






AI or Human? Evaluating Student Feedback Perceptions in Higher Education

Tanya Nazaretsky , Paola Mejia-Domenzain , Vinitra Swamy ,
Jibril Frej , and Tanja Käser 

EPFL, Switzerland
{tanya.nazaretsky, paola.mejia, vinitra.swamy,
jibril.frej, tanja.kaeser}@epfl.ch

Preprint: Accepted to Nineteenth European Conference on Technology Enhanced Learning (ECTEL 2024)

Abstract. Feedback plays a crucial role in learning by helping individuals understand and improve their performance. Yet, providing timely, personalized feedback in higher education presents a challenge due to the large and diverse student population, often resulting in delayed and generic feedback. Recent advances in generative Artificial Intelligence (AI) offer a solution for delivering timely and scalable feedback. However, little is known about students’ perceptions of AI feedback. In this paper, we investigate how the identity of the feedback provider affects students’ perception, focusing on the comparison between AI-generated and human-created feedback. Our approach involves students evaluating feedback in authentic educational settings both before and after disclosing the feedback provider’s identity, aiming to assess the influence of this knowledge on their perception. Our study with 457 students across diverse academic programs and levels reveals that students’ ability to differentiate between AI and human feedback depends on the task at hand. Disclosing the identity of the feedback provider affects students’ preferences, leading to a greater preference for human-created feedback and a decreased evaluation of AI-generated feedback. Moreover, students who failed to identify the feedback provider correctly tended to rate AI feedback higher, whereas those who succeeded preferred human feedback. These tendencies are similar across academic levels, genders, and fields of study. Our results highlight the complexity of integrating AI into educational feedback systems and underline the importance of considering student perceptions in AI-generated feedback adoption in higher education.

Keywords: Generative AI · Formative Feedback · Human Factors · Higher Education.

1 Introduction

Feedback has always been a crucial part of the learning process. Properly formulated and timely feedback helps individuals understand and improve their

performance [14,15,40]. In 2015, Hattie conducted a thorough synthesis of 1200 meta-analyses related to the influences of teaching activities and methods (e.g., traditional lectures, flipped classrooms, problem-based learning, using online and Internet-based solutions, etc.) on student learning and concluded that most likely “the method of teaching is less critical than the attributes of the teaching within the methods” [13]. In particular, this study strongly emphasized evidence-based teaching practices and the importance of feedback to enhance student learning outcomes. However, managing the feedback process is the most challenging part of teaching [5,38,39]. According to the 2022 Student Experience Survey [1], among all “Teaching quality items,” the item aimed to evaluate the quality of feedback (formulated as “Teachers commented on your work in ways that help you learn”) got the lowest positive rates ($< 63\%$ of postgraduates and $< 56\%$ of undergraduate students). Similarly, according to the Annual Report and Accounts 2022-23 [34], the ratings for “Assessment and Feedback” items were the lowest overall “Teaching and Assessment” items ($< 69\%$ of students). Indeed, providing meaningful, personalized feedback requires significant time and effort. This is especially true in higher education contexts, where personal attention from the instructors is limited [3]. At the same time, students often come from diverse backgrounds and have varying levels of prior skills and knowledge. As a result, the feedback delivery is often delayed, reducing its effectiveness, and the provided feedback is not specific to the concrete student performance [12].

Automated feedback produced by AI can help speed up and scale the feedback delivery process [6,17]. Generative AI, especially Large Language Models (LLM), is rapidly gaining momentum for automating feedback provision in educational contexts. A recent systematic scoping review by [41] examined 118 peer-reviewed papers since 2017 on applying LLMs in educational settings. Among these, 54 focused explicitly on student-oriented tasks such as automated feedback provision and resource recommendation. Notable examples include GPT-based models that provide feedback on student performance in data science open-ended tasks [8], source based argument essays in history [33], inquiry-based learning [2], algebra [27] and programming tasks [26,10], achieving high performance in some cases even comparable to well-trained human evaluators.

However, most existing studies primarily target technological aspects like model accuracy and often miss the social-emotional aspects of AI’s acceptance. As shown in other fields, the preference for human over AI advice, known as “algorithm aversion” [9], persists despite AI’s proven efficiency. For example, in medical radiology, where AI outperforms humans [29], previous research demonstrated that expert opinions were less favorable towards recommendations framed as AI-generated compared to the same recommendations framed as humans’ [11], highlighting the importance of addressing algorithm aversion in AI-based system implementation. Although recent studies in education have started considering this issue, they mainly focus on educators’ attitudes [7,19,25], leaving student perspectives on AI feedback provision largely unexplored [18,36].

In this study, we aim to examine how students’ awareness of the feedback provider’s identity might influence their evaluation of feedback content, partic-

ularly in the context of algorithm aversion and preference for human expertise. With these goals in mind, the study seeks to address the following research questions: First, can students distinguish between AI-generated and human-created feedback (simplified Turing Test), and what factors influence their ability to make this distinction (**RQ1**)? Second, how do students’ perceptions of the same feedback content change after revealing the feedback provider’s identity (**RQ2**)? And third, do students hold a negative bias towards AI as a feedback provider (**RQ3**)?

To achieve our goals, we conduct an extensive user study by analyzing the responses of 457 higher education students from various courses, fields of study, and academic levels (Bachelor’s and Master’s). Specifically, we ask students to evaluate two feedback options (human-created and AI-generated) in their authentic learning settings twice: initially without (blind condition) and then with (informed condition) knowledge of the feedback provider’s identity. This deliberate priming and reveal is designed to uncover changes in students’ evaluative judgments, thus helping to evaluate the impact of the provider’s identity on educational feedback perception. Our results reveal that students’ ability to distinguish between AI-generated and human-created feedback depends on the underlying course and task. We also detect a change in feedback perception upon disclosure of the provider’s identity, with a notable trend of increased favorability towards feedback identified as human-created and a decline in the evaluation of AI-generated feedback. Moreover, Turing Test outcomes are significantly correlated with feedback perception. Students failing the test favor AI-generated feedback, while those passing prefer human feedback. In addition, there is a negative correlation between participants’ age and their perceptions of provider credibility for AI and human feedback providers. We, therefore, can confirm a bias against AI as a feedback provider.

2 Motivational Aspects of Effective Formative Feedback

A considerable body of research exists in the area of design and implementation of effective formative human and AI feedback in the context of task-level performance [4,5,14,23,32]. This research advocates for a multi-dimensional approach to feedback, including feedback content and provider characteristics [23], and recommends a learner-centric approach to the feedback considering the learners as active participants of the feedback process rather than passive receivers of the information [3,20,40]. To ensure the uptake of the feedback for effective enhancing learning process, feedback should be task-focused [31,32], provide objective performance evaluations [3], offer actionable advice on how to proceed [31], and motivate learners to adopt it to improve their response [32].

Based on the previous research the following motivational factors influence learners’ willingness to act upon feedback: learners’ perceptions of feedback content objectivity [30], usefulness [15,31], genuineness [16,32] and feedback provider credibility [30]. With respect to objectivity, according to the theories of procedural and informational justice, the feedback should be perceived by learners as fair

and factual alignment between its content provided and the actual performance of the recipient [3,30] and should be precise and specific to the response [32,37]. To act upon receiving feedback, it should be useful [31], meaning informative, relevant, and applicable [23,32]. Applicability can be achieved by error-flagging and providing cues and hints on how to improve an answer while avoiding giving the correct solution itself [23]. With respect to feedback genuineness, it should use a sincere, positive tone [14] and be authentic, meaning honest and earned, not repeated often [16] and focused on the learning process and not student abilities [32]. With respect to feedback provider credibility, the feedback source should be considered by the learners as safe [35] and ethical [41]. Safety refers to trustworthy, well-intentioned, and reliable feedback providers fostering confidence [22,35], while ethics relates to qualities such as accountability, commitment to privacy, equality, and absence of biases [41].

In this research, we explored how students perceive both AI and human feedback quality in terms of objectivity, usefulness, and genuineness, as well as their views on the competence of the feedback provider. These elements, informed by prior research, are shown to predict a student’s inclination to apply the feedback constructively to improve their responses.

3 Methodology

To evaluate how students’ feedback perceptions depend on their knowledge of the feedback provider, we developed two perception instruments and incorporated them into the experimental procedure illustrated in Fig. 1. Our study with 457 students employed a within-subject experimental design with priming effect aimed to blur the information about the feedback identity [28]. The study consisted of three steps (Fig. 1). During Step 1, we presented students with a routine assessment task from their course, their authentic answer, and two variants of feedback generated by human experts and the OpenAI GPT4 Interface (Section 3.2) and asked them to evaluate the feedback content. In Step 2, we revealed the existence of primer manipulation and asked students to guess the origin of each feedback (a simplified Turing Test based on one attempt only). In Step 3, we revealed the actual identity of each feedback provider and asked students if they wanted to change their initial opinions regarding the feedback content. In addition, we asked participants to evaluate the feedback providers. During the study, the students had no access to the correct solutions.

3.1 Participants

We conducted an online survey with 457 students from EPFL who were above the legal age and had voluntarily agreed to participate. These students were enrolled in five different courses: Discrete Mathematics (ICC), Information Theory (AICC), Applied Software Engineering (ASE), Applied Data Analysis (ADA), and Mathematical Analysis (MA). The courses spanned various academic programs, including Computer Science, Communication Systems, Data Science,

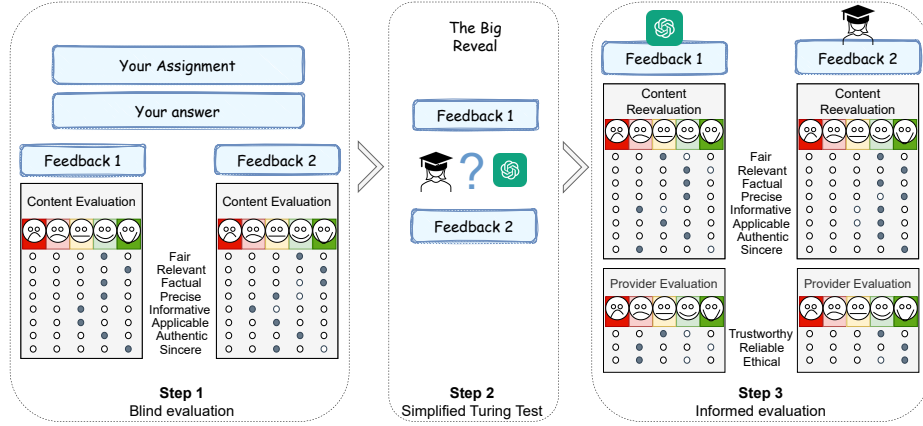


Fig. 1. Experimental design consisting of three steps: 1) blind feedback content evaluation, 2) Turing test, and 3) informed feedback and feedback provider evaluation.

Table 1. Demographics of participants by course and field of study, including degree, year, gender distribution (male/female/non-binary/other), and mean age (standard deviation).

Course	Field of Study	Degree	Year	Participants	Age
ICC	Mathematics, Physics	BA	1st	66 (53/11/0/2)	19.1 (3.0)
AICC	Computer Science	BA	1st	187 (139/37/1/10)	19.3 (2.0)
MA	Mechanical Engineering	BA	1st	28 (22/5/0/1)	18.9 (1.2)
ASE	Life Sciences Engineering	BA	2nd	70 (31/37/0/2)	20.3 (1.7)
ADA	Multiple	MA	1st	106 (72/31/1/2)	23.1 (2.9)

Civil, Mechanical, Life Sciences, Electrical and Electronic, and Financial Engineering, and more. The data collection for this study occurred in fall 2023, and detailed socio-demographic information of the participants is presented in Table 1. In further analyses, we analyzed differences between two genders: males vs. minorities, while minorities referred to students who identify as women and non-binary. This categorization was necessary since a separate statistical analysis for non-binary students alone was not feasible due to their small representation in the sample (only two students). The study was conducted with the approval of the Human Research Ethics Committee at EPFL (HREC 017-2023).

3.2 Feedback Generation

For each course in the study, we selected a pre-existing assignment and gathered genuine student responses. We adopted a positive formative feedback strategy, focusing on encouragement and support, recognizing effort accurately, and highlighting improvement areas. We created two variants of feedback for each student, one from the teaching team and the other from AI. An illustration of human and

Assignment
Let x be a real number. Prove that if x^2 is irrational, then x is irrational using:
a) a proof by contraposition
b) a proof by contradiction
Detail the two proofs in a way that shows that you understand the difference between a proof by contraposition and a proof by contradiction.

Student Answer
1. CONTRAPOSITION
1) x is rational $\rightarrow x^2$ is rational [Premise]
2) $x = p/q$ and p, q are integers and $q \neq 0$ [Definition]
3) $x^2 = p^2/q^2$ [Algebra]
4) Therefore, assuming that the square of an integers is an integer, x^2 is rational [Definition]
2. CONTRADICTION
1) Let's assume that x^2 is rational and x irrational [Premise]
2) $x^2 = p/q$ and p, q are integers and $q \neq 0$ [Definition]
3) $x = r/s$ such as $r^2 = p$ and $s^2 = q$
4) There are r and s which are integers and verify the equation: for example $q = 4$ and $p = 4$ [Counterexample]
5) The demonstration is false so for x irrational, x^2 must be irrational

Feedback (Human-created)
The contraposition is true but you have to review the contradiction. You have to know that proving $p \rightarrow q$ by contradiction: we show that assuming p and not q leads to a contradiction.

Feedback (AI-generated)
You've done a good job on the contraposition proof. However, for the contradiction part, you misunderstood the premise. You should assume x^2 is irrational and x is rational, then derive a contradiction from there. Also, step 4 is not clear and seems unnecessary; it's important to focus on deriving contradictions rather than counterexamples in this instance. Keep practicing your proof writing skills!

Fig. 2. An AICC course assignment, a corresponding student answer, and two variants of feedback: human-created and AI-generated.

Table 2. Overview of course assignments, human assessment routines, and additional feedback provision for various courses, detailing the number of human graders involved in the study.

Course	Assignment	Routine human assessment	Additional human feedback	Number of human graders
ICC	C++ project	Grade	Code review	8
AICC	Logical reasoning proof	Grade	Feedback	9
MA	Logical reasoning proof	Feedback	-	1
ASE	Python project	Code review	-	17
ADA	Statistical inference task	Grade & Feedback	-	3

AI-generated feedback, taken from a logical reasoning task in the AICC course, can be found in Fig. 2.

Human-Generated Feedback For each course, human feedback was provided by the original teaching teams. For MA, ASE, and ADA, existing authentic feedback was used as is, while in ICC and AICC, formative feedback was not provided to the students as part of the routine. For these courses teaching teams were trained through a workshop on creating positive, formative feedback, emphasizing feedback consistency and quality. The workshop included guidelines, feedback examples, and a communication channel for graders to ensure uniformity of feedback. More details about the types of assignments and number of human graders per course are presented in Table 2.

AI-Generated Feedback We asked OpenAI GPT interface¹ to provide elaborated formative feedback using the prompt illustrated in Fig. 3.

¹ Parameters: model = gpt-4, top_p = 1.0, temperature = 0.8, frequency_penalty = 0.8, presence_penalty = 0.5

You are an excellent instructor teaching a course called [COURSE NAME].
 You gave the students the following assignment: [ASSIGNMENT].
 The student submission was [STUDENT ANSWER].
 The correct solution is [SOLUTION].
 Please evaluate the student’s answer and provide elaborated formative feedback.
 Please follow the following instructions:
 The feedback should be addressed directly to the student as is.
 It should be no more than [NUMBER] lines.
 Please provide one sentence of the overall evaluation at the end.

Fig. 3. Standardized prompt template used to create AI-feedback for student assignments, detailing the necessary components and instructions.

While a grading rubric existed for all the courses, we decided not to provide it in the prompt, as we wanted GPT to focus less on the grades and more on formative feedback. The correct solution was provided for the AICC, MA, and ADA courses only. We did not provide solutions for the coding projects in the ASE and ICC courses, as there is a wide range of perfectly correct answers. Our goal was to have GPT focus on enhancing the student’s response instead of making comparisons to a single correct solution provided in the prompt. The GPT-generated feedback was provided to the students as is, except for removing the “Dear student” greeting at the beginning and “Your instructor” at the end (in some cases, GPT added them to the feedback text).

Feedback Quality This study did not aim at quantitatively assessing the objective quality of AI-generated feedback or its comparison to human-created feedback. Instead, our focus was on exploring how the knowledge of the feedback provider’s identity affects students’ perceptions of feedback. As a result, we did not conduct a statistical analysis of the quality of AI feedback as evaluated by experts. However, the study involved a qualitative assessment through a quality control test with 5-10 randomly chosen feedback samples presented to course professors and their teaching teams, who deemed the quality of feedback as high before agreeing to present it to their students. Moreover, a comprehensive review of all generated feedback by the authors revealed no instances of GPT hallucinations or contextually irrelevant content.

3.3 Perception Instrument

Based on the previous research presented in Section 2, we identified three dimensions and eight terms to evaluate the feedback content. More specifically, we defined the *Objectivity* dimension using the terms *Precise*, *Fair*, *Factual*; the *Usefulness* dimension consisting of the three terms *Relevant*, *Informative*, *Applicable*; and the *Genuineness* dimension with the terms *Authentic*, and *Sincere* [16]. In the context of the feedback provider, we identified one dimension, *Credibility*, measured using the following three terms: *Trustworthy*, *Reliable*, and *Ethical*.

Following the methodology proposed in [35], we evaluated students' feedback and feedback provider perceptions with one question per instrument. The questions were formulated as follows: "To what extent do you associate this feedback/feedback provider with the following term?" The students were asked to choose a Likert scale answer for each term (represented by aforementioned adjectives). The Likert items were evaluated on a scale of five (Strongly disagree = 0, Disagree = 1, Neutral = 2, Agree = 3, Strongly agree = 4).

4 Results

We analyzed the students' answers to the blind evaluation, the TT, and their informed evaluations to investigate whether students can distinguish between human and AI-feedback (RQ1), how their perception changes when the feedback provider identity is revealed (RQ2), and whether they hold biases against AI as a feedback provider (RQ3).

We first evaluated the reliability of the proposed instruments by calculating Cronbach's α measure for each dimension separately. The internal consistency was good ($\alpha \geq 0.80$) for the *Objectivity*, *Usefulness*, and *Credibility* dimensions and excellent ($\alpha \geq 0.90$) for the *Genuineness* dimension. Next, we calculated the scores of the factors by averaging over all items associated with the factor and evaluated the responses of the 457 students.

4.1 Turing Test for AI and Human Feedback (RQ1)

In total, 274 out of 457 students could guess the feedback provider correctly. To study the possible correlation of demographics (age, gender, course) with students' ability to guess the feedback provider identity correctly, we fit a Logistic Regression Model². The correlation was not significant for age and gender. However, there was a significant correlation with course topics. Therefore, we used Chi-Square Tests of Independence to evaluate the ability of participants of different courses in guessing the feedback provider identity. For the ASE and ICC courses, the results were significant ($\chi^2_{ASE} = 25.2$, $p < .001$; $\chi^2_{ICC} = 21.9$, $p < .001$), indicating that participants could effectively differentiate between the two types of feedback. In contrast, the AICC, MA and ADA courses did not show significant differences ($\chi^2_{AICC} = 1.20$, $p = .273$; $\chi^2_{MA} = 3.57$, $p = .059$; $\chi^2_{ADA} = 1.85$, $p = .174$), suggesting participants in these courses were not able to identify the feedback provider identity.

4.2 Impact of Identity Disclosure on Perceptions (RQ2)

To analyze how the knowledge of the feedback provider identity influences students' perceptions towards feedback content quality, we utilized a Mixed Linear Model (MLM)³ approach. More specifically, we fit the MLM separately for each

² *stats* R package

³ *lme4* and *afex* R packages

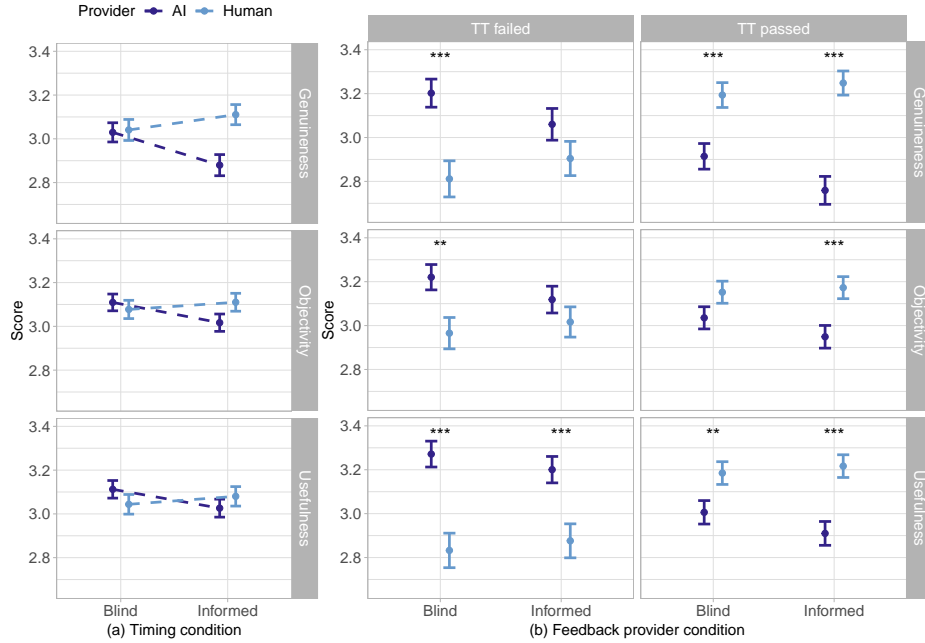


Fig. 4. Student perceptions towards feedback content per dimension (*Genuineness*, *Objectivity*, *Usefulness*) by interactions with Feedback provider (Human/AI), Timing (Blind/Informed) and Turing Test results (TT failed/TT passed). Scores: Strongly disagree = 0, Disagree = 1, Neutral = 2, Agree = 3, Strongly agree = 4. Significance codes: *** < 0.001, ** < 0.01, * < 0.05.

content dimension (*Objectivity*, *Usefulness*, and *Genuineness*) to analyze how the interaction between feedback provider (Human/AI), timing (Blind/Informed) and Turing Test results (TT Passed/TT Failed), and possible covariates (course, age, gender) influence the dimension scores. The individual variation between students was modeled as a random effect. The analysis of the models detected no effect of course, age and gender on the students' perceptions across all the dimensions.

Interestingly, in the follow-up analysis of the means⁴, distinct patterns emerged between perceptions towards AI-generated and human feedback. Students tended to increase their scores for the feedback provided by human evaluators after being informed about the feedback provider's identity. However, the change was not statistically significant in all the content evaluation dimensions (Fig. 4(a), Human condition). At the same time, for AI-generated feedback, the students tended to decrease their scores in all the categories evaluating feedback content after being informed that AI-generated it. The change was statistically significant for the *Genuineness* dimension ($p < 0.01$), but not for the *Objectivity* and

⁴ *emmeans* R package. Multiple comparisons were adjusted using the Tukey method.

Usefulness dimensions (Fig. 4(a), AI condition). These results did not correlate with the Turing Test results.

Subsequently, we analyzed the differences between content evaluation of the human-created and AI-generated feedback for both timing conditions (Blind/Informed) separately. The MLM analysis indicated a correlation with the Turing Test results in this case. Based on the follow-up analysis of differences in the means, we found distinct patterns that emerged between perceptions towards AI-generated feedback and human feedback for students who succeeded and failed to guess the feedback provider identity correctly (TT passed/TT failed). More specifically, in blind conditions, students who failed the TT evaluated AI feedback significantly higher than humans among all dimensions. However, after being informed about the provider identity, they reduced their scores for AI and increased their scores for human feedback, such that the difference was not significant anymore for the *Genuineness* and *Objectivity* dimensions (Fig. 4(b), TT failed condition). Conversely, in blind conditions, students who passed the TT rated AI feedback lower than human’s. However, the difference was significant for the *Genuineness* dimension only. After being informed about the provider identity, they reduced their scores for AI and increased their scores for human feedback, such that the difference became significant in all dimensions (Fig. 4(b), TT passed condition).

4.3 Bias towards AI as a feedback provider (RQ3)

To explore how the feedback provider identity influences students’ perceptions of the provider *Credibility*, we fit a mixed linear model³. More specifically, we modeled an interaction between feedback provider *Credibility* and feedback provider identity (Human/AI), along with possible covariates (course, age, gender, and TT results), while representing the individual variation between students as a random effect. The results revealed a significant positive correlation of human identity with feedback provider *Credibility* ($p < .001$) and a negative correlation with participants’ age ($p < .01$). We found no significant correlation with the course, gender, and ability to guess the feedback provider identity. The results are presented in Fig. 5. The follow-up analysis indicated a substantial effect (Cohen’s $d = 0.88$) of the feedback provider identity on feedback provider *Credibility*, indicating that feedback from human providers ($\mu = 3.29, \sigma = 0.77$) is perceived as significantly more credible than feedback generated by AI ($\mu = 2.25, \sigma = 0.86$).

5 Discussion and Implications

AI has the potential to generate personalized and timely feedback at scale. Consequently, there has been an increase in research on this topic, mainly focusing on assessing feedback quality, missing the social-emotional aspects of AI adoption. In our study, we investigated students’ perceptions of AI-generated feedback, focusing on discovering potential biases against AI as a feedback provider.

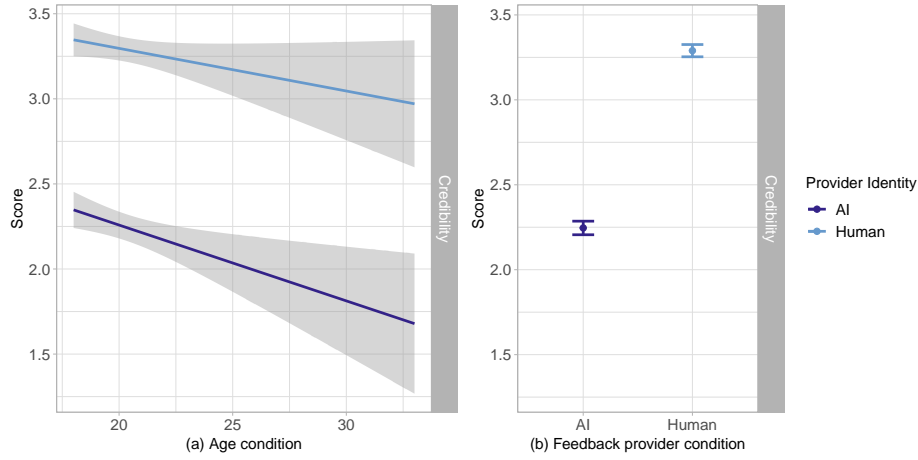


Fig. 5. Student perceptions towards feedback provider *Credibility* by interactions with provider identity (Human/AI) and age. (a) Older students exhibited lower perceived *Credibility* for both feedback providers, yet human providers consistently got significantly higher evaluations. (b) Students evaluated humans as significantly more credible feedback providers than AI. Scores: Strongly disagree = 0, Disagree = 1, Neutral = 2, Agree = 3, Strongly agree = 4.

Quality of AI-generated feedback There is no doubt that feedback quality is a significant factor in students’ perceptions of AI-generated feedback. Our results, however, emphasize that some factors influencing students’ perceptions of AI-generated feedback are not directly related to the feedback content quality (based on RQ2 results). Indeed, students chose to decrease their evaluation of the same feedback content after being informed it was AI-generated. The decrease was significant for the *Genuineness* dimension, measured by *Authentic* and *Sincere* terms, and can be explained by evaluating qualities traditionally attributed to humans. However, the trend of decreased evaluations in the *Objectivity* and *Usefulness* dimensions is more complex to interpret. One possible explanation is that students struggled to assess the feedback’s correctness. In such situations of uncertainty and vulnerability, human evaluation of objectivity and usefulness of the information depends mainly on their trust in the information provider [24]. In our study, the students were initially manipulated to believe that the teaching team they trusted was a feedback provider, so they linked their high trust in educators (evidenced by the results of RQ3) to their evaluation of the feedback content. Once the true AI identity of the feedback was disclosed, their lower trust in AI as a feedback provider influenced their feedback assessment, leading to reduced scores in these dimensions.

The observed correlation between high evaluations of AI-generated feedback and students failing the Turing Test might be attributable to the quality of the feedback itself. In this scenario, the superior quality of the AI-generated feedback might have influenced both the students’ inability to correctly identify the

feedback provider and their positive evaluation of the feedback content. Alternatively, the capability of models like GPT to produce human-like text could have impacted the students’ Turing Test outcomes, reinforcing their belief that the feedback was generated by the teaching team, and leading to higher feedback content evaluations independent of actual feedback quality.

Our future research aims to explore the agreement between human and AI-generated feedback and how each feedback type expert evaluation correlates with students’ perceptions of feedback content. To this end, our next step is labeling the feedback corpus we collected by educational experts with respect to similarities between AI-generated and human feedback and their compliance with elaborated feedback components [14].

Transparency and Ethics Echoing the findings of [25], our study also reveals a strong preference for human guidance over AI-generated suggestions (based on RQ3 findings), indicating a fundamental human inclination for personal interaction and judgment. In this light, AI-powered system designers might be tempted to conceal or blur the AI nature of feedback providers, justifying this by the system’s ability to deliver high-quality feedback. However, we contend that such practices are unethical. Transparency regarding AI involvement in feedback generation is crucial. One reason is the variability in feedback quality between controlled test environments and actual educational settings [42]. Another reason is the educational value of teaching students to recognize and handle potential AI errors. Educating students to critically assess feedback from accessible but imperfect AI technologies like ChatGPT prepares them to continue leveraging such tools for lifelong learning and their future professional environments, enhancing their critical thinking and self-regulated learning skills (see AI literacy [21] and student feedback literacy [5]).

Human-Mediated Automated Feedback (Human-in-the-Loop) A promising approach to addressing algorithm aversion and potential AI inaccuracies involves integrating human educators’ input into AI-based systems. This method entails AI systems initially generating feedback, which human experts review and validate before releasing it to students. Such a hybrid model could efficiently combine AI’s speed and scalability with human expertise and ability to consider the broader learning environment (e.g., social, emotional, and psychological aspects of the learning process) and ensure error-free feedback, while saving teachers time and effort. However, this strategy necessitates new competencies for instructors, broadly categorized under the umbrella of “teacher feedback literacy” [3,5]. The evolving landscape of educational technology, particularly with freely available tools, demands educators develop skills to effectively integrate automated feedback into their pedagogical practices [4]. This includes critically assessing and tailoring AI-generated feedback to meet the specific needs of individual students or student groups and comprehending the potential and limitations of AI tools, especially LLM, e.g., ChatGPT. [43,44]. This comprehension includes critically evaluating the reliability and accuracy of the information provided by AI, being aware of possible model biases, and understanding how these factors might impact educational outcomes. Moreover, recent studies have highlighted the im-

portance of enhancing professional development programs for teachers, focusing on increasing their epistemic understanding of AI [24,19].

Conclusion To summarize, our study analyzing 457 student responses in actual learning contexts gave us a detailed and accurate understanding of student responses to human and AI-generated feedback, a depth that synthetic scenarios might not achieve. We found that students’ feedback evaluations were influenced by their knowledge of the feedback provider’s identity. Students tended to rate human feedback slightly higher after being informed about the provider, whereas AI-generated feedback was rated lower, especially regarding *Genuineness* where the decrease was significant. Furthermore, the results of the Turing Test had a notable correlation with feedback perception. Students who failed the Turing Test rated AI-generated feedback higher than human feedback, while those who passed the test preferred human-generated feedback. A significant finding of the study was the influence of feedback provider identity on the perceived credibility of the feedback. Humans as feedback providers were consistently rated as more credible compared to AI. This underscores the prevailing preference for human feedback in educational settings and highlights the complexities of integrating AI tools into educational environments. Overall, our study provides valuable insights into how students perceive and evaluate AI versus human feedback, which is crucial for the effective integration of Generative AI in education.

Acknowledgments. This project was funded by the Swiss State Secretariat for Education, Research and Innovation (SERI).

References

1. Australian Government Department of Education: Student experience survey. [https://www.qilt.edu.au/surveys/student-experience-survey-\(ses\)](https://www.qilt.edu.au/surveys/student-experience-survey-(ses)) (2022)
2. Bewersdorff, A., Seßler, K., Baur, A., Kasneci, E., Nerdel, C.: Assessing student errors in experimentation using artificial intelligence and large language models: A comparative study with human raters. *Computers and Education: Artificial Intelligence* **5**, 100177 (2023)
3. Boud, D., Molloy, E.: Rethinking models of feedback for learning: the challenge of design. *Assessment & Evaluation in Higher Education* **38**, 698–712 (9 2013)
4. Buckingham Shum, S., Lim, L.A., Boud, D., Bearman, M., Dawson, P.: A comparative analysis of the skilled use of automated feedback tools through the lens of teacher feedback literacy. *International Journal of Educational Technology in Higher Education* **20**, 40 (7 2023)
5. Carless, D., Winstone, N.: Teacher feedback literacy and its interplay with student feedback literacy. *Teaching in Higher Education* **28**, 150–163 (1 2023)
6. Cavalcanti, A.P., Barbosa, A., Carvalho, R., Freitas, F., Tsai, Y.S., Gašević, D., Mello, R.F.: Automatic feedback in online learning environments: A systematic literature review. *Computers and Education: Artificial Intelligence* **2**, 100027 (1 2021)
7. Cukurova, M., Luckin, R., Kent, C.: Impact of an Artificial Intelligence Research Frame on the Perceived Credibility of Educational Research Evidence. *International Journal of Artificial Intelligence in Education* **30**(2), 205–235 (6 2020)

8. Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y.S., Gašević, D., Chen, G.: Can Large Language Models Provide Feedback to Students? A Case Study on ChatGPT. In: 2023 IEEE International Conference on Advanced Learning Technologies (ICALT). pp. 323–325 (2023)
9. Dietvorst, B.J., Simmons, J.P., Massey, C.: Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General* **144**(1), 114–126 (2015)
10. Gabbay, H., Cohen, A.: Combining LLM-Generated and Test-Based Feedback in a MOOC for Programming. In: Proceedings of the Eleventh ACM Conference on Learning @ Scale (2024), to be published
11. Gaube, S., Suresh, H., Raue, M., Merritt, A., Berkowitz, S.J., Lerner, E., Coughlin, J.F., Gutttag, J.V., Colak, E., Ghassemi, M.: Do as AI say: susceptibility in deployment of clinical decision-aids. *NPJ digital medicine* **4**(1), 31 (2021)
12. Gibbs, G., Simpson, C.: Conditions Under Which Assessment Supports Students’ Learning. *Learning and teaching in higher education* **1**, 3–31 (2005)
13. Hattie, J.: The applicability of Visible Learning to higher education. *Scholarship of Teaching and Learning in Psychology* **1**, 79–91 (3 2015)
14. Hattie, J., Timperley, H.: The Power of Feedback. *Review of Educational Research* **77**, 81–112 (3 2007)
15. Henderson, M., Ryan, T., Boud, D., Dawson, P., Phillips, M., Molloy, E., Mahoney, P.: The usefulness of feedback. *Active Learning in Higher Education* **22**, 229–243 (11 2021)
16. Hirunyasiri, D., Thomas, D.R., Lin, J., Koedinger, K.R., Alev, V.: Comparative Analysis of GPT-4 and Human Graders in Evaluating Praise Given to Students in Synthetic Dialogues. *arXiv preprint arXiv:2307.02018* (2023)
17. Keuning, H., Jeuring, J., Heeren, B.: A Systematic Literature Review of Automated Feedback Generation for Programming Exercises. *ACM Transactions on Computing Education* **19**, 1–43 (3 2019)
18. Kim, J., Merrill, K., Xu, K., Sellnow, D.D.: My Teacher Is a Machine: Understanding Students’ Perceptions of AI Teaching Assistants in Online Education. *International Journal of Human–Computer Interaction* **36**(20), 1902–1911 (12 2020)
19. Kizilcec, R.F.: To advance ai use in education, focus on understanding educators. *International Journal of Artificial Intelligence in Education* **34**(1), 12–19 (2024)
20. Lin, J., Dai, W., Lim, L.A., Tsai, Y.S., Mello, R.F., Khosravi, H., Gasevic, D., Chen, G.: Learner-centred Analytics of Feedback Content in Higher Education. In: LAK23: 13th International Learning Analytics and Knowledge Conference. pp. 100–110 (2023)
21. Long, D., Magerko, B.: What is AI Literacy? Competencies and Design Considerations. In: Proceedings of the 2020 CHI conference on human factors in computing systems. pp. 1–16. ACM (4 2020)
22. McCroskey, J.C., Teven, J.J.: Goodwill: A reexamination of the construct and its measurement. *Communications Monographs* **66**(1), 90–103 (1999)
23. Narciss, S.: Designing and Evaluating Tutoring Feedback Strategies for digital learning environments on the basis of the Interactive Tutoring Feedback Model. *Digital Education Review* **13**, 7–26 (2013)
24. Nazaretsky, T., Ariely, M., Cukurova, M., Alexandron, G.: Teachers’ Trust in AI-powered Educational Technology and a Professional Development Program to Improve It. *British Journal of Educational Technology* **53**, 914–931 (7 2022)
25. Nazaretsky, T., Cukurova, M., Alexandron, G.: An Instrument for Measuring Teachers’ Trust in AI-Based Educational Technology. In: LAK22: 12th International Learning Analytics and Knowledge Conference. pp. 56–66. ACM (3 2022)

26. Pankiewicz, M., Baker, R.S.: Large Language Models (GPT) for automating feedback on programming assignments. arXiv preprint arXiv:2307.00150 (2023)
27. Pardos, Z.A., Bhandari, S.: Learning gain differences between ChatGPT and human tutor generated algebra hints. arXiv preprint arXiv:2302.06871 (2023)
28. Ragot, M., Martin, N., Cojean, S.: AI-generated vs. Human Artworks. A Perception Bias Towards Artificial Intelligence? In: Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems. pp. 1–10. ACM (4 2020)
29. Rajpurkar, P., Irvin, J., Zhu, K., Yang, B., Mehta, H., Duan, T., Ding, D., Bagul, A., Langlotz, C., Shpanskaya, K., et al.: CheXNet: Radiologist-Level Pneumonia Detection on Chest X-Rays with Deep Learning. arXiv preprint arXiv:1711.05225 (2017)
30. Roberson, Q.M., Stewart, M.M.: Understanding the motivational effects of procedural and informational justice in feedback processes. *British Journal of Psychology* **97**, 281–298 (8 2006)
31. Ryan, T., Henderson, M., Phillips, M.: Feedback modes matter: Comparing student perceptions of digital and non-digital feedback modes in higher education. *British Journal of Educational Technology* **50**(3), 1507–1523 (2019)
32. Shute, V.J.: Focus on Formative Feedback. *Review of Educational Research* **78**, 153–189 (3 2008)
33. Steiss, J., Tate, T., Graham, S., Cruz, J., Hebert, M., Wang, J., Moon, Y., Tseng, W., Warschauer, M., Olson, C.B.: Comparing the quality of human and ChatGPT feedback of students’ writing. *Learning and Instruction* **91**, 101894 (2024)
34. The Office for Students: Annual report and accounts 2022-23. <https://www.officeforstudents.org.uk/publications/annual-report-and-accounts-2022-23/> (2023)
35. Tormey, R.: Rethinking student-teacher relationships in higher education: a multidimensional approach. *Higher Education* **82**, 993–1011 (11 2021)
36. Tsai, Y.S., Rates, D., Moreno-Marcos, P.M., Muñoz-Merino, P.J., Jivet, I., Scheffel, M., Drachsler, H., Kloos, C.D., Gašević, D.: Learning analytics in European higher education—Trends and barriers. *Computers & Education* **155**, 103933 (2020)
37. Van De Ridder, J.M., Stokking, K.M., McGaghie, W.C., Ten Cate, O.T.J.: What is feedback in clinical education? *Medical Education* **42**, 189–197 (1 2008)
38. Winstone, N., Carless, D.: *Designing Effective Feedback Processes in Higher Education: A Learning-Focused Approach*. Routledge (2019)
39. Winstone, N.E., Nash, R.A., Rowntree, J., Parker, M.: ‘It’d be useful, but I wouldn’t use it’: barriers to university students’ feedback seeking and recipience. *Studies in Higher Education* **42**, 2026–2041 (11 2017)
40. Wisniewski, B., Zierer, K., Hattie, J.: The Power of Feedback Revisited: A Meta-Analysis of Educational Feedback Research. *Frontiers in Psychology* **10** (1 2020)
41. Yan, L., Sha, L., Zhao, L., Li, Y., Martinez-Maldonado, R., Chen, G., Li, X., Jin, Y., Gašević, D.: Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology* **55**(1), 90–112 (2024)
42. Yin, M., Vaughan, J.W., Wallach, H.: Understanding the Effect of Accuracy on Trust in Machine Learning Models. In: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems. pp. 1–12. ACM (5 2019)
43. Zhai, X.: ChatGPT for Next Generation Science Learning. *XRDS: Crossroads, The ACM Magazine for Students* **29**, 42–46 (3 2023)
44. Zhai, X., Nehm, R.H.: AI and formative assessment: The train has left the station. *Journal of Research in Science Teaching* **60**, 1390–1398 (8 2023)