Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

Do student perceptions mirror reality? A mixed-methods study to evaluate the accuracy, utility and efficiency of AI-assisted feedback in introductory mathematics courses

Therese Romslo Saltskår

therese.saltskar@uib.no

University of Bergen

Christian Bru

University of Bergen

Antonella Zanna Munthe-Kaas

University of Bergen

Kasper Troøyen

University of Bergen

Sehoya Cotner

University of Bergen

Research Article

Keywords: Formative assessment, introductory mathematics, Large Language Models, AI-assisted assessment, AI, teaching assistants, perceptions of AI, student trust

Posted Date: November 14th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-11455953/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: Competing interest reported. Author CB represents a commercial start-up focusing on educational applications of AI. This start-up provided the AI tool used in the studies described here. Conflict of interest was mitigated by removing CB from data access and formal analysis and only involving CB's input after a full draft of the manuscript was developed. CB's role in writing the manuscript was creating Figure 1 and writing the "The AI Tool" part of the methodology section. CB had

no role in our interpretation of the data. No updates to our interpretation of the data were driven by CB's comments.

Do student perceptions mirror reality? A mixed-methods study to evaluate the accuracy, utility and efficiency of AI-assisted feedback in introductory mathematics courses

Abstract

Background

Giving meaningful feedback on assignments in large introductory-level mathematics courses is a challenge. The emergence of "out of the box" Large Language Models (e.g., ChatGPT, Gemini) has created a potential solution to this challenge, where the AI tool can be trained to give constructive feedback on student work. However, it is unclear whether these tools are sufficiently accurate, efficient, and trustworthy. The current study investigates the use of a bespoke AI tool to assist in giving feedback to students in two introductory mathematics courses that enroll over 400 students each and are characterized by high attrition.

Results

We evaluated the similarities in grades assigned by the AI tool and human graders, as well as student and grader perceptions of the feedback given by both the AI tool and human graders. We found that the tool was fairly accurate, but that perceptions of the utility of its feedback varied by condition. When evaluated as a stand-alone grading agent, the AI tool was significantly undervalued relative to human graders—both student peers and the professor.

Conclusion

Three themes that characterize our findings: student trust is key, the professor is irreplaceable, and graders and AI tools must be carefully trained to ensure accuracy, transparency, and trust. We conclude with a call to exercise caution in using these tools in grading assistance, particularly with respect to maintaining student trust and transparency in grading.

Keywords

Formative assessment, introductory mathematics, Large Language Models, AI-assisted assessment, AI, teaching assistants, perceptions of AI, student trust

Introduction

Many STEM faculties in Norway struggle to retain students, especially between years one and two of the bachelor program (Haugen et al., 2025). For example, it is typical for 30% or more of incoming students to leave their study program before year two. Further, at some STEM faculties fewer than 50% of incoming students complete a three-year bachelor's degree in four years. This reflects a national challenge and the added concern that Norway is not educating enough scientists to meet current and projected needs (NITO 2025, Research Council of Norway 2026, Ministry of Education and Research 2023). While several reasons have been cited for this problem, one that seems to be at the root of student attrition is how we *assess* our students—especially in the high-enrollment, introductory-level courses that are the focus of this proposal (Ralph et al., 2022, Vyas and Reid 2023, Cotner and Ballen 2017).

Large introductory courses, especially those that meet the curriculum requirements of many study programs, face several challenges. The depersonalized atmosphere of the classroom environment has been associated with negative student affect (e.g., lower sense of belonging, lower self-efficacy; Wilson et al., 2015, Velez et al., 2013), disengagement (Wang and Calvano 2022, Finn et al., 2003, Ballen et al., 2019), and an exacerbation of existing inequities in STEM (Ballen et al., 2018). Further, in a large, mixed-majors class, students are likely to vary in the extent and quality of their prior experience with the topic, making some students feel disadvantaged relative to their peers (Salehi et

al., 2019, Meaders et al., 2020). Giving meaningful and timely formative feedback to 100 or more students is difficult, so students have fewer tools for assessing their own learning in the course and may enter testing situations unprepared (Enu 2021, Yeh 2009).

Some investigators (Hopfenbeck et al., 2023, Vittorini et al., 2021) have suggested that this challenge of providing formative feedback could be mitigated by strategic use of Large Language Models (LLMs) such as ChatGPT or Copilot. These tools can process in seconds what a human might do in days and, unlike humans, theoretically maintain objectivity after grading 5, 50, or 500 assignments. Further, free from the constraints of time, they can generate more detailed feedback than human graders (Dai et al., 2023). LLMs seem especially useful in assessment contexts in which there is a single correct answer, or in which students are evaluated based on a set of well-defined rules (e.g., in mathematics, Bhandari and Pardos 2025, Yu et al 2026).

Others (Wangdi et al., 2025) have challenged the use of LLMs in assessing student work, noting, for example, the tendency of these tools to fabricate output (Gamlem et al., 2026, Kizilcec et al., 2024) or limitations with evaluating creativity or skills acquisition (Elshall and Badir 2025). In mathematics assignments (Luzano 2024), for example, LLMs may assign correct scores to work that was done incorrectly, missing the point of constructive feedback and leading some students to have a false sense of confidence entering a high-stakes testing situation.

Further, student perceptions of these tools will likely be a factor in their overall utility. If students do not trust LLM feedback, the feedback may not be helpful in their learning (Alamäki et al., 2024, Riahi and Cateté 2026). To ensure accuracy and promote trust, several investigators (Vanhoyweghen et al., 2026, Riahi and Cateté 2026) endorse the use of "human in the loop" grading protocols in which humans check, assist, or otherwise curate AI-generated feedback.

Clearly, there are disagreements concerning the use of AI tools in providing formative feedback to students. Rapid, useful feedback is desirable, but not at the cost of student distrust, or incorrect or misleading information. Thus, we conducted two separate, but related, studies to evaluate both student perceptions of the feedback provided by these tools and the extent to which AI feedback aligned with that of human graders.

We studied two large introductory mathematics courses that are characterized by large enrollment (>400 students per course), high-stakes testing (whereby a single final exam constitutes the majority of a student's grade in the course), and limited formative assessment. Both courses have struggled to provide meaningful and timely feedback on mandatory assignments, and both tested the use of AI-assisted grading on these assignments.

Generally, we were interested in the effectiveness of available AI tools at giving constructive feedback on mathematics assignments (in comparison to human-only feedback). We asked the following research questions:

- RQ1: Do AI tools assign similar grades to humans? (both studies)
- RQ2: How do graders perceive AI-generated feedback? (Study One only)
- RQ3: How do students perceive the feedback they receive from AI versus that they receive from humans? (both studies)

Our hope was to both inform local practice and contribute to the developing understanding of the potential of AI tools in education.

Methodology

The AI tool

Both studies used XYZ (developed under the working name XYZ Canvas Integration), an AI-assisted grading and feedback application for the Canvas learning management system, developed in Norway by XYZ AS. Development began in July 2025, and the version used here was an early pilot build deployed for the autumn 2025 semester at the University of XYZ. The tool was designed around a single premise: a large language model drafts formative feedback and a suggested score, and a human instructor remains responsible

for what reaches the student. It is not a chatbot, and it does not act autonomously; it produces a draft that a course instructor can read, edit, and approve.

The tool connects directly to the course's Learning Management System (i.e., Canvas) space and reads the assignment description and the students' submitted files. For each assignment, the instructor may designate up to two reference documents that the model receives together with every submission:

1. a solution manual, used as the reference for what a correct answer looks like, and
2. a grading guide or rubric, containing the instructor's grading criteria, point allocations, and instructions on the form the feedback should take.

When the instructor starts a grading run, the tool creates one independent grading task per submission and processes them concurrently in a background queue, so that a several-hundred-submission assignment completes in minutes rather than sequentially. Student work is converted into formats the model can read before grading: PDFs and images (including photographs of handwritten work) are passed to the model directly, and text documents are converted to a structured text representation. A conversion failure for one file does not stop the rest of the submission from being graded, and a failure on one submission does not affect the others.

For each submission, the model receives the assignment description, the student's work, and the designated reference documents, and returns a response constrained to a fixed

structure: a numeric score on the scale the instructor has defined for the assignment, a written feedback text, a summary of the reasoning behind the score, and lists of specific strengths, weaknesses, and suggested corrections. Constraining the output in this way means every submission is scored on the same scale and receives feedback in the same shape, which is a prerequisite for the comparisons reported here. Feedback is generated in the language of the assignment, which in both courses was Norwegian.

The underlying model in both studies was Google Gemini 2.5 Pro, accessed through Google's application programming interface with extended reasoning enabled. The model was used as supplied by the vendor. It was not fine-tuned, and no student work was used to train or adapt it; all course-specific information reached the model through the assignment description and the reference documents described above. Results were stored in the tool's own database, where the instructor could review and edit both the score and the feedback text before publishing them back to Canvas as a grade, a feedback comment, or both. Published feedback carried a standard notice informing students that it had been generated with the assistance of a large language model (Appendix A). Figure 1 summarizes this flow.

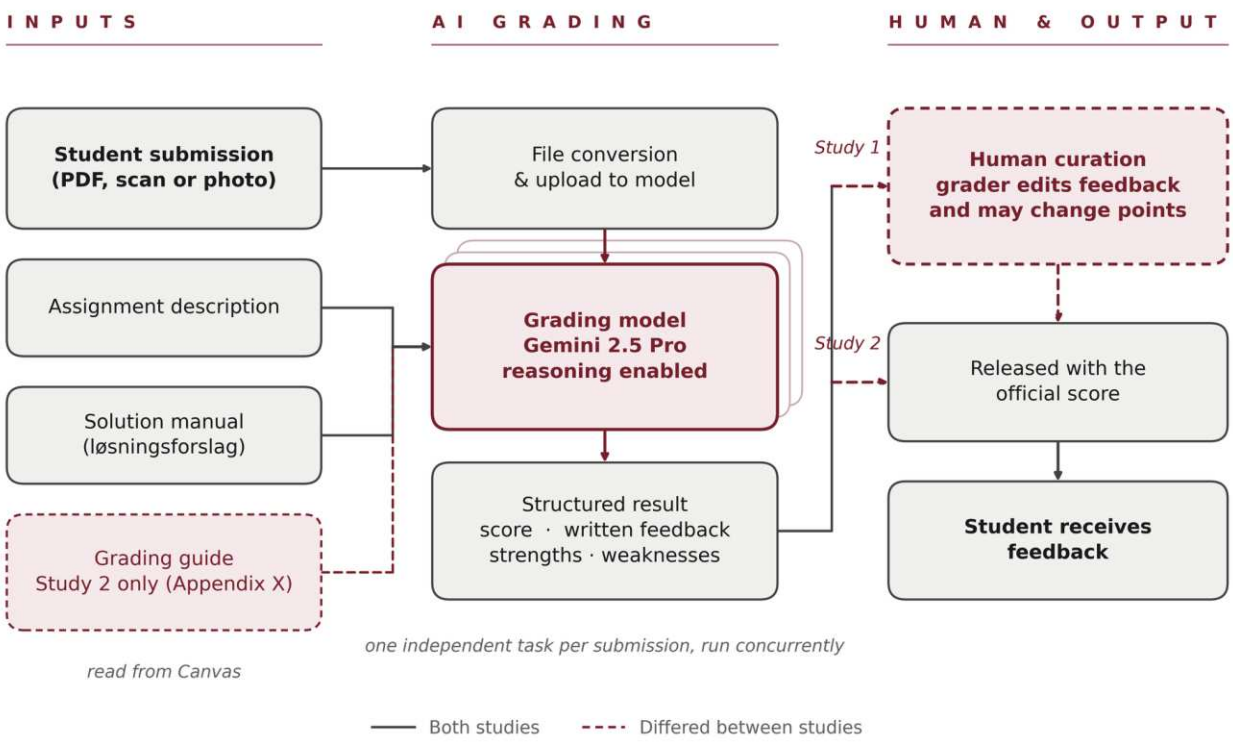


Figure 1. Workflow for AI-assisted feedback on a student submission. The tool reads the assignment and student submissions from Canvas and combines each submission with the instructor's reference documents — a solution manual in both studies, and, in Study Two, a detailed grading guide (Appendix B). Submissions are graded concurrently, one independent task each. The model returns a structured result — a score on the instructor-defined scale plus written feedback — which in Study One was curated by a human grader before release and in Study Two was released to students as generated. Dashed elements indicate components that differed between the two studies.

In Study One, no instructions had been given to the AI tool except for the general instruction of grading the assignment using the solution manual as a reference. The AI model thus gave number grades from 0-17 as the instructor had decided.

In Study Two, based on grader feedback after the first study, the AI tool had been given more thorough instructions. A traffic light model (green, yellow, red) was used to provide information on whether an exercise was done correctly, partially correctly, or wrong. The model was also instructed to not repeat steps that had been done correctly, and to refer to the solution manual instead of providing calculations showing alternative methods. The template used in the prompt in Study Two can be found in Appendix B.

Study One

Course context

Course One is a 10 ECTS (European Credit Transfer and Accumulation System) Calculus 1 course, taken by first year students from mathematics-intensive study programs such as mathematics, physics and engineering, among others. Approximately 500 students are enrolled each semester, with around 450 continuing to the final exam.

Although the final exam determines 100% of the course grade, students are required to pass three mandatory assignments to be allowed to take the final exam. Mandatory assignments are normally graded by teaching assistants (TAs) employed in the course, and the lecturer is normally not involved in grading and hence do not see students' results.

Students pass if they have earned at least 50% of the possible points. Points are normally not shared with the students, who only receive Pass/Fail as their results.

Study design

The current study involved the first mandatory assignment. Eleven student assistants were recruited to take part in a two-day grading retreat away from campus. The student assistants were either current or former teaching assistants in mathematics, or students who recently took Course One and scored either an A or a B.

Assignments graded at the retreat were randomly selected for one of two treatments:

1. Graded by a Human Only
2. Graded using an AI tool, then curated by a human (Human + AI)

A total of 403 students handed in the first mandatory assignment. Of those, a sample were graded by the 11 student assistants at the grading retreat. 159 students were graded by a Human Only and 89 students graded by a Human + AI. The remaining assignments were graded by the teaching assistants in the course.

For assignments in the first treatment (Human Only), the graders were instructed to use a solution manual provided by the lecturer to award each exercise either 0, 0.5 or 1 point. 17 points could be awarded in total. Graders were asked to provide brief feedback where

necessary, but they could also refer to the solution manual. A grade-norming activity was done at the start of the retreat to ensure that graders gave similar scores.

In the second treatment (Human + AI), an AI tool had generated feedback for all assignments. Graders were instructed to read this feedback and modify as necessary, and they were able to change the points awarded by the AI tool if they found the points to be wrong. We also noted the scores assigned by the AI tool, prior to any human curation (for a third, AI Only, condition that was not reported to the students).

Student assistants were then randomly assigned a form with 31 or 32 assignments, where it was denoted whether they were to be graded using the Human Only or Human + AI treatment.

Initially, 20 random assignments were graded four times by four different student assistants, twice using treatment 1 (Human Only) and twice using treatment 2 (Human + AI). This allowed us to measure variability both within and between treatments. Afterwards, the graders worked through their given assignments, pausing often for group discussion and mental breaks.

Students were given written feedback (either human-generated or human-curated from AI) together with a Pass/Fail grade, however we noted points (0-17) in our calculations of grader accuracy. These points were not shared with the students.

At the conclusion of the two-day grading retreat, we held a focus group with the graders to solicit their input on the experience, their impressions of AI-assisted grading, and suggestions for future implementation. We took notes during the discussion but made no recordings. Any findings reported from this focus group are based on general impressions rather than direct statements.

Following the first assignment, students were asked to fill out a brief survey about the usefulness of the feedback they received. Specifically, they were asked whether they had read the feedback (yes or no), how useful the feedback was (on a Likert scale from 1 to 7), and then they were asked to comment freely on the utility of the feedback.

In the beginning of the semester, students received information about the nature of the study, its voluntary nature, and their ability to withdraw at any time. The information and consent form have been translated from the original Norwegian and are included in Appendix F.

Data analysis: Grading

Students graded four times

Due to one missing grade in our initial pool of 20, 19 students were graded 4 times, and 1 student was graded only 3 times. Each assignment was graded by 4 (or 3) graders, drawn randomly from a sample of 11 graders.

To estimate inter-rater reliability, we used Intraclass Correlation Coefficients (ICC) to detect what proportion of variance in grades was caused by the students. The model that was used was for a single grader, absolute agreement, two-way random effects (ICC(2,1) (two-way, single rater)). This model is most appropriate in this case because we wanted to treat the graders as random effects while we measured variance caused by the students separately for each condition: Human + AI and Human Only.

Specifically, we used the "lmer" function from the "lme4" package in R (Bates et al., 2015) with student and grader as random effects. Then we calculated the variance attributable to between-student differences by dividing variance between students by total variance in grades.

The effect of the grading method on mean score was tested using a likelihood ratio test (ANOVA) in R (R Core Team, 2025), comparing two linear mixed-effects models, with and without the method effect, and fitted using maximum likelihood.

Students graded with one method

248 students were graded once, either by Human Only or by AI with human curation. To test for a difference in mean score we used the “lmer” function from the “lme4” package in R (Bates et al., 2015) to fit a model with and without the method effect, fitted using maximum likelihood. Then we compared them using a likelihood ratio test (ANOVA).

To assess whether the variability of grades differed between the two grading methods (Human Only vs. Human + AI), we used two complementary approaches. First, we applied a Brown–Forsythe version of Levene’s test from the car package (Fox & Weisberg, 2019) with the median as the center, which tests the null hypothesis of equal variance across groups and is relatively robust to non-normality compared with the classical F-test.

The grades may be dependent on which grader graded the specific assignment. Therefore, we also performed a cluster-aware, robust variability analysis to correct for potential clustering around single graders. To do this, we fit a linear model for the mean grade including method and grader fixed effects (lm(Grade ~ Method + factor(GraderID))) and extracted the residuals. We calculated absolute deviation of residuals from the method-specific median: $Z = |r - \text{median}(r|\text{Method})|$ We then fit a linear model on z with Method as the only predictor to test for differences in dispersion between methods, and computed cluster-robust standard errors using a CR2 adjustment with Satterthwaite degrees of freedom with the coef_test function from the ClubSandwich package (Pustejovsky JE, 2026).

Data Analysis: Quantitative methods for usefulness

First, we estimated the mean score of usefulness for all students who answered the survey and indicated that they read the feedback. To test whether reported usefulness differed between the two conditions (Human Only vs. Human + AI), we fitted an ordinal logistic regression model, as usefulness was measured on a 1–7 ordinal scale. The analysis was conducted in R using the clm() function from the ordinal package (Christensen R, 2025). Condition was included as the predictor, and statistical significance was assessed with a Wald chi-square test.

Data analysis: Qualitative methods for usefulness

To explore graders' experience with AI-assisted assessment, a group discussion was conducted at the end of the two-day grading retreat. All participating graders (n=11) took part in the discussion, which focused on their experiences of grading assignments manually and grading assignments with AI-generated feedback available as a starting point. The discussion was intentionally informal and exploratory in nature, and no structured interview guide was used. Participants were encouraged to lead the conversation and raise issues that they considered relevant to the grading process.

Notes were taken during the discussion and used as the qualitative data source. The notes were reviewed and subsequently used to write a summary presented in the results section, showing graders' reflections of topics such as efficiency, feedback quality, trust and grading accuracy. Due to the exploratory purpose of this discussion and the use of notes rather than transcript of an audio file, the findings should be interpreted as a descriptive account of grader experiences rather than a comprehensive qualitative analysis.

Study 2 – Course Two

Course context

Course Two is a first-year 10 ECTS service course in mathematics, aimed at students from less mathematics-intensive study programs (e.g., biology, chemistry, geology). Approximately 500 students enroll in this course each fall, with around 450 students continuing to the final exam.

The syllabus in this course is divided into six modules. While 70% of the final grade in this course is based on the final written exam, 30% is based on components done throughout the semester and distributed throughout the modules. 10% of the final grade is determined by performance on milestone checks after each module, and these are the focus of the current study.

The milestone check is a short, written test taken in class at the end of each module. These tests are not mandatory, but many students take them as they count towards their final grade. To get credit for taking the test, students also need to provide feedback to two of their randomly selected peers.

Study design

In this study we focused on milestone checks 2, 3 and 4.

On each milestone check, students were given three types of feedback:

1. Feedback from the course coordinator
2. Feedback from two peers
3. Feedback from an AI model

Students who consented to take part in the study were surveyed after milestone check 4 and were asked, using items identical to those described for Study One, to report whether they had read the different types of feedback, as well as how useful they found each type of feedback.

Data analysis: grading

For Course Two, we compared the score given by the AI model with the score given by the professor. We did this by treating the score given by the professor as correct and calculating the differences between AI and the professor. First, we did a Wilcoxon test on

the paired scores (R Core Team, 2025), which tests if there is a difference in means between the two groups. We calculated the share of score given by AI within 0.25, 0.5, and 1 point difference from the score given by the professor. Then we calculated the mean error, median error, and mean absolute error.

We also compared the scores given by peers to those given by the professor. Specifically, we calculated the mean difference, mean absolute difference and share within one point of the professor's score.

Data analysis: Quantitative methods for survey data

We wanted to test if there were any differences in how students perceived the usefulness of feedback from the professor, their peers, or AI. To do this, we calculated the mean value of reported usefulness from the students who said they read each type of feedback. We also estimated the variance of the reported usefulness, and the proportion of students who said they read the feedback. We used a Kruskal test (R core team, 2025) to check if there is a significant difference in mean between at least two of the groups. After that we did a pair-wise Wilcoxon rank sum test (R core team, 2025) between each group with Bonferroni correction to reduce the probability of type 1 error.

Data analysis: Qualitative methods for survey data

The survey also included an open-ended question in which students were invited to elaborate on their perceptions of the feedback they had received, indicate which feedback type they found most useful, and describe how they had used or intended to use the feedback. Responses were analyzed using an inductive thematic analysis approach.

The analysis proceeded in four main stages. First, all responses were read in full to obtain an overall understanding of the data. During this process, comments relating to the three feedback sources (AI-generated feedback, peer feedback and professor feedback) were identified. Because a single response could discuss multiple feedback sources, comments relating to each feedback type were separated and organized into corresponding datasets.

Second, an initial coding process was conducted in which each comment was summarized into a brief descriptive statement capturing the central meaning of the student's response. These summaries were then reviewed to identify recurring patterns and themes within each feedback category. Microsoft Copilot (2026) was used as an analytical support tool during this stage to assist in identifying candidate themes across the response summaries.

Third, the preliminary themes were reviewed and refined by the researchers through an iterative process. Themes were merged or reorganized where necessary to ensure that they

were conceptually distinct and accurately represented the underlying responses. This process resulted in a final set of themes for each feedback type.

Finally, all comments were re-read and assigned to one or more themes. Multiple coding was permitted, so individual comments could contribute to more than one theme where appropriate. Microsoft Copilot was again used as a support tool during this stage to assist with theme assignment, while the researchers reviewed and finalized all categorizations. Frequencies were subsequently calculated by counting the number of comments associated with each theme. These frequencies indicate the prevalence of themes within the dataset and should not be interpreted as quantitative measures, as individual responses could be associated with multiple themes. An illustrative quote was chosen for each theme and translated by a bilingual member of the research team.

We note that while Microsoft Copilot was used to support identification of themes and the assignment of comments to themes, the researchers remained responsible for reviewing the general themes, determining the final coding framework and validating all thematic categorizations prior to reporting the results.

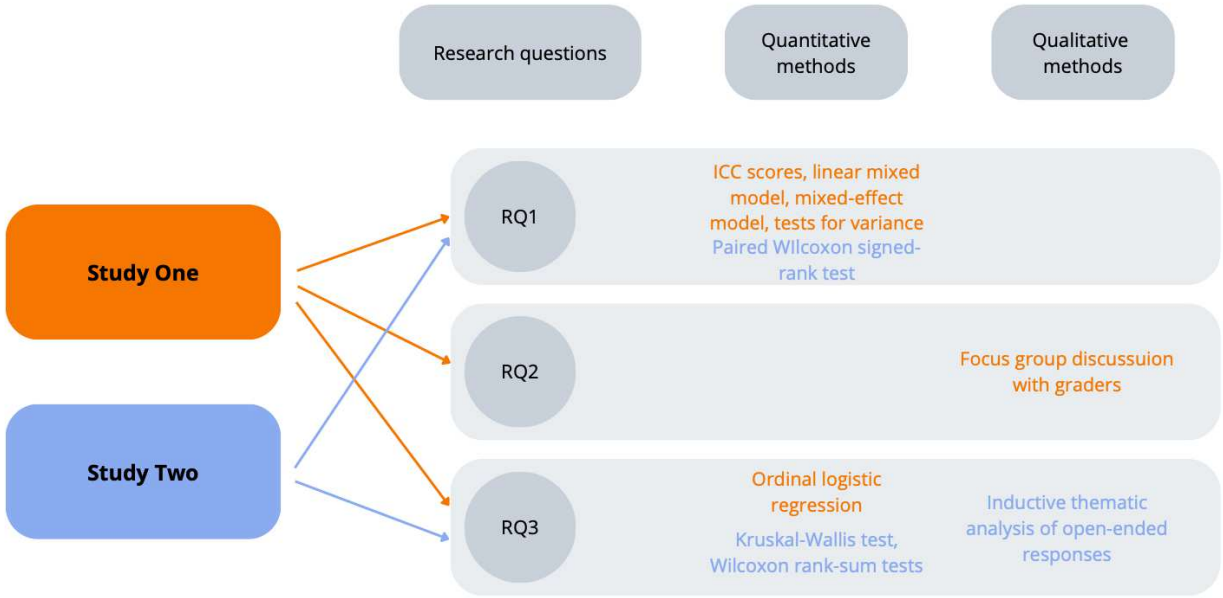


Figure 2. Logic model representing the two studies, our research questions and associated methodology.

Results

Study one

RQ1: Do AI tools assign similar grades to humans?

Students graded multiple times

Nineteen assignments were graded twice by humans only and twice by humans curating AI feedback. One assignment was graded twice by humans only and once with human

curation after AI feedback. Splitting the data into two groups, Human + AI and Human Only, the Human Only condition has 20 assignments with two grading scores each. The Human + AI condition has 20 assignments; 19 of these have two scores and one has one score.

The ICC was 0.874 for the Human Only scores, and 0.606 for the Human + AI scores. When fitting a linear mixed model by restricted maximum likelihood with "Method" as a fixed effect, the Human + AI coefficient was 0.26, indicating that the Human + AI graders gave on average 0.26 more points than the Human Only group. However, the likelihood ratio test between the two models with and without the method effect indicated no statistically significant effect of method on grade ($p = 0.183$). In Figure 3, we also include findings for the AI Only feedback, prior to human curation, even though this was never reported to the students.

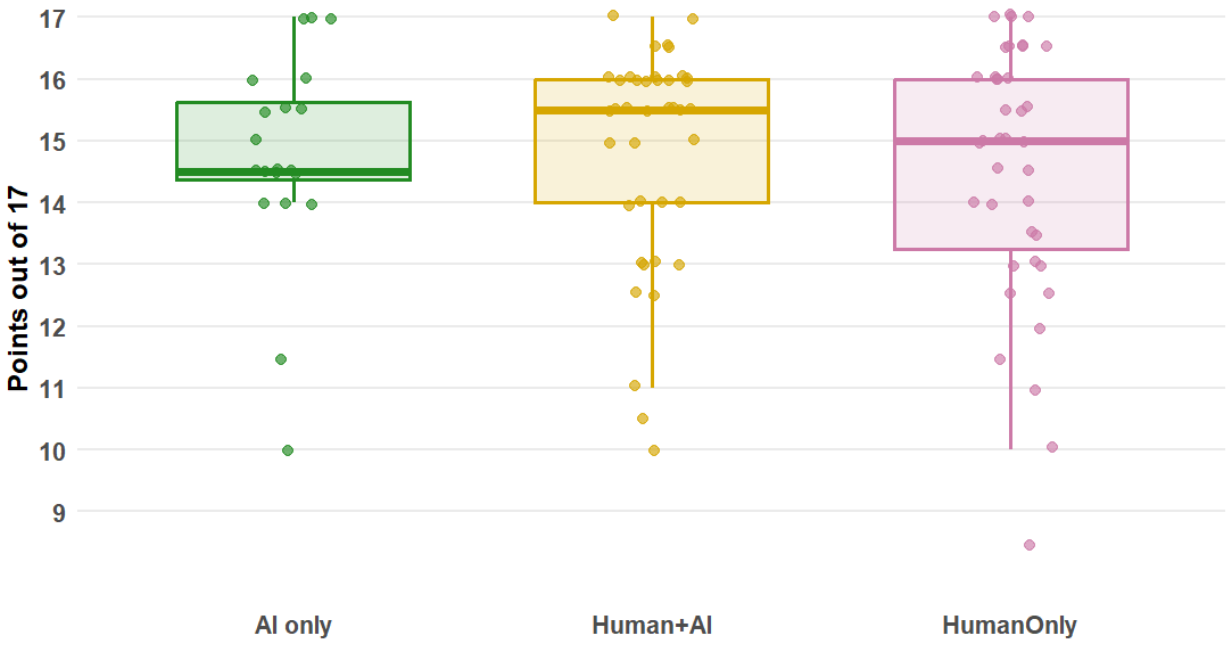


Figure 3. A comparison of grades assigned according to condition. Each of 20 assignments was graded once by the AI tool (AI Only), twice by a Human Only, and twice by a human using the AI output as a starting point. We detected no statistically significant effect of condition on grade distribution, however the variance explained by the student (versus the grader) was lowest in the AI Only condition.

Students graded once

After doing the between- and within-subjects comparison reported above, 159 students were graded by a Human Only and 89 students graded by a Human + AI. This allocation does not represent an even distribution because after the grading retreat, course teaching assistants graded any remaining assignments without AI input. The average score in the Human Only condition was 14.9 (out of 17), and the mean for the Human + AI condition was 14.4. In a mixed-effects model controlling for grader, the Human + AI condition was associated with grades that were on average 0.54 points lower than the Human Only condition, but this effect was not statistically significant ($p = .063$, chi squared (1) = 3.30). We found no significant difference in the variance of grades between the Human Only and Human + AI methods. This was supported both by Levene's test, with $F(1, 246) = 2.09$ ($p = 0.149$) and by the Brown–Forsythe-style residual analysis with grader-clustered standard errors ($b = 0.28$, $p = 0.302$).

RQ2. How do graders perceive AI-generated feedback?

At the end of the grading retreat, the student graders were gathered for a focus group session to explore the graders' experience with AI-assisted assessment. The participants unanimously reported a preference for manual grading over AI-assisted grading. When asked to raise their hand if they would prefer to do all the grading without AI assistance, *every single grader raised their hand*. Most graders indicated that reviewing and correcting AI-generated evaluations required more time and effort than grading assignments independently. As a result, AI assistance was generally perceived as *reducing* rather than *increasing* grader efficiency.

A central concern among graders was the reliability of the AI-generated grades. Participants reported several cases where the AI assigned incorrect scores, including awarding high marks to incomplete or incorrect submissions and failing to accurately interpret student work. These issues were particularly apparent when the AI encountered badly handwritten responses, although some graders attributed the inaccuracies to the AI's reasoning rather than handwriting recognition alone. Graders generally reported that they trusted their own professional judgement better than the AI model.

Participants also identified several strengths of the AI-generated feedback. The AI often produced comments that were well-written, grammatically correct and academically phrased. Particularly, graders viewed the overall summary comments and the automatic

generation of mathematical notation as useful. However, participants frequently found the feedback excessively verbose and often focused on issues that they considered less important. Graders consequently reported retaining only a portion of the AI-generated feedback and thus spending time rewriting much of it themselves.

The discussion suggested that graders viewed AI as more suitable for supporting feedback generation than for assigning grades. Several participants proposed using AI to draft summary feedback from the class, identify common student mistakes, generate standardized comment banks or assist with quality assurance processes. Others suggested that AI could be used as a preliminary screening tool for potentially failing students, thus leaving all high-stakes assessment decisions subject to human review.

RQ3. How do students perceive the feedback they receive from AI versus that they receive from humans?

Considering only those students that were evaluated in this study, 144 students answered the survey. Of these students, 132 said they read the feedback. 88 were in the Human Only treatment and 44 in the Human + AI treatment.

Table 1

Student perceptions of feedback on mandatory assignments in Course One

Condition	Answered survey (*n*)	Read feedback (*n*)	Average utility
Human Only	92	88	5.09
Human + AI	52	44	4.99

As shown in Table 1, students in both conditions reported high perceived utility of feedback, with slightly higher ratings in the human-only condition (M = 5.09 on a seven-point Likert scale) than in the human + AI condition (M = 4.99). We also note how many of them said they read the feedback from the grader. The reported means are based on students who said they read the feedback.

Ordinal logistic regression showed no significant difference in usefulness ratings between the two conditions (chi = 0.018, p = 0.89). The coefficient comparison for the Human + AI condition relative to the Human Only condition was not significant (β=0.045, p = 0.89). This indicates that the likelihood of reporting a higher usefulness rating did not differ between the two conditions.

Study Two

RQ1: Do AI tools assign similar grades to humans?

We compared points given by the AI model with points given by the professor on three different milestone checks. In total, we assessed 1051 different assignments (from 356 students). On each assignment, the students could earn 6 points in total. Not all students handed in all three milestone checks; thus the total number of assignments does not add up to $356 \cdot 3 = 1068$.

Table 2

AI-assigned points relative to those assigned by the professor in Course Two

N	Average deviation	Mean absolute deviation	Share within ±1 point
1,051	-0.24	0.64	85%

As shown in Table 2, AI-assigned scores closely approximated professor-assigned scores. Across 1051 assessments, the average deviation was -0.24 points, the mean absolute deviation was 0.64 points, and 85% of AI-assigned scores were within ±1 point of the professor-assigned score.

A paired Wilcoxon signed-rank test showed that the difference between the professor grade and AI grade was statistically significant ($V = 92340$, $p < 0.01$).

Table 3

Comparison of professor-assigned and AI-assigned scores by score category.

Professor score category	*N*	Mean deviation	Mean absolute deviation	Share within ±1 point
[0, 1]	54	0.53	0.75	76%
(1, 2]	92	0.30	0.68	80%
(2, 3]	156	-0.11	0.67	84%
(3, 4]	246	-0.27	0.65	86%
(4, 5]	271	-0.43	0.66	84%
(5, 6]	232	-0.47	0.55	89%

Note*:* Mean deviation represents the average difference between AI-assigned and professor-assigned scores, with positive values indicating that the AI assigned higher scores than the professor and negative values indicating lower scores. Mean absolute deviation represents the average magnitude of disagreement regardless of direction. Share within ±1 point indicates the percentage of AI-assigned scores that differed from professor-assigned scores by no more than one point.

Dividing the assignments based on how many points the professor gave (Table 3), we can see that most of the students completed most of the tasks correctly. When the professor

gave a low score, the AI tool tended to give higher scores than the professor, while the converse was the case when higher scores were awarded by the professor.

We compared the scores of the 58 students who said they read all types of feedback assigned by peers, AI, and the professor for specifically the 4th milestone check. In this sample, peer-assigned scores were more similar to the professor's scores than AI-assigned scores. Note that some students were graded by two peers, as planned, but some only received one peer assessment; thus, the number of peer graded assignments (87) does not sum to 58 x 2, or 116.

Table 4

Comparison of peer- and AI-assigned grades with professor-assigned grades.

Grader	*N*	Mean deviation from professor	Mean absolute deviation	Share within ±1 grade
Peers	87	-0.14	0.52	86%
AI	58	-0.50	0.74	83%

Note: Mean deviation represents the average difference between the assigned peer/AI grade and the professor-assigned grade, with negative values indicating lower grades than those assigned by the professor. Mean absolute deviation represents the average magnitude of disagreement regardless of direction. Share within ±1 grade indicates the percentage of assigned grades that differed from the professor-assigned grade by no more than one grade point.

Table 4 compares grades assigned by peers and the AI tool to that of the professor for the same assignments. We include mean and absolute deviation from the professor's scores. A negative mean deviation shows that the professor on average awarded higher scores than what the peers and the AI model did.

RQ3. How do students perceive the feedback they receive from AI versus that they receive from humans?

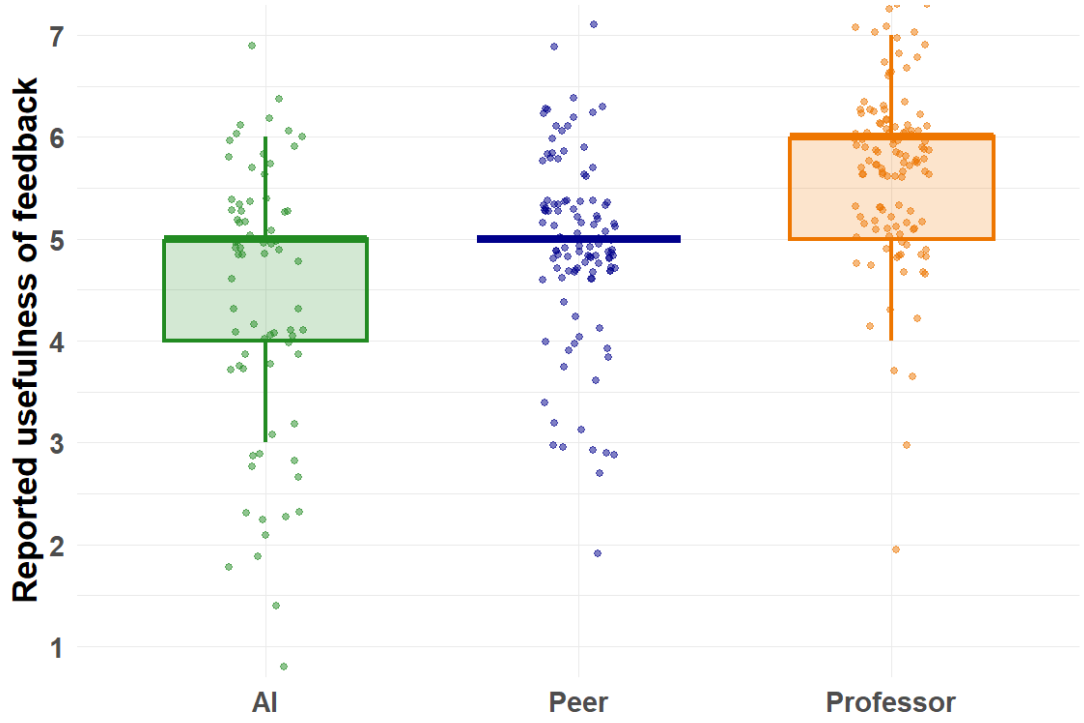
Study Two differed from Study One in that in Study Two, students were able to compare feedback from three different sources. Students perceived the AI feedback as being significantly less useful than feedback from their peers or the professor (Table 5 and Figure 4). Additionally, student impressions of the AI tool were more variable than they were for either peers or the professor. A Kruskal–Wallis test showed a significant difference in perceived usefulness of feedback across the three groups (AI, professor, and peers; chi-squared = 72.64, p value < 0.001). Post hoc pairwise Wilcoxon rank-sum tests with Bonferroni correction showed that all groups differed significantly from each other: AI vs. professor $p < 0.001$, AI vs. peers, $p < 0.001$ and professor vs. peers, $p < 0.01$.

Table 5: Mean usefulness scores for the three different feedback types, with associated variance and number of students who reported reading the feedback.

Feedback source	Mean usefulness	Variance	Number of students
AI model	4,38	1,80	73 (57%)
Peers	4,96	0,79	113 (89%)
Professor	5,72	0,73	120 (94%)

595

596



597

598 Figure 4. Perceived usefulness of the different types of feedback students received from their

599 milestone checks, whereby 1 = not useful at all and 7 = extremely useful.

600

601 Appendices C, D and E shows examples of the three types of feedback, translated into

602 English from the original Norwegian version.

What did students say about different types of feedback?

Of the 127 students who consented to take part in the follow-up survey, 72 students answered an open-ended question elaborating on their previous responses. Students were given the opportunity to state which type of feedback they found the most useful, how they were planning to use the feedback, or any other comments of their choice.

Students generally expressed mixed perceptions of AI-generated feedback (Table 6). The most frequently occurring theme was *limited trust in AI-generated feedback (A1)*, with several students questioning the accuracy and consistency of the feedback and the associated scores. Students also reported that the AI tool seemed to have *difficulties evaluating mathematical reasoning* and interpret solutions (A2), as well as concerns about *inconsistency and misalignment with assessment criteria* (A3). At the same time, some students highlighted benefits of AI-generated feedback, such as *valuable feedback* from detailed explanations that supported understanding (A4) as well as the *quick availability of feedback* (A5). Some students reported they did *not use the AI feedback* at all (A6).

Table 6. Summary of open-ended answers on AI feedback

Theme	**Number of responses**	**Illustrative quote**

A1: Limited trust in AI-generated feedback	12 (36,3%)	«AI feedback is the least useful because I have received varying responses and cannot be completely confident in its reliability, so I tend to compare it with other forms of feedback.»
A2: Difficulties evaluating mathematical reasoning	6 (18,2%)	«The AI did not capture all the steps and did not understand all the steps I did, and therefore provided feedback that was not relevant.»
A3: Inconsistency and misalignment with assessment criteria	5 (15,2%)	«I do not trust the AI-generated feedback. It did not take the assessment criteria into account, and therefore gave me a lower score.»
A4: Valuable feedback	9 (27,3%)	«I think that the AI-generated feedback is the most useful for me, because it provides in-depth and throrough explanations.»
A5: Quick availability of feedback	2 (6,1%)	«The AI responses are also arrive quickly.»
A6: Not using the AI feedback	6 (18,2%)	«I have not used the feedback from the AI model.»

623

Peer feedback was generally viewed positively, and it was frequently described as *useful for learning and reflection* (S1), making this the most common theme identified in student responses (Table 7). Several students also emphasized that reviewing the work of others was a valuable learning activity, in that *learning occurs through giving feedback* (S2). However, students also noted *strong variation in the quality* of peer feedback (S3), with some comments described as detailed and constructive and others as *less detailed than desired* (S4). As with the AI-generated feedback, students found *rapid feedback valuable* (S5). Although some respondents brought up *concerns regarding the accuracy of peer grading* and the competence of individual reviewers (S6), these concerns were often accompanied by recognition that many peers provided helpful feedback. Finaly, a small group of students emphasized the value of receiving feedback from a peer—someone at a similar academic level (S7).

Table 7. Summary of open-ended answers on peer feedback (student feedback)

Theme	**Number of responses**	**Illustrative quote**
S1: Peer feedback is useful for learning and reflection	22 (50%)	«Either I have been lucky, or all of my fellow students are very skilled and invest time in providing good and constructive feedback.»
S2: Learning occurs through giving feedback	3 (6,8%)	«I also found it useful to provide feedback, as it gave me a different

		perspective on the solution manual than by simply looking at it.»
S3: Strong variation in quality	16 (36,4%)	«I think that the feedback from fellow students is highly variable. Some students write detailed explanations, while others only state how many points were awarded for each exercise. There is also considerable variation in how strict fellow students are, which makes it difficult to know which score you actually deserve.»
S4: Feedback is often less detailed than desired	10 (22,7%)	«The peer feedback is often brief comments like «this is right, this is wrong», not very useful.»
S5: Rapid feedback is valuable	6 (13,6%)	«It is nice to get feedback from a student relatively soon after the milestone check has been completed.»
S6: Concerns regarding accuracy and reliability of peer feedback	8 (18,2%)	«Fellow students sometimes provide helpful and detailed feedback, but occasionally they do not know the mathematics well enough and give the wrong answers or do not understand an immediate step in the solution.»

S7: Feedback from a peer is valuable	5 (11,4%)	«Feedback from fellow students is valuable, because someone on your own level is grading, so you may encounter more empathy.»

Finally, professor feedback was viewed most favorably overall (Table 8). The most frequent theme was that professor feedback was *the most useful source of feedback* (P1), with many students explicitly identifying it as their preferred feedback type. Students frequently emphasized the *trustworthiness and expertise* associated with professor feedback (P2), and many described using it to *guide future learning and exam preparation* (P3). Several students also noted the *detailed and explanatory quality of the feedback* (P4). As a contrast to the two previous forms of feedback, the primary limitation identified by respondents was the *slow turnaround time in feedback* from the instructor (P5). Only two students mentioned *insufficient detail* (P6), while a few respondents suggested that professor feedback was *most beneficial when combined with feedback from other sources* (P7).

Table 8. Summary of open-ended answers on professor feedback

Theme	**Number of responses**	**Illustrative quote**
P1: Professor feedback is the most useful source of feedback	24 (63,2%)	«I usually find the professor's feedback the most useful. This is probably largely because they know the subject much better than we students do.»

P2: Trustworthiness and expertise	11 (28,9%)	«The feedback is valuable because you get an indication of how you are performing in the course and its topics from a more advanced academic perspective.»
P3: Guides learning and exam preparation	12 (31,6%)	«Through this feedback I can see what the professor emphasizes and what is required for an answer to receive credit.»
P4: Detailed and explanatory feedback	9 (23,7%)	«The professor's feedback is well explaned and shows me what I have doen wrong and why.»
P5: Slow turnaround time	4 (10,5%)	«We receive feedback from [professor name] so long after the milestone check has been completed that it becomes irrelevant.»
P6: Insufficient detail	2 (5,3%)	«[Professor name] has to grade an extremely large number of assignments, so it is understandable that they cannot spend a great deal of time and provide detailed feedback.»
P7: Best when combined with other feedback sources	3 (7,9%)	«A combination of feedback from fellow students and the professor is best.»

651

In summary, the open-ended responses show that each feedback type have distinct strengths and weaknesses. AI-generated feedback was valued for its speed and at time for the detail provided, but most students had concerns regarding accuracy and trustworthiness. Peer feedback was viewed as useful for learning, but variation in quality was a dominant theme for this type of feedback. Finally, professor feedback was viewed as the most trustworthy and useful source of feedback, although some students saw delayed timing as a disadvantage. These findings show a trade-off between timeliness, opportunities for peer learning and perceived trustworthiness.

Discussion

Summary of findings

Our study was motivated by an interest in the effectiveness of available AI tools at giving constructive feedback on mathematics assignments (in comparison to human-only feedback). Prior to drawing conclusions from our findings, we provide data-informed responses to our initial research questions:

RQ1: Do AI tools assign similar grades as humans?

In both studies, we found that AI tools assign similar grades to those assigned by humans. However, there is nuance to this finding. There was no statistically significant difference between AI-assigned and human-assigned (with or without AI) grades in Study One, and in

Study Two the AI gave on average 0.24 points (out of six total, or 4%) less than the professor, which was a statistically significant trend, if not *practically* relevant. Of course, the relevance of this difference depends on the course grading scheme; in this situation, the combined total of points earned on these milestone checks contributed 10% to the final course grade, so the practical difference—to a student's overall grade—is actually 0.4%. Further, in Study One the variance in grades assigned by humans was higher than that of the AI tool.

The downward-skew of the AI-assigned grades in Study Two may reflect the fact that the tool was instructed to assign grades between 0 and 6, or it could represent an incorrect interpretation of student work. Most AI grading systems (including LLM-based graders) are trained to predict scores that human raters have previously assigned. Because they are optimized to minimize average prediction error, they often perform best on typical, middle-range responses and can be less accurate on exceptionally strong or exceptionally weak work. The result can be a tendency to score outstanding work somewhat lower than expert human raters would, score very poor work somewhat higher than expert human raters would, and compress the score distribution toward the center. In psychometrics and machine learning, this is often discussed as score shrinkage, range restriction, attenuation at the extremes, or regression toward the mean (Deepsheeka 2025, Ramesh and Sanampudi 2022). Regardless of name, this phenomenon may explain our findings from study two.

RQ2: How do graders perceive AI-generated feedback?

Overall, the findings from the focus group revealed skepticism regarding the use of AI for grading in introductory mathematics. AI-generated feedback was viewed as a potentially useful support tool, but not a reliable replacement for human assessment. Graders acknowledged several strengths of the AI model, including its ability to generate detailed and grammatically correct feedback and summarize student performance. However, the dominant sentiment was concern regarding the accuracy and reliability of AI-generated grades, as well as excessive verbosity in feedback.

At the time of Study One, AI assistance did not improve grading efficiency. Graders reported spending considerable time reviewing, correcting and rewriting AI-generated feedback, making AI-assisted grading more time-consuming than grading manually. Graders expressed a clear preference for human-only grading and reported greater trust in their own judgement than in the AI model. While the graders recognized benefits of AI assistance, the focus group discussion highlighted challenges which could be fixed through adjustment and training of the AI model.

RQ3: How do students perceive the feedback they receive from AI versus that they receive from humans?

The two studies diverged in how students perceived AI feedback relative to human feedback. In Study One, there was no significant difference in student perceptions of the feedback given by the two conditions: Human Only or Human + AI. However, we note that

there was no stand-alone AI condition, as all feedback had some form of human input. Thus, there was not a true equivalence in comparison between the two studies.

In Study Two, the conditions were different: AI only, peer feedback, and professor feedback. Here, student perceptions varied significantly and meaningfully, with students finding the feedback from their professors most useful, and that from the AI least useful. Further, the variance in perceptions of the AI tool was over twice that of the variance in perceptions of feedback from peers or the professor. Finally, the percent of students who reportedly read the AI feedback was noticeably lower (57%) than those who claimed to have read feedback from their peers (89%) or the professor (94%). This pattern could suggest an *a priori* lack of trust in AI feedback, or it could reflect student concerns that *any* use of AI will appear to instructors (and the investigators of this work) as potentially problematic (Lund et al., 2025, Tossell et al., 2024).

Key themes

Evaluating these results, we conclude three key themes related to the use of AI tools in assessing mathematics higher education assignments: student trust is key, the professor is irreplaceable, and graders and AI tools must be carefully trained to ensure accuracy, transparency, and trust.

Student trust is key

Whenever a new technology arises, educators must deal with the challenge of maintaining student trust in light of what may appear to be (or simply is) a delegation of teaching duties. This tension is not unique to contemporary artificial intelligence. Similar concerns accompanied the adoption of internet-based learning environments (Howland and Moore 2002, Wang 2014), learning management systems (Khalil et al., 2023), and earlier forms of computer-assisted instruction, each of which was promoted as a means of increasing efficiency, personalization, and scalability while raising questions about the changing role of the instructor. Likewise, AI-assisted grading systems may improve consistency and reduce instructor workload, yet students—here and in other work—have expressed concerns regarding bias, accountability, and accuracy (Alamäki et al., 2024, Riahi and Cateté 2026). Research suggests that transparency regarding how such systems operate and the retention of meaningful human oversight are particularly important for sustaining perceptions of fairness and trust (Vanhoyweghen et al., 2026). As with previous waves of educational technology, the central challenge is therefore not whether educators will use technological tools, but how they can do so in ways that preserve students' confidence that educational decisions remain fair and ultimately guided by human expertise.

The professor is irreplaceable

Following the above theme, we note that trust is best maintained by the instructional choices made by the instructor. In our analysis of student open-ended responses, it is clear that students trust the professor the most and look to the professor as the ultimate authority in the course. This role becomes even more important when there is the risk of

perceiving an abdication of responsibility. Following the above recommendations, we reiterate the importance of communication between the professor and students, as well as the need for transparency regarding how and when AI tools are used. More broadly, our findings suggest that AI-assisted grading should be framed as a tool that supports, rather than replaces, instructor judgment. Students appear willing to accept technological assistance when their instructors remain visibly accountable for grading outcomes and retain responsibility for interpreting, reviewing, and, when necessary, overriding AI-generated assessments. Such an approach is consistent with emerging research showing that trust in educational technologies depends not only on perceived accuracy, but also on transparency, fairness, and the presence of meaningful human oversight. The professor therefore serves an irreplaceable role as both subject-matter expert and accountable decision-maker, providing contextual judgment, empathy, and individualized consideration that students may not readily attribute to automated systems. AI may improve efficiency, but students ultimately appear to place their trust in the instructor who remains responsible for educational decisions.

From a practical perspective, several suggestions emerge. First, instructors should clearly disclose the role of AI in assessment, including which tasks are assisted by AI and which remain under direct instructor control. Second, students should be informed of the criteria by which work is evaluated (e.g., are rubrics supplied? Is there a grade-norming process?) and how AI-generated recommendations are reviewed before grades are assigned. Third, opportunities for appeal, clarification, and discussion with the instructor should remain

available, reinforcing that final responsibility lies with a human educator. Finally, instructors should communicate that AI serves as a decision-support tool rather than an autonomous evaluator, emphasizing the continued importance of professional judgement in interpreting student work. These practices align with recommendations from the learning analytics and AI-assessment literature, which consistently identifies transparency, fairness, and human oversight as critical to maintaining student trust (Đerić et al., 2025).

Graders and AI tools must be carefully trained

A large body of work has pointed to the value of training graders to be consistent, non-biased, and consequently fair. However, with the introduction of human graders in Study One, the variance in assigned grades increased, even though we spent considerable time developing a rubric and leading grade-norming activities to maintain accuracy and fairness. It makes sense that AI tools must be similarly trained with rubrics, feedback, and norming exercises. This conclusion also mirrors that of others (Tan et al., 2025, Liu et al., 2026). Research on both human and automated scoring has consistently demonstrated that reliable assessment depends not only on the quality of the scoring instrument but also on the calibration process used to align evaluators with intended standards (Tan et al., 2025). Automated scoring systems are generally trained on examples previously scored by human experts and evaluated according to their agreement with those experts, making them, in effect, participants in the same norming process that human graders undergo. Just as untrained human graders may apply a rubric inconsistently, poorly calibrated AI

systems risk producing judgements that diverge from instructor expectations, overemphasize superficial features, or fail to recognize unusual but high-quality responses.

Our findings therefore suggest that institutions should approach AI use in assessment following procedures similar to those employed for human raters. Prior to operational use, AI systems should be calibrated against instructor-scored exemplars, tested on representative student work, and evaluated for consistency across performance levels. Periodic re-norming may also be necessary as assignments, student populations, and instructional goals change with time. Such practices align with emerging recommendations for responsible AI-assisted grading, and emphasize not only transparency and human oversight, but continual quality assurance.

Limitations

Several features of the above studies limit our ability to extrapolate too far beyond the study context. For example, AI tools vary. We worked with a tool anchored in a specific LLM "backbone" (Google Gemini 2.5 Pro) and we cannot make claims about performance or perceptions of other systems. In addition, student perceptions are highly dependent on additional factors such as their current performance in a course, their perceptions of their peers and the professor, and past experiences with technology, the subject matter, or higher education in general. However, there is little ambiguity regarding our main claims,

and future research, after a greater societal adjustment to LLMs, may only add nuance rather than dramatically different conclusions.

Conclusion

We conclude by advocating caution in how we approach LLMs in assessment. These tools can be useful in assessing certain types of student work. However, even if these tools were infallible—which they are not—student perceptions are important. Instructional decisions can undermine or support student trust in how they are assessed, and key decisions about communication, transparency, and accountability should accompany any decision to use an AI tool as a grading partner.

List of abbreviatons

AI = Artificial Intelligence
ECTS = European Credit Transfer and Accumulation System
ICC = Inter-Correlation Coefficient
LLM = Large Language Model
RQ = Research Question
TA = Teaching Assistant

References

Alamäki, A., Khan, U. A., Kauttonen, J., & Schlögl, S. (2024). An experiment of AI-based assessment: Perspectives of learning preferences, benefits, intention, technology affinity, and trust. *Education Sciences, 14*(12), 1386. https://doi.org/10.3390/educsci14121386

Ballen, C. J., Aguillon, S. M., Awwad, A., Bjune, A. E., Challou, D., Drake, A. G., George-Williams, S. R., Horgan, K., Knight, J. K., McGaugh, S., Michelangeli, M., Mosher, B., Soneral, P. A. G., & Cotner, S. (2019). Smaller classes promote equitable student participation in STEM. *BioScience, 69*(8), 669-680.

Ballen, C. J., Aguillon, S. M., Brunelli, R., Drake, A. G., Wassenberg, D., Weiss, S. L., Zamudio, K. R., & Cotner, S. (2018). Do small classes in higher education reduce performance gaps in STEM? *BioScience, 68*(8), 593-600.

Bates, D., Mächler, M., Bolker, B., & Walker, S. (2015). Fitting linear mixed-effects models using lme4. *Journal of Statistical Software, 67*(1), 1-48. https://doi.org/10.18637/jss.v067.i01

Bhandari, S., & Pardos, Z. (2025). *Can language models grade algebra worked solutions? Evaluating LLM-based autograders against human grading*. In *Proceedings of the International Educational Data Mining Conference*. https://educationaldatamining.org/EDM2025/proceedings/2025.EDM.poster-demo-papers.290/

Cedefop. (2023). *Skills anticipation in Norway (2023).* European Centre for the Development of Vocational Training. https://www.cedefop.europa.eu/en/data-insights/skills-anticipation-norway-2023

Christensen, R. (2025). *ordinal: Regression models for ordinal data* (Version 2025.12-29) [R package]. https://CRAN.R-project.org/package=ordinal

Cotner, S., & Ballen, C. J. (2017). Can mixed assessment methods make biology classes more equitable? *PLOS ONE, 12*(12), e0189610.

Dai, W., Tsai, Y. S., Lin, J., Aldino, A., Jin, H., Li, T., ... & Chen, G. (2024). Assessing the proficiency of large language models in automatic feedback generation: An evaluation study. *Computers and Education: Artificial Intelligence*, *7*, 100299.

Deepshikha. (2025). A comprehensive review of AI-powered grading and tailored feedback in universities. *Discover Artificial Intelligence, 5*, 251. https://doi.org/10.1007/s44163-025-00517-0

Đerić, E., Frank, D., & Milković, M. (2025). Trust in generative AI tools: A comparative study of higher education students, teachers, and researchers. Information, 16(7), 622.

Elshall, A. S., & Badir, A. (2025). Balancing AI-assisted learning and traditional assessment: The FACT assessment in environmental data science education. *Frontiers in Education, 10*, Article 1596462.

Enu, J. (2021). Factors affecting teacher educators adoption of formative assessment strategies in the mathematics classroom. *Journal of Education and Learning (EduLearn), 15*(4), 483-489.

Finn, J. D., Pannozzo, G. M., & Achilles, C. M. (2003). The “why’s” of class size: Student behavior in small classes. *Review of Educational Research, 73*(3), 321-368.

Fox, J., & Weisberg, S. (2019). *An R companion to applied regression* (3rd ed.). Sage. https://www.john-fox.ca/Companion/

Gamlem, S. M., McGrane, J., Brandmo, C., Moltudal, S., Sun, S. Z., & Hopfenbeck, T. N. (2026). Exploring pre-service teachers' attitudes and experiences with generative AI: A mixed methods study in Norwegian teacher education. *Educational Psychology, 46*(1), 27-51.

Haugen, A., Holme, T., & Rabben, Å. K. (2025). *Tilstandsrapport for høyere utdanning 2025*. Direktoratet for høyere utdanning og kompetanse.

Hopfenbeck, T. N., Zhang, Z., Sun, S. Z., Robertson, P., & McGrane, J. A. (2023). Challenges and opportunities for classroom-based formative assessment and AI: A perspective article. *Frontiers in Education, 8*, Article 1270700.

Howland, J. L., & Moore, J. L. (2002). Student perceptions as distance learners in internet-based courses. *Distance Education, 23*(2), 183-195. https://doi.org/10.1080/0158791022000009196

Khalil, M., Prinsloo, P., & Slade, S. (2023). Fairness, trust, transparency, equity, and responsibility in learning analytics. *Journal of Learning Analytics, 10*(1), 1-7. https://doi.org/10.18608/jla.2023.7983

Kizilcec, R. F., Huber, E., Papanastasiou, E. C., Cram, A., Makridis, C. A., Smolansky, A., Raduescu, C., et al. (2024). Perceived impact of generative AI on assessments: Comparing educator and student perspectives in Australia, Cyprus, and the United States. *Computers and Education: Artificial Intelligence, 7*, 100269.

Liu, T., Ye, L., & Yan, W. (2026). A framework for evaluation of large language models in essay assessment: Reliability, alignment, and causal reasoning. *Computers and Education: Artificial Intelligence, 10*, 100565.

Lund, B. D., Lee, T. H., Mannuru, N. R., & Arutla, N. (2025). AI and academic integrity: Exploring student perceptions and implications for higher education. *Journal of Academic Ethics, 23*(3), 1545-1565.

Luzano, J. F. (2024). *Assessment in mathematics education in the sphere of artificial intelligence: A systematic review on its threats and opportunities*. SSRN. https://ssrn.com/abstract=5221121

Meaders, C. L., Lane, A. K., Morozov, A. I., Shuman, J. K., Toth, E. S., Stains, M., Couch, B. A., Smith, B. I., LaMar, M. D., Cronin, S. J., Cooper, M. M., & Smith, M. K. (2020). Undergraduate student concerns in introductory STEM courses: What they are, how they change, and what influences them. *Journal for STEM Education Research, 3*(2), 195-216.

Ministry of Education and Research. (2023a). *Long-term plan for research and higher education 2023-2032 (Meld. St. 5 (2022-2023))*. Government of Norway.

Ministry of Education and Research. (2023b). *Outlook on the skills needs in Norway: Report to the Storting (2022-2023)*. Government of Norway.

Ministry of Education and Research. (2025). *Skilled workers for a new era: With higher vocational education (Report No. 11 to the Storting (2024-2025))*. Government of Norway.

NITO (Norwegian Society of Engineers and Technologists). (2025, October 20). *Demands for new government: Norway needs a science boost*. https://nye.nito.no/en/News/Nito-in-society/Feature-article-Norway-needs-a-science-boost/

Pustejovsky, J. E. (2026). *clubSandwich: Cluster-robust (sandwich) variance estimators with small-sample corrections* (Version 0.6.2) [R package]. https://CRAN.R-project.org/package=clubSandwich

R Core Team. (2025). *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. https://www.R-project.org/

Ralph, V. R., Scharlott, L. J., Schafer, A. G., Deshaye, M. Y., Becker, N. M., & Stowe, R. L. (2022). Advancing equity in STEM: The impact assessment design has on who succeeds in undergraduate introductory chemistry. *JACS AU, 2*(8), 1869-1880.

Ramesh, D., & Sanampudi, S. K. (2022). Automated essay scoring systems: A systematic literature review. *Artificial Intelligence Review, 55*, 2495-2527. https://doi.org/10.1007/s10462-021-10068-2

Research Council of Norway. (2025). *Science and technology indicators for Norway 2024/2025*.

Research Council of Norway. (2026). *The Research Council of Norway's strategy*.

Riahi, B., Storozhevykh, V., & Cateté, V. (2026). *Humanizing AI grading: Student-centered insights on fairness, trust, consistency and transparency*. arXiv. https://arxiv.org/abs/2602.07754

Salehi, S., Cotner, S., Azarin, S. M., Carlson, E. E., Driessen, M., Ferry, V. E., Foltz, L., Fortenberry, N., Liu, J., & Ballen, C. J. (2019). Gender performance gaps across different assessment methods and the underlying mechanisms: The case of incoming preparation and test anxiety. *Frontiers in Education, 4*, Article 107.

Tan, L. Y., Hu, S., Yeo, D. J., & Cheong, K. H. (2025). A comprehensive review on automated grading systems in STEM using AI techniques. *Mathematics, 13*(17), 2828. https://doi.org/10.3390/math13172828

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

Vanhoyweghen, A., Holst, V., Mobini, M., Van de Voorde, L., Vanleke, T., Verbruggen, B., Verbeken, B., Algaba, A., Verboven, S., Guerry, M.-A., Van Droogenbroeck, F., & Ginis, V. (2026). *Human-in-the-loop LLM grading for handwritten mathematics assessments*. arXiv. https://arxiv.org/abs/2603.13083

Velez, J. J., Sorenson, T., McKim, A., & Cano, J. (2013). Self-efficacy and task value motivation of students based on classroom, instructor and student variables. *NACTA Journal, 57*(4), 65-71.

Vittorini, P., Menini, S., & Tonelli, S. (2021). An AI-based system for formative and summative assessment in data science courses. *International Journal of Artificial Intelligence in Education, 31*(2), 159-185.

Vyas, V. S., & Reid, S. A. (2023). What moves the needle on DFW rates and student success in general chemistry? A quarter-century perspective. *Journal of Chemical Education, 100*(4), 1547-1556.

Wang, L., & Calvano, L. (2022). Class size, student behaviors and educational outcomes. *Organization Management Journal, 19*(4), 126-142.

Wang, Y. D. (2014). Building student trust in online learning environments. *Distance Education, 35*(3), 345-359.

Wangdi, T., Rigdel, K. S., Dawa, T., & Tshering, K. (2025). Using ChatGPT as an assessment tool in education: A systematic literature review of practices and limitations. *Issues in Educational Research, 35*(2), 818-837.

Wilson, D., Jones, D., Bocell, F., Crawford, J., Kim, M. J., Veilleux, N., Floyd-Smith, T., Bates, R., & Plett, M. (2015). Belonging and academic engagement among undergraduate STEM students: A multi-institutional study. *Research in Higher Education, 56*(7), 750-776.

Yeh, S. S. (2009). Class size reduction or rapid formative assessment? A comparison of cost-effectiveness. *Educational Research Review, 4*(1), 7-15.

Yu, Z., Liu, X., Mao, H., Liu, M., Chen, L., Xin, J., & Yu, Y. (2026). *Evaluating AI grading on real-world handwritten college mathematics: A large-scale study toward a benchmark*. arXiv. https://arxiv.org/abs/2603.00895

Appendix

Appendix A: Message given to students in the Human + AI treatment in Study One

This semester, we are piloting the use of OpenAI as an assessment partner. We used a large language model to generate feedback on your assignment, and this feedback was reviewed, edited where necessary, and ultimately approved by a human grader. To be clear: your grade was assigned by a human. The feedback may be incomplete, and we recommend that you also consult the official solutions published on the course learning platform. If you have any questions about the feedback or your assessment, please contact your teaching assistant in Course One or one of the assistants at the [name of mathematics tutoring room]. We hope you find this feedback useful!

Appendix B: Template for AI-generated feedback in Study Two, as specified in the prompt to the AI model

Feedback on [Assignment Name] – [Course Code] [Semester, year]

This is automatically generated feedback. This semester, we are testing the use of artificial intelligence as an assessment partner, with the aim of providing you, as a student, with formative feedback on submitted assignments. We used a large language model to generate feedback on your work. The feedback may be incomplete, and we recommend that you also consult the official solutions published on the course learning platform. If you have any questions about the feedback, please contact your teaching assistant in [Course Code] or one of the assistants in the [name of mathematics tutoring room]. The final assessment (score or pass/fail decision) will be determined by an instructor or teaching assistant. We hope you find this feedback useful!

General Feedback:

Total score: [X out of Y possible points]

[Write an introductory sentence describing the overall quality of the submission. Examples of aspects that may be mentioned include: choice of method, intermediate calculations, algebraic manipulations, justification of solutions, and whether the student demonstrates understanding and mastery of the material.]

Mention some strengths of the submission (up to 3 points):

- Example 1

- Example 2
- Example 3

Mention some areas for improvement (up to 3 points), based on the official solutions:

- Example 4
- Example 5
- Example 6

Below is an overview of the feedback for each problem.

Overview of Individual Problems

[Use a traffic-light model to categorize each problem. Each comment should begin with an emoji: ☑ if the solution matches the official answer, ⚠ if the solution is partially correct, and ✗ if the solution is incorrect or missing. After the emoji, state the score for that problem.]

If ☑: Write: *"The problem is solved correctly."* No further comments are needed.

If ⚠: Briefly explain what is correct and what could be improved or done differently. Do not provide a step-by-step mathematical solution. Instead, give general comments on the method and approach. Computational errors may be mentioned, but avoid including lengthy calculations. Refer the student to the official solutions for more details. Keep the feedback concise.

If ✗: If the problem is missing, write: *"The problem is missing."* If the solution is incorrect, briefly explain how the problem could be approached without going into detailed calculations. Refer the student to the official solutions.

Problem 1: [emoji] [X out of Y points]

[Write comment based on the guidelines above.]

Problem 2: [emoji] [X out of Y points]

[Same as above.]

...

[Repeat for all problems]

Summary

[Write a brief concluding paragraph highlighting both the strengths of the submission and the areas where improvement is needed. End with a positive and encouraging statement for the student.]

Appendix C: Example of professor feedback given in Study Two

Ex1: a) you should repeat the change of variables ... 0p
b) good, 2p
Ex2: a) you are on to something, but the integral of cos is sin. 0.5p
b) 0p

Appendix D: Examples of peer feedback given in Study Two

Examples of thorough peer feedback:

Exercise 1:
a): Good computations. A small error in the final answer. Here you have forgotten to include that it is raised to the power of 8. Otherwise very good!
b) Nice! You could have expanded your answer as shown in the solution manual, but your answer is fully correct!
Total: 4 points for exercise 1.

Exercise 2:
a) You need to look a little closer on this one. You have used the correct values and arrived at the correct expression, but you have not integrated cos x. Here you should have used sin x. The answer looks to be correct, but your sign is wrong. The answer is positive, not negative.
b) Everything is correct here!
Total: 1,5 points for exercise 2.

In total: 5,5 points

Examples of less thorough peer feedback:

1a) 2 p
b) 1 p, not fully correct, but you are on to something .

2a) correct wrt what you have put two lines under. But here it was a little difficult to keep track of what you are doing. 2p
b) 0p

Appendix E: Example of AI-generated feedback given in Study Two, using the template given in Appendix A

Automatically Generated Feedback

This is automatically generated feedback. This semester, we are testing the use of artificial intelligence as an assessment partner, with the aim of providing students with formative feedback on submitted assignments. We used a large language model to generate feedback on your work. The feedback may be incomplete, and we recommend that you also review the official solutions published on the course learning platform.

If you have questions about the feedback, please contact your teaching assistant or one of the course support staff. The final assessment (score or pass/fail decision) will be determined by an instructor or teaching assistant.

We hope you find this feedback useful.

General Feedback

Total score: 4.0 out of 6 possible points

You demonstrate a good understanding of several central topics in the course, particularly in vector calculus, where you solved both problems completely correctly. In the calculus problems, you show that you understand the methods involved, such as the product rule and solving separable differential equations. However, it is important to be more careful with calculations and to ensure that all parts of a problem are answered. With a bit more attention to detail and accuracy, your solutions will become even stronger.

Strengths

- Very good and correct work on the entirety of Problem 3 (both the dot product and velocity calculations).
- Correct use of the product rule to find the derivative in Problem 1b.
- Good method for solving the differential equation in Problem 2a, including correctly determining the integration constant.

Areas for Improvement

- Make sure to answer all parts of a problem. In Problem 1b, you did not find and classify the critical point.
- Be very careful with signs. A small sign error in Problem 2a changed the resulting function.

- Double-check calculations involving definite integrals. In Problem 2b, there were errors in both the antiderivative and the evaluation, which led to an incorrect answer.

Overview of Individual Problems

Problem 1a: ☑ 1/1 point

The problem is solved correctly.

Problem 1b: ⚠ 0.5/1 point

You correctly calculated the derivative using the product rule, which is very good.

However, the problem also asks you to determine any critical points. This requires setting the derivative equal to zero, solving for the variable, and using a sign chart to classify the critical point. This part of your solution is missing.

Problem 2a: ⚠ 0.5/1 point

The method used to solve the separable differential equation is correct, and you correctly determined the integration constant. Unfortunately, there is a sign error in the final expression for the function. Please refer to the official solution for details.

Problem 2b: ✗ 0/1 point

You set up the correct integral based on the function found in Problem 2a, but there are several errors in the calculation. Both the antiderivative and the evaluation of the limits appear to contain mistakes, resulting in an incorrect final answer.

It is a good idea to go through calculations like these carefully, step by step, to avoid errors. See the official solution for the correct approach.

Problem 3a: ☑ 1/1 point

The problem is solved correctly.

Problem 3b: ☑ 1/1 point

The problem is solved correctly.

Summary

You demonstrate a very strong understanding of vector calculus. In calculus-related problems, it is important to focus on accuracy in calculations and to read the problem statements carefully to ensure that all required parts are addressed.

You have a solid foundation. Keep up the good work, and your results will continue to improve. Good luck with your studies!

Appendix F: Consent form given to students for participation in Studies One and Two

Consent form

Thank you for considering participation in this study on students' perceptions of mathematics and science courses at [faculty and university name]. The survey takes approximately 10–15 minutes to complete. Your responses are highly valuable and will help us improve both teaching and student assessment.

This research project is conducted by [name of research center] at [name of university].

The purpose of the study is to evaluate your opinions and perceptions in order to help us identify opportunities for improving the learning environment. We are interested in your feedback on various aspects of your background and experiences in this course, as well as in higher education more generally. Your responses will be used to evaluate how we teach our courses and how we can improve the student experience. The data may also be used for academic publications.

By providing your consent (at the bottom of this page) and submitting the survey, you grant us permission to:

1. Use and analyze the survey data; and
2. Retrieve information from the student administration system related to your academic performance (e.g., grades from this course and grades from upper secondary/high school).

We will only use your personal data, such as your student ID number, name, age, gender, and grades, for the purposes described above. All personal data will be processed in accordance with the General Data Protection Regulation (GDPR).

Your personally identifying information (such as your name and student ID number) will only be used to link your survey responses with academic performance data. Your responses will be stored [name of university website] for a maximum of 24 hours before

being deleted once they have been downloaded and stored on a secure computer that course instructors and teaching staff cannot access. The data will then be anonymized and stored on a secure, password-protected computer with two-factor authentication before being analyzed.

You may discontinue the survey at any time. Participation in this project is entirely voluntary. If you choose to participate, you may withdraw your consent at any time without providing a reason. If you withdraw, all information associated with you will be anonymized. There will be no negative consequences if you choose not to participate or if you decide to withdraw at a later stage. After the initial 24-hour period, the course coordinator and other teaching staff involved in the course will not have access to your responses. Your participation will not affect your standing in the course or your status at the university in any way. You have the right to submit a complaint to the relevant national data protection service for the higher education sector (Sikt). Throughout the project period, you may request that all data you have submitted be deleted.

The project will run from 1 August 2023 to 31 December 2028. After this period, all data will be deleted.

For more information about any aspect of the project, please contact [name and email of project leader]. You may also contact [name and email of university data protection officer] regarding privacy-related questions.

We greatly appreciate your time and assistance. Thank you for your participation!

I have received and understood the information about the project [name of project] and am participating voluntarily. I consent to participate in the survey and to the processing of my personal data and survey responses until the conclusion of the project (approximately 31 December 2028). I also consent to researchers accessing my upper secondary/high school grades and my final grade in this course for research purposes.

☐ **I agree to participate in this study.**

☐ **I do not agree to participate in this study.**