (Information and Communication Technologies) in learning, with emphasis on science education.
2. Thematic Unit Analysis
    - Select a science-related topic appropriate for preschoolers. Examples include:
        - Living organisms and their characteristics
        - Physical phenomena and changes of matter
        - Earth and the planetary system
    - Justify your choice of topic (e.g., children's common alternative ideas, originality, or gaps in current teaching) using sources from the internet. List all sources at the end of the document under "References."
3. Learning Objectives
    - Identify the competencies to be developed, such as observation, classification, and understanding human–nature interactions.
    - Define clear learning objectives for the activity. Objectives should be specific, measurable, achievable, realistic, and time-bound (SMART). Examples:
        - Children will identify the basic characteristics of living organisms.
        - Children will understand fundamental principles of physical phenomena.
        - Children will recognize Earth's position in the planetary system.
4. ScratchJr Activity Design
    - Provide a detailed description of the educational scenario.
    - Include clear development instructions with the following elements:
        - Central idea and story (e.g., an animal moving through environments to explain survival conditions).
        - Connection between the story and the learning objectives.
        - Description of characters' movements, interactions, and roles using ScratchJr programming commands.
        - Use of colors, sounds, and movements to enhance understanding and engagement.
5. Classroom Implementation
    - Plan how the activity will be applied in the classroom.
    - Suggest pedagogical techniques to support children's engagement.
    - Propose adaptations for children with diverse learning needs.
6. Evaluation of the Activity
    - Suggest ways to assess children's understanding after implementation. This may include:
        - Comprehension questions during the activity.
        - Small exercises or games to confirm understanding of concepts.
    - Define specific evaluation criteria (e.g., children's ability to identify natural phenomena or describe cause–effect relationships).

Submission Guidelines

- Submit the assignment as a written document, accompanied by screenshots of the ScratchJr activity.
- The written report should be 1,500–2,000 words.
- The assignment will be assessed based on clarity, creativity, alignment with learning objectives, and pedagogical approach.

Assessment Criteria

- Pedagogical relevance and alignment with the curriculum
- Originality and creativity of the activity
- Use of ScratchJr tools to support understanding of science concepts
- Practical applicability with preschool children

Useful Links

New Curriculum Programs: https://www.iep.edu.gr/provoli-neon-programmaton-spoudon/

Appendix II.

Prompt template
Evaluation instructions for the assignment

Read the full assignment brief, which includes the evaluation rubric, as well as the submitted work file and any accompanying materials (e.g., screenshots of the ScratchJr activity). Based on the evaluation rubric provided, grade the work criterion by criterion and provide both overall feedback and a final grade.

1. Review all uploaded files thoroughly, including the assignment brief, the evaluation rubric, and the submitted work/materials, before beginning the assessment.
2. Evaluate each rubric criterion in detail, noting strengths, weaknesses, and realistic suggestions for improvement.
3. Assign a score for each criterion and provide a final overall score (out of 90), clearly justifying your choices.
4. Check whether the required screenshots, multimedia elements, or implementation evidence requested in the assignment brief have been included, and take them into account in your evaluation.
5. Provide overall feedback in a professional, encouraging, and pedagogical tone. Highlight the value of the student's work while offering concrete suggestions for improvement without discouragement.
6. Maintain a consistent and accessible academic style throughout the evaluation, avoiding unnecessary detail or vague comments.

Appendix III.

Rubric – Evaluation guidelines
Assignment evaluation rubric

Evaluation Criterion	Excellent (9–10)	Very Good (7–8)	Good (5–6)	Needs Improvement (1–4)	Score
1. Pedagogical Relevance & Learning Objectives	The activity is fully aligned with the learning objectives of the Preschool Education Curriculum and strongly promotes the development of science-related skills.	The activity aligns with the curriculum objectives, but with slightly less clarity.	The activity is relevant, but the learning objectives are unclear or incomplete.	The activity is not clearly connected to science learning objectives or the curriculum.	/10
2. Originality and Creativity	The activity is highly creative and original, making smart use of ScratchJr tools.	The activity is creative and introduces new ideas, though in a more limited way.	The activity is adequate but follows common approaches without much originality.	The activity lacks creativity or new ideas.	/10
3. Technical Implementation in ScratchJr	Excellent use of ScratchJr functions (movement, interactions, sequences), with full achievement of activity objectives.	Good use of ScratchJr functions, with minor shortcomings in achieving objectives.	Basic use of ScratchJr, with significant weaknesses in technical implementation.	Very limited use of ScratchJr, with major errors or missing functions.	/10
4. Connection to Science Content	The activity clearly and accurately integrates science concepts, supporting meaningful understanding of the topic.	The activity integrates basic science concepts, though with some room for improvement.	The activity relates to science, but with weak connections or limited depth.	The activity does not clearly integrate science concepts or presents them inadequately.	/10
5. Use of ScratchJr	The ScratchJr activity is fully aligned with learning objectives and science content. Characters, movements, and	The ScratchJr activity adequately supports the learning objectives, though some interactions could	The ScratchJr activity is limited; characters or movements do not meaningfully support	The ScratchJr activity is unclear or does not sufficiently support science learning.	/10

	interactions support conceptual understanding.	be clearer or more age-appropriate.	understanding of science concepts.		
6. Coherence and Organization	The activity is well organized, with clear instructions and a flow that helps children understand the objectives. Steps and activities are age-appropriate.	The activity is reasonably organized, though some instructions or steps require more clarity or adaptation for preschoolers.	The organization of the activity shows weaknesses, and instructions are not sufficiently clear for children.	The activity is unclear, lacks flow and organization, making classroom use difficult.	/10
7. Classroom Applicability	The activity is fully adaptable for preschool classroom use, with clear guidance for teachers.	The activity can be implemented in class with minor modifications.	The activity requires improvement to be effectively used in class.	The activity is not easily applicable in class or requires substantial revision.	/10
8. Learning Assessment	Comprehensive methods for assessing children's learning are proposed (e.g., quizzes, feedback activities, evaluations with clear criteria).	Assessment methods are included but could be more detailed or age-appropriate.	Learning assessment is basic, without clear criteria or feedback methods.	No clear assessment method is provided, or it is inadequate for preschool learning.	/10
9. Written Presentation of the Assignment	The written report is excellently structured, with clarity, complete descriptions, and proper referencing. No grammatical or syntactic errors.	The report is well structured, with minor weaknesses in organization or some grammatical errors.	The report is basic, with structural or grammatical issues and unclear descriptions.	The report is poorly written, unclear, or contains many errors that hinder understanding.	/10