The Role of Performance Feedback in Enhancing Human-AI Collaboration: Evidence from A Randomized Classroom Experiment*

Antonio Dávila[1]
Antonio.davila@unil.ch
ORCID ID: 0000-0002-8815-4201

Ismail El Fassi[2]
Ismail.elfassi@unisg.ch
ORCID ID: 0009-0000-5883-9667

Daniel Oyon[1]
Daniel.oyon@unil.ch
ORCID ID: 0009-0004-0102-6832

Nicolas Rudolf[1]
Nicolas.rudolf@unil.ch
ORCID ID: 0000-0002-3717-5555

[1]University of Lausanne
[2]University of St. Gallen

August 2024

Abstract: We conduct a randomized field experiment in which participants solve numerical and conceptual management questions under time constraints and strong incentives to outperform their peers. We find that an autonomous human-AI collaboration strategy—where AI provides the solution to a task—initially outperforms by 12 percent an interactive human-AI collaboration strategy—where humans co-find the solution to a task, indicating that an autonomous use of AI increases performance for objective and quantitative tasks under time constraints and high data processing requirements. However, the performance gap disappears after subjects receive relative performance feedback (RPF) on their overall performance. Furthermore, after an additional round of feedback that includes more detailed performance analysis including task-specific RPFs, we find another incremental increase in performance for the interactive strategy similar in magnitude to the first RPF. These findings suggest that feedback helps participants using an interactive strategy to learn how to better use AI tools. In contrast, an autonomous strategy does not benefit from RPF.

JEL classification: C93, D83, M41
Key Words: Generative Artificial Intelligence, ChatGPT, Human-AI Collaboration, Relative Performance Feedback.

* We gratefully acknowledge financial support from the Swiss National Science Foundation (SNF Grant 10.002.887). We appreciate the helpful comments of Klaus Moller, Mael Schnegg, and seminar participants at the University of St. Gallen and at European Accounting Symposium for Young Scholars (EASYS). We also thank Deborah Dulex for providing excellent research assistance.

1. Introduction

Artificial intelligence (AI) is rapidly becoming an important contributor to individual and organizational performance. Thus, understanding how strategies and processes to enhance human-AI collaborations affect performance are important research and management questions (Berente et al. 2021; Maslej et al. 2024).[1] This paper explores how human-AI collaboration strategies affect performance in managerial tasks, and how this performance evolves as users receive feedback.

AI functionality for processing and generating text, images, videos, and other types of data is driving the integration of AI tools into organizational processes to autonomously perform specific activities and support people across a wide range of tasks (Murray, Rhymer, & Sirmon, 2021; Shavit et al., 2023). The fast-increasing power and accuracy[2] of generative AI tools, such as ChatGPT, Gemini, or Mistral, can be integrated into organizations' processes beyond the automation of simple tasks and augment human capabilities that translate into competitive advantages (Krakowski, Luger, & Raisch, 2023).

Not surprisingly, organizations are integrating AI tools into workflows to enhance human-AI collaboration (Estep, Griffith, & MacKenzie, 2023; Raisch & Krakowski, 2021; Wilson & Daugherty, 2018). Prior research has explored various collaboration strategies (Brynjolfsson & McAfee, 2014; Davenport & Kirby, 2016; Dell'Acqua et al., 2023; Dwivedi et al., 2023; Raisch & Krakowski, 2021; Wilson & Daugherty, 2018). The choice between these strategies depends on the specific organizational context and the nature of the tasks involved. Raisch and Fomina (2024)

[1] Berente et al. (2021: 1440) suggest that "the interaction between humans and AI is perhaps the key managerial issue of our time".

[2] This increasing functionality has spurred growing interest in the research community to assess the performance of AI tools across various domains, including auditing (Bertomeu, Cheynel, Floyd, & Pan, 2021; Christ, Emett, Summers, & Wood, 2021; Fedyk et al., 2022), accounting (Bertomeu, 2020; Ding et al., 2020; Eulerich et al., 2022; Peng et al., 2023; Wood et al., 2023), education (Extance, 2023), human resources (Campion et al. 2024), law (Choi, Monahan, & Schwarcz, 2023; Kleinberg, Lakkaraju, Leskovec, Ludwig, & Mullainathan, 2018), marketing (Castelo, Bos, & Lehmann, 2019), medical diagnosis (Esteva, Chou, Yeung, Naik, et al., 2021; Lebovitz, Lifshitz-Assaf, & Levina, 2022), psychology (Fan, Sun et al. 2023) and more.

theorize on potential merits and outcomes of human-AI collaboration strategies, confronting the interactive strategy—where AI systems augment human capabilities through close interaction that leverages the strengths of both human and AI tools and promotes mutual learning—to an autonomous strategy—where tasks are delegated to AI tools with minimal human intervention.

Organizations can rely on existing performance management tools, such as performance feedback to effectively motivate and enhance human-AI collaborations. However, the relevance of these traditional tools in human-AI collaborations is uncertain (Harrison & Dossinger, 2018). Existing evidence on the impact of performance feedback on task performance focuses solely on human-performed tasks (Allen et al., 2017; Azmat et al., 2019; Azmat & Iriberri, 2010; Eyring, Ferguson, & Koppers, 2021; Eyring & Narayanan, 2018; Hannan et al., 2019). However, feedback directed at humans may also be relevant to shape human-AI collaboration. Yet, the effect of this traditional management tool on AI-collaboration strategies is unknown. We provide empirical evidence on two research questions: (1) Which type of human-AI collaboration strategy benefits performance in management accounting tasks? and (2) Does performance feedback affect the performance of human-AI collaboration strategies differently?

To address these questions, we conduct a controlled randomized field experiment within an undergraduate quantitative management accounting course. Students perform weekly quizzes with conceptual and numerical questions under time constraints for ten weeks. To motivate students to participate in the experiment a bonus is given to them depending on their performance; in particular, students ranking in the top 50% of the quizzes test receive a significant bonus on their final grade.[3] We create a controlled classroom environment with a safe-exam browser where

[3] For equal treatment purposes, we also gave the bonus to those in the top 50% of their respective randomly assigned group.

students can only use a generative AI tool (ChatGPT-3.5) during the quizzes, blocking any other documentation or communication and allowing no other tools.

We employ a mixed model design with treatments at both the question and subject levels.[4] The main treatment in our study operates at the question level (within subjects). We guide participants toward interactive or autonomous human-AI collaboration strategies (Raisch & Fomina 2024). To implement this, every question is displayed in two formats: an image (screenshot) or a text format. The image format encourages an interactive collaboration strategy as the AI tool cannot interpret the text of the screenshot and students are forced to "translate" the image into queries to the AI tool. Students have on average 90 to 120 seconds to answer each question. This time constraint deliberately reduces the possibility of students deviating from an interactive collaboration strategy when exposed to the image format, because doing so would require manually typing the entire question into the AI tool.[5] Therefore, the image format encourages participants to create shorter prompts that distill the main elements of the question. Conversely, the text format allows participants to copy and paste the entire question into the AI interface, enabling the AI to generate solutions with minimal participant interaction. This format facilitates an autonomous collaboration strategy.

This within subject treatment, by duplicating each question in two formats, is necessary to keep the experiment hidden from participants until its conclusion. It also allows us to increase the power of the model in case the number of users, which is unknown in the beginning of the experiment, turns out to be low. Finally, it allows us to control for unobserved subject characteristics using subject-fixed effects. In addition, we create two additional randomly assigned

[4] This design is optimal in the context of our study. First, it keeps the treatment hidden to subjects until the end of the experiment. Second, because the number of AI users is unknown ex-ante, this ensures that we observe within-subject differences even if this number turns out to be small.

[5] Students could not use images as inputs in ChatGPT because screenshots were disabled during the exam through the Safe Exam Browser Moreover, by the time of the experiment, OpenAI had not yet implemented this option.

groups of students at the beginning of the experiment that either receive 80% or 20% of the questions as an image. This between-subject treatment allows us to capture the effect of feedback on AI usage management. Before the start of each quiz, subjects can choose to use the AI tool. While this choice may lead to self-selection into treatment in our setting, we cannot compel students to use AI to solve questions. However, within each group (AI users versus non-AI users), students are randomly exposed to questions that facilitate different human-AI collaboration strategies. Therefore, our setting allows selection for using AI tools but not for receiving questions in a specific format. Furthermore, we address potential self-selection bias by employing subject-fixed effects. If the choice to use the AI is driven by time-invariant student characteristics, subject-fixed effects control for this effect. In addition, we provide robustness tests that explicitly model the selection using a Heckman procedure.

After one-third of the quizzes have been completed, we provide students with relative performance feedback (RPF) reporting their performance relative to the median score as well as the score attributed to the AI tool without human intervention, which is equivalent to the 75-percentile performance of the class. The supplementary information on AI performance provides incentives to students to further interact with AI.[6] After completing two-thirds of the quizzes, we provide students with a second and more detailed RPF. This feedback includes the same information as the first feedback, but it also separately details students' relative performance on conceptual and numerical questions.

We find that the performance of subjects using an interactive collaboration strategy is initially 12 percent (6.1 percentage points) lower than the performance of subjects using an autonomous collaboration strategy. This finding suggests that, under time constraints and for tasks requiring

[6] This was labeled as ChatGPT's performance (Appendix D). This is interpreted as the AI score on the quizzes if the task was automated without human interaction. The score is set at the third quartile of the performance distribution, therefore encouraging the interaction with the AI for most subjects.

data processing rather than creativity, autonomous collaboration enhances performance compared to an interactive collaboration strategy. However, the performance gap disappears after subjects receive relative performance feedback (RPF) on their overall performance, with the interactive strategy increasing its performance by 15 percent (7.6 percentage points) compared to the autonomous strategy. This finding is consistent with feedback provision encouraging subjects who use an interactive strategy to improve their performance through higher efficacy and higher effort. In contrast, our results suggest that subjects using an autonomous collaboration strategy may over-rely on AI in time-sensitive contexts. In fact, we observe that RPF not only increases performance of the interactive strategy, but students spend also more time per question post-feedback when they use AI interactively. This difference was not significant before feedback was provided. Furthermore, recent literature argues that more frequent and more detailed feedback does not necessarily improve performance (Casas-Arce et al., 2017; Kohler et al., 2023; Lam et al., 2011; Lu, 2022; Lurie and Swaminathan 2009). We explore the effect of a second more detailed feedback that includes not only information on overall performance but also on individual task performance. Our results indicate that providing additional and more detailed feedback increases performance for an interactive collaboration strategy. Interestingly, the performance effect of this second feedback is similar in magnitude to the first RPF, indicating that the second feedback is as relevant and informative as the first one. Finally, prior studies indicate that the impact of feedback on performance varies depending on pre-feedback performance relative to peers (Eyring & Narayanan, 2018; Kluger & DeNisi, 1996). In our setting, we find no significant differences in the RPF effects across subjects that are below or above median performance nor across performance quartiles.

We further explore factors influencing students' decisions to rely on the AI tool. While prior research suggests that individuals with lower performance are more inclined to adopt AI tools, we

find no statistically significant relationship between AI adoption and prior performance (Allen & Choudhury, 2022; Commerford, Dennis, Joe, & Ulla, 2022; Logg, Minson, & Moore, 2019). Previous findings also suggest gender differences in behaviors and preferences including trust and technology adoption (Croson & Gneezy, 2009; Gefen & Straub, 1997; Reuben, Sapienza and Zingales, 2024). However, we find no significant effect of gender on AI adoption.[7]

Finally, we use a placebo test to verify that the difference in performance across the two formats of displaying questions indeed captures the differences in human-AI collaboration strategies instead of other confounding factors. In particular, we examine performance differences between the two display formats for subjects not using the AI tool. The absence of any performance disparity in this subgroup of non-users suggests that the observed effects in the AI user group are attributable to changes in human-AI collaboration strategies, not to unobserved behavioral differences related to the question display format. ← How is this useful?

Our study contributes to the expanding field of human-AI collaboration in management and the effect of traditional management tools. Contrasting interactive versus autonomous collaboration strategies and the use of feedback, we provide empirical evidence on their impact on quantitative and conceptual managerial tasks as well as the effect of feedback provision, enhancing our understanding of effective human-AI collaboration strategies and their interaction with traditional management tools (Brynjolfsson & McAfee, 2014; Murray et al., 2021; Raisch & Krakowski, 2021). Our findings are also relevant to practitioners as technology providers increasingly offer services to facilitate human-AI collaborations (e.g., Cloud software, IBM Watson, Microsoft Co-Pilot for PowerBI). Accounting is at the forefront of AI augmentation because of the large amounts of data and structured processes that support well-defined decisions

[7] This may be due to a lack of statistical power. ← Yeah, drop that part.

(Bertomeu, 2020; Bertomeu et al., 2021; Ding et al., 2020) making the effective selection and management of AI collaboration strategies highly relevant for accounting practice.

Additionally, our research contributes to the management accounting literature on feedback. While existing feedback studies focus on the impact of relative performance feedback on tasks performed solely by humans (e.g., Eyring et al., 2021; Eyring & Narayanan, 2018; Hannan, Krishnan, & Newman, 2008; Hannan et al., 2013; Kohler et al., 2023), our study examines how feedback influences task performance when humans collaborate with AI tools. Our findings indicate that performance feedback can help to close the performance gap between interactive and autonomous collaboration strategies, as participants increase effort or learn how to collaborate more effectively with AI tools. Conversely, providing performance feedback when AI is used autonomously does not improve performance and could even be counterproductive if it encourages over-reliance on AI.

1. Literature Review and Hypothesis Development

1.1. Human-AI Collaboration Strategies and Task Performance

Raisch and Fomina (2024) propose two distinct strategies for collaboration between humans and generative AI: autonomous and interactive collaboration strategies.[8] An autonomous collaboration strategy involves the generative AI tool independently generating solutions to a given problem or task. It does not require significant guidance from the human collaborator once the task has been introduced to the AI tool as the AI tool independently generates and communicates the solution. The human then reviews and decides whether the solution(s) that the AI tool provides is (are) appropriate. This strategy leverages the AI's capability to rapidly explore a wide variety of

[8] The third human-AI collaboration introduced in their framework (sequential) does not involve the use of generative AI and is therefore not of interest in our paper. "A sequential search uses predictive AI for the problem definition, but a human subsequently conducts a solution search without the use of generative AI" (Raisch and Fomina, 2024: 15)

solutions, leaving the final decision-making and judgment to the human. In contrast, an interactive strategy involves a continuous, real-time collaboration between the human and the AI tool. Rather than working independently, the human engages the AI tool in a joint search for solutions, with the human providing input, feedback, and adjustments throughout the process, refining the parameters of the stated task and suggesting modifications. The interactive strategy emphasizes joint work where both the human's judgment and the AI's computational power collaborate to work towards a solution.

In essence, an interactive strategy fosters a partnership where both human and AI insights are integrated throughout the problem-solving process, while an autonomous strategy capitalizes on the AI's ability to generate diverse solutions quickly, with human oversight for final selection. Hence, autonomous collaboration is particularly useful when decision-making speed is critical with generative AI providing answers almost instantaneously, making the trade-off between speed and accuracy almost inexistent (Shrestha, Ben-Menahem, & von Krogh, 2019).

Prior literature documents that AI tools that make sense of large unstructured datasets are superior to humans in processing data comprehensively (Murray et al., 2021; Raisch & Fomina, 2024), while humans outperform in tasks requiring generating creative ideas (Brynjolfsson & McAfee, 2014; Raisch & Fomina, 2024). Therefore, we predict that an autonomous collaboration strategy increases performance for tasks within the scope of AI capabilities (Dell'Acqua et al., 2023), thus tasks requiring more data processing than creativity (Raisch & Fomina, 2024) such as quantitative management problem-solving.

H1. An autonomous human-AI collaboration strategy leads to higher performance than an interactive human-AI collaboration strategy in time-sensitive quantitative management tasks.

1.2. Relative Performance Feedback and Task Performance

Traditional management tools such as performance feedback can affect human-AI interaction. Prior accounting research indicates that feedback often improves task performance through learning and motivation, but its effectiveness can vary with task context and the nature of the task itself (e.g., Eyring et al., 2021; Eyring & Narayanan, 2018; Hannan et al., 2013). Certain context and task-specific factors, including time constraints, task ambiguity, complexity, and uncertainty can reduce the effectiveness of feedback (Brehmer, 1980; Frederickson, 1992; Gevers and Demerouti, 2013; Hannan et al. 2019; Harrison & Dossinger, 2018; Hoffman, Earle, & Slovic, 1981; Kruger & DeNisi, 1996). Therefore, as human-AI collaboration involves not only humans' analytical thinking but also human interaction and reliance on AI tools' opaque intelligence process, it is unclear whether and under which circumstances feedback can be beneficial to human-AI collaborations' performance.

Feedback may improve interactive human-AI collaborations and enhance task performance, encouraging deeper human-AI collaboration to jointly analyze data and look for solutions through increased cognitive and analytical engagement. For instance, in complex and ambiguous tasks where feedback can potentially reduce the focus on learning (Brehmer, 1980; Hannan et al., 2019; Harrison & Dossinger, 2018; Hoffman, Earle, & Slovic, 1981; Kruger & DeNisi, 1996), interactive human-AI collaboration may compensate this drop in learning motivation by encouraging humans to rely on AI to support their analytical thinking and comprehensive understanding of the task. In addition, feedback informs individuals engaged in interactive human-AI collaboration about the effectiveness of their time utilization, guiding them to learn how to best leverage AI tools to meet time constraints (Gevers and Demerouti, 2013). Relative feedback also compares performance to peers, which can increase motivation to improve performance (Festinger, 1954; Kohler et al., 2023). This form of social comparison is likely more pronounced in interactive human-AI

collaborations where subjects are more engaged in the problem-solving process than in autonomous AI collaborations.

In autonomous human-AI collaborations, the AI tool provides solutions with minimal human engagement. Feedback in this setting allows users to evaluate AI-generated outcomes, but it does not directly motivate them to get further involved in the problem-solving process, limiting the potential of feedback to enhance performance. Particularly in settings where time constraints and outcome uncertainty play significant roles, the instantaneous nature of autonomous AI solutions means that feedback might only serve to confirm the effectiveness of these solutions rather than improving human-AI collaboration performance (e.g., Frederickson 1992; Kruger & DeNisi 1996). Moreover, because an autonomous strategy reduces the need for analytical thinking, feedback in this context may fail to motivate significant behavioral changes or encourage learning (Shrestha, Ben-Menahem, & von Krogh, 2019).[9] This argument is in line with the dual-process theory of cognition (Kahneman, 2013), which suggests that when subjects rely on AI for analytical thinking, they tend to increase their reliance on intuitive decision-making, especially in time-sensitive situations. Furthermore, the ecological rationality perspective posits that individuals are likely to adhere to the default option – in this case, the solution generated autonomously by AI– especially when time constraints limit their ability to deliberate (Broder and Schiffer, 2003; Todd and Gigerenzer, 2007). In contrast, interactive collaboration encourages participants to actively engage in the analytical thinking process alongside AI. Thus, feedback encourages them to reflect on how they process information and engage with AI. Our second hypothesis captures these arguments.

H2. The extent to which RPF enhances human-AI performance will be greater under an interactive than an autonomous human-AI collaboration.

[9] According to Mollick "people "fall asleep at the wheel" when faced with "good-enough" AI content. They become less critical, and less likely to fact-check or thoroughly edit the AI's output." See his Financial Times opinion article https://www.ft.com/content/389e505c-a1cc-4176-a592-dd1d0fa171b8.

2. Methodology

2.1. Field Setting

We conducted the field experiment in an undergraduate management class covering quantitative and conceptual material. Students participated in weekly quizzes to assess their understanding of the course material. Quiz participation was strongly incentivized, rewarding students above the median performance with a 0.5-point bonus on their grade (over a maximum grade of 6). The grade bonus encouraged students to participate in the quizzes. On average 310 out of 371 students answered the quizzes in each session, with the majority of students answering more than 95% of the questions.[10] The quizzes covered strategic control topics, such as financial planning, designing business units, transfer prices, and translating strategies into performance targets. Specific tasks assessed include calculating break-even points under various scenarios, using different cost systems to measure product cost and profitability, and calculating sales mix and contribution margins.

To create a controlled environment for the experiment, students were required to use a safe exam browser that restricted access to external information and communication sources, except for ChatGPT-3.5. At the start of each quiz, students could choose to access ChatGPT through a direct link provided on the first page of the quiz. We observe on the quiz platform whether students clicked on this link, allowing us to track ChatGPT usage for each individual quiz.

The course spanned 12 weeks, with quizzes commencing in week three. The experiment began in week six, providing a pre-experimental phase to gather pre-experimental data on student performance without access to ChatGPT. During this initial phase, questions were displayed in only one format (text or image). From week six onwards, students could access ChatGPT on the test platform. All quizzes covered course-related material, incorporating a mix of conceptual and

[10] Two subjects did not consent to the usage of their data and therefore were excluded from the sample.

numerical questions. Numerical and conceptual questions accounted for 76% and 24% of the observations, respectively. In week seven, students received feedback on their individual scores and the median score. In week nine, we provided students with another round of feedback, including more detailed information on their performance in conceptual and analytical tasks.

2.2. Experiment Design

The main treatment in our study is at the question level (within-subjects). Each question is presented either as text, corresponding to an autonomous collaboration strategy, or as image, corresponding to an interactive collaboration strategy. Both formats provide the subject with the same information (see example in appendices B and C). At the beginning of the experiment, we randomly assign students to two groups kept constant throughout the experiment. Each group receives either 80% or 20% of the questions as images. In our setting, this combination of within-subject and between-subject treatment offers two main advantages over a pure between-subject design. First, before the start of the experiment, an important unknown parameter is the number of subjects that will decide to use the AI to solve the questions. The mixed-model design increases statistical power and the likelihood of capturing a within-subject effect, even if the number of AI users is small (Charness, Gneezy, & Kuhn, 2012). Second, the between-subject part of our design is necessary to keep the existence of the two different types of question display formats hidden from participants until the conclusion of the study. Third, the within-subject design allows us to include subject fixed effects that absorb time-invariant subject characteristics, such as gender or general ability.

However, the mixed-treatment design is also subject to some limitations. Exposing participants to both image and text could influence their AI usage behavior depending on the order of the questions. We address this issue by randomly ordering the questions to avoid ordering bias. Given the time restrictions, questions in image format implicitly require subjects to use short queries with

the AI tool. Therefore, the image format deliberately constrains subjects' ability to deviate from an interactive collaboration strategy. In contrast, the text format allows for the entire question to be copied and pasted into a single AI query, enabling the AI to generate solutions with minimal user interaction, thus, encouraging an autonomous collaboration strategy.

One week after the start of the experiment and following one-third of the quizzes, we provide students with the first relative performance feedback (RPF) on their overall scores. This feedback includes their individual score, the median score required to receive the bonus, and the "AI score" – the hypothetical score if the AI had answered the questions without human intervention set at the third quartile (see the feedback template in Appendix D). This allows for further examination of students' behavior if they were above and below the median, as well as if they find themselves above or below the AI performance.

After completing two-thirds of the quizzes in week nine, we provide subjects with a second round of feedback. This feedback introduces an additional layer of information by providing students with information on their performance relative to peers for both conceptual and numerical questions separately. In addition, we provide the same information as in the first feedback regarding students' overall performance relative to peers (see Appendix D).

3. Statistical Models

We first explore subjects that are more likely to use the AI tool. Previous findings indicate that less knowledgeable and lower performing subjects trust more AI compared to their peers (Allen & Choudhury, 2022; Logg et al., 2019), which may lead to higher adoption rates among less knowledgeable individuals. Furthermore, prior literature indicates that gender differences result in different preferences and levels of trust in technology that could affect the probability of AI usage (Gneezy, Niederle, and Rustichini, 2003; Gneezy and Rustichini, 2004; Croson & Gneezy, 2009;

Gefen & Straub, 1997; Reuben, Sapienza and Zingales, 2024). To explore whether these subject characteristics influence AI adoption, we estimate the following probit model:

(1) $$AI\ User_{i,j} = Subject\ Characteristics_{i,j} + e_{i,j}$$

Where AI User is a binary variable that takes the value of one if the subject opted to use the AI in at least three out of 11 sessions (see Table 1c). AI usage is defined as whether the subject accessed ChatGPT at the beginning of the quiz using the direct link provided in our secure test environment. AI users represent 20.5% of all students in our setting. Our results remain robust across various definitions of the AI-user group, including thresholds of using the AI at least three, four or five times.

Subject Characteristics includes two observable characteristics, gender and subject knowledge. First, *Female* takes the value of one if the subject is female and zero if the subject is male. Second, we use two different proxies for subject knowledge. First, we use a binary variable *High Knowledge* which is equal to one if the subject was performing above peers in the pre-experimental quizzes between week four and six and zero otherwise. Second, we employ *Subject knowledge* as a continuous measure of knowledge based on a subject's average pre-experimental score on quiz questions between week 3 and 5.

Next, we estimate the effect of the two human-AI collaboration strategies on performance during the pre-feedback period. Specifically, we estimate the following mixed-effect model for the pre-feedback period (week 6) within the subset of students who selected to use the AI in order to compare the two human-AI collaboration strategies:

(2) $$Score_{i,j} = Interactive\ Collaboration_{i,j} + Fixed\ Effects + e_{i,j}$$

Where *Score* takes the value of one if the answer to a question is correct and zero if wrong. Interactive Collaboration is a dummy variable set to one if the question appears as an image (interactive collaboration) and zero if it appears in text format (autonomous collaboration).

Given the hierarchical nature of our dataset, with students nested within questions, we employ a multilevel mixed-effects linear regression model (Tabachnick, Fidell, & Ullman, 2013). This approach accounts for the dependence of responses within questions and students, considering the nested structure of the data. Our results are robust to alternative estimator choices, such as OLS or Logit.

Furthermore, we include subject, question, and quiz fixed effects. Subject fixed effects capture time-invariant differences within subjects, such as personality attributes, gender, or general ability that could affect a subject's interaction with the AI tool. Question fixed effects capture differences in the difficulty and the type of questions (conceptual and numerical). Quiz fixed effects control for time effects, and similarities between questions within a given quiz (similar topic coverage, and difficulty and type of questions). In alternative specifications we relax the fixed effects and instead control for subject and task characteristics including a proxy for the difficulty of the question measured using the *Average score per question*, *Subject knowledge*, and *Female.* In specifications without subject fixed effects, we include the inverse mills ratio estimated in equation (1) to account for potential self-selection.

Next, we analyze changes in the students' performance after receiving the first performance feedback depending on the human-AI collaboration strategy. We include only the AI users' sample. We estimate the following model from week six until the second feedback in week nine:

(3) $$Score_{i,j} = Interactive\ Collaboration_{.j} + Interactive\ Collaboration_{.j} \times RPF\_1_{i,j} + Fixed\ Effects + e_{i,j}$$

Where RPF_1 is a binary variable that is equal to one for the period between the first and second feedback (week 7-9) and zero otherwise. All other variables are defined as in equation (2).

In the last step of our analysis, we examine how students' performance changes after the second feedback. Therefore, we estimate the following model over the entire experimental period, from week six until week 12:

(4) $Score_{i,j} = Interactive\ Collaboration_{.j} + Interactive\ Collaboration_{.j} \times RPF\_1_{i,j} + Interactive\ Collaboration_{.j} \times RPF\_2_{i,j} + Fixed\ Effects + e_{i,j}$

Where RPF_2 takes the value of one for periods after the second feedback and zero before it. All other variables are defined as in equation (2) and (3).

4. Results

4.1. Descriptive statistics

Table 1 presents descriptive statistics for the variables in our sample spanning the entire duration of the experiment fromweek six. We observe a pronounced variability in the difficulty level of the questions. The average score is 0.41 points per question, with a standard deviation of 0.24. Conceptual questions represent 26% and numerical questions 74% of all observations.

Additionally, there is diversity in subjects' knowledge as evidenced by their performance on pre-experimental questions, with subjects' individual average scores ranging from 0 to 0.82 during week 3-5. The average score of subjects in pre-experimental questions is 0.44 with an associated standard deviation of 0.16, indicating a moderate dispersion around the mean. In our sample, 37% of students are female.

Table 2 illustrates the evolution of AI users over time. AI users represent 20.5% of the subjects and 22% of the observations. We observe that 66 subjects used the AI tool in the week preceding the feedback, 59 in the first week following the feedback, and 79 in the second week after the feedback. We note that these usage patterns are only descriptive and may be influenced by other factors than feedback, including quiz length, question type, and the number of quiz participants. For example, in week eight, we gave subjects a long recapitulative quiz that may explain the surge in AI adoption. Moreover, in the last two sessions, quizzes involved complex problems with multiple parameters, reflecting the class material, which may have discouraged AI usage. We account for these variations by including question and quiz fixed effects in our main analyses.

Insert TABLE 1 about here

Insert TABLE 2 about here

4.2. Determinants of AI adoption

Table 3 reports the effects of gender and knowledge on the decision to adopt the AI tool. Overall, our results indicate no statistically significant differences in the adoption behavior related to either knowledge or gender.

Insert TABLE 3 about here

4.3. Human-AI Collaboration Strategies, Feedback, and Performance

Our main analysis begins testing the effect of human-AI collaboration (interactive vs autonomous collaboration strategy) on performance. Table 4 indicates that an interactive collaboration strategy is associated with lower performance. In our most stringent specification, which includes subject and question fixed effects in Column 5, the coefficient on Interactive collaboration is -0.07, (p-value < 0.01) suggesting that an interactive collaboration strategy decreases the question score by 7 percentage points relative to autonomous collaboration. The magnitude of this effect is substantial, representing a 19 percent reduction in performance compared to the average score of 38 percent achieved with an autonomous collaboration strategy (constant term is 0.38, $p < 0.01$). Results remain quantitatively and qualitatively similar when replacing subject and question fixed effects with subject and questions characteristics, including task difficulty, subjects' knowledge, and gender. The results are consistent with hypothesis one and an human-AI autonomous collaboration strategy is superior to an interactive strategy.

Insert TABLE 4 about here

Table 5 explores the effect of RPF on the interactive collaboration strategy relative to the

autonomous collaboration strategy. The coefficient for the interaction between Interactive Collaboration and RPF in our most stringent specification, including quiz, subject and question fixed effects in Column 5 is 0.06, (p-value < 0.05) suggesting that RPF is incrementally beneficial for subjects' performance under an interactive collaboration strategy compared to an autonomous strategy. Our results indicate that interactive collaboration initially reduces performance by 6.1 percentage points before RPF and enhances performance by 7.6 percentage points after RPF, compared to autonomous collaboration. This leads to a non-significant average difference between both collaboration strategies over the full sample period as documented in Column 1 without the interaction effect of RPF.

Insert TABLE 5 about here

Next, we test the effect of a second and more detailed feedback. Increasing the feedback's level of detail and frequency does not always increase performance (Casas-Arce et al., 2017). Lam et al. (2011) argue that more frequent feedback may saturate an individual's cognitive resource capacity, consequently reducing task effort. This effect can be particularly pronounced when subjects delegate tasks to AI under an autonomous collaboration strategy, fostering overreliance on the AI tool.

Table 6 analyzes the entire period of the experiment to examine the effect of the second RPF. Consistent with Table 5, Table 6, Column 1 documents that the average effect of the interactive collaboration strategy compared to autonomous collaboration is insignificant. Similarly, in the pre-RPF period the interactive collaboration strategy is associated with an approximately 5.2-6.0 percentage points lower performance. However, the first RPF is associated with an increase in performance of the interactive collaboration strategy of 5.9-7.4 percentage points compared to the autonomous collaboration. The second feedback also positively affects the interactive collaboration

relative to the autonomous collaboration, with an effect size nearly equivalent to the first RPF (increase in performance of 5.4-7.4 percentage points). However, we note that the effect of the second RPF becomes insignificant at conventional statistical levels after including question fixed effects in Column 5.

Insert TABLE 6 about here

4.4. Validity of the display format as a proxy for the human-AI collaboration strategy

To examine whether image and text display formats accurately capture differences in human-AI collaboration strategies, we test for performance differences between the two display formats among subjects that do not use the AI tool. This test allows us to rule out that the question format itself influences subject performance other than through the collaboration strategy. Table 7 reports the results. Panel A presents the results for the pre-RPF period and Panel B for the full sample period including the RPF. Our findings suggest that the question format has no effect in the sample of non-AI users, indicating that the question format only influences performance through the way subjects use the AI tool.

Insert TABLE 7 about here

4.5. Additional analysis

Table 8 analyzes the influence of the two different AI collaboration strategies on the time spent in quizzes. It regresses the duration of the quiz in seconds on whether the subject received interactive or autonomous questions. To capture time usage differences between subjects, we use only observations from quizzes where each subject received the same format for all questions.[11]

[11] To achieve a 20%-80% split in question format across subjects, depending on the quiz, subjects in the 80% interactive group receive for example 3, 4, or 5 image questions and 0 or 1 text question, so that, on average, they receive 80% image and 20% text questions. In this between-subject test, we only use quizzes where the 80% interactive group receives 0 text questions, and the 80% autonomous group receives 100% text questions to observe differences in time management at the quiz level.

We find no significant differences in time utilization between the interactive and autonomous collaboration strategy before the RPF. However, our findings show that an interactive AI usage results in significantly more time spent on problem-solving after the first and second RPF, compared to the autonomous group. The interactive group spends on average 44 seconds more after the first feedback and 46 seconds more after the second feedback than the autonomous group. This is equivalent to an increase of 8% in total time spent on each quiz **(**average time spent on each quiz: 562 seconds). This indicates that feedback improves time management for an optimal interactive AI usage.

Insert TABLE 8 about here

Prior work highlights the discouragement effect of negative feedback (e.g., Goulas and Megalokonomou, 2021). Appendix E examines the effect of receiving below-peer performance feedback in the second RPF and subsequent outcomes in the final exam, controlling for factors such as quiz performance and AI usage during the quizzes. We find a negative relationship between below-peer performance in the second RPF and final exam performance. However, we find no significant relationship between the first RPF and performance in final exam. This finding indicates that, as the final exam deadline approaches, negative feedback appears to demotivate rather than incentivize further effort. We find similar results when we use an abnormal measure of performance in the final exam, measured by comparing subject performance in the exam to their average performance in the quizzes.

Prior studies indicate that feedback impacts motivation, learning, and ultimately performance differently depending on individuals' performance relative to their peers (e.g., Eyring & Narayanan, 2018; Kluger & DeNisi, 1996). In non-tabulated analysis, we test whether RPF affects subjects differently based on whether their performance is below or above peers at the time of

feedback, using our entire sample. We find no significant difference in the effect of feedback relative to subjects' prior performance. This observation remains consistent across our entire sample, for both AI user and AI non-user subgroups. Similarly, we find no differences in the effect of feedback when we divide our subjects based on their performance quartile at the time of the first feedback.

4.6. Robustness tests

In Appendix F, we assess the robustness of our main results in Table 4 using different estimators, including Ordinary Least Squares (OLS), Probit, and Logit, with subject and question fixed effects in all models. Column 1 shows the results using our baseline mixed-effects model for comparison. In Columns 2-4 the coefficients for Interactive Collaboration are consistently negative and statistically significant. The coefficient estimate of the OLS regression in Column 2 is the same as in our main model and is significant at the 5% level. The coefficients of the Probit and Logit regressions in Columns 3 and 4 respectively, are also significantly negative at the 1% level.

Appendix G replicates Table 5 using OLS, Probit and Logit estimators. Again, we find that our results appear to be robust to these alternative estimators.

Appendix H interacts control variables including subject and task characteristics with the RPF dummy variable to rule out that the feedback effect is driven by specific subgroups or particularities inherent to certain tasks or individuals. We also interact RPF with a dummy variable for the group receiving 80% of the questions in an image format (interactive group). This allows us to compare the effect across those who received 20% images and those who received 80% images to see whether the observed RPF effect is driven by only one of the two groups. The findings support hypotheses 1 and 2, namely that the interactive strategy has a significantly negative effect on task performance at first, and the RPF has a more positive effect on the interactive strategy. We find that none of interactions of RPF with the control variables is significant, suggesting that our main

results are not driven by any subgroups.

5. Discussion and Conclusions

Our study contributes to the expanding field of human-AI collaboration in management and accounting analyzing performance differences across human-AI collaboration strategies in a competitive, time-sensitive environment. We find that an autonomous collaboration strategy initially enhances performance. However, the introduction of feedback significantly improves the performance of the interactive strategy, pointing at the importance of feedback mechanisms in optimizing human-AI collaboration. Therefore, our study offers insights for organizations seeking to maximize the benefits of human-AI collaboration by implementing the appropriate process strategies and control mechanisms.

Our study is subject to some limitations. Specifically, we cannot observe how subjects modified their prompting behavior to the AI tool post-feedback, which could provide more detailed insights into the feedback mechanism. In addition, despite the opportunity to switch their behavior, subjects seem to not deviate from the autonomous strategy post-feedback, suggesting that autonomous AI strategies may reduce engagement and decision-making efforts (see also Ahmad et al., 2023). Future research could explore the long-term effects of feedback on AI collaboration strategies and investigate the detailed changes in cognitive processes underlying these behaviors.

Furthermore, our findings can help organizations to maximize the benefits of human-AI collaboration. Implementing feedback mechanisms can enhance the effectiveness of interactive human-AI strategies, leading to improved performance. Firms that use interactive human-AI collaborations could consider integrating feedback processes to encourage active engagement with AI tools, while integrating checks and balances that reduce the potential for over-reliance on AI in autonomous settings.

References

Ahmad, S. F., Han, H., Alam, M. M., Rehmat, M., Irshad, M., Arraño-Muñoz, M., & Ariza-Montes, A. 2023. Impact of artificial intelligence on human loss in decision making, laziness and safety in education. ***Humanities and Social Sciences Communications***, 10(1): 1-14.

Allen, E. J., Dechow, P. M., Pope, D. G., & Wu, G. 2017. Reference-Dependent Preferences: Evidence from Marathon Runners. ***Management Science***, 63(6): 1657–1672.

Allen, R. T., & Choudhury, P. 2022. Algorithm-Augmented Work and Domain Experience: The Countervailing Forces of Ability and Aversion. ***Organization Science***, 33(1): 149–169.

Azmat, G., Bagues, M., Cabrales, A., & Iriberri, N. 2019. What You Don't Know…Can't Hurt You? A Natural Field Experiment on Relative Performance Feedback in Higher Education. ***Management Science***, 65(8): 3714–3736.

Azmat, G., & Iriberri, N. 2010. The importance of relative performance feedback information: Evidence from a natural experiment using high school students. ***Journal of Public Economics***, 94(7–8): 435–452.

Berente, N., Gu, B., Recker, J., & Santhanam, R. 2021. Managing artificial intelligence. ***MIS quarterly***, 45(3): 1433-1450.

Bertomeu, J. 2020. Machine learning improves accounting: discussion, implementation and research opportunities. ***Review of Accounting Studies***, 25(3): 1135–1155.

Bertomeu, J., Cheynel, E., Floyd, E., & Pan, W. 2021. Using machine learning to detect misstatements. ***Review of Accounting Studies***, 26(2): 468–519.

Brehmer, B. 1980. In one word: Not from experience. ***Acta Psychologica***, 45: 223.

Bröder, A., & Schiffer, S. 2003. Take The Best versus simultaneous feature matching: Probabilistic inferences from memory and effects of reprensentation format. ***Journal of Experimental Psychology: General***, 132(2), 277.

Brynjolfsson, E., & McAfee, A. 2014. ***The second machine age: Work, progress, and prosperity in a time of brilliant technologies.*** (W. W. Norton.). New York, NY.

Casas-Arce, P. A. B. L. O., Lourenço, S. M., & Martínez-Jerez, F. A. 2017. The performance effect of feedback frequency and detail: Evidence from a field experiment in customer satisfaction. ***Journal of Accounting Research***, 55(5), 1051-1088.

Castelo, N., Bos, M. W., & Lehmann, D. R. 2019. Task-Dependent Algorithm Aversion. ***Journal of Marketing Research***, 56(5): 809–825.

Charness, G., Gneezy, U., & Kuhn, M. A. 2012. Experimental methods: Between-subject and within-subject design. ***Journal of Economic Behavior & Organization***, 81(1): 1–8.

Chen, W., & Srinivasan, S. 2023. Going digital: implications for firm value and performance. ***Review of Accounting Studies***, 1–47.

Choi, J. H., Monahan, A., & Schwarcz, D. B. 2023. Lawyering in the Age of Artificial Intelligence. ***SSRN Electronic Journal***. https://doi.org/10.2139/SSRN.4626276.

Christ, M. H., Emett, S. A., Summers, S. L., & Wood, D. A. 2021. Prepare for takeoff: improving asset measurement and audit quality with drone-enabled inventory audit procedures. ***Review of Accounting Studies***, 26(4): 1323–1343.

Campion, E. D., Campion, M. A., Johnson, J., Carretta, T. R., Romay, S., Dirr, B., Deregla, A., & Mouton, A. 2024. Using natural language processing to increase prediction and reduce subgroup differences in personnel selection decisions. ***Journal of Applied Psychology***, 109(3), 307–338.

Commerford, B. P., Dennis, S. A., Joe, J. R., & Ulla, J. W. 2022. Man Versus Machine: Complex Estimates and Auditor Reliance on Artificial Intelligence. ***Journal of Accounting Research***, 60(1): 171–201.

Croson, R., & Gneezy, U. 2009. Gender Differences in Preferences. ***Journal of Economic Literature***, 47(2): 448–74.

Davenport, T. H., & Kirby, J. 2016. ***Only humans need apply: Winners and losers in the age of smart machines.*** . New York, NY: HarperCollins.

Dell'Acqua, F., McFowland, E., Mollick, E. R., Lifshitz-Assaf, H., Kellogg, K., et al. 2023. Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality. ***SSRN Electronic Journal***. https://doi.org/10.2139/SSRN.4573321.

Ding, K., Lev, B., Peng, X., Sun, T., & Vasarhelyi, M. A. 2020. Machine learning improves accounting estimates: evidence from insurance payments. ***Review of Accounting Studies***, 25(3): 1098–1134.

Dwivedi, Y. K., Kshetri, N., Hughes, L., Slade, E. L., Jeyaraj, A., Kar, A. K., ... & Wright, R. 2023. "So what if ChatGPT wrote it?" Multidisciplinary perspectives on opportunities, challenges and implications of generative

conversational AI for research, practice and policy. ***International Journal of Information Management***, 71, 102642.

Estep, C., Griffith, E. E., & MacKenzie, N. L. 2023. How do financial executives respond to the use of artificial intelligence in financial reporting and auditing? ***Review of Accounting Studies***, 1–34 https://doi.org/10.1007/s11142-023-09771-y

Eulerich, M., Waddoups, N. J., Wood, D. A., Burton, G., Cooper, L., et al. 2022. A Framework for Using Robotic Process Automation for Audit Tasks*. ***Contemporary Accounting Research***, 39(1): 691–720.

Eulerich, M, Sanatizadeh, A, Vakilzadeh, H, and Wood, D A. 2023. Is it All Hype? ChatGPT's Performance and Disruptive Potential in the Accounting and Auditing Industries. ***SSRN Electronic Journal***. http://dx.doi.org/10.2139/ssrn.4452175

Extance, A. 2023. ChatGPT has entered the classroom: how LLMs could transform education. ***Nature***, 623(7987): 474–477.

Eyring, H., Ferguson, P. J., & Koppers, S. 2021. Less Information, More Comparison, and Better Performance: Evidence from a Field Experiment. ***Journal of Accounting Research***, 59(2): 657–711.

Eyring, H., & Narayanan, V. G. 2018. Performance Effects of Setting a High Reference Point for Peer-Performance Comparison. ***Journal of Accounting Research***, 56(2): 581–615.

Fan, J., Sun, T., Liu, J., Zhao, T., Zhang, B., Chen, Z., Glorioso, M., & Hack, E. 2023. How well can an AI chatbot infer personality? Examining psychometric properties of machine-inferred personality scores. ***Journal of Applied Psychology***, 108(8): 1277–1299.

Fedyk, A., Hodson, J., Khimich, N., & Fedyk, T. 2022. Is artificial intelligence improving the audit process? ***Review of Accounting Studies***, 27(3): 938–985.

Festinger, L. 1954. A theory of social comparison processes. ***Human relations***, 7(2): 117-140.

Frederickson, J. R. 1992. Relative performance information: The effects of common uncertainty and contract type on agent effort. ***The Accounting Review***, 647-669.

Gefen, D., & Straub, D. W. 1997. Gender differences in the perception and use of e-mail: An extension to the technology acceptance model. ***MIS Quarterly: Management Information Systems***, 21(4): 389–400.

Gevers, J. M., & Demerouti, E. 2013. How supervisors' reminders relate to subordinates' absorption and creativity. ***Journal of Managerial Psychology***, 28(6), 677-698.

Goulas, S., & Megalokonomou, R. (2021). Knowing who you actually are: The effect of feedback on short-and longer-term outcomes. ***Journal of Economic Behavior & Organization***, 183: 589-615.

Hannan, R. L., Krishnan, R., & Newman, A. H. 2008. The Effects of Disseminating Relative Performance Feedback in Tournament and Individual Performance Compensation Plans. ***The Accounting Review***, 83(4): 893–913.

Hannan, R. L., McPhee, G. P., Newman, A. H., & Tafkov, I. D. 2013. The Effect of Relative Performance Information on Performance and Effort Allocation in a Multi-Task Environment. ***The Accounting Review***, 88(2): 553–575.

Hannan, R. L., McPhee, G. P., Newman, A. H., Tafkov, I. D., & Kachelmeier, S. J. 2019. The Informativeness of Relative Performance Information and Its Effect on Effort Allocation in a Multitask Environment. ***Contemporary Accounting Research***, 36(3): 1607–1633.

Harrison, S. H., & Dossinger, K. 2018. Pliable Guidance: A Multilevel Model of Curiosity, Feedback Seeking, and Feedback Giving in Creative Work. ***Academy of Management Journal***, 60(6): 2051–2072.

Heckman, J. J. 1979. Sample selection bias as a specification error. ***Econometrica***, 47(1): 153–161.

Hoffman, P. J., Earle, T. C., & Slovic, P. 1981. Multidimensional functional learning (MFL) and some new conceptions of feedback. ***Organizational behavior and Human performance***, 27(1), 75-102.

Kanheman D. Thinking fast and slow. Farrar, Straus and Giroux; New York: 2013. Reprint edition.

Karaevli, A. 2007. Performance consequences of new CEO 'Outsiderness': Moderating effects of pre- and post-succession contexts. ***Strategic Management Journal***, 28(7): 681–706.

Kleinberg, J., Lakkaraju, H., Leskovec, J., Ludwig, J., & Mullainathan, S. 2018. Human Decisions and Machine Predictions. ***The Quarterly Journal of Economics***, 133(1): 237–293.

Kluger, A. N., & DeNisi, A. 1996. The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. ***Psychological Bulletin***, 119(2): 254.

Kohler, M., Mahlendorf, M. D., Seiter, M., Vogelsang, T. 2024. Social Comparison on Multiple Tasks: Sacrificing Overall Performance for Local Excellence? ***Journal of Accounting Research***, In-press.

Krakowski, S., Luger, J., & Raisch, S. 2023. Artificial intelligence and the changing sources of competitive advantage. ***Strategic Management Journal***, 44(6): 1425–1452.

Lam, C. F., DeRue, D. S., Karam, E. P., & Hollenbeck, J. R. 2011. The impact of feedback frequency on learning and task performance: Challenging the "more is better" assumption. ***Organizational Behavior and Human Decision Processes***, 116(2), 217-228.

Lebovitz, S., Lifshitz-Assaf, H., & Levina, N. 2022. To Engage or Not to Engage with AI for Critical Judgments: How Professionals Deal with Opacity When Using AI for Medical Diagnosis. ***Organization Science***, 33(1): 126–148.

Logg, J. M., Minson, J. A., & Moore, D. A. 2019. Algorithm appreciation: People prefer algorithmic to human judgment. ***Organizational Behavior and Human Decision Processes***, 151: 90–103.

Lu, J. 2022. Limited attention: Implications for financial reporting. ***Journal of Accounting Research***, 60(5), 1991-2027.

Lurie, N. H., & Swaminathan, J. M. 2009. Is timely information always better? The effect of feedback frequency on decision making. ***Organizational Behavior and Human decisión processes***, 108(2), 315-329.

Madsen, P. M., & Rodgers, Z. J. 2015. Looking good by doing good: The antecedents and consequences of stakeholder attention to corporate disaster relief. ***Strategic Management Journal***, 36(5): 776–794.

Maslej, N., Fattorini, L., Perrault, R., Parli, V., Reuel, A., Brynjolfsson, E., Etchemendy, J., Ligett, K., Lyons, T., Manyika, J., Niebles, J. C., Shoham, Y., Wald, R., & Clark, J. ***The AI Index 2024 Annual Report.*** AI Index Steering Committee, Institute for Human-Centered AI, Stanford University, Stanford, CA, April 2024.

Murray, A., Rhymer, J., & Sirmon, D. G. 2021. Humans and Technology: Forms of Conjoined Agency in Organizations. ***Academy of Management Review***, 46(3): 552–571.

Peng, Y ;, Ahmad, S. F. ;, Ahmad, A. Y. A. B. ;, Al, S. ;, Daoud, M. K. ;, et al. 2023. Riding the Waves of Artificial Intelligence in Advancing Accounting and Its Implications for Sustainable Development Goals. ***Sustainability,*** (19): 14165.

Raisch, S., & Fomina, K. 2024. Combining Human and Artificial Intelligence: Hybrid Problem-Solving in Organizations. ***Academy of Management Review***, In-Press. https://doi.org/10.5465/AMR.2021.0421.

Raisch, S., & Krakowski, S. 2021. Artificial Intelligence and Management: The Automation–Augmentation Paradox. ***Academy of Management Review***, 46(1): 192–210.

Reuben, E., Sapienza, P., & Zingales, L. 2024. Overconfidence and preferences for competition. ***The Journal of Finance***, 79(2): 1087-1121.

Reyna, V. F. 2004. How people make decisions that involve risk: A dual-processes approach. ***Current directions in psychological science***, 13(2), 60-66.

Semadeni, M., Cannella, A. A., Fraser, D. R., & Lee, D. S. 2008. Fight or flight: managing stigma in executive careers. ***Strategic Management Journal***, 29(5): 557–567.

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., et al. 2023. Practices for Governing Agentic AI Systems. OpenAI Technical Report. ***Https://Cdn.Openai.Com/Papers/Practices-for-Governing-Agentic-Ai-Systems.Pdf***.

Shrestha, Y. R., Ben-Menahem, S. M., & von Krogh, G. 2019. Organizational Decision-Making Structures in the Age of Artificial Intelligence. ***California Management Review***, 61(4): 66–83.

Sundaram, S., Schwarz, A., Jones, E., & Chin, W. W. 2007. Technology use on the front line: how information technology enhances individual performance. ***Journal of the Academy of Marketing Science***, 35, 101-112.

Tabachnick, B. G., Fidell, L. S., & Ullman, J. B. 2013. ***Using multivariate statistics*** , 6h edition. Boston, MA: Pearson.

Todd, P. M., & Gigerenzer, G. (2007). Environments that make us smart: Ecological rationality. ***Current directions in psychological science***, 16(3), 167-171.

Wilson, H. J., & Daugherty, P. R. 2018. Collaborative Intelligence: Humans and AI Are Joining Forces. ***Harvard Business Review***, 1–11.

Wood, D. A., Achhpilia, M. P., Adams, M. T., Aghazadeh, S., Akinyele, K., et al. 2023. The ChatGPT Artificial Intelligence Chatbot: How Well Does It Answer Accounting Assessment Questions? ***Issues in Accounting Education***, 38(4): 81–108.

Zhao, J., & Wang, X. 2024. Unleashing efficiency and insights: Exploring the potential applications and challenges of ChatGPT in accounting. ***Journal of Corporate Accounting & Finance***, 35(1): 269–276.

Table 1. Descriptive Statistics

	Stats	N	Mean	SD	Min	p25	p50	p75	Max	1	2	3	4
1	AI user	25340	0.22	0.42	0.00	0.00	0.00	0.00	1.00	1			
2	Avg. score per question	25340	0.38	0.24	0.00	0.18	0.36	0.57	1.00	-0.00	1		
3	Conceptual question	25340	0.28	0.45	0.00	0.00	0.00	1.00	1.00	-0.01	0.29*	1	
4	Subject knowledge	25314	0.44	0.16	0.00	0.33	0.44	0.56	0.81	-0.13*	-0.01	-0.03*	1
5	Female	24811	0.37	0.48	0.00	0.00	0.00	1.00	1.00	-0.03*	-0.00	-0.00	-0.02*

Table 1 shows descriptive statistics for our baseline sample over the entire experimental period (stage 1 to 3, from week 6 to 12). All variables are defined in Appendix A.

Table 2. Distribution of Questions and AI Usage

Stage	Description	Week	# Observations	# Total Questions	# Numerical questions	# Sessions	# Quizzes	# AI Users (in % of Subjects)	# Subjects
0	Pre-experiment	3	4,442	14	8	2	3	-	357
0	Pre-experiment	4	4,620	14	8	2	3	-	351
0	Pre-experiment	5	2,540	8	4	2	2	-	345
1	Pre-RPF	6	6,174	22	16	2	5	20.6%	321
2	Post-RPF_1	7	3,794	14	8	2	3	19.5%	303
2	Post-RPF_1	8	3,705	13	10	1	1	27.7%	285
2	Post-RPF_1	9	3,487	13	13	2	3	16.2%	296
3	Post-RPF_2	10	4,168	16	10	2	3	15.9%	289
3	Post-RPF_2	11	2,016	8	8	1	6	8.6%	255
3	Post-RPF_2	12	1,996	8	4	1	2	13.5%	251

Table 2 shows the distribution of questions over the stages of the experiment and the evolution of the number of users over the weeks. The quizzes started in the third week of the course and the experiment started in the sixth week.

Table 3. AI Adoption

	(1) AI User	(2) AI User	(3) AI User	(4) AI User	(5) AI User
High Knowledge	-0.027			-0.028	
	(-0.64)			(-0.66)	
Subject knowledge		-0.023			-0.028
		(-0.21)			(-0.25)
Female			-0.033	-0.034	-0.034
			(-0.77)	(-0.79)	(-0.79)
Constant	0.202***	0.200***	0.204***	0.215***	0.215***
	(7.56)	(4.13)	(7.77)	(6.84)	(4.14)
Number of observations	361	361	361	361	361

Table 3 shows regressions results for the effect of different subject characteristics on AI adoption. *AI User* on gender and knowledge. We employ two different proxies for knowledge. In Columns (1) and (4), we use a binary variable *High Knowledge* which equals one if the subject was performing above peers in the pre-experimental quizzes, and zero otherwise. In Columns (2) and (5), we use a continuous measure of knowledge based on subjects' pre-experimental scores in quiz questions: *Subject knowledge.* All other variables are defined as in Appendix A. The table reports OLS coefficient estimates and (in parentheses) t-statistics based on robust standard errors. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels (two-tailed), respectively.

Table 4. Hypothesis 1: Autonomous versus Interactive Collaboration Strategies (Pre-RPF period)

		(1) Score	(2) Score	(3) Score	(4) Score	(5) Score
Interactive Collaboration	H1 (-)	-0.083***	-0.054**	-0.067***	-0.061**	-0.073***
		(-3.16)	(-2.31)	(-2.67)	(-2.34)	(-2.70)
Avg score per question			0.903***	0.907***		
			(15.82)	(15.83)		
Subject knowledge			0.385***		0.330**	
			(2.71)		(2.29)	
Female			-0.005		-0.004	
			(-0.09)		(-0.07)	
Inverse Mills Ratio			-0.202		-0.121	
			(-0.42)		(-0.25)	
Constant		0.422***	0.164	-0.061	0.496	0.383***
		(15.23)	(0.26)	(-0.58)	(0.79)	(3.45)
Subject Fixed Effects		NO	NO	YES	NO	YES
Question Fixed Effects		NO	NO	NO	YES	YES
Number of observations		1363	1363	1363	1363	1363

Table 4 shows regressions results for the effect of the collaboration strategy on question scores. Score is a binary variable that is equal to 1 if the answer to a question is correct and 0 otherwise. Interactive Collaboration is a binary variable that is equal to 1 if the question is displayed in image format, and 0 if it is in text format. All other variables are defined as in Appendix A. The sample includes only AI users and covers the first stage of the experiment (pre-feedback). The table reports mixed-effects coefficient estimates and (in parentheses) t-statistics based on robust standard errors two-way clustered by subject and question. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels (two-tailed), respectively.

Table 5. Hypothesis 2: RPF Effect on Autonomous and Interactive Collaboration Strategies

		(1) Score	(2) Score	(3) Score	(4) Score	(5) Score
Interactive Collaboration	H1 (-)	-0.021	-0.053**	-0.058**	-0.060**	-0.061**
		(-1.38)	(-2.38)	(-2.56)	(-2.34)	(-2.39)
Interactive Collaboration x RPF 1	H2 (+)		0.060**	0.058**	0.078**	0.076**
			(2.14)	(2.12)	(2.44)	(2.44)
Avg score per question			0.868***	0.868***		
			(20.82)	(21.51)		
Subject knowledge			0.414***		0.391***	
			(4.89)		(4.58)	
Female			-0.024		-0.026	
			(-0.72)		(-0.80)	
Inverse Mills Ratio			-0.204		-0.172	
			(-0.72)		(-0.60)	
Constant		0.344***	0.105	-0.095	0.680*	0.511***
		(17.83)	(0.28)	(-1.44)	(1.83)	(6.76)
Quiz Fixed Effects		NO	YES	YES	YES	YES
Subject Fixed Effects		NO	NO	YES	NO	YES
Question Fixed Effects		NO	NO	NO	YES	YES
Number of observations		3762	3762	3762	3762	3762

Table 5 shows regression results for the differential effect of the collaboration strategy on question scores after the first performance feedback. Score is a binary variable that is equal to 1 if the answer to a question is correct and 0 otherwise. Interactive Collaboration is a binary variable that is equal to 1 if the question is displayed in image format, and 0 if it is in text format. RPF 1 a binary variable that is equal to 1 for the period between the first and second feedback (week 7-9) and zero otherwise. All other variables are defined as in Appendix A. The sample includes only AI users and covers the first two stages of the experiment: periods pre-feedback and after the first feedback. The table reports mixed-effects coefficient estimates and (in parentheses) t-statistics based on robust standard errors two-way clustered by subject and question. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels (two-tailed), respectively.

Table 6. Effect of the Second RPF on Autonomous and Interactive Collaboration Strategies

		(1) Score	(2) Score	(3) Score	(4) Score	(5) Score
Interactive Collaboration	H1 (-)	-0.014	-0.052**	-0.055**	-0.060**	-0.060**
		(-1.13)	(-2.31)	(-2.40)	(-2.30)	(-2.29)
Interactive Collaboration x RPF 1	H2 (+)		0.059**	0.054**	0.078**	0.074**
			(2.10)	(1.96)	(2.42)	(2.33)
Interactive Collaboration x RPF 2	H2 (+)		0.074**	0.059*	0.063*	0.054
			(2.49)	(1.93)	(1.82)	(1.53)
Avg score per question			0.895***	0.896***		
			(26.93)	(27.71)		
Subject knowledge			0.469***		0.456***	
			(6.81)		(6.61)	
Female			-0.002		-0.003	
			(-0.08)		(-0.10)	
Inverse Mills Ratio			-0.542**		-0.536**	
			(-2.33)		(-2.31)	
Constant		0.352***	0.539*	-0.052	1.165***	0.572***
		(20.20)	(1.77)	(-0.92)	(3.80)	(8.08)
Quiz Fixed Effects		NO	YES	YES	YES	YES
Subject Fixed Effects		NO	NO	YES	NO	YES
Question Fixed Effects		NO	NO	NO	YES	YES
Number of observations		5,638	5,638	5638	5638	5638

Table 6 shows regression results for the differential effect of the collaboration strategy on question scores after the first and the second performance feedback. *Score* is a binary variable that is equal to 1 if the answer to a question is correct and 0 otherwise. *Interactive Collaboration is* a binary variable that is equal to 1 if the question is displayed in image format, and 0 if it is in text format. *RPF 1* a binary variable that is equal to 1 for the period between the first and second feedback (week 7-9) and zero otherwise. *RPF 2* is a binary variable that takes the value of 1 for periods after the second feedback (week 10-12) and 0 before it. All other variables are defined as in Appendix A. The sample includes only AI users and covers all three stages of the experiment: periods pre-feedback and after the first and the second feedback. The table reports mixed-effects coefficient estimates and (in parentheses) t-statistics based on robust standard errors two-way clustered by subject and question. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels (two-tailed), respectively.

Table 7. Placebo Test Using the AI Non-Users Subsample

Panel A. Pre-RPF period

	(1) Score	(2) Score	(3) Score	(4) Score	(5) Score
Interactive Collaboration	-0.009	-0.001	-0.005	-0.006	-0.006
	(-0.62)	(-0.11)	(-0.35)	(-0.44)	(-0.44)
Avg. score per question		1.032***	1.021***		
		(35.43)	(36.65)		
Subject knowledge		0.565***		0.582***	
		(8.51)		(8.63)	
Female		0.020		0.024	
		(0.67)		(0.80)	
Inverse Mills Ratio		-0.500**		-0.524**	
		(-1.98)		(-2.04)	
Constant	0.416***	0.157	-0.065	0.942***	0.724***
	(26.52)	(0.83)	(-1.15)	(4.89)	(11.99)
Subject Fixed Effects	NO	NO	YES	NO	YES
Question Fixed Effects	NO	NO	NO	YES	YES
Number of observations	4,811	4,811	4,811	4,811	4,811

Panel B. Pre- and post-RPF

	(1) Score	(2) Score	(3) Score	(4) Score	(5) Score
Interactive Collaboration	0.007	-0.002	-0.006	-0.007	-0.006
	(0.86)	(-0.14)	(-0.50)	(-0.48)	(-0.45)
Interactive Collaboration x RPF 1		0.011	0.010	0.017	0.009
		(0.75)	(0.68)	(0.98)	(0.51)
Avg score per question		1.035***	1.027***		
		(46.10)	(48.27)		
Subject knowledge		0.532***		0.540***	
		(14.24)		(14.36)	
Female		0.005		0.006	
		(0.31)		(0.33)	
Inverse Mills Ratio		-0.264*		-0.263*	
		(-1.83)		(-1.81)	
Constant	0.416***	0.157	-0.065	0.942***	0.724***
	(26.52)	(0.83)	(-1.15)	(4.89)	(11.99)
Quiz Fixed Effects	NO	YES	YES	YES	YES
Subject Fixed Effects	NO	NO	YES	NO	YES
Question Fixed Effects	NO	NO	NO	YES	YES
Number of observations	13,398	13,372	13,398	13,372	13,398

Table 7 replicates Table 3 and 4 including only non-AI-users in the sample. Panel A replicates Table 3. The sample includes only the first stage of the experiment: pre-feedback. Panel B replicates Table 4. The sample includes the first two stages of the experiment: period pre-feedback and after the first feedback. All variables are defined as in Appendix A. The table reports mixed-effects coefficient estimates and (in parentheses) t-statistics based on robust standard errors two-way clustered by subject and question. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels (two-tailed), respectively.

Table 8. Time Management in Human-AI Collaborations

	(1) Duration (in sec)	(2) Duration (in sec)	(3) Duration (in sec)
	Stage 1	*Stages 1 & 2*	*Stages 1 to 3*
80% Interactive group	-10.055	-7.311	-6.996
	(-0.56)	(-0.42)	(-0.40)
80% Interactive group x RPF 1		44.029*	44.205*
		(1.91)	(1.93)
80% Interactive group x RPF 2			46.606*
			(1.85)
Subject knowledge	-89.830	62.310	124.360
	(-1.04)	(0.54)	(1.05)
Female	25.595	12.066	17.426
	(0.60)	(0.36)	(0.56)
Inverse Mills Ratio	66.813	-62.131	-169.967
	(0.19)	(-0.22)	(-0.59)
Constant	429.673	669.674*	744.304*
	(0.93)	(1.84)	(2.10)
Number of observations	309	719	1127

Table 10 shows regression results of the effect of different AI collaboration strategies on the time used in the quizzes (in seconds). The sample includes only quizzes where subjects in the 80% Interactive group received all questions in an image format. Therefore *80% Interactive group* indicates that the subject received all questions in image format (interactive human-AI collaboration). All other variables are defined as in Appendix A. The table reports OLS coefficient estimates and (in parentheses) t-statistics. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels (two-tailed), respectively.

Appendix A. Variable Definitions

AI User is a binary variable that takes the value of one if the subject opted to use the AI in at least three out of 11 sessions (see Table 1c). AI usage is defined as whether the subject accessed ChatGPT at the beginning of the quiz using the direct link provided in our secure test environment.

Score in the question, a binary variable that is equal to 1 if the answer is correct and 0 otherwise.

Interactive collaboration is a binary variable that is equal to 1 if the question is displayed in image format, and 0 if it is in text format.

RPF 1 is a binary variable that is equal to 1 for the period between the first and second feedback (week 7-9) and 0 otherwise.

RPF 2 is a binary variable that takes the value of 1 for periods after the second feedback (week 10-12) and 0 before it.

Avg. score per question is the average score of all subjects on a given question.

Conceptual question is a binary variable that is equal to 1 if the question is conceptual and 0 if it is numerical.

Subject knowledge is measured using the average score of the subject on the questions before the start of the experiment (quizzes in stage 0 in weeks 3 to 5).

Female is a binary variable that is equal to 1 if the subject is a female and 0 if the subject is male.

Below Peers in RPF 1 is a binary variable that is equal to 1 if the subject score is below peers' median score before RPF 1 and 0 otherwise.

Below Peers in RPF 2 is a binary variable that is equal to 1 if the subject score is below peers' median score before RPF 2 and 0 otherwise.

80% Interactive group is a binary variable that is equal to 1 if the subject is in the randomly assigned group that receives 80% image format questions and 0 if the subject is in the group that receives 20% image format questions.

Inverse Mills Ratio is generated based on the Heckman selection model (Heckman, 1979). Therefore, we estimate the probability of selection to use the AI from our entire sample of AI users and non-users based on observed subject characteristics (gender and knowledge).

Appendix B. Example of a Conceptual Question in Text and Image Modes.[12]

Text format:

Some statements on competitive efficiency variance

	True	False
A favourable or unfavourable difference in competitive efficiency may be entirely due to factors beyond the control of management.	○	○
Variations in the efficiency of inputs/resources cannot have any influence on the competitive efficiency gap.	○	○
Variations in the cost of inputs/resources cannot influence the competitive efficiency gap.	○	○
The static budget gap is the sum of the competitive efficiency gaps.	○	○

Image format:

Some statements on competitive efficiency variance

a. A favourable or unfavourable difference in competitive efficiency may be entirely due to factors beyond the control of management.

b. Variations in the efficiency of inputs/resources cannot have any influence on the competitive efficiency gap.

c. Variations in the cost of inputs/resources cannot influence the competitive efficiency gap.

d. The static budget gap is the sum of the competitive efficiency gaps.

	True	False
a.	○	○
b.	○	○
c.	○	○
d.	○	○

[12] Translated text from French to English using DeepL.

Appendix C. Example of a Numerical Question in Text and Image Formats.[13]

Text format:

Illustration 8.7: Information on materials for the fifth season

	Planned cost per unit	Fashion jacket	X-jacket	Total
Materials per unit	€7.00	3	6	
Total materials (units)		90,000	60,000	150,000
Total material cost (€)		€630,000	€420,000	€1,050,000

Illustration 8.8: Direct manufacturing labor hours (DLH) for the fifth season

	Fashion fleece	X-jacket	Total
Units	30,000	10,000	
Direct labor hours	1	1.5	
Total hours	30,000	15,000	45,000
Cost per hour	20	20	
Total cost of direct labor	€600,000	€300,000	€900,000

Based on the above standards, assuming MFD plans to produce 33,600 units of fashion jackets and 11,200 units of technical jackets (X), what would be the expected total cost of materials for the fashion jacket?

(Give your answer without a thousand separator)

Answer:

[13] Translated question from French to English using DeepL.

Image format:

Illustration 8.7: Information on materials for the fifth season

	Planned cost per unit	Fashion jacket	X-jacket	Total
Materials per unit	€7.00	3	6	
Total materials (units)		90,000	60,000	150,000
Total material cost (€)		€630,000	€420,000	€1,050,000

Illustration 8.8: Direct manufacturing labor hours (DLH) for the fifth season

	Fashion fleece	X-jacket	Total
Units	30,000	10,000	
Direct labor hours	1	1.5	
Total hours	30,000	15,000	45,000
Cost per hour	20	20	
Total cost of direct labor	€600,000	€300,000	€900,000

Based on the above standards, assuming MFD plans to produce 33,600 units of fashion jackets and 11,200 units of technical jackets (X), what would be the expected total cost of materials for the fashion jacket?

(Give your answer without a thousand separator)

Answer:

Appendix D. Feedback Given to Students on Their Quiz Performance.[14]

First feedback after one-third of the quizzes:

Dear [student name],
I hope the semester is going well for you.
Here's a little feedback regarding your progress on the Quizzes.
Your current Quiz score is [student points] points out of a total of 58. The median is 21/58.
We've only completed a third of the quizzes so far. Any student can still get the 0.5-point bonus!
For your information, we submitted the Quizzes to ChatGPT, which scored 27/58. As a reminder, you can use ChatGPT's assistance by clicking on its link once the Safe Exam Browser (SEB) is launched.
We wish you all the best.
Best regards,

Second feedback after two-thirds of the quizzes:

Dear [student name],
I hope the semester is going well for you.
Here's a little feedback regarding your progress on the Quizzes.
Your current Quiz score is [student points] points out of a total of 98. The median is 32/98.
For the numerical questions, your score is [student points] /70 and the median is 19/70.
For conceptual questions, your score is [student points] /28 and the median is 13/28.
We've only completed two-thirds of the quizzes so far.
For your information, we submitted the Quizzes to ChatGPT, which scored 28/70 on the numerical questions and 18/28 on the conceptual questions.
We wish you all the best.
Best regards,

[14] Translated from French to English using DeepL.

Appendix E. RPF and Performance in the Final Exam

	(1) Exam Score	(2) Exam Score	(3) Abnormal Score	(4) Abnormal Score
Below Peers in RPF 1	0.031	0.060	0.031	0.060
	(0.78)	(1.27)	(0.78)	(1.27)
Below Peers in RPF 2	-0.081**	-0.119***	-0.081**	-0.119***
	(-2.20)	(-2.74)	(-2.20)	(-2.74)
AI User	0.000	-0.052	0.000	-0.052
	(0.01)	(-1.09)	(0.01)	(-1.09)
AI User x Below Peers in RPF 1		-0.079		-0.079
		(-0.99)		(-0.99)
AI User x Below Peers in RPF 2		0.121		0.121
		(1.51)		(1.51)
80% Interactive group	-0.011	-0.013	-0.011	-0.013
	(-0.48)	(-0.56)	(-0.48)	(-0.56)
AI User x 80% Interactive group	0.049	0.053	0.049	0.053
	(0.93)	(1.00)	(0.93)	(1.00)
Subject knowledge	0.281***	0.273***	-0.719***	-0.727***
	(3.36)	(3.26)	(-8.59)	(-8.67)
Female	-0.088***	-0.104***	-0.088***	-0.104***
	(-4.08)	(-4.39)	(-4.08)	(-4.39)
AI User x Female		0.081		0.081
		(1.46)		(1.46)
Constant	0.507***	0.522***	0.507***	0.522***
	(10.50)	(10.60)	(10.50)	(10.60)
R^2	0.165	0.178	0.298	0.308
Adjusted R^2	0.148	0.154	0.284	0.288
Number of observations	354	354	354	354

Table 9 shows regression results of for the effect of AI usage on the final exam score. Columns (1) and (2) employ the final exam score in absolute terms as dependent variable. Columns (3) and (4) employ Abnormal Exam Score which is the exam score minus the subject's average performance in the quizzes. All other variables are defined as in Appendix A. The table reports OLS coefficient estimates and (in parentheses) t-statistics. ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels (two-tailed), respectively.

Appendix F. OLS, Probit and Logit models to Test for Hypothesis 1

	(1) Score	(2) Score	(3) Score	(4) Score
Statistical Model	*Multilevel*	*OLS*	*Probit*	*Logit*
Interactive Collaboration	-0.073***	-0.073**	-0.279***	-0.466***
	(-2.70)	(-2.33)	(-2.78)	(-2.71)
Constant	0.383***	0.383***	-0.388**	-0.599**
	(3.45)	(10.62)	(-2.21)	(-2.12)
Subject Fixed Effects	YES	YES	YES	YES
Question Fixed Effects	YES	YES	YES	YES
Number of observations	1363	1363	1206	1206

Table 11 replicates Table 3 using different estimators. Column (1) presents the results of the multilevel mixed-effects linear regression with random intercepts. In Column (2) we use a standard OLS estimator. Column (3) shows the results using a probit estimator, and Column (4) presents results for a logit estimator. All other variables are defined as in Appendix A. The table reports coefficient estimates and (in parentheses) t-statistics based on robust standard errors two-way clustered by subject and question in Columns (1) and (2) and clustered by subject in Columns (3) and (4). ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels (two-tailed), respectively.

Appendix G. OLS, Probit and Logit models to Test for Hypothesis 2

	(1) Score	(2) Score	(3) Score	(4) Score
Statistical Model	*Multilevel*	*OLS*	*Probit*	*Logit*
Interactive Collaboration	-0.061**	-0.061**	-0.234**	-0.402**
	(-2.39)	(-2.00)	(-2.44)	(-2.44)
Interactive Collaboration x RPF	0.076**	0.076*	0.299**	0.528**
	(2.44)	(1.74)	(2.19)	(2.27)
Constant	0.511***	0.505***	-0.085	-0.175
	(6.76)	(16.98)	(-0.39)	(-0.47)
Quiz Fixed Effects	YES	YES	YES	YES
Subject Fixed Effects	YES	YES	YES	YES
Question Fixed Effects	YES	YES	YES	YES
Number of observations	1363	1363	1206	1206

Table 12 replicates Table 4 using different estimators. Column (1) presents the results of the multilevel mixed-effects linear regression with random intercepts. In Column (2) we use a standard OLS estimator. Column (3) shows the results using a probit estimator, and Column (4) presents results for a logit estimator. All other variables are defined as in Appendix A. The table reports coefficient estimates and (in parentheses) t-statistics based on robust standard errors two-way clustered by subject and question in Columns (1) and (2) and clustered by subject in Columns (3) and (4). ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels (two-tailed), respectively.

Appendix H. Including Interactions between RPF and Control Variables

	(1) Score	(2) Score	(3) Score	(4) Score
Statistical Model	*Multilevel*	*OLS*	*Probit*	*Logit*
Interactive Collaboration	-0.067**	-0.065**	-0.179*	-0.307*
	(-2.37)	(-2.35)	(-1.84)	(-1.85)
RPF	0.016	-0.071	-0.238	-0.250
	(0.02)	(-0.08)	(-0.08)	(-0.05)
Interactive Collaboration x RPF 1	0.061**	0.058*	0.179*	0.303*
	(2.13)	(1.87)	(1.84)	(1.85)
80% Interactive group	-0.016	-0.015	-0.026	-0.046
	(-0.67)	(-0.52)	(-0.27)	(-0.27)
Interactive Collaboration x 80% Interactive group	0.026	0.025	0.060	0.100
	(0.73)	(0.78)	(0.51)	(0.49)
Avg score per question	0.893***	0.896***	2.885***	4.824***
	(16.41)	(16.80)	(14.31)	(13.87)
Avg score per question x RPF 1	0.011	0.004	0.112	0.224
	(0.16)	(0.06)	(0.39)	(0.44)
Female	-0.009	-0.006	-0.020	-0.028
	(-0.17)	(-0.07)	(-0.07)	(-0.06)
Female x RPF 1	-0.021	-0.024	-0.085	-0.143
	(-0.32)	(-0.32)	(-0.30)	(-0.29)
Inverse Mills Ratio	-0.167	-0.202	-0.614	-1.133
	(-0.36)	(-0.33)	(-0.29)	(-0.32)
Inverse Mills Ratio x RPF 1	-0.065	-0.001	-0.095	-0.311
	(-0.11)	(-0.00)	(-0.04)	(-0.07)
Subject knowledge	0.372***	0.385**	1.237**	2.151**
	(2.71)	(2.12)	(2.14)	(2.22)
Subject knowledge x RPF 1	0.073	0.069	0.387	0.744
	(0.42)	(0.34)	(0.57)	(0.64)
Constant	0.134	0.175	-1.094	-1.711
	(0.22)	(0.22)	(-0.41)	(-0.38)
Number of observations	3,762	3,762	3,762	3,762

Table 13 shows regression results for the differential effect of the collaboration strategy on question scores after the first performance feedback. We interact all control variables with RPF to analyze whether associations between performance and our control variables change through the feedback. The sample includes only AI users and covers stage one and two of the experiment: periods pre-feedback and after the first feedback. All other variables are defined as in Appendix A. The table reports coefficient estimates and (in parentheses) t-statistics based on robust standard errors two-way clustered by subject and question in Columns (1) and (2) and clustered by subject in Columns (3) and (4). ***, **, and * indicate statistical significance at the 1%, 5%, and 10% levels (two-tailed), respectively.