Comparing Student Preferences for AI-Generated and Peer-Generated Feedback in AI-driven Formative Peer Assessment

Insub Shin*
Seoul National University
Seoul, Republic of Korea
iss9802@snu.ac.kr

Su Bhin Hwang*
Seoul National University
Seoul, Republic of Korea
sbhwangt@snu.ac.kr

Yun Joo Yoo
Seoul National University
Seoul, Republic of Korea
yyoo@snu.ac.kr

Sooan Bae
Seoul National University
Seoul, Republic of Korea
snubbb@snu.ac.kr

Rae Yeong Kim
Seoul National University
Seoul, Republic of Korea
pianist9430@snu.ac.kr

Abstract

Formative assessment can enhance student learning and improve teaching practices by identifying areas for growth and providing feedback. However, practical obstacles remain, such as time constraints and students' passive participation and the low quality of peer feedback. Artificial intelligence (AI) has been explored for its potential to automate grading and provide timely feedback, making it a valuable tool in formative assessment. Nevertheless, there is still limited research on how AI can be used effectively in the context of formative peer assessment. In this study, we conducted an AI-driven formative peer assessment with 108 secondary school students. During the peer assessment process, students not only evaluated peers' responses and received peer-generated feedback, but also evaluated AI-generated responses and received AI-generated feedback. This research focused on analyzing the differences in preference between AI-generated and peer-generated feedback using trace data and dispositional data. In scenarios where student participation was low or the quality of peer feedback was insufficient, students showed a higher preference for AI-generated feedback, demonstrating its potential utility. However, students with high *Math Confidence* and *AI Interest* preferred peer-generated feedback. Based on these findings, we will propose practical strategies for implementing AI-driven formative peer assessment.

CCS Concepts

• Human-centered computing → Human computer interaction (HCI); **• Applied computing → Interactive learning environments**; **• Computing methodologies → Artificial Intelligence**.

*Both authors contributed equally to this research.

LAK 2025, Dublin, Ireland

ACM ISBN 979-8-4007-0701-8/25/03
https://doi.org/10.1145/3706468.3706488

Keywords

Formative Peer Assessment, Feedback, Generative Artificial Intelligence, Trace Data, Dispositional Data

ACM Reference Format:
Insub Shin, Su Bhin Hwang, Yun Joo Yoo, Sooan Bae, and Rae Yeong Kim. 2025. Comparing Student Preferences for AI-Generated and Peer-Generated Feedback in AI-driven Formative Peer Assessment. In *LAK25: The 15th International Learning Analytics and Knowledge Conference (LAK 2025), March 03–07, 2025, Dublin, Ireland.* ACM, New York, NY, USA, 11 pages. https://doi.org/10.1145/3706468.3706488

1 Introduction

Formative assessment helps students enhance learning and teachers adjust teaching by providing feedback based on the information gathered during the learning process. Feedback in formative assessment is multifaceted, originating from various sources such as teachers, peers, textbooks, and parents, thereby creating a support system for student learning and comprehension [17]. Among these, peer feedback plays a crucial role, as it allows students to engage simultaneously as evaluators and learners. This dual role fosters a collaborative learning environment where learning is enhanced through mutual assessment and the development of critical thinking skills [12, 48]. Additionally, the process of formative peer assessment allows students to deepen their learning [31, 49], develop a clearer understanding of assessment requirements, and enhance the quality of their work [44].

However, the effective implementation of formative peer assessment faces challenges: constraints of time, low participation, and poor quality feedback [30]. Given the limited school hours, managing time-consuming formative peer assessments can be arduous task for teachers. Additionally, low participation malfunctions activities within formative peer assessment, for example, insufficient engagement often leads to inadequate discussions between students [30]. Furthermore, poor quality feedback can hinder rather than enhance learning through feedback.

Artificial Intelligence (AI) can be leveraged to address these challenges. The development of AI, especially large language models like GPT, has accelerated the integration of AI into education in various contexts, such as tutoring, assessment, and personalized learning. In formative assessment, AI can evaluate students' responses and provide feedback based on the evaluation [15, 18, 43, 50]. The greatest advantage of using AI is its assistance to teachers in conducting assessments within limited time constraints and the provision of timely and valid feedback.

To effectively integrate AI into education, it is essential to focus not only on the technological capabilities of AI but also on its educational impact. However, students' perceptions of feedback received from AI in formative peer assessments have not been studied well. Therefore, the aim of this study is to analyze students' preferences for AI-generated feedback in formative peer assessment, including differences in preference based on classroom environments and students' disposition.

In this paper, we conducted AI-driven formative peer assessments in four distinct math classes, involving a total of 108 students. In the context of AI-driven formative peer assessment, AI also took on the role of a peer in the formative assessment process. Specifically, while students evaluated their peers' responses and received feedback from them, they also evaluated AI-generated responses and received feedback from AI. Since the assessment was conducted anonymously, students could not determine who provided the feedback, whether it was from AI or a peer. Each class carried out 2 to 3 formative peer assessments, resulting in the collection of 856 feedback consisting of AI-generated and peer-generated. Trace data collected systematically during the process and dispositional data from self-report surveys were used to analyze the preference gap between AI-generated and peer-generated feedback and to construct a preference prediction model. Through this research, we aim to propose practical recommendations for implementing formative peer assessment using AI, grounded in Learning Analytics. To this end, our goal is to answer the following research questions (RQ):

- **(RQ1)** How do preferences for AI-generated feedback and peer-generated feedback relate to classroom environment?
- **(RQ2)** How do preferences for AI-generated feedback and peer-generated feedback relate to dispositions?
- **(RQ3)** What recommendations can be made for the implementation of AI-driven formative peer assessment based on analysis using trace data and dispositional data?

2 BACKGROUND and RELATED WORK

2.1 Formative Peer Assessment

Formative evaluation, first introduced by Scriven [34], focused on curriculum development through information gathered during a course, in contrast to summative evaluation, which assessed overall outcomes of course [4, 7, 25, 48]. Bloom [9] used the same term but conceptualized the distinction between formative and summative in relation to student learning [4, 7, 48]. Similarly, Sadler [33] described this distinction in terms of purpose and effect, noting that the purpose of formative assessment was to leverage judgements of student performance to enhance competency. In the current literature, formative assessment encompasses all activities conducted by teachers, learners, and peers, from which evidence is elicited, interpreted, and used as feedback to adjust their instruction and learning [6, 8].

From the aforementioned definition of formative assessment, peers, like teachers and learners, are also considered agents of the assessment, which implies that their decision is included in formative assessment [8]. Activities such as assessing, reviewing, and providing feedback based on the responses of peers can be included in formative assessment, specifically within the context of formative peer assessment. Formative peer assessment offers the advantage of enhancing learning through the experiences of both the assessor and the assessee. For instance, Prins et al. [30] reported that students highlighted that receiving feedback is valuable because the provider has undergone the same learning process, and that giving feedback helps develop clarity and structure in writing. In [31, 42, 49], the potential advantages of peer assessment are described as follows:

- Giving a sense of autonomy and ownership of the assessment process and improving motivation.
- Encouraging students to take responsibility for their own learning and development.
- Treating assessment as part of learning so that mistakes are seen as opportunities rather than failures.
- Practicing the transferable skills needed for life-long learning particularly related to evaluation skills.
- Using external evaluation to provide a model for internal self-assessment of a student's own learning (metacognition).
- Encouraging deep rather than surface learning.

However, formative peer assessment faces difficulties in implementation. For example, negative attitudes and opposition to peer assessment tackles successful implementation [21, 41, 44]. Monitoring every student in real time during formative peer assessments is both challenging and time-consuming for a teacher. Prins et al. [30] demonstrated that, during formative peer assessment, the discussion of assessment criteria was hampered by low participation, which stemmed from students' lack of experience. Also, students may perceive peer assessment as an act of helping peers who are seen as competitors, which can lead to difficulties in participation [2]. Additionally, a lack of experience with peer assessment can lead to poor quality feedback. In [30], students reported having limited experience with receiving critical feedback from peers, even though criteria and feedback rules were provided. Peer assessments are often considered inaccurate and unreliable, based on the assumption that students are not only inexperienced in the discipline but also unfamiliar with academic practices in general [2]. To ensure the successful implementation of formative peer assessment, it is essential to address challenges.

2.2 AI in Formative Assessment

Integrating AI into education is pervasive in addressing educational challenges. AI supports both teachers and students across various contexts, contributing to the reform of traditional teaching methods. Especially, the text generation capability of AI is a significant asset for providing text-based feedback. For instance, AI-driven feedback systems have assisted teachers in creating learning designs [29], AI has collaborated with humans in tutoring [40], and AI has generated human tutor-style hints in programming education [28]. The applicability of AI in feedback practices is growing, without restriction to specific contexts or disciplines.

In the realm of formative assessment, AI is effectively utilized for tasks such as automatic grading and providing feedback [50]. For example, Vittorini, Menini, and Tonelli [43] developed an AI-based system for formative and summative assessments in a Data Science Course. The system used AI to evaluate assignments composed of R commands and provided feedback in natural language [43]. They found that the AI-based assessment system reduced grading time, minimized errors, and automatically provided feedback to support students. Additionally, students did not report any usability issues [43]. Zhai and Nehm [50] proposed that AI is already extensively utilized in formative assessment across diverse educational settings, so that the focus should shift from questioning whether AI *can* or *should* be used to exploring *how* to effectively integrate AI to enhance formative assessment practices. Despite this, there is still limited research on *how* to apply AI in formative peer assessment. This study aims to investigate how AI-generated feedback can be effectively utilized within formative peer assessment processes.

There are concerns about applying AI to formative assessment. According to Li et al. [23], formative assessment should consider students' emergent sensemaking practices, which AI may struggle. Therefore, Li et al. [23] suggests for researching human-in-the-loop approaches to address these challenges. In this study, AI-driven formative peer assessment did not rely solely on AI-generated feedback but utilized both AI-generated feedback and peer (human)-generated feedback. Investigating the differences between AI-generated and peer-generated feedback in this context will contribute to understanding the complementary role of AI and human in formative assessment.

2.3 Learning Analytics in Formative Assessment

Learning Analytics refers to the data-driven process of measuring, gathering, analyzing, and reporting on learners and educational contexts to better understand and enhance both learning and the environments where it takes place [22, 32]. Learning Analytics combines the technical and the social/pedagogical aspects of learning by utilizing predictive models inferred from data [32]. Through Learning Analytics, it is possible to predict student performance [16, 26], identify at-risk students [14], and recommend feedback strategies [20].

Formative assessments in face-to-face courses have limitations in capturing and analyzing all learning interactions and outcomes [13]. However, advancements in technology have enabled the capture and analysis of performance and assessment data [5], leading to the application of learning analytics in formative assessments. Data derived from continuous learner activities in formative assessment is suitable for Learning Analytics [38]. Andriamiseza et al. [3] demonstrated the effectiveness of an interactive voting system in face-to-face classrooms to support formative peer assessment by analyzing consecutive data such as votes, confidence levels, and rationales from a formative peer assessment process. Therefore, Learning Analytics can be used to design formative assessment practices using data obtained from the formative assessment process.

While early educational research relied on observations, self-report surveys, or think-aloud protocols to develop theories, recent Learning Analytics research employs systematically collected data from platforms [35]. Dispositional Learning Analytics merges traditional educational research methods with Learning Analytics techniques by combining learning data from Learning Management Systems with learner data, such as student dispositions, values, and attitudes collected through self-report surveys [10]. For example, previous researches [27, 36, 37] encompasses a range of data types, including trace data (e.g., total number of attempts, number of worked examples, correct/incorrect attempts, proportion of correct answers, time on task), dispositional data (e.g., attitudes, learning emotions, epistemic emotions, goal setting, help-seeking behaviors, motivation, self-regulation), e-tutorial trace data, and formative data for Dispositional Learning Analytics. By leveraging the diverse data mentioned above, at-risk groups can be characterized using dispositional data [36], different profiles of learning strategies can be compared [37], and indirect path between learning dispositions and academic performance through feedback preferences can be analyzed [27].

Disposition Learning Analytics have the potential to stimulate students' intrinsic motivation and lead to changes in learning [10]. Therefore, in this study, we will investigate practices for implementation by leveraging both the trace data collected from AI-driven formative peer assessment and students' dispositional data.

3 METHOD

3.1 AI-Driven Formative Peer Assessment Process

Technology-enhanced formative assessment process consists of four stages, with interactions occurring among four key actors: teachers, students, peers, and technology, throughout each stage [12]. At the ask stage, a question is elicited by the teacher, peers, or technology, marking the beginning of the formative assessment process. In the answer stage, the student provides a complete or partial response to the question, which offers information about their learning. During the analyze stage, the submitted response is reviewed by the teacher, peers, and technology, and these three actors interpret the response and provide feedback. The final stage, the adapt stage, uses the results from the previous stages to optimize teaching and learning. We structured the AI-driven formative peer assessment process into four stages as outlined by [12]. Details are provided in the following subsections. In this process, AI participates as an anonymous peer.

3.1.1 Ask and Answer Stage. A problem related to the concepts covered earlier in the lesson is posed. Students are asked to submit their responses in two parts: the solution, which includes a detailed explanation of the mathematical process, and the answer.

3.1.2 Analyze Stage. In the Analyze stage, the main feedback procedures take place, where students, peers, and AI provide feedback to each other. AI classifies each student's response according to the mathematical process outlined in the solution submitted from the previous stage, based on the predefined categories established by the teacher according to the expected responses. With the classification results, each student provides feedback on every category of responses. Examples of the response categories are provided in Table 1. If a particular category is underrepresented, the AI-generated response is adjusted accordingly during the peer assessment process. In the feedback process of the Analyze stage, students receive the problem along with answers from assigned peers, as illustrated in Figure 1 (c). Since most students are not skilled at providing feedback, evaluation criteria are provided to guide them on how

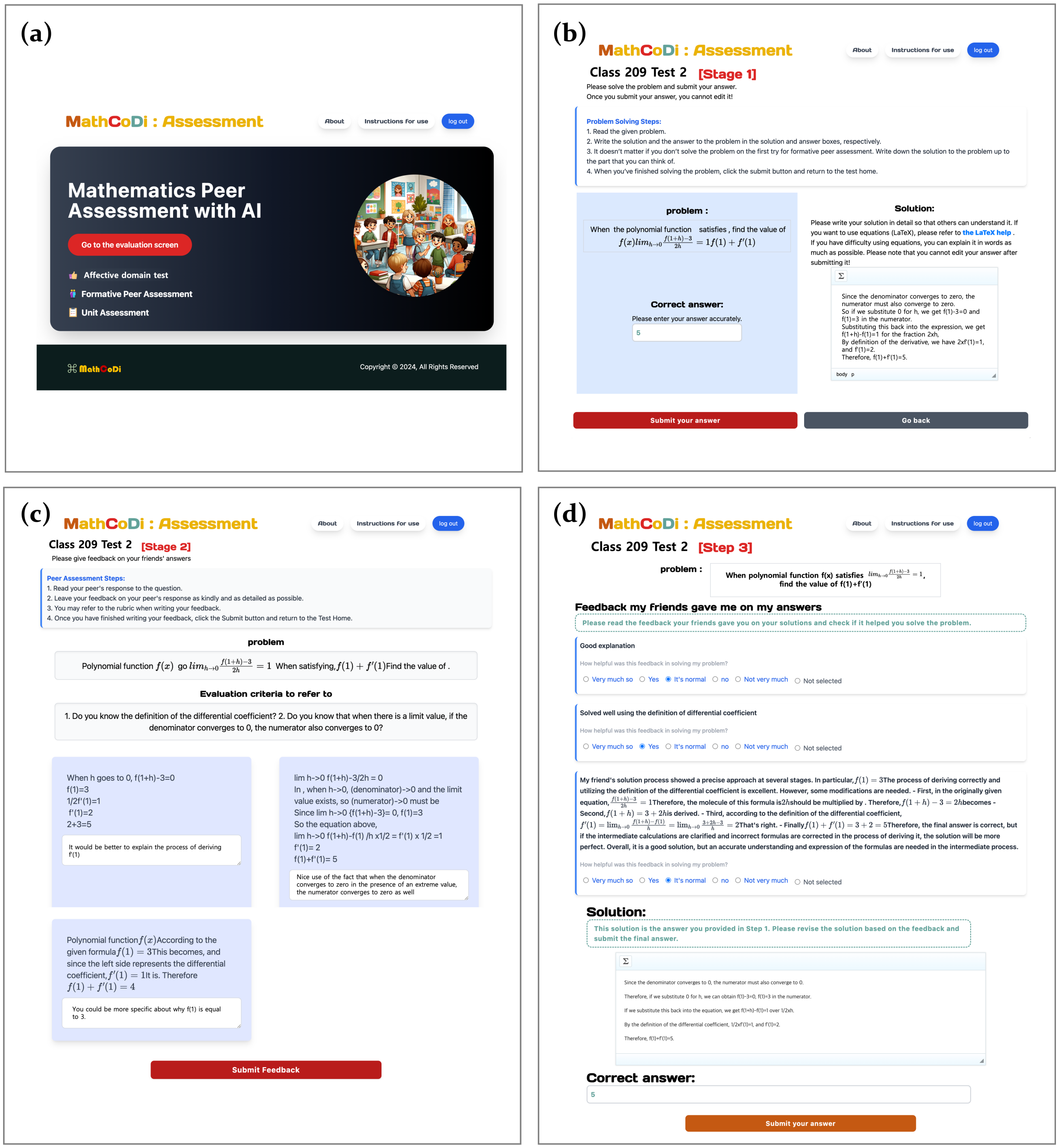

Figure 1: Student view of the three phases, illustrated from (a) to (d) :(a) Landing page of the AIFPA system, (b) Stage 1 - Answer and Answer Stage, (c) Stage 2 - Analyze Stage, (d) Stage 3 - Adapt Stage

to give feedback. While students provide feedback to each other, AI also participates in the peer assessment process as a peer and offers feedback. Feedback generation utilizes OpenAI's GPT-4 [1]. The prompts for generating feedback include strategies such as persona and output customization [45]. Also, we instructed the AI to deliver feedback at both the task and process levels, following the four levels of effective feedback proposed by Hattie [17]. For details on prompt engineering, see Table 2.

3.1.3 Adapt Stage. In the Adapt stage, students receive both AI-generated and peer-generated feedback simultaneously. Since the peer assessment process is conducted anonymously, students are not informed about the identity of the feedback providers, nor are they aware whether the feedback comes from AI or a peer, as shown

Table 1: Example of a problem and corresponding response categories designed by the teacher

	Example
Problem	*Consider a polynomial function* $f(x)$*. Given that* $\lim_{h\to 0} \frac{f(1+h)-3}{2h} = 1$*, find the value of* $f(1) + f'(1)$*.*
Category 1 : Correct Answer	First, we use the given limit to find that $f(1) = 3$. Since $\lim_{h\to 0} \frac{f(1+h)-3}{2h} = 1$ and $\lim_{h\to 0} 2h = 0$, $\lim_{h\to 0} f(1+h) - 3 = 0$. This expression represents the derivative of $f(x)$ at $x = 1$. By applying the limit definition of the derivative, $f'(1) = \lim_{h\to 0} \frac{f(1+h)-f(1)}{h}$, and we find that $f'(1) = 2$, as the limit converges to 2. Therefore, $f(1) + f'(1) = 3 + 2 = 5$.
Category 2 : Inadequate Answer	Since the denominator approaches zero, we substitute 0 into the denominator, finding that $f(1) = 3$. The left-hand side is related to the derivative, so we set $f'(1)/2 = 1$, yielding $f'(1) = 2$. Thus, $f(1) + f'(1) = 3 + 2 = 5$. This answer applies some correct steps, but does not fully utilize the definition of the derivative.
Category 3 : Answer with Misconception	Given the indeterminate form $0/0$, we substitute zero into the denominator, finding $f(1) = 3$. Since the left-hand side is the derivative, we assume $f'(1) = 1$. Thus, $f(1) + f'(1) = 3 + 1 = 4$. This answer incorrectly handles the derivative and fails to recognize the correct limit process.
Category 4 : Answer with Failure	I cannot solve the problem as it seems too complex.

Table 2: Example of Prompts Used for Feedback Generation

Category	**Prompt**
Persona	You are a secondary school student in South Korea taking a mathematics class.
Feedback Strategy	Provide positive feedback for correctly solved parts. For incorrectly solved parts, explain the mistakes, their reasons, and how to properly correct them.
Output Format	Read the <problem> below and provide feedback on your <classmate's answer> in approximately 50 words. Avoid using phrases that refer to your friend, such as "your friend is-". <problem>{problem} <classmate's answer>{classmate's answer}

in Figure 1 (c). Students then rate the effectiveness of each feedback using a 5-point Likert scale. These ratings are used as feedback preferences in data analysis. After rating all received feedback, students submit their final attempt in two forms: a solution and an answer, similar to the ask and answer stages.

3.2 Participants

Participants were 108 secondary school sophomores attending a mathematics course in South Korea. In the course, students were provided with personal electronic devices (e.g., smartphones, laptops, tablets) to access web-based content. Only 37% of the students had prior learning experience with AI. The course was conducted face-to-face, and an AI-driven formative peer assessment process was implemented in the middle of the course, following the students' learning of a new mathematical concept. However, most of the students had no prior experience with peer assessment. To support the peer assessment process, the teacher provided explanations in class and posted instructions online, offering real-time assistance to students who were struggling. Additionally, since submitting mathematical processes required LaTeXformatting, basic grammar related to the processes handled in the posed problem was taught before the assessment began. Participants were informed about the research process, and consent was obtained from both participants and their parents. Also, this study was conducted with prior approval from the Seoul National University Institutional Review Board (IRB No. 2406/004-010).

A total of 108 students are divided into four different classes with varying levels of mathematics study intensity. Class A ($N = 27$) is a class with lower intensity of mathematics learning, while Class B ($N = 26$) and Class C ($N = 29$) exhibit higher intensity of mathematics learning. Class D ($N = 26$), organized around a science-focused curriculum, shows the highest engagement in mathematics classes and the strongest intensity of mathematics learning compared to the other classes. The goal is to analyze how preferences for AI-generated feedback and peer-generated feedback differ by classroom environment, in order to address **RQ1**.

3.3 Data Processing

In this study, trace data from the AI-driven formative peer assessment process and self-report survey data on learning dispositions were combined for analysis. The trace data includes representative variables such as correctness of responses to problems (*FirstAttempt-Correct*), time spent on each phase (*FirstAttemptTime*) [36, 37, 39], and feedback preferences collected during adapt stage (*Preference*). Additionally, because the brief length of feedback can be a drawback for students [11], feedback length was also included as a variable (*FeedbackLength*).

In analyzing the preference for AI-generated feedback, we assumed that the quality of peer feedback could influence students' preference for AI-generated feedback. Thus, additional variables were included in the analysis, such as the preference and length of peer feedback(*PeerPreferences*, *PeerFeedbackLength*), as well as the proportion of peers who submitted a correct first attempt (*CorrectRatio*).

From dispositional data used to predict math performance in [16], *Math Anxiety*, *Math Confidence*, *Math Interest*, and *Math Value* were included in this study. *Math Anxiety* refers to the anxiety associated with mathematical achievement and feelings of unease related to mathematics, which is negatively correlated with achievement. It was measured using 9 items from the scale developed by [47].

Math Confidence refers to the confidence students have in their mathematical abilities and is positively correlated with academic achievement [16]. *Math Value*, on the other hand, measures the perceived importance of mathematics, which is associated with continued enrollment in math-related courses [16]. These two domains were measured using 5 items and 4 items, respectively, from the scale developed by [24].

Math Interest refers to the interest in mathematics, which enhances engagement and is positively related to achievement [16]. *Math Interest* was measured with 2 items from [46].

Additionally, items from [19], adaption of the Fennema-Sherman Mathematics Attitude Sclaes (FSMAS) for educational technology, were revised to focus specifically on AI. *AI Relevance* refers to how familiar students feel with AI. *AI Relevance* included items such as "I think the AI is the area that I use often in my life." *AI Interest* similarly assesses interest but specifically in relation to AI, akin to *Math Interest*. *AI Relevance* and *AI Interest* were measured with 4 items and 3 items, respectively.

After collecting dispositional data through a self-report survey on a 5-point Likert scale, items with low loading values were removed using confirmatory factor analysis. The remaining items were then reassigned to their respective variables. Following these adjustments, the variables were measured with the following number of items: *Math Anxiety* with 3 items, *Math Confidence* with 7 items, *Math Value* with 3 items, *Math Interest* with 2 items, *AI Relevance* with 3 items, and *AI Interest* with 3 items. The internal consistency (α) for the scales were 0.84, 0.91, 0.75, 0.85, 0.81, and 0.85, respectively.

The AI-driven formative peer assessment was conducted twice in Class A and three times in Class B, Class C, and Class D. The same set of problems was used across all classes. Feedback was categorized into four groups based on the identity of giver and receiver. Group 1 (N = 259) includes feedback offered by the AI, Group 4 (N = 235) consists of feedback received by the AI, Group 2 (N = 188) encompasses feedback provided by students who answered correctly on their first attempt, and Group 3 (N = 174) contains feedback from students who answered incorrectly on their first attempt. For feedback in Group 4, since AI was the receiver, feedback preference could not be measured. Additionally, feedback without preference in Group 1, Group 2, and Group 3 were excluded from the analysis. After preprocessing, only feedback in Group 1 (N = 182), Group 2 (N = 153), and Group 3 (N = 154) with preference was used in analysis. To explore **RQ1** and **RQ2**, ANOVA was used to analyze the differences in preferences among Group 1, Group 2, and Group 3. Based on the analysis results, we will investigate the practices for implementing AI-driven formative peer assessment (**RQ3**).

To propose practices for the implementation of AI-driven formative peer assessment, a prediction model was also constructed. Feedback preference was used as the dependent variable, while previously mentioned trace data and dispositional data were used as the independent variable. Since formative peer assessments were conducted 2 to 3 times in each class, an initial attempt was made to create a prediction model using a Linear Mixed Model with assessment as a random effect. However, in the fitted model, the influence of the random effect was not statistically significant. Therefore, to investigate **RQ3**, Linear Regression was used for the prediction model.

4 Results

4.1 Differences in Preference for AI-Generated and Peer-Generated Feedback in Relation to Classroom Environment

Table 3 and Table 4 display the average values of trace data and dispositional data by class. Class A, compared to other classes, shows lower levels of *Math Confidence* and *Math Interest*. The trace data also reveals that Class A had the lowest accuracy in the first attempt. Class C, on the other hand, exhibited the highest *Math Anxiety* and had notably high *FirstAttemptTime* compared to other classes. Class D demonstrated the highest levels of *Math Confidence*, *Math Interest*, and *AI Interest* and also had the highest *FirstAttemptCorrect*, but it had the lowest *FeedbackLength*.

Figure 2 illustrates the distribution of average preference ratings for feedback given by each student from the giver's perspective. The red horizontal line represents the average preference rating for feedback offered by AI. In Class A and Class C, the average preference rating for AI-generated feedback is higher than the third quartile (Q3), while in Class D, it is lower than the first quartile (Q1).

Table 5 presents the results of ANOVA analyzing the differences in preference for feedback among Group 1, Group 2, and Group 3 across different classes. Dunnett's Test was used as the post-hoc analysis to compare preference ratings between peer-generated feedback (Group 2 and Group 3) relative to AI-generated feedback (Group 1). In Class A, difference in feedback preferences between groups was significant at the 0.0001 significance level. In Class C and Class D, the differences were significant at the 0.05 and 0.01 significance levels, respectively. However, while Class A and Class C showed a higher preference for AI-generated feedback (Group 1) compared to peer-generated feedback (Group 2 and Group 3), Class D demonstrated a higher preference for peer-generated feedback (Group 2 and Group 3) over AI-generated feedback (Group 1).

Table 3: Average of dispositional data by class

Class	Math Anxiety	Math Confidence	Math Value	Math Interest	AI Relevance	AI Interest
Class A	3.15	2.04	3.42	2.45	3.78	2.83
Class B	3.30	2.77	3.58	3.38	3.68	2.85
Class C	3.76	2.65	4.12	3.10	4.04	2.85
Class D	3.51	2.88	4.07	3.42	4.00	3.25

Table 4: Average of trace data by class

Class	First Attempt Correct	First Attempt Time	Feedback Length
Class A	0.417	211.858	205
Class B	0.518	178.779	234
Class C	0.500	736.910	228
Class D	0.534	255.299	192

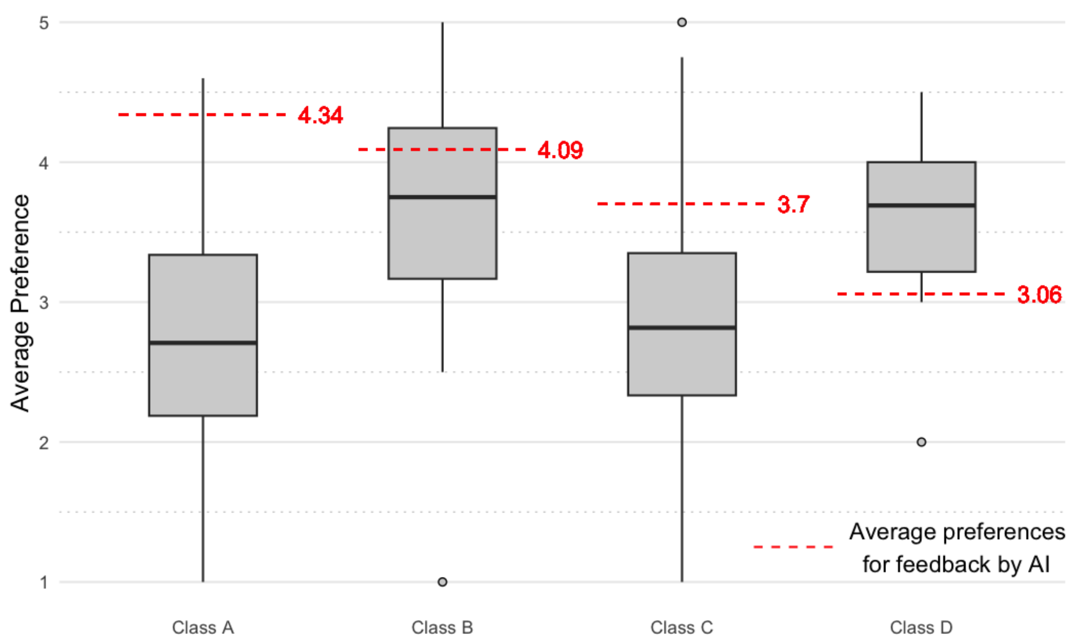

Figure 2: Boxplots of average preference of feedback from the perspective of the giver

4.2 Differences in Preference for AI-Generated and Peer-Generated Feedback in Relation to Disposition Level

The dispositional data, measured through a self-report survey, was divided into high and low level based on the median values of each disposition. The preference difference between AI-generated feedback (Group 1) and peer-generated feedback (Group 2 and Group 3) was compared within each level using ANOVA, with the results summarized in Table 6. As shown in Table 6, statistically significant differences were not found for *Math Anxiety* and *Math Interest*. However, in the low *Math Confidence* level, AI-generated feedback (Group 1) was significantly preferred over the feedback provided by students who got the first attempt wrong (Group 3) at the 0.1 significance level. In the low *Math Value* level, Group 1 feedback was significantly preferred over Group 3 feedback at the 0.01 significance level. In the low *AI Relevance* group, Group 1 feedback was significantly preferred over Group 3 feedback at the 0.1 significance level. Although the initial ANOVA analysis did not yield statistically significant results for *AI Interest*, the subsequent post-hoc analysis using Dunnett's test identified a significant preference for AI-generated feedback over Group 3 feedback in the low *AI Interest* level at the 0.1 significance level.

Table 5: ANOVA results and post-hoc analysis with Dunnett's Test for feedback preferences in different classroom environment (. : $p \leq 0.1$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$)

Class	F value	p-value	Contrast	Estimate	p-value
Class A	12.040	<.0001***	**group2 - group1**	-1.35	0.0030**
			group3 - group1	-1.72	<.0001***
Class B	1.391	0.2540	**group2 - group1**	0.11	0.9058
			group3 - group1	-0.46	0.2990
Class C	4.194	0.0169*	**group2 - group1**	-0.58	0.0629.
			group3 - group1	-0.71	0.0135*
Class D	4.773	0.0095**	**group2 - group1**	0.61	0.0133*
			group3 - group1	0.59	0.0219*

Table 6: ANOVA results and post-hoc analysis with Dunnett's test: comparison of feedback preferences at high and low disposition levels(. : $p \leq 0.1$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$)

Disposition	level	F value	p-value	Contrast	Estimate	p-value
Math Anxiety	low	1.451	0.2370	**group2 - group1**	0.24	0.4493
				group3 - group1	-0.15	0.7245
	high	1.574	0.2100	**group2 - group1**	-0.42	0.1480
				group3 - group1	-0.24	0.4834
Math Confidence	low	2.400	0.0937 .	**group2 - group1**	-0.45	0.1527
				group3 - group1	-0.49	0.0893 .
	high	1.573	0.2090	**group2 - group1**	0.22	0.4419
				group3 - group1	-0.15	0.6747
Math Value	low	7.095	0.0012**	**group2 - group1**	-0.04	0.9812
				group3 - group1	-0.93	0.0015**
	high	0.380	0.6840	**group2 - group1**	0.10	0.8725
				group3 - group1	0.22	0.5876
Math Interest	low	2.261	0.1090	**group2 - group1**	-0.60	0.1759
				group3 - group1	-0.56	0.1065
	high	0.245	0.7830	**group2 - group1**	0.16	0.6996
				group3 - group1	0.08	0.9024
AI Relevance	low	2.490	0.0867 .	**group2 - group1**	-0.02	0.9934
				group3 - group1	-0.45	0.0841 .
	high	1.354	0.2610	**group2 - group1**	-0.46	0.1920
				group3 - group1	-0.29	0.4974
AI Interest	low	2.287	0.1050	**group2 - group1**	-0.36	0.2347
				group3 - group1	-0.48	0.0807 .
	high	1.276	0.2810	**group2 - group1**	-0.08	0.8948
				group3 - group1	-0.36	0.2182

4.3 Prediction Model of AI-Generated and Peer-Generated Feedback Preference

A regression analysis was conducted to develop models predicting preference for AI-generated feedback and peer-generated feedback. All variables were standardized before the analysis. The models with the highest R^2 values for the combination of trace data and dispositional data are reported in Table 7 and Table 8.

The best combination of predictors of AI-generated feedback preference was *Math Confidence*, *Math Value*, *PeerPreferences*, and *CorrectRatio*, accounting for 43% of the variance in preference for AI-generated feedback. The coefficient for *Math Confidence* was -0.277, suggesting a significant negative relationship at the 0.01 significance level. *Math Value* had a coefficient of -0.165, which is marginally significant and indicates a negative relationship at the 0.1 significance level. Conversely, *PeerPreferences* showed a strong positive coefficient of 0.855, suggesting a significant positive

Table 7: Linear regression results: predicting AI-generated feedback preference(. : $p \leq 0.1$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$)

Predictor	Estimate	Std. Error	t-value	p-value
(Intercept)	3.609	0.085	42.295	< 0.0001 ***
Math Confidence	-0.277	0.085	-3.265	0.0013 **
Math Value	-0.165	0.095	-1.736	0.0843 .
Peer Preferences	0.855	0.087	9.823	< 0.0001 ***
Correct Ratio	-0.205	0.082	-2.491	0.0137 *
Model Summary				
Multiple R^2				0.4410
Adjusted R^2				0.4277
p-value				< 0.0001

Table 8: Linear regression results: predicting peer-generated feedback preference(. : $p \leq 0.1$, *: $p \leq 0.05$, **: $p \leq 0.01$, ***: $p \leq 0.001$)

Predictor	Estimate	Std. Error	t-value	p-value
(Intercept)	3.365	0.074	45.042	< 0.0001 ***
Math Anxiety	-0.133	0.077	-1.730	0.0847 .
Math Interest	0.252	0.084	2.967	0.0032 **
Math Value	-0.176	0.101	-1.745	0.0821 .
AI Relevance	-0.160	0.080	-1.996	0.0469 *
Feedback Length	0.194	0.071	2.737	0.0066 **
Model Summary				
Multiple R^2				0.1063
Adjusted R^2				0.0905
p-value				<0.0001

relationship at the 0.001 significance level. Finally, the *CorrectRatio* coefficient was −0.205, revealing a significant negative association at 0.05 significance level.

The best combination of predictors for peer-generated feedback preference included *Math Anxiety*, *Math Interest*, *Math Value*, *AI Relevance*, and *Feedback Length*, accounting for 10% of the variance in peer-generated feedback preference. The coefficients for *Math Anxiety*, *Math Value*, and *AI Relevance* were −0.133, −0.176, and −0.160, respectively, suggesting marginally significant negative relationships at the 0.10, 0.10, and 0.05 significance levels. In contrast, *Math Interest* and *Feedback Length* had positive coefficients of 0.252 and 0.194, indicating significant positive associations at the 0.01 significance level.

5 Discussion on Implementation Practices

In this section, we will discuss practical recommendations for the implementation of AI-driven formative peer assessment by integrating insights from **RQ1**, **RQ2**, and **RQ3**, based on the analysis results presented in the previous section.

ANOVA analysis of Class A revealed a statistically significant preference for AI-generated feedback (Group 1) over peer-generated feedback (Group 2 and Group 3). Based on trace data and dispositional data, the classroom environment in Class A can be described as "low engagement and low achievement." *Math Confidence* and *Math Interest* were the lowest among the four classes, and the proportion of answering correctly on the first attempt was also the lowest. Low engagement and poor quality of feedback from peers decreased satisfaction with peer-generated feedback, which likely led to a higher preference for feedback from AI. In an environment characterized by low engagement, AI can play a complementary role in supporting feedback process. Therefore, when issues such as low participation or poor feedback quality are expected, proactively planning for the integration of AI can be advantageous.

ANOVA analysis of Class C also demonstrated a statistically significant preference for AI-generated feedback (Group 1) over peer-generated feedback (Groups 2 and Group 3). Among the four classes, *Math Confidence* and *Math Interest* were the second lowest. Additionally, the proportion of answering correctly on the first attempt was also the second lowest. However, students in Class C spent the longest time on their first attempts and wrote the second longest feedback. Therefore, Class C can be described as "low achievement but high engagement". Similar with discussion about Class A result, AI-generated feedback appears to be beneficial when implementing formative assessments in low achievement class. Furthermore, the results suggest that AI can integrate effectively even in high engagement contexts, fostering constructive interactions with students.

In contrast, the ANOVA analysis for Class D revealed a statistically significant preference for peer-generated feedback (Group 2 and Group 3) over AI-generated feedback (Group 1). Based on trace data and dispositional data, the classroom environment in Class D can be described as "high achievement and high interest." Class D exhibited the highest levels of *Math Confidence*, *Math Interest*, and *AI Interest* among the four classes, and the ratio of answering correctly on the first attempt was also the highest. In this environment, students were able to receive quality feedback from their peers, which likely contributed to their lower preference for feedback from AI. Additionally, according to the teacher who conducted the assessments, students, due to their high *AI Interest*, were able to easily identify AI-generated feedback among the anonymous responses and viewed it more critically compared to peer feedback. This suggests that they were more sensitive to the hallucination than students in other classes, which may explain their lower preference for AI-generated feedback. Therefore, in a group demonstrating high achievement and interest in AI, it is essential to establish and introduce appropriate evaluation design to ensure the validity of AI-generated feedback in formative peer assessment.

The *Math Confidence* and *Math Interest* levels of Class B were similar to those of Class D, but its *AI Interest* level was lower in Class B compared to Class D. Additionally, the proportion of answering correctly on the first attempt was second highest, following Class D. Class B can be characterized as "high achievement but low AI Interest". In contrast, students in Class C were less likely to detect and criticize AI-generated feedback compared to Class D. As a result, ANOVA analysis of Class B did not reveal a statistically significant preference for peer-generated feedback over AI-generated feedback. As discussed in the result of Class D, students' interest in AI may serve as a crucial factor in determining the effectiveness of AI-generated feedback in formative assessments.

An analysis was conducted to examine the differences in preference between AI-generated feedback and peer-generated feedback based on levels of disposition. However, most results were not statistically significant. Low *Math Value* level group showed a tendency to prefer AI-generated feedback (Group 1) over peer feedback from students who made wrong first attempt (Group 3). While this may suggest that students with lower value placed on math might perceive AI feedback as more reliable and supportive, potentially due to its structured nature, further investigation is necessary to fully understand this preference.

In the AI-generated feedback preference prediction model, the predictors used were *Math Confidence*, *Math Value*, *PeerPreferences*, and *CorrectRatio*. The strong positive relationship with *PeerPreferences* suggests that students who are satisfied with peer-generated feedback are also likely to be satisfied with AI-generated feedback. Therefore, AI-generated feedback can be offered as additional content to students who actively engage in the feedback process and receive it positively. Conversely, the negative relationship with *CorrectRatio* indicates that as the number of students answering correctly increases, their preference for AI-generated feedback may decrease. Thus, when there is a high correct response rate on the first attempt, AI-generated feedback may be used more conservatively, while it can be utilized more actively when the correct response rate is low. Additionally, the negative relationship between *Math Confidence*, *Math Value*, and AI-generated feedback preference suggests that for students with low levels of *Math Confidence* and *Math Value*, it may be more beneficial to prioritize AI over human peers in the peer assessment pairing process.

In the prediction model for peer-generated feedback preference, the predictors used were *Math Anxiety*, *Math Interest*, *Math Value*, *AI Relevance*, and *Feedback Length*. The positive relationship between *Feedback Length* and peer-generated feedback preference suggests that, similar to the findings of [11], students tend to prefer longer feedback. Since providing lengthy feedback can be a challenging task for students, it is essential to offer adequate training on how to give effective feedback. Additionally, since *Math Interest* enhances engagement [16], the positive relationship between *Math Interest* and preference indicates the importance of active engagement in the peer assessment process.

In summary, the following practices for the implementation of AI-driven formative peer assessment can be recommended:

(1) AI-generated feedback may help activate the peer assessment process in classes with low participation rates.
(2) AI-generated feedback may be beneficial when conducting formative peer assessments in classes experiencing difficulties in math learning.
(3) AI-generated feedback may be useful during peer assessments on problems with low correct response rates.
(4) Students with high *AI Interest* are sensitive to hallucination, therefore an appropriate evaluation design is necessary to ensure the validity of AI-generated feedback.
(5) A pairing algorithm can be designed to prioritize the assignment of human peers over AI for students with high *Math Confidence* and *Math Value* level.

6 Limitation and Future Work

In this study, we conducted an AI-driven formative peer assessment in which AI participated in the peer evaluation process. Using trace data and dispositional data, we analyzed preferences for AI-generated feedback and provided recommendations for implementation practices. The research was carried out in a face-to-face classroom setting with a sample of 108 students. Future studies should explore the implementation of AI-driven formative peer assessments in diverse contexts, such as distance learning, higher education, and MOOCs. Dispositional data included *Math Anxiety*, *Math Confidence*, *Math Value*, *Math Interest*, *AI Relevance*, and *AI Interest*. However, the limited scope of this dispositional data may have constrained a comprehensive understanding of the relationship between students' dispositions and their preferences for AI-generated feedback. Future research should incorporate analyses of self-regulated learning, goal-setting, help-seeking behaviors, and motivation, as outlined by [36, 39]. Although hallucinations of AI did not introduce misconceptions significant enough to mislead students' learning, the hallucination rate varied considerably across the items. Future work could address this by optimizing a model, which would likely reduce errors and enhance feedback accuracy. In this research, AI was able to facilitate the evaluation process despite a low participation rate but was unable to increase the participation rate itself. Therefore, future research should explore methods to leverage AI to enhance student participation.

Acknowledgments

This research was partially funded by the Seoul National University Learning Science Research Institute (Project Number: 0767-20240008).

References

[1] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).
[2] Chie Adachi, Joanna Hong-Meng Tai, and Phillip Dawson. 2018. Academics' perceptions of the benefits and challenges of self and peer assessment in higher education. *Assessment & Evaluation in Higher Education* 43, 2 (2018), 294–306.
[3] Rialy Andriamiseza, Franck Silvestre, Jean-François Parmentier, and Julien Broisin. 2023. How Learning Analytics can help orchestration of formative assessment? Data-driven recommendations for Technology-enhanced learning. *IEEE Transactions on Learning Technologies* 16, 5 (2023), 804–819.
[4] Randy Elliot Bennett. 2011. Formative assessment: A critical review. *Assessment in education: principles, policy & practice* 18, 1 (2011), 5–25.
[5] Jacqueline Bichsel. 2012. *Analytics in higher education: Benefits, barriers, progress, and recommendations.* EDUCAUSE Center for Applied Research.
[6] Paul Black and Dylan Wiliam. 1998. Assessment and classroom learning. *Assessment in Education: principles, policy & practice* 5, 1 (1998), 7–74.
[7] Paul Black and Dylan Wiliam. 2003. 'In praise of educational research': Formative assessment. *British educational research journal* 29, 5 (2003), 623–637.
[8] Paul Black and Dylan Wiliam. 2009. Developing the theory of formative assessment. *Educational Assessment, Evaluation and Accountability (formerly: Journal of personnel evaluation in education)* 21 (2009), 5–31.
[9] Benjamin S Bloom. 1969. Some theoretical issues relating to educational evaluation. *Teachers College Record* 70, 10 (1969), 26–50.
[10] Simon Buckingham Shum and Ruth Deakin Crick. 2012. Learning dispositions and transferable competencies: pedagogy, modelling and learning analytics. In *Proceedings of the 2nd international conference on learning analytics and knowledge.* 92–101.
[11] Deirdre Burke. 2009. Strategies for using feedback students bring to higher education. *Assessment & Evaluation in Higher Education* 34, 1 (2009), 41–50.
[12] Diane Dalby and Malcolm Swan. 2019. Using digital technology to enhance formative assessment in mathematics classrooms. *British journal of educational technology* 50, 2 (2019), 832–845.

[13] Cath Ellis. 2013. Broadening the scope and increasing the usefulness of learning analytics: The case for assessment analytics. *British Journal of Educational Technology* 44, 4 (2013), 662–664.

[14] Ed Foster and Rebecca Siddle. 2020. The effectiveness of learning analytics for identifying at-risk students in higher education. *Assessment & Evaluation in Higher Education* 45, 6 (2020), 842–854.

[15] Libby Gerard, Ady Kidron, and Marcia C Linn. 2019. Guiding collaborative revision of science explanations. *International Journal of Computer-Supported Collaborative Learning* 14, 3 (2019), 291–324.

[16] Sara Hart, Mia Daucourt, and Colleen Ganley. 2017. Individual differences related to college students' course performance in calculus II. *Journal of Learning Analytics* 4, 2 (2017), 129–153.

[17] John Hattie and Helen Timperley. 2007. The power of feedback. *Review of educational research* 77, 1 (2007), 81–112.

[18] Monika Hooda, Chhavi Rana, Omdev Dahiya, Ali Rizwan, and Md Shamim Hossain. 2022. Artificial intelligence for assessment and feedback to enhance student success in higher education. *Mathematical Problems in Engineering* 2022, 1 (2022), 5215722.

[19] Murat Kahveci. 2010. Students' Perceptions to Use Technology for Learning: Measurement Integrity of the Modified Fennema-Sherman Attitudes Scales. *Turkish Online Journal of Educational Technology-TOJET* 9, 1 (2010), 185–201.

[20] Fatma Gizem Karaoglan Yilmaz and Ramazan Yilmaz. 2020. Student opinions about personalized recommendation and feedback based on learning analytics. *Technology, knowledge and learning* 25 (2020), 753–768.

[21] Julia H Kaufman and Christian D Schunn. 2011. Students' perceptions about peer assessment for writing: Their origin and impact on revision work. *Instructional science* 39 (2011), 387–406.

[22] Charles Lang, George Siemens, Alyssa Wise, and Dragan Gasevic. 2017. *Handbook of learning analytics*. SOLAR, Society for Learning Analytics and Research New York.

[23] Tingting Li, Emily Reigh, Peng He, and Emily Adah Miller. 2023. Can we and should we use artificial intelligence for formative assessment in science. *Journal of Research in Science Teaching* 60, 6 (2023), 1385–1389.

[24] Siew Yee Lim and Elaine Chapman. 2013. Development of a short form of the attitudes toward mathematics inventory. *Educational studies in mathematics* 82 (2013), 145–164.

[25] Marina Marchisio, Alice Barana, Michele Fioravera, Sergio Rabellino, and Alberto Conte. 2018. A model of formative automatic assessment and interactive feedback for STEM. In *2018 IEEE 42nd Annual Computer Software and Applications Conference (COMPSAC)*, Vol. 1. IEEE, 1016–1025.

[26] Pedro Manuel Moreno-Marcos, Ting-Chuen Pong, Pedro J Munoz-Merino, and Carlos Delgado Kloos. 2020. Analysis of the factors influencing learners' performance prediction with learning analytics. *IEEE Access* 8 (2020), 5264–5282.

[27] Quan Nguyen, Dirk Tempelaar, Bart Rienties, and Bas Giesbers. 2016. What learning analytics based prediction models tell us about feedback preferences of students. *Quarterly Review of Distance Education* 17, 3 (2016), 13–33.

[28] Tung Phung, Victor-Alexandru Pădurean, Anjali Singh, Christopher Brooks, José Cambronero, Sumit Gulwani, Adish Singla, and Gustavo Soares. 2024. Automating human tutor-style programming feedback: Leveraging gpt-4 tutor model for hint generation and gpt-3.5 student model for hint validation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 12–23.

[29] Gerti Pishtari, Edna Sarmiento-Márquez, María Jesús Rodríguez-Triana, Marlene Wagner, and Tobias Ley. 2024. Mirror mirror on the wall, what is missing in my pedagogical goals? The Impact of an AI-Driven Feedback System on the Quality of Teacher-Created Learning Designs. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 145–156.

[30] Frans J Prins, Dominique MA Sluijsmans, Paul A Kirschner, and Jan-Willem Strijbos. 2005. Formative peer assessment in a CSCL environment: A case study. *Assessment & Evaluation in Higher Education* 30, 4 (2005), 417–444.

[31] P Race. 1998. Practical pointers on peer assessment. *SEDA PAPER* (1998), 113–122.

[32] Cristobal Romero and Sebastian Ventura. 2020. Educational data mining and learning analytics: An updated survey. *Wiley interdisciplinary reviews: Data mining and knowledge discovery* 10, 3 (2020), e1355.

[33] D Royce Sadler. 1989. Formative assessment and the design of instructional systems. *Instructional science* 18, 2 (1989), 119–144.

[34] Michael Scriven. 1967. The Methodology of Evaluation. In *Perspectives of Curriculum Evaluation*, R.W. Tyler, R.M. Gagne, and M. Scriven (Eds.). Rand McNally, Chicago, IL, 39–83.

[35] Dirkjan Tempelaar, Quang Nguyen, and Bart Rienties. 2019. Learning Feedback Based on Dispositional Learning Analytics. In *Machine Learning Paradigms*, Maria Virvou, Efthimios Alepis, George A. Tsihrintzis, and Lakhmi C. Jain (Eds.). Vol. 158. Springer, 69–89.

[36] Dirk Tempelaar, Bart Rienties, Jenna Mittelmeier, and Quan Nguyen. 2018. Student profiling in a dispositional learning analytics application using formative assessment. *Computers in Human Behavior* 78 (2018), 408–420.

[37] Dirk Tempelaar, Bart Rienties, and Quan Nguyen. 2018. Investigating learning strategies in a dispositional learning analytics context: the case of worked examples. In *Proceedings of the 8th International Conference on Learning Analytics and Knowledge*. 201–205.

[38] Dirk T Tempelaar, André Heck, Hans Cuypers, Henk van der Kooij, and Evert van de Vrie. 2013. Formative assessment and learning analytics. In *Proceedings of the third international conference on learning analytics and knowledge*. 205–209.

[39] Dirk T Tempelaar, Bart Rienties, and Bas Giesbers. 2015. In search for the most informative data for feedback generation: Learning analytics in a data-rich context. *Computers in Human Behavior* 47 (2015), 157–167.

[40] Danielle R Thomas, Jionghao Lin, Erin Gatz, Ashish Gurung, Shivang Gupta, Kole Norberg, Stephen E Fancsali, Vincent Aleven, Lee Branstetter, Emma Brunskill, et al. 2024. Improving student learning with hybrid human-AI tutoring: A three-study quasi-experimental investigation. In *Proceedings of the 14th Learning Analytics and Knowledge Conference*. 404–415.

[41] Marjo Van Zundert, Dominique Sluijsmans, and Jeroen Van Merriënboer. 2010. Effective peer assessment processes: Research findings and future directions. *Learning and instruction* 20, 4 (2010), 270–279.

[42] Philip Vickerman. 2009. Student perspectives on formative peer assessment: an attempt to deepen learning? *Assessment & Evaluation in Higher Education* 34, 2 (2009), 221–230.

[43] Pierpaolo Vittorini, Stefano Menini, and Sara Tonelli. 2021. An AI-based system for formative and summative assessment in data science courses. *International Journal of Artificial Intelligence in Education* 31, 2 (2021), 159–185.

[44] Thomas Wanner and Edward Palmer. 2018. Formative self-and peer assessment for improved student learning: the crucial factors of design, teacher participation and feedback. *Assessment & Evaluation in Higher Education* 43, 7 (2018), 1032–1047.

[45] Jules White, Quchen Fu, Sam Hays, Michael Sandborn, Carlos Olea, Henry Gilbert, Ashraf Elnashar, Jesse Spencer-Smith, and Douglas C Schmidt. 2023. A prompt pattern catalog to enhance prompt engineering with chatgpt. *arXiv preprint arXiv:2302.11382* (2023).

[46] Allan Wigfield and Jacquelynne S Eccles. 2000. Expectancy–value theory of achievement motivation. *Contemporary educational psychology* 25, 1 (2000), 68–81.

[47] Allan Wigfield and Judith L Meece. 1988. Math anxiety in elementary and secondary school students. *Journal of educational Psychology* 80, 2 (1988), 210.

[48] Dylan Wiliam and Marnie Thompson. 2017. Integrating Assessment with Learning: What Will It Take to Make It Work? In *The Future of Assessment*. Routledge, 53–82.

[49] Archie Zariski. 1996. Student peer assessment in tertiary education: Promise, perils and practice. In *Teaching and Learning Within and Across Disciplines, p189-200. Proceedings of the 5th Annual Teaching and Learning Forum, Murdoch University*.

[50] Xiaoming Zhai and Ross H Nehm. 2023. AI and formative assessment: The train has left the station. *Journal of Research in Science Teaching* 60, 6 (2023), 1390–1398.