How does the ‘source’ affect the ‘message’? Looking at feedback perception and uptake for perceived feedback from Teachers, Peers, and Generative AI

Rik D.W.H. Helder
Department of Psychology, University of Twente
Educational Psychology: 202000384
Dr. Mohammadreza Farrokhnia & Dr. Pantelis Papadopoulos
January 22nd, 2025

Abstract

Feedback is an integral part of our educational system, and its effectiveness depends on a number of factors, the way feedback is perceived, the way it is incorporated, and the personal stance of the individual receiving the feedback. Generative AI (GenAI) is a tool that is considered more and more when talking about new feedback sources for its ability to provide near instant and personalised feedback.

The objective of this research is to investigate the effects of perceived feedback source- being either a peer, a teacher, or GenAI -on feedback perception and uptake, when taking feedback and AI literacy into account. The target group is students of higher education.

An experiment was set up with a between-subjects mixed methods design to investigate the influence of perceived feedback source on feedback perception and uptake. The sample consisted of 31 (recently graduated) students of higher education (mean age = 21.9). Participants wrote an argumentative essay, on which they were told they would receive feedback from one of three sources, either GenAI, a teacher, or a peer. The participants were randomly divided amongst the perceived feedback sources, but in actuality all feedback was generated by AI using different prompts. After the participants received their feedback, they revised their essay using said feedback and filled in a Feedback Perception Questionnaire (FPQ). The essays were coded using a coding scheme and the data was analysed using various methods.

Overall, there seems to be no relationship between feedback source and feedback uptake, this is possibly due to the fact that the feedback source was the same (GenAI) for each condition, which would result in a similar level of quality of feedback for all participants. The results showed significant effects of feedback source on feedback perception, these differences in feedback perception seem to indicate a bias against GenAI feedback. Interesting to note is that the bias against GenAI feedback does not lead to differences in essay revision quality.

Introduction

Providing feedback to students is an integral part of our education system, and effective feedback is key to the learning process (Poulos & Mahony, 2008). Feedback allows students to compare their performance to the set learning goal and to see whether they have achieved their own goals. Receiving regular and effective feedback also improves learning outcomes, stimulates self-regulation, and allows for an environment which fosters academic growth. Without effective feedback, errors or mistakes remain uncorrected, the desired performance is not reinforced, and progress slows down (Schartel, 2012).

Teachers are commonly the primary source for providing feedback to students. This has the advantage of a teacher usually having more knowledge on the subject they are teaching, and their feedback leads to large improvements in writing (Yang et al., 2006). For a teacher, providing feedback is quite time consuming, as they usually have many students to provide feedback to, which leads to a reduced net amount of feedback which each student receives from a teacher. To combat this, different sources for feedback have been explored and two feedback sources that show promise are peers and Generative AI (GenAI). Peer feedback allows students to give feedback on each other, leading to increased student autonomy (Yang et al., 2006). GenAI tools have risen to popularity in the last few years for their ability to answer and respond to complex tasks or questions, whilst allowing the user to pose their questions in layman's terms. An example of such a GenAI tool is ChatGPT, which is a GenAI tool that uses a GPT model. Within the context of education, GenAI has started a conversation about the possible value it may have for the current academic system (Lee & Song, 2024). GenAI can provide feedback to student work within seconds, which can then be checked by a teacher to ensure the quality of the feedback. The idea behind this is that it may reduce the amount of time required to provide adequate feedback per student (Mizumoto & Eguchi, 2023), though teachers would still need to take the time to read through the student work to check if the feedback is applicable.

Research has been done comparing the different sources of feedback amongst each other, and it has been found that the sources provide different levels of feedback and focus on different aspects of a piece of work. Banihashem et al. (2024) found that peers focus their feedback on the content of essays and the identification of the problem, whilst GenAI provided more descriptive feedback on how the essay was written. Mizumoto & Eguchi (2023) found that GenAI has a certain level of accuracy and reliability compared to teacher evaluations and suggests that GenAI could be of value when used as support for providing feedback. Reynolds et al. (2021) found that students performed better on some essays when they were told the feedback came from an automated writing evaluation (AWE) website called *PaperRater* (PaperRater, 2025), rather than the group who was told that the feedback came from a teacher, even though both groups received AWE feedback. This result could be due to a novelty effect of the AI, which could explain why the latest essay showed the opposite result. Though there are differences in feedback provided by teachers and GenAI, there does not seem to be a

significant difference in the learning process when comparing these feedback sources. There seems to be a split preference amongst students between AI provided feedback or teacher provided feedback (Escalante et al., 2023).

Whilst the previously mentioned studies have compared the effectiveness of different feedback sources amongst each other, a gap remains where all three of the sources, their effectiveness, and the perception of feedback, are compared in a single study. Comparing these three feedback sources in the same study opens up the possibility to also look at new impacting factors. Two factors which can be looked at are how the perceived feedback source impacts feedback perception and uptake, or whether factors such as feedback literacy or AI literacy play a mediating role on feedback perception and uptake. Feedback literacy can be understood as a student's ability to optimise the benefits of feedback opportunities (Nieminen & Carless, 2023), which differs per person and can influence feedback uptake regardless of the feedback content. AI literacy, as defined by Ng et al. (2021), refers to a person's ability to live, learn, and work in our digital world through AI-driven technologies. When someone has too much trust in AI or lacks trust in AI it leads to risks in the human AI interaction (Amoozadeh, 2024). Some research has been done to whether people are able to discern between AI generated and human made content, whether this is feedback, or chat messages or artworks (Reynolds et al., 2021; Ramu et al., 2024; Bellaiche et al., 2023). The results of these studies show that most of the time humans are not able to consistently identify which content is made by humans or AI. Yet people still prefer either AI generated feedback or teacher provided feedback (Escalante et al., 2023), which suggests that there is a different reason than the content of feedback to explain why these preferences exist. A bias towards a certain source of feedback may influence the uptake of said feedback, which then might be influenced by AI literacy, when the feedback source is GenAI. Thus, feedback literacy and AI literacy are expected to influence feedback uptake and perception and will be considered within this research.

This research paper will try to answer the research question: how does the perceived feedback source impact feedback perception and uptake in university students, regarding perceived teacher, perceived peer, or GenAI feedback, when simultaneously accounting for Feedback and AI literacy? This study will be accomplished by asking university students to write a short argumentative essay, on which they will be told they received feedback from either a teacher, another student, or AI, whilst in reality all receiving feedback from AI. Afterwards they get the opportunity to process the feedback and will be asked survey questions regarding their perception of the feedback. They will also receive a questionnaire regarding their AI literacy and feedback literacy level. The essays and survey answers will be analysed to answer the research question.

Theoretical framework

Feedback perception

Feedback perception is classified as the way a person interprets, evaluates, and makes sense of feedback they received. It encompasses emotional and cognitive responses to feedback and plays a crucial role in determining whether feedback is accepted, understood, and utilized effectively (Strijbos et al., 2021). The way feedback is perceived directly influences the way it is incorporated in the work of a student. For instance, feedback that is perceived as fair and from a trustworthy source is more likely to be considered and used, while feedback that is perceived as unfair and from an unreliable source the opposite effect is true (Skagerberg et al., 2008). The setting in which feedback is provided also influences the perception of said feedback, when feedback is provided to a group of students, the message of the feedback may be lost due to the perception of relevance of the feedback to an individual student in the group (Hattie & Timperley, 2007). Expectation of feedback also influences how it is seen, feedback that is similar to the expected feedback is seen as clearer and leads to higher satisfaction (Albright & Levy, 1995).

Feedback uptake

Feedback uptake refers to the process by which individuals engage with and apply the feedback they receive to improve their performance or learning process (Carless & Boud, 2018). This process is influenced by several different factors, such as feedback literacy, feedback quality, acceptance of feedback, and feedback perception. Feedback uptake depends on individual factors of a student, their previous experiences and personal characteristics, like feedback literacy. The way students respond to feedback varies within disciplines, curricula and contextual settings (Carless & Boud, 2018). Successful feedback uptake is critical for achieving learning and performance goals, as it transforms feedback into actual change in behaviour. Confidence, motivation, and feedback literacy play an important role in the success of feedback uptake. Feedback uptake can be measured by comparing the work before and after a feedback intervention has taken place (Nozoori et al., 2016), but this does not show the complete picture, as feedback can be accepted, modified, or rejected by a student while still being considered “uptaken” (Hattie & Temperley, 2007).

Feedback literacy

Feedback Literacy, as defined by Sutton (2012) is a set of generic practices, skills, and attributes in a series of learning practices. According to Sutton (2012), individuals experience and respond differently to the dimensions of feedback literacy, which makes becoming feedback literate a complex and challenging process. Carless and Boud (2018) define feedback literacy as the understandings, capacities, and dispositions needed to make sense of information and use it to enhance work or learning strategies. Feedback literacy can also be considered as the capacity of teachers and students to optimise the benefits of feedback opportunities (Nieminen & Carless, 2022). In this research, the

focus of feedback literacy lies on the ability to understand, interpret, and effectively use feedback to improve one's own educational or learning process. Feedback literacy is a crucial element when it comes to perceiving and uptaking feedback as it indicates a person's ability to make use of received feedback. When a person has a high feedback literacy they tend to realise and actively work towards their own active position in the feedback process, are constantly working to improve their own judgements when it comes to education, and manage affect in positive ways (Carless & Boud, 2018). Provided that AI is an upcoming potential source of feedback, people with higher feedback literacy are more likely to be able to use new and unfamiliar feedback more effectively as they take a more active stance and work to improve their own judgements (Carless & Boud, 2018). In this research, besides viewing taking an active stance, improving own judgements, and managing affect as indicators for feedback literacy, using feedback information and providing feedback are also considered to be a part of feedback literacy (Dawson et al., 2023). Use feedback information indicates how well a person uses the feedback in practice and provide feedback indicates how well a person can provide feedback on others.

AI literacy

AI Literacy can be defined as the ability to live, learn, and work in the digital world using AI-driven technologies (Ng et al., 2021). Ng et al. (2021) divide this definition into four subcategories, namely: know and understand AI, use and apply AI, evaluate and create AI, and AI ethics. The way someone scores on these categories gives an insight into how well they understand AI and to what extent they know how to properly use its content. These four subcategories can be ranked from lowest literacy level to the highest in a Bloom's taxonomy type figure (Carolus et al., 2023). Where knowing and understanding AI indicate the ability to understand and remember information, being able to use it in a new scenario and in different contexts. Use and apply AI signify the ability to organise, compare, decompose, and abstract an AI problem. AI Ethics signifies the ability to take a stance and justify the position when it comes to AI statements. Evaluate and Create AI signify the highest level of ability, signifying the ability to produce new or original AI programs.

Methods

Sample

In total 31 participants finished this study with a mean age of 21.90 (sd = 1.92). Participants were found based on voluntary responses and was a convenience sample consisting of students the researcher knew. The participants were contacted via different channels of student networks and by snowball sampling. Besides this the participants were informed that they could win one of five €10.00 gift cards if they finished the entire experiment. Other participants were promised to receive a number of Sona credits, a form of credit which is required for Psychology students at the University of Twente to gather in order to receive their diploma. The inclusion criteria for the participants were being a current student of higher education or having graduated from higher education within the past year, able to provide consent, and being over the age of 18. The only exclusion criteria was if the participant was unable to write and understand English to a near fluent level. The participants were randomly divided into the three conditions, 9 perceived the feedback to come from a teacher (P-Teacher), 11 perceived the feedback to come from a peer (P-Peer), and 11 received feedback from GenAI (F-GenAI). In Table 1 an overview of the demographics, namely gender, nationality, and educational level can be found.

Table 1

Demographic data of the sample

Demographic	Total sample		P-Teacher		P-Peer		F-GenAI	
	n	%	n	%	n	%	n	%
Gender								
Female	18	58.06	6	66.67	6	54.55	6	54.55
Male	13	41.94	3	33.33	5	45.45	5	45.45
Other	0	0	0	0	0	0	0	0
Nationality								
Dutch	23	74.19	5	55.56	10	90.91	8	72.73
German	5	16.13	2	22.22	1	9.09	2	18.18
Other	3	9.68	2	22.22	0	0	1	9.09
Educational level								
University Bachelor	17	54.84	5	55.56	7	63.64	5	45.45
University Master	12	38.71	3	33.33	3	27.27	6	54.55
Higher education (Dutch HBO)	2	6.45	1	11.11	1	9.09	0	0

Note. Groups are perceived teacher feedback (P-Teacher), perceived peer feedback (P-Peer), and GenAI feedback (F-GenAI). Average age for condition P-teacher was 21.78 (sd: 1.39), for condition P-Peer it was 21.91 (sd: 2.39), and for condition P-AI it was 22.00 (sd: 1.95)

Design

The current study employed a pre-post, mixed-methods, between subjects approach to investigate the impact of the perceived source of feedback on feedback perception and uptake. The dependent variables were perception and uptake, and the independent variable was the perceived source of feedback (Teacher, Peer, or GenAI). Possible moderating variables that were considered were feedback literacy and AI literacy. All feedback was provided by GenAI and for the peer condition the GenAI was tasked with changing the tone of the feedback to resemble that of a peer. This was done to assure a similar level of quality between the feedback sources.

Materials

Information sheet and informed consent

An information sheet and informed consent form were created based on the Informed Research Participation Procedure by domain Humanities & Social Sciences of the Faculty of Behavioural, Management and Social Sciences at the University of Twente. The information sheet and consent form can be found in Appendix A.

Literacy surveys

In order to identify the literacy levels of participants in both the territory of Feedback and AI two scales were found and used. The Feedback Literacy Behaviour Scale (FLBS) by Dawson et al. (2023) was used to identify the level of feedback literacy in each participant. The FLBS contains five scales with each four or five items. The participants answer each item on a six-point likert scale (never, almost never, rarely, sometimes, almost always, and always). The scales are *Seek feedback information*, with questions like "I seek out examples of good work to improve my work."; *Make sense of information*, with items such as "When deciding what to do with comments, I consider the credibility of their sources."; *Use feedback information*, with for example "I check whether my work is better after I have acted on comments."; *Provide feedback information*, which contains questions such as "When commenting on the work of others, I provide constructive criticism."; and finally, *Manage effect*, which includes items like "I make use of critical comments even if they are difficult to receive.".

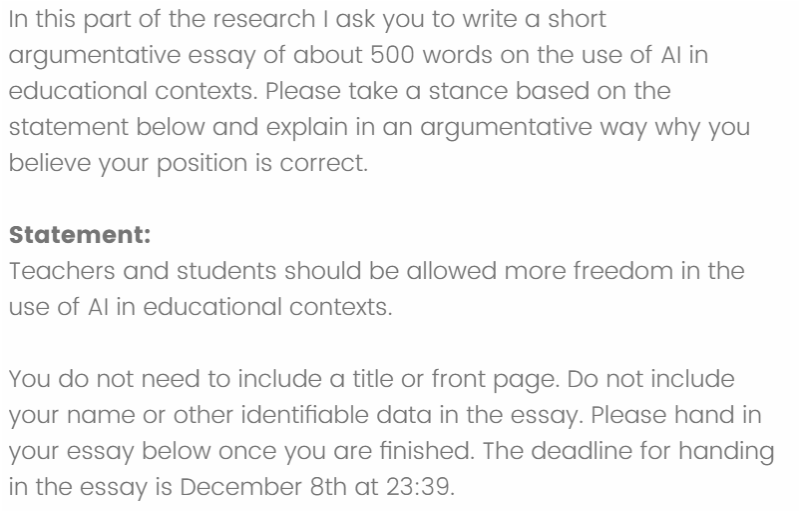
The second literacy survey was on AI literacy. For this the Meta AI Literacy Scale (MAILS) was used (Carolus et al. 2023). The MAILS contains a total of 34 items, of which only the first 22 were used as those questions pertain to AI literacy, the final 12 items measure AI self-efficacy and AI self-competency. The remaining items were adjusted to specify GenAI use and understanding, rather than AI in general. For this, everywhere the original scale mentioned AI was replaced with GenAI. The final scale consists of five subscales with each three to six items. For each item participants indicate how well they can use and understand GenAI on a 11-point Likert scale ranging from 0 (not at all) to 10 (near perfection). The subscales are, *Apply GenAI*, which contains items like "In everyday life, I can interact with GenAI in a way that makes my tasks easier"; *Understand GenAI*, which has items such as "I can assess what advantages and disadvantages the use of GenAI entails"; *Detect GenAI*, which contains items such as "I can tell if I am dealing with an application based on GenAI"; *GenAI Ethics*, which contains, for example "I can weigh the consequences of using GenAI for society"; and finally, *Create GenAI*, which contains items like "I can develop new GenAI applications". All of these items together provide some insight into the level of GenAI Literacy of the participants.

Essay instruction

Participants were provided with a short instruction on how to write their essay. The instruction was deliberately kept short and somewhat vague so there would be differences between participants and how they filled in the essay, and thus what kind of feedback they would receive. Participants were informed that they were requested to write a 500-word argumentative essay on the use of AI in education and were provided a statement toward which they had to take a stance. The exact instruction can be found below in Figure 1.

Figure 1

Essay instruction



Feedback

All feedback was acquired by ChatGPT (using ChatGPT-4o) after providing it with a single Chain of Thought (CoT) prompt. After which half of the essays were sent to receive feedback. This was repeated for the second half of the essays. The exact prompt can be found in Appendix B1. The prompt itself was designed for an (at the time of writing) unpublished study, in which an automatic Chain of Thought (Auto-CoT) approach was used. An Auto-CoT approach explains what the model is supposed to do by providing an extensive step-by-step explanation of how the model is supposed to handle the incoming assignment, whilst also providing an example of expected input and output. This enhances the reasoning capabilities of sufficiently large language models (LLM) by guiding the model to take smaller steps to solve complex questions (Wei et al., 2022). This increases the ability of the model to accurately solve complex tasks such as maths problems, commonsense reasoning and symbolic manipulation. According to Wei et al. (2022) it should, in principle, be applicable to any task that humans can solve through language. Therefore, the choice was made to use an Auto-CoT prompt for the feedback generation. The reason the pre-existing prompt (Appendix B1) was used over designing a new prompt was that this prompt was designed for a very similar context, providing short

feedback to undergraduate students and grading ~500-word argumentative essays on a similar AI related topic and designing a new prompt was deemed outside of the scope of this research.

In this case the prompt starts with defining the role ChatGPT is supposed to take and the context of the assignment, the role was that of a professor of academic writing at the undergraduate level in the context of providing feedback to students who wrote a 500-word argumentative essay. The prompt provides an overview of critique points which ChatGPT is supposed to focus the feedback on, those being a relevant introduction, a clear position in regard to the statement, reasons in support of their position, counterarguments against their position with arguments or evidence to refute those counterarguments, and an effective conclusion. After the critique points ChatGPT is instructed to follow a step-by-step approach to respond. Then an example of an essay is provided, after which each step that needs to be taken is explained in a way that a person would go about providing feedback. Each critique point is mentioned and explained that ChatGPT needs to provide suggestions for its improvement if that part of the essay is lacking. An example of how the feedback could look is provided, in which each critique point is mentioned, and plausible feedback is written based on the example essay.

After putting in the prompt, each essay was sent one by one to ChatGPT and the responses documented. One-third of the essays was randomly selected to receive peer feedback and ChatGPT was provided with both prompts from Appendix B1 and Appendix B2 and with two documents, containing the essay itself and the feedback which was already generated. The prompt of Appendix B2 was chosen from three prompts, this final prompt was used because it was more similar in writing style and kept the structure of the original feedback, which the other prompts did not do to the same extent. The extra prompt was created to change the tone to be more peer-like. The “teacher” feedback and GenAI feedback were not changed.

Feedback perception questionnaire

To investigate feedback perception the Feedback Perception Questionnaire (FPQ) by Strijbos et al. (2021) was used. The FPQ contains seven subscales with each three or six items on how the feedback is perceived. Participants answered each item on a 11-point Likert scale ranging from 0 (fully disagree) to 10 (fully agree). The items on the original FPQ were formulated in the future subjunctive tense, e.g. “I would be willing to improve my performance”, for this research most statements were changed to the present continuous tense, e.g. “I am willing to improve my performance”. The subscale *Acceptance* was forgotten to be included in the final survey which was sent to participants. The used subscales of the FPQ included, *Fairness*, with items such as “I am satisfied with this feedback”; *Usefulness*, with items such as “I consider this feedback helpful”; *Willingness to improve*, with items such as “I am willing to improve my performance”; And *Affect*, which included both positive and negative statements such as “I feel successful because of this feedback” or “I feel offended because of this feedback”.

Debrief

In accordance with the Informed Research Participation Procedure by domain Humanities & Social Sciences of the Faculty of Behavioural, Management and Social Sciences at the University of Twente a debriefing note was made which informed participants that they had been deceived during the experiment. It explained that all feedback was generated using GenAI and explicitly states that participants are still allowed to have their data excluded from the final sample and that they need to reconfirm their consent. The exact debriefing document can be found in Appendix C.

Coding scheme

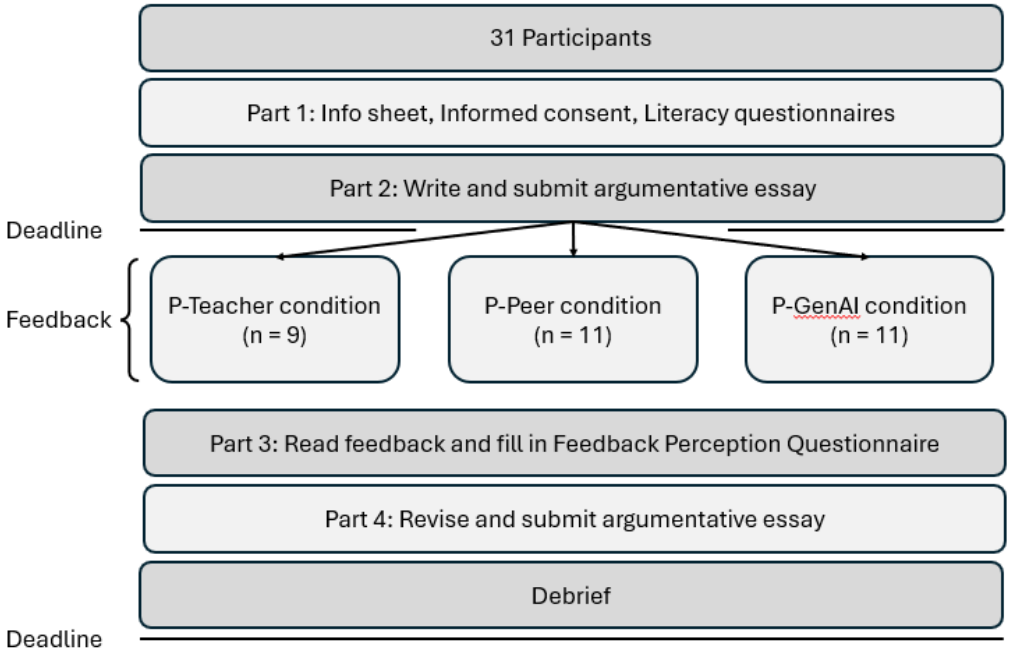
To grade the essays and revised essays the coding scheme designed by Nozoori et al. (2016) was used. The coding scheme provided seven items, for each item an essay could score from zero to two points, thus a perfect essay could receive a maximum of 14 points. The items were, intuitive opinion, claims in favour of the topic, justification of claim(s) in favour of the topic, claims against the topic, justification of claim(s) against the topic, integration of pros and cons, and conclusion.

Procedure

The entire study took place online, participants could decide themselves when they would finish each part if it was before the deadline. Participants were contacted via various platforms and networks, for example via private messages on WhatsApp, WhatsApp group chats, or via LinkedIn. Participants were sent an invitation where they could read more information on the experiment, how much time it would approximately take, and where they could fill in their email address to sign-up for the study. Contact with participants was done via email and to assure anonymity the email addresses filled in the sign-up form were not coupled to the data, instead the participants created a unique id-code which was used to couple the data from different surveys. In Figure 2 an overview of the procedure for participants can be found.

Figure 2

Overview of procedure for participants



After signing up participants would receive an email with the first two parts. Part one consisted of a survey which included the information sheet, which explained the experiment in further detail, see Appendix A1 for the information sheet. Part one also included an informed consent form (Appendix A2), demographic questions, and Literacy questionnaires. After finishing part one, participants were guided to part two, which included the essay instruction and a hand-in form. Participants had up to eight days to finish the first two parts, depending on which they signed up for the experiment.

Once the deadline for parts one and two had passed, all participants received an email stating that they had not been selected to provide peer feedback and were instructed to wait for their feedback, which would take three days before they got it. In the meantime, the essays were fed to ChatGPT to receive feedback using the prompt (Appendix B1). The participants were randomly divided into three groups, P-teacher, P-Peer, and F-GenAI. For the P-Peer group the feedback was revised by ChatGPT based on the prompt in Appendix B2, which changed the tone of the feedback to be more peer-like. The feedback was placed in password protected files on OneDrive with the title being the unique id-code of the participant for whom the feedback was. At the top of each file, before the feedback a note was added which mentioned the type of feedback it was supposed to be (i.e. Teacher, Peer, or GenAI). The password of the feedback file was created based on the demographic data provided by participants during part one.

Participants were informed via email that they could continue with the experiment and finish parts three and four. They were clearly instructed to only read the feedback at first, then do part three

(the Feedback Perception Questionnaire), and only afterwards finish part four (revise and hand-in their essay). After the participants had handed in the revised version of their essay, they were directed to the debriefing document (Appendix C) and were requested to reconfirm their consent. Participants had nine days to finish the final parts.

Once all revised essays had been received the essays were coded using the coding scheme. Both the original essay and the revised versions were analysed separately by one coder. The essays were decoupled from the feedback received so the coder would not be influenced by the feedback source. First all revised essays were coded and then the original essays, the grades for the revised essays were hidden while coding the original versions to reduce as much bias as possible. Once all essays had been coded the results were put in an excel file which could be exported for data analysis.

Data reduction

In total 42 people signed up for the experiment. Out of those 42, one person already had to be excluded as they were not a student of higher education for over a year. Out of the 41 participants that signed up, 32 finished the entire experiment and gave consent to the use of their data. One person handed in the same essay for the revised version, as this person did fill in relatively high scores on the FPQ, it was assumed that they handed in the wrong version of the essay for the revised version, and therefore the data from this participant was excluded from the final data set. The data from the other nice participants were excluded from the final sample as they did not finish the experiment.

Data analysis

To assess how the main independent variable, perceived feedback conditions (P-Peer, P-Teacher, F-GenAI), influenced the dependent variables, feedback perception and uptake, whilst attempting to account for Feedback and AI literacy, multiple analyses were conducted.

Feedback perception was calculated using the scores of the FPQ and averaging them per condition. For feedback perception, the FPQ scores are taken and based on the average score per item (items on a 11-Point Likert scale from 0 to 10), the mean score is then calculated per feedback condition. The items on the subscale *Affect Negative* were negatively phrased, meaning that low scores indicated higher perception. Thus, the scores from this subscale were reverse coded for data analysis. For feedback uptake, or Improvement, the coding scheme scores per essay are taken and the total score of the first essay is subtracted from the total score of the revised essay. The possible scores of the essays range from 0 to 14 and the improvement is also in this range. Originally, some Improvement scores were negative. These negative improvement scores were replaced with 0 improvement scores, as the assumption is that changes in the essays would only yield positive improvement scores and that negative improvement scores were because of errors in coding. For the AI literacy scale, the scores are calculated based on the average score per item (items on a 11-point

Likert scale from 0 to 10). For the feedback literacy scale the scores are calculated based on the average score per item (items on a 6-point Likert scale from 1 to 6).

Descriptive statistics were calculated for the demographic data (age, gender, nationality, and level of education), feedback perception, feedback uptake, feedback literacy, and AI literacy for the total sample and each subgroup. To see which statistical methods would be most suitable for analysing feedback perception and feedback uptake, Shapiro-Wilk tests of normality and Levene's tests of Homogeneity were performed on each group (P-Peer, P-Teacher, and F-GenAI). The Shapiro-Wilk test showed that the distribution of FPQ scores ($W = .97$, $p > .05$) and Improvement scores ($W = .97$, $p > .05$) significantly differed from a normal distribution, which indicates that both variables violate the assumption of normality in the groups. The Levene's test was performed on the groups, which showed that the variances of FPQ ($F(2, 28) = 1.98$, $p = .16$) and Improvement scores ($F(2, 28) = .25$, $p = .78$) between the groups were not equal. This means that the assumption of homogeneity is not violated. Based on the violated assumption of normality, the use of an ANOVA or ANCOVA analysis was ruled out for either variable. Because the assumption of normality was violated in the conditions a non-parametric test needed to be used. For this a Kruskal-Wallis test was selected and a Dunn's test for pairwise comparison was used as a post-hoc analysis. The Kurskal-Wallis test is able to indicate whether there are significant differences in the dependent variable between the conditions and the Dunn's test for pairwise comparison is able to indicate which groups differ if the Kruskal-Wallis test returns a significant difference. The Kruskal-Wallis test does not take covariate variables into account, as feedback literacy and AI literacy are covariate variables in this research, another test needed to be conducted. A Generalized Linear Model with gamma distribution and log link was selected for this, as it was able to take numeric covariate variables into account and their influence on the differences in feedback perception between the conditions.

To analyse feedback uptake, or improvement and its items, the ANOVA assumptions of normality and homogeneity were tested. Because the assumption of normality was violated again, also a non-parametric test was required. The Kruskal-Wallis test was used and a Dunn's test for pairwise comparison was selected again if the Kruskal-Wallis test yielded significant results. Though the Kruskal-Wallis test still cannot take covariate variables into account; a Generalized Linear Model with gamma distribution and log link was again selected to analyse the impact of the feedback source on feedback uptake when accounting for feedback literacy and AI literacy. These analyses could be used to make a conclusion.

Results

First some descriptive statistics of important variables are presented. Table 2 provides an overview of the mean scores and standard deviations of AI Literacy, Feedback Literacy, Feedback Perception Questionnaire (FPQ) results, and Feedback uptake or Improvement for the different conditions. Improvement Original is the original scored improvement for the conditions, before the negative improvement items were replaced with improvement scores of zero.

Table 2

Descriptive data of Literacies, Feedback Perception, and Improvement

Items	**Total sample**		**P-Teacher**		**P-Peer**		**F-GenAI**	
	Mean	*sd*	*Mean*	*sd*	*Mean*	*sd*	*Mean*	*sd*
AI Literacy	5.47	1.11	5.07	.99	6.64	1.32	5.62	.97
Feedback Literacy	4.66	.31	4.69	.23	4.63	.33	4.67	.37
FPQ	7.30	1.25	7.47	1.08	8.19	.60	6.27	1.16
Improvement	2.55	1.59	2.33	1.66	2.82	1.83	2.46	1.37
Original Improvement	2.00	1.93	1.78	2.11	2.73	2.24	1.91	1.58

Note. AI Literacy is the average score of participants over all AI Literacy items. Feedback Literacy is the average score of participants over all Feedback Literacy items. FPQ is the average score of participants over all FPQ items. Improvement is the difference in score between the original essay and the revised version. Original Improvement is the improvement before data modification.

The Kruskal-Wallis test was performed to see whether conditions P-Peer, P-Teacher, and F-GenAI had an influence on FPQ scores. The results showed a significant difference between the conditions $\chi^2(2, n = 32) = 13.109$, $p = .0014$. Which means that there are significant differences between the three groups when it comes to FPQ scores, but it is still unclear between which exact groups the difference lies. A Dunn's test for pairwise comparison was performed as post-hoc analysis, which is supposed to show which specific groups significantly differ from each other. The Dunn's test revealed that the F-GenAI group had significantly lower scores on the FPQ than the P-Peer group (mean: -3.61, $p = .0005$). No significant differences were found between the P-Peer and P-Teacher group ($p > .05$) or between the F-GenAI and P-Teacher group ($p = .085$). Running the same Kruskal-Wallis test on the subscales of the FPQ results can be found in Table 3. The subscale items *Fairness* and *Affect Positive* indicated significant differences between the groups (p's $< .05$). For these items the Dunn's test for pairwise comparison was performed as post-hoc analysis to see which groups differed. For *Fairness* the F-GenAI group scored significantly lower than the P-Peer group (mean: -2.43, $p = .022$); the F-GenAI and P-Teacher groups did not significantly differ ($p = .105$) and the P-

Peer and P-Teacher groups also did not significantly differ (p = .928). For *Affect Positive* the F-GenAI group scored significantly lower than the P-Peer group (mean: -3.12, p = .002); the F-GenAI and P-Teacher groups did not significantly differ (p = .098), and neither did the P-Peer and P-Teacher groups (.398).

Table 3

FPQ subscales Kruskal-Wallis results

FPQ Subscale	**$\chi^2(2, n = 31)$**	***p* value**
Fairness	6.48	.039 *
Usefulness	5.65	.059
Willingness to improve	4.16	.125
Affect Positive	9.89	.007 *
Affect Negative	4.91	.086

Note. Items with an asterisk (*) indicate significant p-values ($p < .05$)

For Improvement a Kruskal-Wallis test was also conducted to check whether the conditions (P-Peer, P-Teacher, and F-GenAI) had an influence on Improvement scores. The Kruskal-Wallis test reported no significant differences between the groups $\chi^2(2, n = 31) = .345$, $p > .05$. The Kruskal-Wallis test was also run on each individual item of Improvement, for those results, see Table 4. None of the items showed significant differences between the groups, indicating that there are no significant differences in improvement when comparing the three different groups F-GenAI, P-Peer, and P-Teacher.

Table 4

Improvement items Kruskal-Wallis results

Improvement item	**χ^2(2, n = 31)**	***p* value**
IntuitiveOpinion	4.03	.133
ClaimsFavour	.25	.881
JustificationClaimsFavour	3.21	.201
ClaimsAgainst	3.03	.220
JustificationClaimsAgainst	1.82	.403
ProsCons	.52	.769
Conclusion	3.49	.174

Besides the Kruskal-Wallis tests, a Generalized Linear Model (GLM) with a Gamma distribution and log link was used to predict the effects of Feedback Literacy, AI Literacy, and feedback group (P-Peer, P-Teacher, F-GenAI) on FPQ scores and Improvement scores. The model showed that when accounting for Feedback Literacy and AI Literacy, feedback group significantly predicted FPQ scores (p's < .05). Because the link used for the GLM was a log function, to accurately portray the results it needs to be 'undone', therefore the exponential of the results is calculated. The FPQ scores of the P-Peer group were relatively about 1.311 times higher than the F-GenAI group (p = .00015). The FPQ scores of the P-Teacher group were relatively about 1.214 times higher than the scores in the F-GenAI group (p = .0067). When accounting for feedback group, neither Feedback Literacy nor AI Literacy significantly predicted FPQ scores (p's > .05).

When running the same model for Improvement, scores of 0, or no improvement, had to be excluded. The model showed that when accounting for Feedback Literacy and AI Literacy, feedback group did not significantly predict Improvement scores (p's > .05). Using the same model, when accounting for feedback groups, neither Feedback Literacy nor AI Literacy were able to significantly predict Improvement scores (p's > .05) either.

Discussion

Summary and interpretation of results

The goal of this research was to investigate the influences of three different perceived feedback sources (P-Peer, P-Teacher, and F-GenAI) on feedback perception and uptake, whilst accounting for differing levels of Feedback and AI Literacy in students of higher education. As represented by the research question, how does the perceived feedback source impact feedback perception and uptake in university students, regarding perceived teacher, perceived peer, or GenAI feedback, when simultaneously accounting for Feedback and AI literacy? The results showed that the different perceived feedback sources seem to have an influence on feedback perception, but not on feedback uptake. This cannot be stated for certain, as the feedback prompt for peer feedback was generated in a different way than the teacher and GenAI feedback and there are differences in literacy scores between the groups. In general, the feedback was perceived positively and the essays improved after the participants received feedback. This suggests that receiving and working with AI generated feedback has a positive impact on the quality of argumentative essays regardless of the perceived source of feedback for students of higher education. When controlling literacy scores, the perception (FPQ scores) difference between the groups becomes larger.

In the case of feedback perception, there are differences in the perception of feedback depending on the perceived source. Feedback provided by GenAI is scored lower than GenAI feedback that is posed as peer feedback. This is the case when taking Feedback and AI Literacy into account and when not taking literacy levels into account. When controlling for literacy levels, GenAI as the source falls further behind and even GenAI feedback that is posed as Teacher feedback scores higher on the FPQ scale. This indicates that there is some skepticism towards AI generated feedback, or that GenAI feedback, when posed as human feedback, is perceived more favourably than when GenAI feedback is presented as the actual source. An extra prompt was used to generate the “peer” feedback and therefore it is not possible to definitively say whether the changes observed in feedback perception are due to the difference in prompting or due to the different perceived feedback source. There was no difference in the feedback prompt between perceived teacher feedback and GenAI feedback and there was a significant difference in feedback perception when controlling for the literacy differences. This does suggest that the difference between the perceived teacher feedback and the GenAI feedback is due to the perceived feedback source. These results are similar to what Nazaretsky et al. (2024) found in their study on feedback perception based on the source. They found that students would decrease their evaluation of the feedback after being told that the feedback was AI generated. They suggest that students have a strong preference for human guidance over AI-generated suggestions for improvement, which could be an explanation of the differences in perception based on the source that were found in this study as well.

When taking a deeper look at the FPQ, more importantly, at its subscales, only the *Fairness* and *Affect Positive* subscales showed significant differences between the conditions. The post-hoc

analysis then showed that the F-GenAI group scored significantly lower on both of these subscales compared to the P-Peer group. No differences were found for the P-Teacher group when compared to either P-Peer or F-GenAI. Though the literacies were not taken into account for these analyses it shows that participants perceived the F-GenAI feedback as less fair and felt less positive because of the feedback than participants in the P-Peer group. Nazaretsky et al. (2024) again found similar results in differences in perception, but not for the same sources. They told students that feedback came from a team of teachers and when they were told the feedback came from AI students would decrease their evaluations in different categories, namely, *Genuineness*, *Objectivity*, and *Usefulness*. Though in the present study no significant differences in perception were found between the P-Teacher and F-GenAI groups. This may be explained by the setup of the different studies, in the present study students were not able to change their perception of the feedback after being told that the feedback came from GenAI.

In terms of feedback uptake, there were no significant differences between the three groups. On average all essays increased their Improvement score by 2 or 2.55 points based on whether data modification was performed. This indicates that even though the perception of the feedback significantly changed based on the perceived feedback source, the actual improvement showed no significant differences between the perceived feedback sources. Even when taking feedback literacy or AI literacy into account no significant relationships between the perceived feedback source and the level of improvement were found. It is also interesting to note that this lack of differences can also be found in the individual items of the coding scheme for improvement. Though this could be explained by the fact that GenAI feedback is more descriptive of how the essay was written, rather than focused on the content of the essay (Banihashem et al., 2024). This could result in broader feedback which makes it difficult for students to implement in a specific part of the essay.

Plenty of research has been done comparing AI generated feedback with peer or teacher feedback (e.g. Dai et al., 2023; Lee & Song, 2024; Banihashem et al., 2024), though few studies have examined the effects, perception, or uptake of feedback when the source of origin was obscured. Reynolds et al. (2021) researched the effects of perceived feedback source and compared a perceived teacher feedback group with a perceived automated writing evaluation (AWE) group. They used a different software based on AI to provide the feedback for both groups but told one group they received feedback from their teacher. They found that the perceived feedback source had an influence on writing performance. They suggest that the attitude of students towards the feedback source either facilitated or impeded their writing performance. The present study is somewhat conflicting with the findings of Reynolds et al. (2021), as here it seems that the perceived feedback source had no influence on the improvement of the essay. An explanation of these differences is the difference in tool which was used to provide feedback, as they used PaperRater (PaperRater, 2025), an online AI based AWE tool and ChatGPT was used for this study.

Limitations

While the results show strong evidence for the effects of the perceived feedback sources, there are some limitations of this study which need to be mentioned. First, a sample size of 31 does come with extra challenges. With such a small sample, it is difficult to extrapolate these results to a more general population of students of higher education. Besides, due to the sample size, the assumption of normality was more likely to be violated, which meant that non-parametric tests needed to be done. In general, non-parametric tests have less statistical power than parametric tests. At the end of the study, it was realised that a Quade's non-parametric ANCOVA could have been performed instead of the current non-parametric tests, which could have yielded different insights. Another issue which could have resulted from the small sample size was that the AI literacy scores differed between the groups, which should not have been the case as the participants were randomly divided into those groups.

Besides the small sample size there was only one coder for the analysis of essays and this coder seemed to have coded inconsistently. Because there was only one coder, there was no interrater reliability and no check on the coding quality. The interpretation of the coding scheme and the strictness of coding seemed to have decreased over the course of coding all essays. At first all revised essays were coded and afterwards all original essays were coded. When coding was finished there were quite a few essays which seemed to have gotten worse on some parts after the revision, though most participants only added more information to their revised essays and did not remove much. This then seems to be impossible as the content of the essay would not reduce when more information is added. This then seems to be caused by the coder and their biases or changes in coding throughout the process. Due to time constraints the coding could not be redone with a second coder to assure a level of interrater reliability. Instead, an attempt was made to reduce the effect of the coding bias by changing the improvement scores of the essays to be zero (no improvement) if they were originally negative (negative improvement), though this is not scientifically backed to reduce or remove the introduced bias.

There are some issues which emerged from the study being completely executed online. First, there was no way of knowing how motivated participants were when they started out versus when they had to revise their essays. This could majorly influence the quality of the revision, as well as the way the participants perceived and incorporated the feedback. Besides this, another issue is that there was no way of knowing whether participants used AI to generate or improve their essay, while they were told not to use AI if they asked the researcher about it, they were not explicitly informed not to use AI in the general instructions. Therefore, some participants could have completed the experiment using AI and there would be no way of excluding their data.

In the end there ended up being two issues regarding the feedback. The first was that several participants reached out to the researcher stating suspicion that they believed that their feedback was AI generated regardless of what the feedback source was. Another set of participants mentioned that

they had some suspicion that the feedback was AI generated but only mentioned so after they had been debriefed, which could have been caused by bias on their part. This suspicion may have consciously, or subconsciously, influenced the way people reacted to the feedback, regarding both feedback perception and feedback uptake. The second issue regarding the feedback was that the "peer" feedback was generated using an additional prompt, but there was no scientific research or design done to use the best prompt to assure that there would be no difference between the "peer" feedback and the GenAI or "teacher" feedback. This leads to the results regarding the perceived peer feedback being inconclusive and less strong, as the feedback may be vastly different, because the prompt used was different. A quick check was done on several feedback texts to check if the content of the feedback was still the same, and while that did seem to be the case, it was not done in a scientifically backed manner.

When sending the Feedback Perception Questionnaire (FPQ) to participants an error was made and the items pertaining to the subscale *Acceptance* were left out. This error was only recognised after the experiment had already run its course and was therefore not able to be rectified. The items on this subscale could have been valuable to see how the feedback from the different sources would have been accepted or rejected by participants.

Finally, the last limitation of the current study is that there was no clear plan on how to deal with participants who did not keep to the essay instructions. One participant wrote an essay that was over 1000 words long, when the instructions stated that the essay should be approximately 500 words in length. Another participant handed in the same essay twice. There was no plan on what to do with participants who did not follow the essay instructions. In the end the data of the participant who handed in the same essay twice was removed from the sample, but the data from the participant who wrote over 1000 words was kept. Arguments for keeping or removing each participant's data can be made either way, as not changing anything is also a way of handling feedback, or that having twice as many words as the aim leaves a lot of room to include the required content.

Future research

Though there are some limitations which should not be overlooked, this research also provides a broad base for possible further research. In general, for similar research it is advised to take the limitations of this study into account and to learn from the mistakes made here. One possibility for future research, given a larger sample, is to dive deeper into one component of this research. This study had a lot of components which all influenced one another and focusing on one or only a few may provide new deeper insights into the effects of specific different variables on feedback perception or uptake. For instance, a deeper focus could be put on the literacy components, diving deeper into how the level of feedback or AI literacy influences the way someone perceives the feedback from different sources could yield informative results regarding the use and trust of AI and what benefits and risks are associated with the use of and trust in AI. Another avenue of research could be to look

into using actual peer and teacher feedback. In this research feedback quality was one of the things that was attempted to be factored out, but using actual peer and teacher feedback could allow for new research regarding the quality of GenAI feedback when compared to the two alternatives.

This study has implications for the way we look at AI generated feedback in the educational world. Whilst the results for this study suggest that GenAI content is currently still looked down upon and perceived less favourably than peer or teacher feedback, it also shows that the perceived feedback source may not impact the implementation of feedback. Further research could be done on how the perception of feedback influences the implementation and quality of improvement and how the perception of AI influences this process when it is the source of feedback.

Finally, another idea for future research would be to investigate how people recognise AI generated content or feedback specifically. As some participants were able to recognise the feedback to be AI generated, through various means, e.g. “the feedback was too nice”, or “the structure of feedback was a giveaway that it was AI generated”. Investigating the way people recognise AI generated content could help develop AI to become indistinguishable from human writing, or it could allow others to realise the way AI generated content can be recognised, which could allow educators to distinguish between original student work and AI generated work.

Conclusion

This study contributes to the world of education and its view on GenAI as a possible source of feedback by investigating the influence of perceived feedback source on feedback perception and uptake. The research question of "how does the perceived feedback source impact feedback perception and uptake in university students, regarding perceived teacher, perceived peer, or GenAI feedback, when simultaneously accounting for Feedback and AI literacy?" was answered to the best of the present studies ability. An experiment was set up in which 31 participants wrote and revised a short argumentative essay when presented with GenAI feedback posed as one of the three different sources. Feedback perception seems to be influenced by the perceived feedback source when accounting for feedback and AI literacy, while feedback uptake does not seem to be influenced by the perceived feedback source when accounting for feedback and AI literacy. These results highlight the need for further research before GenAI becomes broadly used to provide feedback to students of higher education. Future research into the way GenAI works, what biases there are against GenAI, and how humans recognise GenAI content can still be done to further understand how GenAI might be used responsibly to provide effective and desired feedback.

References

Albright, M. D., & Levy, P. E. (1995). The effects of source credibility and performance rating discrepancy on reactions to multiple raters. *Journal of Applied Social Psychology, 25*(7), 577–600. https://doi.org/10.1111/j.1559-1816.1995.tb01600.x

Amoozadeh, M., Daniels, D., Nam, D., Kumar, A., Chen, S., Hilton, M., Ragavan, S. S., & Alipour, M. A. (2024). Trust in Generative AI among Students: An exploratory study. *SIGCSE 2024.* https://doi.org/10.1145/3626252.3630842

Banihashem, S. K., Kerman, N. T., Noroozi, O., Moon, J., & Drachsler, H. (2024). Feedback sources in essay writing: peer-generated or AI-generated feedback? *International Journal of Educational Technology in Higher Education, 21(1).* https://doi.org/10.1186/s41239-024-00455-4

Bellaiche, L., Shahi, R., Turpin, M. H., Ragnhildstveit, A., Sprockett, S., Barr, N., Christensen, A., & Seli, P. (2023). Humans versus AI: whether and why we prefer human-created compared to AI-created artwork. *Cognitive Research Principles and Implications, 8*(1). https://doi.org/10.1186/s41235-023-00499-6

Carolus, A., Koch, M. J., Straka, S., Latoschik, M. E., & Wienrich, C. (2023). MAILS - Meta AI literacy scale: Development and testing of an AI literacy questionnaire based on well-founded competency models and psychological change- and meta-competencies. *Computers in Human Behavior Artificial Humans, 1*(2), 100014. https://doi.org/10.1016/j.chbah.2023.100014

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

Dai, W., Lin, J., Jin, H., Li, T., Tsai, Y., Gašević, D., & Chen, G. (2023). Can large language models provide feedback to students? A case study on ChatGPT. *Institute of Electrical and Electronics Engineers*. https://doi.org/10.1109/icalt58122.2023.00100

Dawson, P., Yan, Z., Lipnevich, A., Tai, J., Boud, D., & Mahoney, P. (2023). Measuring what learners do in feedback: the feedback literacy behaviour scale. *Assessment & Evaluation in Higher Education, 49*(3), 348–362. https://doi.org/10.1080/02602938.2023.2240983

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

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research,*

77(1), 81–112. https://doi.org/10.3102/003465430298487

Lee, S., & Song, K. (2024). Teachers' and Students' Perceptions of AI-Generated Concept Explanations: Implications for Integrating Generative AI in Computer Science Education. *Computers and Education Artificial Intelligence*, 100283. https://doi.org/10.1016/j.caeai.2024.100283

Mizumoto, A., & Eguchi, M. (2023). Exploring the potential of using an AI language model for automated essay scoring. *Research Methods in Applied Linguistics, 2(2),* 100050. https://doi.org/10.1016/j.rmal.2023.100050

Nazaretsky, T., Mejia-Domenzain, P., Swamy, V., Frej, J., & Käser, T. (2024). AI or Human? Evaluating Student Feedback Perceptions in Higher Education. *Lecture notes in computer science (pp. 284–298)*. https://doi.org/10.1007/978-3-031-72315-5_20

Ng, D. T. K., Leung, J. K. L., Chu, S. K. W., & Qiao, M. S. (2021). Conceptualizing AI literacy: An exploratory review. *Computers and Education Artificial Intelligence, 2,* 100041. https://doi.org/10.1016/j.caeai.2021.100041

Nieminen, J. H., & Carless, D. (2022). Feedback literacy: a critical review of an emerging concept. *Higher Education, 85*(6), 1381–1400. https://doi.org/10.1007/s10734-022-00895-9

Noroozi, O., Biemans, H., & Mulder, M. (2016). Relations between scripted online peer feedback processes and quality of written argumentative essay. *The Internet and Higher Education*, 31, 20–31. https://doi.org/10.1016/j.iheduc.2016.05.002

PaperRater. (2025). PaperRater. Retrieved from, https://www.paperrater.com

Poulos, A., & Mahony, M. J. (2008). Effectiveness of feedback: the students' perspective. *Assessment & Evaluation in Higher Education, 33*(2), 143–154. https://doi.org/10.1080/02602930601127869

Ramu, D., Jain, R., & Jain, A. (2024). Generation Z's ability to discriminate between AI-generated and Human-Authored text on Discord. *arXiv (Cornell University)*. https://doi.org/10.48550/arxiv.2401.04120

Reynolds, B. L., Kao, C., & Huang, Y. (2021). Investigating the effects of perceived feedback source on second language writing performance: A Quasi-Experimental Study. *The Asia-Pacific Education Researcher, 30*(6), 585–595. https://doi.org/10.1007/s40299-021-00597-3

Schartel, S. A. (2012). Giving feedback – An integral part of education. *Best Practice & Research Clinical Anaesthesiology, 26*(1), 77–87.

https://doi.org/10.1016/j.bpa.2012.02.003

Skagerberg, E. M., & Wright, D. B. (2008). Susceptibility to postidentification feedback is affected by source credibility. *Applied Cognitive Psychology, 23*(4), 506–523. https://doi.org/10.1002/acp.1470

Strijbos, J., Pat-El, R., & Narciss, S. (2021). Structural validity and invariance of the Feedback Perceptions Questionnaire. *Studies in Educational Evaluation, 68*, 100980. https://doi.org/10.1016/j.stueduc.2021.100980

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., V., & Zhou, D. (2022). Chain-of-Thought prompting elicits reasoning in large language Models. https://proceedings.neurips.cc/paper_files/paper/2022/hash/9d5609613524ecf4f15af0f7b31abca4-Abstract-Conference.html

Yang, M., Badger, R., & Yu, Z. (2006). A comparative study of peer and teacher feedback in a Chinese EFL writing class. *Journal of Second Language Writing, 15*(3), 179–200. https://doi.org/10.1016/j.jslw.2006.09.004

Appendix A1: Information sheet

Information sheet for "How does the 'source' affect the 'message'? Looking at feedback perception and uptake"

Purpose of the research

The goal of this research is to investigate how different sources of feedback impact feedback uptake and perception. The different sources used in this study are a teacher, peers, and generative AI. You are asked to write a short (max. 500 words) essay on the usage of AI in educational contexts. Afterwards one third of the participants will be asked to provide feedback on another third of the essays, these will be used for the peer provided feedback. A teacher will provide feedback on another third of essays, and finally the rest of the essays will be given to AI. You will receive feedback on your essay, along with the information which source has provided said feedback. You are asked to revise your essay with the feedback and hand it in. Then you are asked to fill in a questionnaire. The essay, received feedback, revised essay, and questionnaire answers will be analysed to try to answer the research question: How does the 'source' affect the 'message'?

Benefits and risks of participating

The benefits of participating in this study are: first, you will receive feedback on an essay you have written. Second, you may be asked to provide feedback on another participants' essay, gaining experience critically reading an essay and providing feedback. Third, you may gain more experience working with AI to improve a piece of work. There are no known risks of participation. This research project has been reviewed and approved by the BMS Ethics Committee (domain Humanities & Social Sciences).

Procedures for withdrawal from the study

If, at any point during or after the study you want to withdraw and have your data excluded from the sample, you can let the researcher know via Email (d.w.h.helder@student.utwente.nl). No further explanation needs to be given.

Personal information

At the end of the study you will be asked to fill in a questionnaire, included in this questionnaire are some questions about personal information, such as your age, gender identity, and level of

education. This is done to be able to analyse any possible demographic influences of the dataset on the results. This data will be pseudo-anonymised, meaning the essays, feedback, and questionnaire answers will be linked via a unique identification number so your name and email address can be decoupled from the dataset. It is not possible to identify you via the data you have provided, but it does allow us to exclude your data if you wish. You are at any time allowed to request access to, rectification of, or erasure of personal data.

The dataset will not be shared beyond the study team. The data will be handled according to the Guideline on Ethics, Privacy & Research Data Management BMS and will not be accessible for commercial or personal use.

Contact Information for Questions about Your Rights as a Research Participant

You can contact the researcher via Email: Rik Helder (d.w.h.helder@student.utwente.nl). You can also contact the supervisor of this research via Email for any questions: Mohammadreza Farrokhnia (m.farrokhnia@utwente.nl)

If you have questions about your rights as a research participant, or wish to obtain information, ask questions, or discuss any concerns about this study with someone other than the researcher(s), please contact the Secretary of the Ethics Committee/domain Humanities & Social Sciences of the Faculty of Behavioural, Management and Social Sciences at the University of Twente by ethicscommittee-hss@utwente.nl**. You can find the informed consent form below.**

Appendix A2: Informed consent form

Consent Form for "How does the 'source' affect the 'message'? Looking at feedback perception and uptake"

Please tick the appropriate boxes	**Yes**	**No**
Taking part in the study		
I have read and understood the study information dated [26/10/2024], or it has been read to me. I have been able to ask questions about the study and my questions have been answered to my satisfaction.	□	□
I consent voluntarily to be a participant in this study and understand that I can refuse to answer questions and I can withdraw from the study at any time, without having to give a reason.	□	□
I understand that taking part in the study involves writing a short essay, possibly providing feedback on another essay, being asked to implement feedback on the essay, and filling in a questionnaire.	□	□
Risks associated with participating in the study		
I understand that taking part in the study involves the following risk: A chance of receiving unexpected, or critical feedback.	□	□
Use of the information in the study		
I understand that information I provide will be used for writing a bachelor thesis report, which might be developed into a published paper.	□	□
I understand that personal information collected about me that can identify me, such as my age, gender identity, or level of education, will not be shared beyond the study team.	□	□
I agree to joint copyright of the written and revised essay to D.W.H. Helder (researcher).	□	□
Future use and reuse of the information by others		

I give permission for the written essay, feedback, revised essay, and questionnaire answers that I provide to be archived anonymously in the University of Twente Student Theses database so it can only be used for future research and learning. The data will not be accessible for commercial or personal use. □ □

Please sign on the final page below.

Signatures

_____	_____	_____
Name of participant	Signature	Date

The potential participant has received the information sheet and I have, to the best of my ability, ensured that the participant has been able to ask questions and understands to what they are freely consenting.

Rik (D.W.H.) Helder		26/10/2024
Name of researcher	Signature	Date

Study contact details for further information or any questions:

Researcher: d.w.h.helder@student.utwente.nl

Supervisor: m.farrokhnia@utwente.nl

Appendix B1: ChatGPT (using ChatGPT-4o) Feedback Prompt

Feedback prompt

You are a professor of academic writing at the undergraduate level. Your students have submitted an argumentative essay as an assignment for this course on the role of artificial intelligence in education, which should be no more than 500 words in length. You now need to provide each student with feedback on the quality of their argumentative writing, identifying problems in their essays and suggesting solutions. Your feedback should be given in a paragraph containing 250 to 300 words and address the following questions:

- Has the student provided an introduction relevant to the topic?
- Has the student presented a clear and definite position regarding the topic?
- Has the student provided reasons to support their position? And have they substantiated these reasons with credible evidence (in the form of examples, personal experiences, statistics, expert opinions, and research evidence)?
- Has the student presented any counterarguments to their position? And have they refuted these counterarguments with credible evidence (in the form of examples, personal experiences, statistics, expert opinions, and research evidence)?
- Has the student concluded their essay effectively?

Please use a step-by-step approach to respond to this request.

For example, a student has sent you the following argumentative essay:

“The role of artificial intelligence in education

Today, artificial intelligence is used extensively in all fields, especially in education. The use of AI tools in education has benefits such as creating smart educational content, learning in any language, creating tests, etc., but alongside such benefits are challenges and limitations such as lack of quality data, cost, and access. Generally, the benefits of using AI tools outweigh the challenges, and in my opinion, these tools can be used in the classroom to great advantage because firstly, using these tools improves learners' education—for example, by making many of the abstract contents of courses like physics and chemistry concrete, learning occurs faster and better. Secondly, teachers can receive ideas and help regarding their teaching methods from these tools and clarify educational material for students. For example, teachers for lower grades can use games and entertainments provided by AI tools. Thirdly, with constant access to artificial intelligence, teachers and learners can use it at any time to get more information and enhance their knowledge. Students can easily have essays and any needed texts in class. Teachers can also create

creative exercises and smart tests through these tools and even use AI for grading and checking tests. Therefore, artificial intelligence can create a major revolution in the education industry and cause an increase in productivity in various aspects of it.”

Now you want to provide feedback to improve this argumentative text. To prepare this feedback, you should act as follows. First, check whether the student has provided a suitable introduction before stating their own position. If the introduction is weak or incomplete, offer suggestions for its improvement. Next, check whether the student has clearly and explicitly stated their position. If the position is vague, offer suggestions for its improvement. Then, it should be checked whether the student has provided valid evidence (in the form of examples, personal experience, statistics, expert opinion, and research evidence) in support of their position. If the evidence is insufficient or not credible, offer suggestions for its improvement. Next, it should be noted whether the student has raised opposing views to their position and refuted them with valid evidence (in the form of examples, personal experience, statistics, expert opinion, and research evidence). If this part is weak, offer suggestions for its improvement. Finally, check whether the student has provided an extensive summary. If the summary is insufficient, offer suggestions for its improvement. Following these considerations, a sample 250 to 300-word feedback for the argumentative text above could be as follows:

"In the beginning of the text (the first four lines) you mention the advantages and challenges and at the end you also state your position in the form 'In my opinion, these tools can be used in the classroom to great advantage.' It would be better instead of stating these advantages and disadvantages to start your text with a brief introduction that clarifies the background of the topic and creates an engaging entry into the subject. After this introduction, you should state your position clearly.

After stating your position, you have discussed one reason, saying that the use of these tools makes many of the abstract contents of courses concrete and then mentioned several courses like physics and chemistry as examples or evidence, which is excellent. In the second reason, you have stated that teachers can use artificial intelligence in their teaching methods. Then you have mentioned an example or evidence that games and entertainments produced by AI tools can be used. As the third reason, you talk about teachers having access to artificial intelligence at any time which can update their knowledge, but for this reason, you have not mentioned evidence in the form of examples, expert opinions, personal experiences, and statistics.

Finally, you have clearly stated your conclusion. However, it would have been better to mention the negative aspects of using AI tools in the classroom as well, after referring to the positive aspects. Providing evidence to refute these negative aspects and then stating

conclusions based on that would strengthen your argument, making it easier for the reader to accept."

Please wait while I send you 36 argumentative essays so that you can provide feedback based on the above instructions.

Appendix B2: ChatGPT (using ChatGPT-4o) Peer Tone Prompt

Peer tone prompt (after receiving the Appendix B1 prompt)

Above you can find a prompt I have given you; I will send you a reply you have created based on an essay which I will send you as well. Please wait until I send both your response and the original essay. Please look at both and change the tone of your reply in a way that a student would write it in a peer-to-peer feedback context.

Appendix C: Debriefing document

Debriefing document for “How does the ‘source’ affect the ‘message’? Looking at feedback perception and uptake”

Dear Participant,

Thank you very much for taking the time to participate in the study called How does the 'source' affect the 'message'? Looking at feedback perception and uptake. I have to inform you that during the study you have not been fully informed about the exact methodology of the experiment. This study aimed to investigate the differences between the perception and uptake of feedback that came from different sources, namely teachers, peers, and generative AI. You were told that the feedback you received on your essay came from one of these sources, and that some of the participants were asked to provide this feedback. This has not been the case, all feedback has been provided by AI. This was necessary to assure a similar level of quality amongst the feedback. Having a similar level of quality for all feedback was essential to investigate the differences in perception and uptake based on the expectations, understanding, and prejudice of the sources of the feedback, rather than the quality of the feedback itself.

Now that you have been made aware of the deception used during this research, know that you are still allowed to request your data to be excluded from the final sample. To do this all you need to do is press 'No' at the end of this survey. Please keep in mind that some of the other participants may nog have finished yet and are still unaware of the deception, so please keep it to yourself until all participants have finished the entire study. You can send an email if you have any questions or remarks about the entirety of this research study.

Warm regards,

Rik Helder
d.w.h.helder@student.utwente.nl

Supervisor
Mohammadreza Farrokhnia
m.farrokhnia@utwente.nl