Id: 1884134

Participatory Educational Research (PER)
Vol.13(4), pp. 251-270, July 2026
Available online at http://www.perjournal.com
ISSN: 2148-6123
http://dx.doi.org/10.17275/per.26.58.13.4

From black box to pedagogical partner: Students' sense-making of LLM-based automated assessment

Abdulkadir Kara*
Dept. of Distance Education Application and Research Center, Bayburt University, Bayburt, Türkiye
ORCID: 0000-0003-3255-1408

Serkan Yıldırım
Dept. of Computer Education and Instructional Technology, Atatürk University, Erzurum, Türkiye
ORCID: 0000-0002-8277-5963

Article history

Received:
07.02.2026

Received in revised form:
19.03.2026

Accepted:
06.04.2026

Key words:
Automated assessment; large language models; AI-generated feedback; student experience; explainable AI

While current educational research on automated assessment heavily emphasizes technical validity, a significant gap remains in understanding students' sustained, real-world experiences with these systems in authentic learning environments. This study examines students' experiences with a large language model-based automated assessment system embedded in the regular flow of a university course. The study employed a mixed-methods design with 47 university students over a seven-week period. Quantitative data were obtained from system interaction logs and student feedback ratings, while qualitative data were collected from focus group interviews with 24 students and 175 written feedback responses. The results reveal that students perceive the LLM-based assessment system as a learning assistant, an impartial evaluator, and a self-assessment tool. The transparency of explanations was identified as a decisive factor in building trust in the algorithmic system by helping students understand the rationale behind scores and feedback. Sustained interaction with the system triggered a shift from high-frequency trial and error to more efficient and strategic participation, indicating that the assessment criteria were gradually internalized. The system appeared to create an environment free from perceived social judgment, providing favorable conditions for productive failure, repeated attempts, and self-directed revision. Overall, the study demonstrates that artificial intelligence can be positioned as a tool that scales pedagogical intent without replacing the teacher and can be effectively integrated within the framework of human–AI complementarity in higher education.

Introduction

Assessment and feedback are central components of effective learning environments, shaping not only what students learn but also how they engage with learning processes (Çınar et al., 2020; Kurbanoğlu & Olcaytürk, 2023). Despite their pedagogical value, formative assessment practices are often constrained by practical limitations, including increased instructor workload (Abdul-Salam et al., 2022; Westera et al., 2018), delayed feedback cycles,

* Correspondency: abdulkadirkara@bayburt.edu.tr

and challenges in providing individualized guidance in large or resource-limited classrooms (Gombert et al., 2024). As a result, feedback, one of the most powerful drivers of learning, frequently becomes episodic, summative, and judgment-oriented rather than continuous and learning-focused. Short-answer evaluations involving natural language responses are valuable approaches for detailed assessment because they provide evidence of students' higher-order thinking skills, such as analysis, synthesis, and evaluation (Ariely et al., 2022; Westera et al., 2018).

Short answers can range from a few sentences to several paragraphs (Nath et al., 2023). Although short answers, due to their structural characteristics, can reveal students' ability to construct and express knowledge in detail (Uto & Uchida, 2020), practical difficulties limit their use in assessment practices (Çınar et al., 2020). Xavier et al. (2025) highlight the difficulty of delivering timely and individualized feedback for open-ended responses. Moreover, inconsistencies stemming from subjective judgments among human raters raise concerns about the reliability and fairness of short-answer evaluations (Zhu et al., 2022), further dissuading their classroom adoption.

Recent advances in artificial intelligence, particularly large language models (LLMs), have renewed interest in the automation of assessment and feedback processes. LLM-based systems offer the potential to generate rapid, scalable, and linguistically rich feedback, addressing long-standing concerns related to timeliness and consistency (Hao et al., 2024). Indeed, Singerin et al. (2025) note that automatic assessment applications stand out in the context of AI use in higher education. Consequently, a growing body of research has examined the technical performance, validity, and reliability of automated scoring systems and LLM-based assessment tools (Hao et al., 2024; Kortemeyer, 2024; Mendonça et al., 2025). These studies have demonstrated that such systems can achieve high levels of agreement with expert raters under controlled or benchmarked conditions.

While technical evaluations dominate the literature, learning technologies cannot be meaningfully evaluated without considering the experiences of the students who interact with them (Boud & Molloy, 2013; Laurillard, 2009; Nicol, 2021). Feedback is informative, but it is also a relational and interpretive assessment process shaped by trust, perceived fairness, and opportunities for action (Carless & Boud, 2018; Shute, 2008; Tossell et al., 2024). When combined with feedback loops, automated assessment systems go beyond being merely assessment tools and serve multifaceted pedagogical functions that influence students' study behaviors, motivational orientations, and self-regulation strategies (Taub et al., 2021; Tossell et al., 2024). Understanding these dynamics requires moving beyond short-term experiments and focusing on sustainable, real-world applications that prioritize student voice (Khosravi et al., 2022; Luo et al., 2025).

However, the increasing technical sophistication of automated assessment systems has not been matched by an equivalent understanding of their pedagogical implications in real learning environments (Hao et al., 2024; Holmes et al., 2022; Zawacki-Richter et al., 2019). Recent review studies indicate that research on automated assessment and LLM-based systems has largely conceptualized success through performance-oriented metrics such as accuracy, agreement coefficients, and reliability indices (Ahmad et al., 2023; Luo et al., 2025). Within this literature, learning contexts are frequently treated as neutral delivery spaces, with limited attention to how assessment technologies are experienced, interpreted, and appropriated by learners over time (Tossell et al., 2024). In particular, it is emphasized that systematic evidence regarding students' sustained interaction with LLM-based feedback

within the framework of placing automated assessment systems as active components of real learning environments remains limited (Luo et al., 2025). Given the limitations in the existing literature and the need for empirical evidence, this study seeks to answer the following research questions:

- RQ1. How do students experience and make sense of an LLM-based automated assessment system when it is embedded as an active component of an authentic learning environment?
- RQ2. In what ways does sustained interaction with an LLM-based automated assessment system influence students' learning behaviors, self-regulation, and perceptions of assessment?

Background and Related Work

Transformer-based language models, particularly Bidirectional Encoder Representations from Transformers (BERT) and its variants, have been widely adopted for automatic short-answer scoring across diverse linguistic and educational settings. Empirical studies report that BERT-based models achieve substantial agreement with human raters on benchmark datasets such as the Semantic Evaluation shared task dataset (SemEval), establishing the technical feasibility of automated scoring systems (Sung et al., 2019; Ghavidel et al., 2020). Multilingual research further demonstrates that transformer-based approaches generalize across languages, including German and Spanish, supporting their applicability beyond English-dominant contexts (Sawatzki et al., 2021; Mardini et al., 2024). However, prior work also indicates that scoring performance may decline as task complexity increases or as scoring granularity expands, suggesting limitations in relying solely on technical performance metrics (Ghavidel et al., 2020). Taken together, these studies suggest that while automated scoring systems can provide consistent and scalable assessments, scoring alone offers limited pedagogical value unless it is complemented by meaningful feedback that supports learning processes.

Building on these developments, recent work highlights the rising potential of LLMs for short-answer evaluation and feedback generation. LLMs have expanded the scope of automated assessment beyond numerical scoring toward the generation of explanatory and linguistically rich feedback. Empirical research indicates that LLM-based systems can generate feedback comparable to that provided by human instructors in specific task contexts. For example, Chang and Ginter (2024) and Latif and Zhai (2024) report that GPT-based models achieve moderate to high agreement with human scoring, while Cheong (2025) shows that generative feedback is particularly effective for lower- to mid-level cognitive tasks, with performance on higher-order tasks remaining sensitive to task design and prompting strategies. Similarly, Lohr et al. (2025) emphasize that LLMs are now capable of providing rich and personalized feedback. Delalibera and Carvalho (2024) further emphasize that AI-supported automated assessment systems can significantly reduce educator workload while supporting learning through timely and comprehensive feedback. Nevertheless, these findings are largely derived from technically oriented evaluations rather than sustained classroom implementations.

When student and instructor perspectives are explicitly examined, learners generally appreciate the clarity, immediacy, and explanatory nature of feedback generated by LLMs (De-Wet et al., 2025; Escalante et al., 2023), yet concerns regarding accuracy, originality, and reliability remain significant (Chiang et al., 2024; Tossell et al., 2024). Er et al. (2025) found

that students perceived instructor feedback as more beneficial and pedagogically supportive than feedback generated by AI, particularly in terms of fairness and developmental value, although these differences were not always statistically significant. Xavier et al. (2025) reported that GPT-4-generated personalized feedback was perceived as credible and reliable by instructors. Focusing on the effectiveness of explanatory feedback, Gombert et al. (2024) evaluated their system using a six-dimensional questionnaire (comprehensibility, usefulness, progress, reflection, self-regulation, and motivation) and observed that students responded positively to detailed feedback.

Taken together, existing research demonstrates the effectiveness of BERT-based models in automated short-answer scoring and highlights the potential of LLMs in feedback provision. However, studies examining students' sustained experiences with such systems, particularly in real learning environments and non-English instructional contexts, remain limited, especially with respect to dimensions such as learning strategies, perceptions, and engagement. To address these gaps, the present study adopts a participatory educational research approach to examine students' sustained experiences with an LLM-based automated scoring and feedback system embedded in an authentic learning environment.

Method

Research Design

This study employed a qualitative-dominant mixed-methods research design to investigate students' sustained experiences with an LLM-based automated assessment system embedded as an active component of an authentic learning environment. In this study, an authentic learning environment refers to the regular flow of a university course in which the AI-supported assessment system was integrated into ongoing coursework over seven weeks, rather than being implemented in a short-term or laboratory-based setting.

The design foregrounds learners' perspectives and meaning-making processes as the primary sources of evidence. Within this design, qualitative data constitute the main analytic focus and are used to explore how students perceive, interpret, and engage with automated scoring and explanatory feedback (RQ1). Complementary quantitative indicators were incorporated to examine patterns of sustained interaction, learning-related behaviors, and self-regulatory tendencies over time (RQ2). Quantitative data are used descriptively to contextualize students' experiences rather than to evaluate learning outcomes or model performance.

Learning Context and Participants

The study was conducted within the regular flow of a higher education course in a non-English instructional context, situated in the social sciences and designed to support conceptual understanding through written short-answer assessments. The course included regular formative assessment activities requiring students to produce open-ended responses.

Before the main implementation, a pilot study was conducted with a separate group of 8 participants to examine the functionality and practical applicability of the AI-supported assessment system. These participants were not included in the main study sample. For the main implementation, participants were selected from a single group of students taking the same university course. Although the course was offered to students from multiple departments, the study focused on one department because these students were expected to have the basic computer literacy required to engage with the AI-supported assessment system.

The selection therefore followed a purposive convenience approach. Participation was voluntary throughout the study. Of the 60 students who initially had access to the system, 47 remained in the final study sample and generated analyzable log data, while 13 did not continue in the study. In addition, post-implementation interviews were conducted with 24 participants.

Students interacted with automated scoring and feedback as a natural part of their learning process across multiple assessment tasks over an extended instructional period. Participation in research-related data collection activities did not influence course grades or academic standing. All participants had prior experience with conventional instructor-led assessment practices, providing a meaningful basis for reflection on automated scoring and feedback.

LLM-Based Assessment System

The assessment system was designed as an integrated automated assessment and feedback mechanism, combining two complementary components: (1) a transformer-based model for automated short-answer scoring and (2) a generative large language model (GPT-4o) for explanatory feedback generation. Scores and feedback were presented concurrently to students as part of a unified formative evaluation experience.

Automated Scoring Component

The automated scoring component was trained using a custom-developed short-answer dataset for a university-level History course. The dataset consisted of student responses paired with expert-assigned scores based on predefined scoring criteria aligned with official course learning outcomes. Content validity was established through expert review (Content Validity Index [CVI] = 0.86), and expert scoring demonstrated high inter-rater reliability (Cohen's Kappa [κ] = 0.93). Together, these procedures provided baseline evidence of content, construct, and scoring validity consistent with argument-based validity frameworks (Kane, 2006; Williamson et al., 2012).

Explanatory Feedback Component and Theoretical Framework

The explanatory feedback component was built on the GPT-4o model and designed to generate personalized, pedagogically meaningful feedback rather than generic responses. Feedback generation was guided by structured prompts incorporating the question text, student response, automated score, and instructional objectives. Feedback content was adapted based on response accuracy, providing explanations for errors, identifying missing elements, or reinforcing correct understanding.

The design of the feedback component was grounded in established educational theories. Shute's (2008) formative feedback principles, including specificity, clarity, instructional relevance, and timeliness, guided the construction of feedback prompts. Moore's (1989) interaction theory informed the system's capacity to simulate learner–content and learner–teacher interaction through responsive explanations. In addition, Nicol and Macfarlane-Dick's (2006) model of formative assessment and self-regulated learning underpinned the assumption that explanatory feedback can support reflection, metacognitive engagement, and learner autonomy. Taken together, these frameworks position the system not only as an automated grading tool but also as a pedagogically informed learning support tool. A conceptual overview of the integrated assessment system and its workflow is presented in Figure 1 to clarify the system's structure and role in classroom implementation.

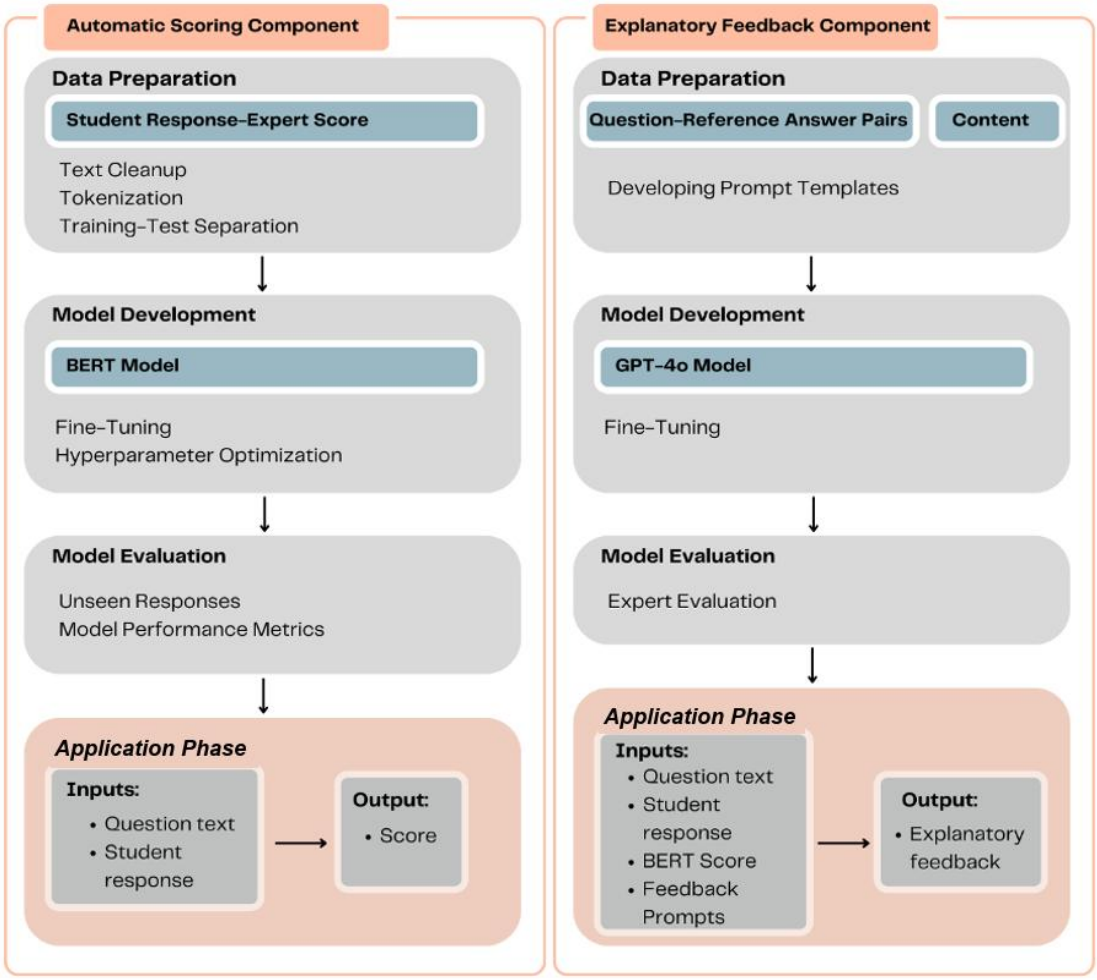

Figure 1. Automated assessment system architecture

Pilot Study

Prior to classroom deployment, a pilot study was conducted to examine the pedagogical suitability and practical applicability of the integrated assessment system. The pilot was carried out with a small group of students (n = 8) using a five-item short-answer assessment. Automated scores were compared with expert judgments to ensure reasonable alignment and to identify potential issues that could hinder classroom use. Agreement indices indicated a high level of consistency between system outputs and expert evaluations ($\kappa \approx .85$-.89; Pearson correlation coefficient [r] ≈ .83-.90), suggesting that the scoring component met an acceptable threshold for formative use.

In addition, explanatory feedback generated by the system was reviewed by five field experts using established principles of effective feedback (Shute, 2008). Expert review showed that some of the system-generated feedback was unnecessarily detailed. Based on this review, the prompts were revised to produce clearer and more concise feedback before classroom deployment.

Classroom Implementation

Following the pilot study, the LLM-based automated assessment system was implemented as a regular component of the course assessment process over a seven-week instructional period. During this phase, the system was integrated into routine formative

assessment activities and used by students as part of their normal coursework rather than as an experimental add-on.

Each week, students completed short-answer assessment tasks aligned with course learning outcomes and received automated scores and explanatory feedback immediately upon submission. Students were able to review feedback, reflect on their responses, and, when applicable, use this information to guide subsequent study and revision.

The instructor did not intervene in the automated scoring or feedback process during the implementation phase, allowing the system to function autonomously while remaining embedded within the regular instructional context of the course. Students' interactions with the system generated log data documenting participation, response activity, and feedback ratings, which were later used for descriptive analysis. This sustained classroom implementation provided the empirical basis for examining students' sense-making processes, learning behaviors, and perceptions of automated assessment over time.

Data Collection and Analysis

Multiple data sources were employed to capture students' sustained experiences with the assessment system. Qualitative data, including students' written reflections, open-ended responses, and semi-structured interviews, served as the primary source of evidence and were analyzed using thematic analysis. An iterative and inductive coding process was used to identify themes related to feedback perception, trust, engagement, self-regulation, and learning strategies.

Quantitative data were derived from system-generated interaction logs and student feedback ratings. These data were analyzed descriptively and, where appropriate, through supplementary longitudinal analyses to examine change over time. Because the number of questions varied across weeks, question-adjusted interaction rates were calculated. Given the unbalanced repeated measures caused by participant attrition, linear mixed-effects models were used with week as a fixed effect and student as a random effect. Satisfaction analyses were based on rated responses only, and the association between adjusted interaction rate and satisfaction was examined using Spearman rank-order correlation. Integration of qualitative and quantitative findings occurred at the interpretation stage to support methodological triangulation.

Ethical Considerations

All procedures adhered to institutional ethical guidelines for research involving human participants. Participation in research-related data collection was voluntary. Informed consent was obtained, and participants could withdraw at any time without penalty. Research participation did not affect course grades or academic standing. All data were anonymized prior to analysis, securely stored, and reported either in aggregate form or in anonymized form. The use of automated scoring and LLM-generated feedback was transparently communicated to students as part of the course assessment process.

Findings

The findings are organized around the study's two research questions. First, students' perceptions of assessment performance and system features are presented (RQ1). Second, findings related to students' learning experiences, emotional processes, and study strategies are presented (RQ2). Qualitative findings are based on thematic analysis of focus group

interviews and are supported by representative participant quotes; quantitative findings are presented through descriptive statistics and frequency distributions.

Students' Experience and Sense-Making of LLM-Based Assessment System (RQ1)

The findings revealed students' perceptions of the automated assessment system.

Perceptions of the System

Analysis of focus group interviews revealed that students conceptualized the automated assessment system through four distinct frames. Table 1 presents the frequency distribution.

Table 1. Students' conceptualizations of the automated assessment system

Code	Frequency	Key Findings
Learning helper	12	System viewed as assistant that facilitated understanding
Impartial evaluator	7	System perceived as objective and fair
Self-assessment tool	6	System enabled instant self-evaluation
Teacher-like entity	5	System compared to instructor feedback

Participants articulated multiple ways of conceptualizing the system's role in the learning process. The majority (f = 12) perceived the system as a learning helper that facilitated understanding and supported knowledge acquisition (e.g., P9: "It made learning easier for me and helped me identify my shortcomings"). Others (f = 7) emphasized its function as an impartial evaluator, highlighting the fairness and objectivity of its assessments (e.g., P18: "It was a very good and impartial system"). In addition, some participants (f = 6) described the system as a self-assessment tool that enabled instant evaluation of their own learning status (e.g., P3: "I can see my own status instantly"). A smaller group (f = 5) compared the system to instructor feedback, characterizing its responses as teacher-like in nature (e.g., P19: "We don't have to wait for instructors; it responds like they do").

Interpreting Instant Score and Explanatory Feedback Integration

Students' interpretations of the combined instant score and explanatory feedback presentation revealed four key themes. Table 2 presents the frequency distribution.

Table 2. Interpreting instant score and explanatory feedback integration

Code	Frequency	Key Findings
Explanatory guidance	17	Feedback explained rationale behind scores
Error correction process	15	Detailed explanations enabled targeted learning
Pathway to correct answers	12	Feedback guided toward correct responses
Integrated system	5	Score and explanation perceived as unified

Participants frequently emphasized the explanatory and guiding nature of the feedback generated by the system. A large proportion of participants (f = 17) highlighted the system's explanatory guidance, noting that it clearly presented correct answers together with underlying reasons (e.g., P2: "It presented the correct answer with reasons through the detailed explanation"; P11: "Even though we gave a short answer, it gives an explanatory response"). In addition, many participants (f = 15) described the feedback as an effective error

correction process, indicating that explanations helped them identify mistakes and prompted further learning efforts (e.g., P9: "The explanations from feedback showed our mistakes"; P24: "Even when wrong, it encouraged me to research the topic again"). Some participants (f = 12) framed the feedback as a pathway to correct answers, emphasizing that clear explanations supported improvement even when initial responses were incomplete or incorrect (e.g., P18: "For my 'I don't know' answers, it gave me a clear explanation. This time, I received full points"). A smaller group of participants (f = 5) explicitly referred to the system as an integrated structure in which automated scoring and explanatory feedback operated together (e.g., "The scoring system and AI work in tandem with each other").

Trust, Fairness, and Credibility of Automated Assessment

The findings are grouped into four dimensions: accuracy, transparency, fairness, and consistency. Table 3 presents the frequency distribution of student perceptions.

Table 3. Student perceptions of system trustworthiness

Dimension	Positive	Negative	Key Findings
Accuracy	20	4	High accuracy perceived; some frustration when brief but correct answers received low scores.
Transparency	18	4	Detailed feedback improved clarity and understanding; long explanations occasionally caused confusion.
Fairness	9	3	Viewed as fair and content-based; some concern that system rewarded lengthy responses.
Consistency	5	-	Consistent scores across attempts enhanced student confidence in system reliability.

Accuracy: The majority of participants (f = 20) perceived the system's evaluations as accurate and reliable (e.g., P21: "I can say the accuracy rate is ninety percent"). However, some participants (f = 4) noted instances where the system penalized brief but correct answers (e.g., P12: "When I wrote it briefly, it said partially correct... it actually wanted longer answers").

Transparency: Most participants (f = 18) found the system's evaluations transparent, noting that explanatory feedback clarified the evaluation criteria (e.g., P2: "With the detailed explanation it gave at the bottom, it presented the correct answer with reasons, it was quite clear and understandable"). Some participants (f = 4) expressed concerns about the length of feedback, which occasionally caused confusion (e.g., P9: "The feedback was too detailed. I couldn't understand the scoring logic").

Fairness: Several participants (f = 9) perceived the system as fair, noting that it evaluated responses based on content and recognized different answer variations (e.g., P19: "I wrote the correct answer in different ways, and it said correct to those too"). However, some participants (f = 3) expressed concerns about the system's tendency to require detailed answers (e.g., P12: "It didn't seem to want only the critical information for the answer").

Consistency: Some participants (f = 5) noted that the system provided consistent evaluations for similar answers across different times (e.g., P24: "I tried similar answers at different times, and it scored them the same way, which gave me confidence"). No negative comments regarding consistency emerged.

Satisfaction with AI Feedback

Quantitative analysis of student satisfaction ratings revealed high satisfaction with AI-generated feedback. Students rated the quality of feedback using a 5-point star rating system after each response. A total of 260 ratings were collected across seven weeks of system use.

Table 4. Distribution of student satisfaction ratings (n = 260)

Rating	Count	Percentage
5 Stars (Excellent)	184	70.8%
4 Stars (Good)	25	9.6%
3 Stars (Moderate)	20	7.7%
2 Stars (Poor)	18	6.9%
1 Star (Very Poor)	13	5.0%

The majority of ratings (70.8%) were 5-star ratings, indicating strong satisfaction with feedback quality. Together, 4- and 5-star ratings accounted for 80.4% of all responses, suggesting that most students perceived the AI-generated feedback as good or excellent in quality. The overall mean rating was 4.34 out of 5.0 (SD = 1.18).

Analysis of weekly rating patterns revealed a trend. Although participation declined over the seven-week period, satisfaction ratings showed a steady upward trajectory. Mean ratings increased from 4.22 in Week 1 to 4.54 in Week 7, corresponding to a 7.6% improvement. This pattern suggests that students who remained engaged with the system developed increasingly positive evaluations of feedback quality over time.

Written comments associated with high ratings frequently emphasize the explanatory nature of the feedback, with students describing the system as clear and supportive of learning (e.g., "The feedback modules are very explanatory, a successful system for learning"; "It provided feedback in a very explanatory manner"). In contrast, lower ratings are typically associated with practical concerns related to feedback length or perceived scoring inconsistencies. Some comments indicate that feedback was perceived as overly long (e.g., "The feedback was a bit lengthy"), while others point to isolated scoring issues (e.g., "The scoring system made a mistake on the last question").

Influence of Sustained Interaction on Learning and Self-Regulation (RQ2)

Perceived Pedagogical Influence of the Automated Assessment System

The analysis revealed the system's perceived influence on various aspects of students' learning experiences. The findings are grouped into three main categories: learning processes, affective processes, and study strategies. Table 5 summarizes the pedagogical influence indicators.

Table 5. Perceived pedagogical influence of the automated assessment system

Category	Code	Positive	Negative	Key Findings
Learning Processes	Learning quality	18	4	Enhanced deeper learning and retention; some cognitive overload from lengthy feedback
	Self-assessment	15	-	Facilitated instant self-evaluation and gap identification
	Ease of learning	14	-	Accelerated error correction and deficiency identification
	Course success	12	-	Contributed to exam preparation and academic achievement
Affective Processes	Motivation	16	5	Increased motivation through immediate feedback and gamified scoring; reduced for some due to lengthy explanations or question difficulty.
	Self-confidence	7	-	Strengthened confidence through successful responses and validation of learning.
	Self-efficacy	5	1	Enhanced perception of competence and mastery; one case of reduced efficacy due to task complexity.
Study Strategies	Research behavior	11	-	Encouraged students to seek detailed information and engage in self-directed inquiry.
	Review strategy	6	-	Promoted review through screenshots and reuse of feedback for revision.
	Regular study	6	-	Fostered weekly study habits and consistent engagement with content.
	Exam preparation	5	-	Supported development of written exam strategies through exposure to realistic question types.

Learning Processes: The majority of participants (f = 18) reported that the system enhanced learning quality by supporting deeper and more durable learning, primarily through its detailed and constructive feedback (e.g., P24: "I really think that the constructive feedback was useful. I can say that I learned the topics better"). In addition, a substantial number of students (f = 15) emphasized that immediate feedback facilitated self-assessment by enabling them to identify their strengths and knowledge gaps in real time (e.g., P19: "By giving immediate responses, it allowed me to see myself. By seeing what I know and what I don't know, I realized the gaps I need to complete"). Many participants (f = 14) also reported that the system made learning easier by guiding error correction and supporting clearer understanding of expected responses (e.g., P17: "It tells you how to give a more correct answer. It helped me easily correct my deficiencies"), while some students (f = 12) perceived a positive contribution to their academic performance and exam preparation (e.g., P16: "I was able to prepare for the exam simply by using the system"). Nevertheless, a small group of participants (f = 4) noted that overly long explanations occasionally hindered learning by creating cognitive overload (e.g., P6: "It provided long explanations even for the correct answers, which was tiring. I felt confused").

Affective Processes: The majority of participants (f = 16) reported increased motivation due to instant feedback, scoring mechanisms, and gamification elements (e.g., P17: "The scoring mechanism in the system motivated me. I started doing more research to get higher scores"). However, some participants (f = 5) reported decreased motivation due to feedback length and question difficulty (e.g., P12: "I hadn't encountered such difficult questions before. I can't say it motivated me much"). Some participants reported increased self-confidence (f = 7) through successful performance (e.g., P13: "Because the questions were difficult and I could give correct answers to these questions... my self-confidence increased in other courses too") and

enhanced self-efficacy (f = 5) through mastery of content (e.g., P10: "Seeing what you know felt good. It helped me master the subjects better").

Study Strategies: Many participants (f = 11) reported that the system encouraged research behavior to provide more detailed answers (e.g., P11: "The existence of such a system makes us start researching much more... I did even more research to get ten points"). Some participants reported using the system's detailed explanations for review by taking screenshots (f = 6) (e.g., P24: "I also took screenshots to study again later"), developing regular study habits (f = 6) (e.g., P14: "I started studying more regularly... Now I understood the importance of studying at regular intervals"), and improving exam preparation strategies (f = 5) (e.g., P21: "I understood how I should prepare for the exam... I prepared based on the important points in the explanations given").

Engagement Patterns Over Time

System interaction logs revealed clear weekly patterns in student participation and satisfaction over the seven-week implementation period. Table 6 presents the longitudinal patterns.

Table 6. Weekly participation and satisfaction patterns (n = 47)

Week	Participants	Questions	Responses	Avg. Rating
1	38	4	431	4.22
2	43	2	165	4.13
3	37	3	197	4.31
4	37	3	175	4.00
5	37	3	128	4.43
6	36	2	81	4.59
7	29	2	70	4.54

The number of active participants remained relatively stable between Weeks 1 and 6, while a decline in participation was observed in the final week. The total number of student responses decreased steadily over time, from 431 responses in Week 1 to 70 responses in Week 7. Despite this decline in participation and response volume, average satisfaction ratings remained consistently high throughout the implementation. Mean ratings ranged from 4.00 to 4.59 on a five-point scale.

Descriptively, weekly average ratings fluctuated across the seven weeks but remained consistently high. The average rating was 4.22 in Week 1, decreased slightly to 4.13 in Week 2, and then ranged between 4.00 and 4.59 in the subsequent weeks. The final week remained high at 4.54.

Because the number of questions varied across weeks, raw response counts were interpreted cautiously. To account for this variation, a question-adjusted interaction rate was calculated by dividing each student's weekly interaction count by the number of questions administered that week. In addition, because the data involved unbalanced repeated measures due to participant attrition, longitudinal change was examined using a linear mixed-effects model with week as a fixed effect and student as a random effect.

The analysis revealed a statistically significant decrease in adjusted interaction rate over time

(β = -0.180, z = -4.68, p < .001). A separate linear mixed-effects model based on rated responses only showed no significant weekly change in satisfaction ratings (β = 0.058, z = 1.39, p = .165). This suggests that although students interacted with the system less frequently per available question over time, their evaluations of the AI-generated feedback remained consistently positive across the seven-week period. In addition, no significant association was observed between students' overall adjusted interaction rate and their overall satisfaction (Spearman's ρ = 0.071, p = .735).

Discussion

This study examined students' ongoing experiences with an LLM-based automated assessment system embedded as an active component of an authentic learning environment. Unlike previous studies that primarily evaluated automated assessment systems based on technical performance metrics (Filighera et al., 2022; Latif & Zhai, 2024), this study focused on students' sustained experiences. By integrating qualitative student perspectives with descriptive log data, the findings provide a nuanced understanding of how automated scoring and explanatory feedback are experienced in practice.

Making Sense of Automated Assessment: Beyond the Black Box

One of the key findings of this study is that students did not perceive the LLM-based assessment system as a "black box" or purely technical grading mechanism. Instead, many participants interpreted the system as a hybrid pedagogical tool that combines evaluative judgment with instructional support. Students frequently described the system as a learning assistant, impartial evaluator, self-assessment tool, or teacher-like entity. This is consistent with earlier research describing feedback as a relational and interpretive process, rather than a one-way transmission of information (Carless & Boud, 2018; Shute, 2008). This differs from the concerns reported by Ruwe and Kuklick (2026), who found that students associated algorithmic grading with potential bias and limited nuance.

The critical mediating factor appeared to be the transparency of explanations. Students trusted the feedback because the system provided detailed, rubric-aligned explanations alongside scores. This is consistent with research showing that perceived transparency supports trust in AI systems (Shin, 2021) and with the Explainable Artificial Intelligence in Education (XAI-ED) framework, which positions explainability in educational AI as a means of supporting learning and metacognitive engagement rather than merely justifying algorithmic decisions (Khosravi et al., 2022). The results also suggest that explanatory feedback helped students compare their current responses with expected standards and use that comparison for revision. In this sense, the system appears to have supported a more active student role in the feedback process, consistent with work on sense-making and internal feedback generation (Leighton, 2019; Nicol, 2021; Panadero & Lipnevich, 2022).

Trust Through Explanation: Accuracy, Transparency, and Fairness

The results show that trust in the system is not unconditional. Across the dimensions of accuracy, transparency, and fairness, students' accounts revealed both strengths and tensions. Most participants perceived the system's assessments as accurate, transparent, and consistent; students valued detailed explanations that "presented the correct answer along with its reasoning," and consistent scores across different attempts reinforced confidence in the system's reliability. However, students' perceptions of fairness, particularly when concise but correct answers are involved, revealed a recurring tension between algorithmic

consistency and pedagogical flexibility. Some students expressed concern that short but correct answers received lower scores, suggesting that the system may have rewarded elaboration over concision.

These concerns have been highlighted in research on algorithmic assessment and fairness. Holmes et al. (2022) identify the accuracy of AI-based judgments about students as a central ethical concern in educational settings. Similarly, Kizilcec and Lee (2022) emphasize that algorithmic systems may produce discriminatory outcomes, while Williamson et al. (2012) note that rigid adherence to scoring criteria can disadvantage concise or stylistically different responses. Kundu and Barbosa (2024) also reported that LLMs tend to assign lower scores than human raters. Similarly, Gao et al. (2024) note that while LLMs show strong correlations with human evaluators on objective metrics, they struggle with nuanced responses. However, the results reveal that rather than rejecting the system entirely, students expressed their concerns in a reflective and critical manner, demonstrating a form of assessment literacy emerging through interaction with AI-based systems. This supports the argument that continuous exposure, rather than one-time use, is necessary for students to meaningfully interpret and evaluate automated assessment tools.

A further tension concerned the length and usability of explanations. For some students, detailed feedback increased clarity and confidence; for others, it became cognitively demanding. In this respect, the results align with cognitive load theory (Sweller, 2011) and with research emphasizing that learners' informational needs vary and that feedback design involves choices about type, detail, and length (Lohr et al., 2025). AI-supported feedback systems may therefore benefit from more adaptive explanation formats, such as shorter initial feedback with the option to access additional detail when needed.

Sustained Interaction and Behavioral Change

The results of the quantitative analysis of interaction logs revealed a significant pattern: while the frequency of AI queries decreased over the seven-week period, satisfaction ratings remained consistently high. Rather than indicating indifference, this pattern may reflect a gradual normalization of the feedback process. In the early weeks, students appeared to engage in repeated trial-and-error attempts, using the system frequently to test, revise, and refine their responses. This may be interpreted as a concrete reflection of Kapur's (2016) concept of "productive failure". This environment, free from social judgment pressure, created a low-risk space where making mistakes contributed to learning. Over time, this pattern seemed to become more selective and strategic, suggesting that students were becoming more familiar with the feedback process and the assessment criteria.

Taken together, these patterns can be interpreted as indicators of self-regulated learning development. Students' reports of instant self-assessment, increased awareness of knowledge gaps, and more purposeful use of feedback suggest that the system supported self-monitoring and revision rather than passive reception of scores. In this sense, Öztürk and Çebi (2025) also highlight that sustained AI-supported feedback may contribute to the development of self-monitoring and learning regulation over time. This aligns with feedback research that identifies self-regulation as a central component of learning (Hattie & Timperley, 2007; Nicol & Macfarlane-Dick, 2006; Zimmerman, 2002).

The Instructor's Role: Complementarity, Not Substitution

The results show that some students explicitly compare the system to their teachers and emphasize teacher-like interaction. This perception suggests that the system was experienced as pedagogically aligned with course expectations rather than as a detached automated tool. This raises important questions about pedagogical dynamics in AI-supported assessment. While Xavier et al. (2025) argue that AI can support teachers by reducing routine workload, Braun et al. (2023) emphasize that human intervention remains necessary to prevent pedagogical drift.

In the present study, these perspectives appeared to converge. Although the system operated autonomously at the point of scoring and feedback, it remained tightly connected to the teacher's assessment criteria and pedagogical framing. In this sense, the teacher was not removed from the loop; rather, the system appeared to extend the teacher's pedagogical intentions into the feedback process. While the instructor could focus on broader instructional design, higher-order learning goals, and more complex student needs, the AI system provided immediate micro-level formative support. This interpretation is also consistent with Tang et al. (2025), who reported that generative AI increased satisfaction but did not improve knowledge mastery to the same extent without teacher supervision. Taken together, these results support a complementary model in which AI provides scalable formative support, while the teacher remains central in shaping pedagogical goals, assessment criteria, and higher-level instructional guidance.

Theoretical and Practical Implications

The results underscore the importance of examining AI-based assessment systems as lived pedagogical experiences. Students' voices revealed nuances, such as trust negotiation, perceived fairness, and adaptive study behaviors, that would remain invisible in performance-only evaluations. These findings have both theoretical and practical implications.

Theoretically, the results offer four key contributions. First, they suggest that explanatory transparency is an important factor in building trust in algorithmic systems. Second, they indicate that AI feedback can support self-regulated learning processes. Third, they suggest that AI may create conditions conducive to "productive failure" by establishing a low-risk learning environment free from perceived social evaluation pressure. Fourth, they demonstrate that AI can be positioned as a tool that scales pedagogical intent without replacing the teacher.

In practice, the findings suggest several recommendations for designers and educators: (a) explanatory feedback mechanisms should be used to support trust, (b) layered or adaptive explanation options should be provided to balance cognitive load, (c) assessment criteria, especially scoring weights, should be shared transparently to improve students' perceptions, (d) low-risk environments for productive failure should be designed by allowing penalty-free repeated attempts, (e) AI systems may be more effective when educators remain involved in the instructional context, and (f) feedback should be calibrated to students' cognitive capacities.

Limitations

This study has several limitations. First, the sample was drawn from a single higher education course, limiting generalizability to other disciplinary settings. Second, the qualitative data relied on self-reported perceptions, which may be subject to social desirability

bias. Third, the study did not include a control or comparison group. As a result, the influence of the AI-supported assessment system cannot be clearly separated from other instructional factors, such as course progression, repeated task exposure, or students' increasing familiarity with the assessment process. Although the purpose of the study was to examine student experiences rather than to test causal effects, this remains an important design limitation. Future research should address this issue through experimental or quasi-experimental designs that compare AI-supported feedback with instructor-only or hybrid feedback conditions. Fourth, the study focused on student experience rather than objective or measured learning outcomes, and future research should examine whether positive perceptions translate into measurable achievement gains.

Conclusion

This study provides empirical evidence on how LLM-based automated assessment systems are experienced in authentic classroom settings beyond controlled laboratory environments. The results suggest that the transparency of explanations plays a critical role in building trust, instant feedback supports the development of self-regulated learning, and human–AI complementarity enhances pedagogical effectiveness. Ultimately, this study suggests that artificial intelligence can be positioned not as a substitute for teachers but as a strategic partner that extends pedagogical support and facilitates students' learning processes.

Declarations

Acknowledgments: *The authors would like to thank Atatürk University Institute of Educational Sciences and Atatürk University Scientific Research Projects Coordination Unit for their support.*
Funding: *The author(s) received no financial support for the research, authorship, and/or publication of this article.*
Ethics Statements: *This study was approved by the Social and Human Sciences Ethics Committee, Educational Sciences Unit Ethics Committee, with the decision dated 03 February 2022 and numbered 02/04.*
Conflict of Interest: *The authors declare no potential conflicts of interest.*
Informed Consent: *All participants provided informed consent prior to participation.*
Data availability: *The dataset used and analyzed in this study has been made publicly available on Kaggle under the title "Turkish History Education ASAG dataset" (https://www.kaggle.com/datasets/kadirkara2013/turkish-historyeducation-asag-dataset/).*
Author Note: *This article was derived from the first author's PhD thesis titled "Automatic Evaluation of Open-Ended Questions with Artificial Intelligence and User Experiences."*

References

Abdul-Salam, M., El-Fatah, M. A., & Hassan, N. F. (2022). Automatic grading for Arabic short answer questions using optimized deep learning model. *Plos one, 17*(8), e0272269. https://doi.org/10.1371/journal.pone.0272269

Ahmad, K., Iqbal, W., El-Hassan, A., Qadir, J., Benhaddou, D., Ayyash, M., & Al-Fuqaha, A. (2023). Data-driven artificial intelligence in education: A comprehensive review. *IEEE Transactions on Learning Technologies*, *17*, 12-31. https://doi.org/10.1109/TLT.2023.3314610

Ariely, M., Nazaretsky, T., & Alexandron, G. (2022). Personalized automated formative feedback can support students in generating causal explanations in biology. In *Proceedings of the 16th International Conference of the Learning Sciences-ICLS 2022, pp. 953-956*. International Society of the Learning Sciences.

Boud, D., & Molloy, E. (2013). Rethinking models of feedback for learning: The challenge of design. *Assessment & Evaluation in Higher Education*, *38*(6), 698-712. https://doi.org/10.1080/02602938.2012.691462

Braun, D., Rogetzer, P., Stoica, E., & Kurzhals, H. (2023, April). Students' perspective on AI-supported assessment of open-ended questions in higher education. In *15th International Conference on Computer Supported Education, CSEDU 2023* (pp. 73-79).

Carless, D., & Boud, D. (2018). The development of student feedback literacy: Enabling uptake of feedback. *Assessment & Evaluation in Higher Education*, *43*(8), 1315-1325. https://doi.org/10.1080/02602938.2018.1463354

Chang, L. H., & Ginter, F. (2024, March). Automatic short answer grading for Finnish with ChatGPT. In *Proceedings of the AAAI Conference on Artificial Intelligence* (Vol. 38, No. 21, pp. 23173-23181).

Cheong, M. (2025). ChatGPT's performance evaluation in spreadsheets modelling to inform assessments redesign. *Journal of Computer Assisted Learning*, *41*(3), e70035. https://doi.org/10.1111/jcal.70035

Chiang, C. H., Chen, W. C., Kuan, C. Y., Yang, C., & Lee, H. Y. (2024). Large language model as an assignment evaluator: Insights, feedback, and challenges in a 1000+ student course. *arXiv preprint arXiv:2407.05216*.

Çınar, A., Ince, E., Gezer, M., & Yilmaz, O. (2020). Machine learning algorithm for grading open-ended physics questions in Turkish. *Education and Information Technologies, 25*(5), 3821-3844. https://doi.org/10.1007/s10639-020-10128-0

De Wet, M., Da Silva, M. O., & Bohnsack, R. (2025). Can AI give good feedback on essay-type assignments? An explorative case study of LLMs in higher education. *Innovations in Education and Teaching International*, *62*(5), 1484-1499. https://doi.org/10.1080/14703297.2025.2536609

Delalibera, D. C. A. R., & Carvalho, D. F. (2024). Emissão de feedbacks pelos docentes no processo de ensino e aprendizagem no curso de graduação em Direito [Emission feedbacks by teachers in the teaching and learning process in the undergraduate Law course]. *Inter-Ação* [*Inter Action*], *49*(1), 104–120. https://doi.org/10.5216/ia.v49i1.75976.

Er, E., Akçapınar, G., Bayazıt, A., Noroozi, O., & Banihashem, S. K. (2025). Assessing student perceptions and use of instructor versus AI-generated feedback. *British Journal of Educational Technology*, *56*(3), 1074-1091. https://doi.org/10.1111/bjet.13558

Escalante, J., Pack, A., & Barrett, A. (2023). AI-generated feedback on writing: Insights into efficacy and ENL student preference. *International Journal of Educational Technology in Higher Education*, *20*(1), 57. https://doi.org/10.1186/s41239-023-00425-2

Filighera, A., Parihar, S., Steuer, T., Meuser, T., & Ochs, S. (2022, May). Your answer is incorrect... would you like to know why? introducing a bilingual short answer feedback dataset. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* (pp. 8577-8591).

Gao, R., Guo, X., Li, X., Narayanan, A. B. L., Thomas, N., & Srinivasa, A. R. (2024). Towards scalable automated grading: Leveraging large language models for conceptual question evaluation in engineering. *arXiv preprint arXiv:2411.03659*.

Ghavidel, H. A., Zouaq, A., & Desmarais, M. C. (2020). Using BERT and XLNET for the automatic short answer grading task. In *CSEDU (1)* (pp. 58-67). https://doi.org/10.5220/0009422400580067

Gombert, S., Di Mitri, D., Karademir, O., Kubsch, M., Kolbe, H., Tautz, S., ... & Drachsler, H. (2023). Coding energy knowledge in constructed responses with explainable NLP models. *Journal of Computer Assisted Learning, 39*(3), 767-786. https://doi.org/10.1111/jcal.12767

Hao, J., von Davier, A. A., Yaneva, V., Lottridge, S., von Davier, M., & Harris, D. J. (2024). Transforming assessment: The impacts and implications of large language models and generative AI. *Educational Measurement: Issues and Practice*, *43*(2), 16-29. https://doi.org/10.1111/emip.12602

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research, 77*(1), 81-112. https://doi.org/10.3102/003465430298487

Holmes, W., Porayska-Pomsta, K., Holstein, K., Sutherland, E., Baker, T., Shum, S. B., ... & Koedinger, K. R. (2022). Ethics of AI in education: Towards a community-wide framework. *International Journal of Artificial Intelligence in Education, 32*(3), 504-526. https://doi.org/10.1007/s40593-021-00239-1

Kane, M. T. (2006). Validation. In R. L. Brennan (Ed.), *Educational Measurement* (4th ed., pp. 17–64). American Council on Education / Praeger.

Kapur, M. (2016). Examining productive failure, productive success, unproductive failure, and unproductive success in learning. *Educational Psychologist*, *51*(2), 289-299. https://doi.org/10.1080/00461520.2016.1155457

Khosravi, H., Shum, S. B., Chen, G., Conati, C., Tsai, Y. S., Kay, J., ... & Gašević, D. (2022). Explainable artificial intelligence in education. *Computers and Education: Artificial intelligence*, *3*, 100074. https://doi.org/10.1016/j.caeai.2022.100074

Kizilcec, R. F., & Lee, H. (2022). Algorithmic fairness in education. In *The ethics of artificial intelligence in education* (pp. 174-202). Routledge.

Kortemeyer, G. (2024). Performance of the pre-trained large language model GPT-4 on automated short answer grading. *Discover Artificial Intelligence*, *4*(1), 47. https://doi.org/10.1007/s44163-024-00147-y

Kundu, A., & Barbosa, D. (2024). Are large language models good essay graders?. arXiv preprint arXiv:2409.13120.

Kurbanoğlu, N. I., & Olcaytürk, M. (2023). Investigation of the exam question types attitude scale for secondary school students: Development, validity, and reliability. *Sakarya University Journal of Education, 13*(2), 191-206. https://doi.org/10.19126/suje.1187470

Latif, E., & Zhai, X. (2024). Fine-tuning ChatGPT for automatic scoring. *Computers and Education: Artificial Intelligence*, *6*, 100210. https://doi.org/10.1016/j.caeai.2024.100210

Laurillard, D. (2009). The pedagogical challenges to collaborative technologies. *International Journal of Computer-supported Collaborative Learning*, *4*(1), 5-20. https://doi.org/10.1007/s11412-008-9056-2

Leighton, J. P. (2019). Students' interpretation of formative assessment feedback: Three claims for why we know so little about something so important. *Journal of Educational Measurement*, *56*(4), 793-814.

Lohr, D., Keuning, H., & Kiesler, N. (2025). You're (not) my type-can LLMs generate feedback of specific types for introductory programming tasks?. *Journal of Computer Assisted Learning*, *41*(1), e13107. https://doi.org/10.1111/jcal.13107

Luo, J., Zheng, C., Yin, J., & Teo, H. H. (2025). Design and assessment of AI-based learning tools in higher education: A systematic review. *International Journal of Educational*

Technology in Higher Education, *22*(1), 42. https://doi.org/10.1186/s41239-025-00540-2

Mardini G, I. D., Quintero M, C. G., Viloria N, C. A., Percybrooks B, W. S., Robles N, H. S., & Villalba R, K. (2024). A deep-learning-based grading system (ASAG) for reading comprehension assessment by using aphorisms as open-answer-questions. *Education and Information Technologies, 29*(4), 4565-4590. https://doi.org/10.1007/s10639-023-11890-7

Mendonça, P. C., Quintal, F., & Mendonça, F. (2025). Evaluating LLMs for automated scoring in formative assessments. *Applied Sciences*, *15*(5), 2787. https://doi.org/10.3390/app15052787

Moore, M. G. (1989). Three types of interaction. *American Journal of Distance Education*, *3*(2), 1-7. https://doi.org/10.1080/08923648909526659

Nath, S., Parsaeifard, B., & Werlen, E. (2023) Automated short answer grading using BERT on German datasets. The *20th biennial EARLI Conference (EARLI 2023).* https://www.researchgate.net/publication/373556564

Nicol, D. (2021). The power of internal feedback: Exploiting natural comparison processes. *Assessment & Evaluation in Higher Education*, *46*(5), 756-778. https://doi.org/10.1080/02602938.2020.1823314

Nicol, D. J., & Macfarlane-Dick, D. (2006). Formative assessment and self-regulated learning: A model and seven principles of good feedback practice. *Studies in Higher Education*, *31*(2), 199-218. https://doi.org/10.1080/03075070600572090

Öztürk, Y., & Çebi, A. (2025). The potential of AI-generated feedback from the students' perspective: A systematic review. *Assessment & Evaluation in Higher Education*, 1-17. https://doi.org/10.1080/02602938.2025.2588385

Panadero, E., & Lipnevich, A. A. (2022). A review of feedback models and typologies: Towards an integrative model of feedback elements. *Educational Research Review*, *35*, 100416. https://doi.org/10.1016/j.edurev.2021.100416

Ruwe, T., & Kuklick, L. (2026). Quality counts? Examining the role of feedback provider and feedback quality on students' feedback perceptions. *British Journal of Educational Technology*, *57*(1), 272-298. https://doi.org/10.1111/bjet.70011

Sawatzki, J., Schlippe, T., & Benner-Wickner, M. (2021). Deep learning techniques for automatic short answer grading: Predicting scores for English and German answers. In *International conference on artificial intelligence in education technology* (pp. 65-75). Singapore: Springer Nature Singapore. https://doi.org/10.1007/978-981-16-7527-0_5

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies*, *146*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Shute, V. J. (2008). Focus on formative feedback. *Review of Educational Research, 78*(1), 153-189. https://doi.org/10.3102/0034654307313795

Singerin, S., Yafie, E., Nugroho, A., Pratiwi, A. P., Krobo, A., & Marhadi, N. (2025). Higher education transformation through AI-based learning innovation: Faculty members' perception, challenges, and adoption in teaching and assessment. *Participatory Educational Research*, *12*(6), 280-299. https://doi.org/10.17275/per.25.90.12.6

Sung, C., Dhamecha, T., Saha, S., Ma, T., Reddy, V., & Arora, R. (2019, November). Pre-training BERT on domain resources for short answer grading. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)* (pp. 6071-6075).

Sweller, J. (2011). Cognitive load theory. In *Psychology of learning and motivation* (Vol. 55, pp. 37-76). Academic Press.

Tang, Q., Deng, W., Huang, Y., Wang, S., & Zhang, H. (2025). Can generative artificial intelligence be a good teaching assistant?—An empirical analysis based on generative AI-assisted teaching. *Journal of Computer Assisted Learning*, *41*(3), e70027. https://doi.org/10.1111/jcal.70027

Taub, M., Azevedo, R., Rajendran, R., Cloude, E. B., Biswas, G., & Price, M. J. (2021). How are students' emotions related to the accuracy of cognitive and metacognitive processes during learning with an intelligent tutoring system?. *Learning and Instruction*, *72*, 101200. https://doi.org/10.1016/j.learninstruc.2019.04.001

Tossell, C. C., Tenhundfeld, N. L., Momen, A., Cooley, K., & de Visser, E. J. (2024). Student perceptions of ChatGPT use in a college essay assignment: Implications for learning, grading, and trust in artificial intelligence. *IEEE Transactions on Learning Technologies*, *17*, 1069-1081. https://doi.org/10.1109/TLT.2024.3355015

Uto, M., & Uchida, Y. (2020). Automated short-answer grading using deep neural networks and item response theory. In *Artificial Intelligence in Education: 21st International Conference, AIED 2020, Ifrane, Morocco, July 6–10, 2020, Proceedings, Part II 21 (pp. 334-339).* Springer International Publishing. https://doi.org/10.1007/978-3-030-52240-7_61

Westera, W., Dascalu, M., Kurvers, H., Ruseti, S., & Trausan-Matu, S. (2018). Automated essay scoring in applied games: Reducing the teacher bandwidth problem in online training. *Computers & Education, 123*, 212-224. https://doi.org/10.1016/j.compedu.2018.05.010

Williamson, D. M., Xi, X., & Breyer, F. J. (2012). A framework for evaluation and use of automated scoring. *Educational measurement: Issues and practice*, *31*(1), 2-13. https://doi.org/10.1111/j.1745-3992.2011.00223.x

Xavier, C., da Costa, N. T., Valdo, A. K., Alves, G., Rodrigues, L., Rodrigues, L. F., ... & Mello, R. F. (2025, September). Human teacher vs. LLM-Generated feedback in secondary education: A Comparative Study on Student Perceptions. In *European Conference on Technology Enhanced Learning* (pp. 534-548). Cham: Springer Nature Switzerland.

Xavier, C., Rodrigues, L., Costa, N., Neto, R., Alves, G., Falcão, T. P., ... & Mello, R. F. (2025). Empowering instructors with AI: Evaluating the impact of an AI-driven feedback tool in learning analytics. *IEEE Transactions on Learning Technologies*. https://doi.org/10.1109/TLT.2025.3562379

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), 1-27. https://doi.org/10.1186/s41239-019-0171-0

Zhu, X., Wu, H., & Zhang, L. (2022). Automatic short-answer grading via BERT-based deep neural networks. *IEEE Transactions on Learning Technologies, 15*(3), 364-375. https://doi.org/10.1109/tlt.2022.3175537

Zimmerman, B. J. (2002). Becoming a self-regulated learner: An overview. Theory into Practice, 41(2), 64-70. https://doi.org/10.1207/s15430421tip4102_2