AI in the Classroom: Barrier or Gateway to Academic and Labor Market Success?

Pre-Analysis Plan

Catalina Franco* Natalie Irmert† Siri Isaksson‡

December 4, 2024

Abstract

Artificial Intelligence (AI) is becoming an increasingly important skill in the labor market, but will it also affect academic success? Recent research shows that current students –who will be facing this rapidly changing labor market– are adopting AI tools at differential rates based on both gender and ability. Whether AI will affect adopters' academic success hinges on whether AI interferes with or enhances learning, which in turn depends on whether AI is being used as a substitute for or complement of own effort. If AI harms learning, students with high adoption rates would be worse prepared for the labor market than those with low adoption rates. If AI enhances learning, students who do not become proficient at AI would be left behind. To assess the impact of AI on learning, we run a controlled lab experiment which allows us to restrict and allow access to AI in different between-subject treatment variations in which students learn about a new topic. We explore whether AI is employed in a way that causally creates a gap in learning and as a consequence payoffs. Our results provide evidence on the important question of whether the documented differential adoption and use of AI by gender and ability is likely to create gender gaps in academic success. In addition, we explore several mechanisms such as AI being a complement or substitute to own effort, confidence, cheating perceptions, and engagement and motivation.

*Center for Applied Research (SNF) at NHH – Norwegian School of Economics.
†Lund University
‡Norwegian School of Economics.

1 Introduction

Recent research shows that AI is poised to reshape different parts of the economy. There is emerging evidence that various factors determine the early adoption of AI, with gender and ability identified as key determinants of its use. For instance, Carvajal, Franco and Isaksson (2024) find that top female students opt out of AI use. At the same time, educational institutions are grappling with how, and whether, to incorporate AI into the formal curriculum. This raises key questions: is AI beneficial or harmful for learning? Can educational institutions ensure that everyone reaps the benefits of AI in their learning by guiding students AI use?

In order to answer these questions, we design a laboratory experiment with three between subject treatment variations. Students in the lab must go over a lecture and work on practice questions to learn about a topic that may be new to them. In the baseline treatment, students learn without access to AI but with access to Google Search, the status quo before the emergence of generative AI tools. In the AI-access treatment, students learn with access to AI, but with no formal guidance on how to use this tool except some general instructions. In the AI-guided treatment, students learn with access to AI, and are given guidance on how to use it best to promote their own learning. This design will allow us to answer the crucial question of whether, and for whom, AI is helpful as a learning tool and whether providing guidance of how to use the tool matters.

2 Design

We will collect data from 600 current students at the University of Nottigham at the CedEX Lab. Students typically come from undergraduate programs across the university and there may be some graduate students as well.

The three treatments are:

- T1: No access to ChatGPT (Control)
- T2: Access to ChatGPT (AI assisted)
- T3: Guided access to ChatGPT (AI guided)

Regardless of treatment, subjects will learn about the same topic: Esperanto. The treatment differences lies in whether or not they have access to ChatGPT/ChatGPT with guidance while learning Esperanto. The topic was selected since it is a language that very few people know or have studied and the point is to see how well students can learn about a new topic (rather than measuring their pre-existing skills).

Each session will adhere to the same procedure:

1. **Stage 1**: Students learn Esperanto for 15 minutes using written learning materials provided by the researchers.

2. **Stage 2**: Students complete practice questions with access to different learning aids depending on treatment for about 20 minutes. This is the stage where the treatment variations take place.

3. **Stage 3**: Students take a test without access to any notes or learning aids.

4. **Stage 4**: Subjects take a post-study survey asking them about demographics as well as their learning experience during the session and attitudes and beliefs about AI.[1]

2.1 *T1: No access to ChatGPT*

T1 was designed to mirror learning before ChatGPT and to work as a baseline against which we can measure the impact of giving access to (guided) AI to see if it helps or hurts learning outcomes. Since before the introduction of AI students would have access to a web browser and online learning, we will provide access to this during stage 2 in T1. We will however block websites providing access to AI sites in order to block the access to AI.

2.2 *T2: Access to ChatGPT*

T2 was designed to mirror learning when ChatGPT is available to students without clear guidelines, structure and training—a situation in which it is up to each student whether and how she or he wants to use AI in the learning process. This mirrors the situation which many universities face as they are grappling with the important task of adjusting their policies to the advent of AI.

Students will access ChatGPT on their browser with a logged in paid account. The account will be provided by the researchers an thus be cleared of any memory. The prompting history will be archived and saved by researchers to be analyzed.

2.3 *T3: Access to Guided ChatGPT*

This treatment is the exact same as T2, except that students now also receive guidance (written by the researchers) on how to use ChatGPT. The idea behind this treatment is to explore the potential of providing not just equal access but also equal training and encouragement in AI use for students.

[1]The basic demographic questions are asked at the beginning of the study, before stage 1.

2.4 Incentives

The structure of incentives is as follows:

- GBP 5 for completing the study
- GBP 7 for correctly answering at least 20 out of over 50 practice questions
- GBP 1 per correct answer in the test (maximum amount GBP 15)

The maximum payment they can earn is GBP 27.

From the beginning of the session students are informed that the most important part of the session (and where they can earn the most) is the test. This is emphasized throughout the session. The intention is that they focus their efforts of trying to learn as much as possible form all the resources they have available.

3 Outcomes of interest

The main outcome of interest is the score on the test in stage 3. The main research question is whether this score differs across treatments. Our secondary research question is whether the answer to this question depends on background variables such as gender and GPA.

We also plan to analyze whether the treatments affect the likelihood of a student being a top or low scorer in the test. A top (low) scorer is someone who obtains more than 10 correct (less than 5 correct) questions in the test.

The preparation for the test during stage 2, when the treatments are implemented, will be assessed by examining the outcomes: number of questions solved, number of questions solved correctly, and time spent working on the test.

4 Econometric Specifications

The main specification will regress the outcomes presented above on indicators for the two AI treatments, keeping the control as the excluded indicator:

$$y_i = \alpha_0 + \alpha_1 \text{AI-assisted}_i + \alpha_2 \text{AI-guided}_i + \varepsilon_i \tag{1}$$

The main specification will not add any baseline covariates, but if we find imbalances in these across treatments, we will present the main results controlling for covariates.

We will asses whether there are any gender differences in responses to the treatment using the specification:

$$y_i = \beta_0 + \beta_1 \text{AI-assisted}_i + \beta_2 \text{AI-guided}_i + \beta_3 \text{Female}_i + \beta_4 \text{AI-assisted}_i \times \text{Female}_i + \beta_5 \text{AI-guided}_i \times \text{Female}_i + \varepsilon_i \quad (2)$$

Where the coefficients β_1 and β_2 measure the effects of the treatments among men, β_3 measures any gender gaps in the control group, and β_4 and β_5 measure the differential effects of the treatments for women.

A similar specification will be presented replacing the Female variable with High GPA, a measure of academic skill equal to 1 if students re in first-class honours (GPA 70% or above) and 0 otherwise.

We may present results split by gender/GPA instead of pooled if the coefficients are easier for interpretation. We may also present results in table or graph form depending on which format delivers the result in the most clear way.

A final specification involves the interaction effects between treatments, gender and GPA. Following the previous literature, this specification may allow us to see differential effects by gender and GPA simultaneously. For example, if the treatments affect top female students differentially.

$$\begin{aligned} y_i = \gamma_0 + \gamma_1 \text{AI-assisted}_i + \gamma_2 \text{AI-guided}_i + \gamma_3 \text{Female}_i + \gamma_4 \text{High GPA}_i + \gamma_5 \text{AI-assisted}_i \times \text{Female}_i + \\ \gamma_6 \text{AI-assisted}_i \times \text{High GPA}_i + \gamma_7 \text{AI-guided}_i \times \text{Female}_i + \gamma_8 \text{AI-guided}_i \times \text{High GPA}_i + \\ \gamma_9 \text{AI-assisted}_i \times \text{Female}_i \times \text{High GPA}_i + \gamma_{10} \text{AI-guided}_i \times \text{Female}_i \times \text{High GPA}_i + \varepsilon_i \end{aligned} \quad (3)$$

The triple interactions will measure whether top female students are differentially affected by the treatments. The double interactions with High GPA will measure if top male students are differentially affected by the treatments. We may opt for presenting these results in the appendix since this specification is quite involved for the number of observations we will collect.

5 Hypotheses

We formulate the following hypotheses regarding the main outcome:

1. Guided AI access results in the best learning outcome: exam score in T3>T1
2. AI access without guidance worsens exam score as related to baseline: exam score in T1>T2

The first hypothesis builds on the idea that receiving guidance on how to make the best use of ChatGPT may help students achieve better results. For example, the guidance prompts students to avoid simply looking for answers but to get examples and ask follow-up questions to ChatGPT.

The second hypothesis relates to the idea that, with no guidance, students may rely too much on getting answers without going deep in their engagement with the materials and questions. This would suggest that students using ChatGPT without guidance may perform worse in the test, when they do not have any external aids available.

A secondary set of hypotheses involve subgroup analysis by gender and high GPA.

1. **Hypothesis 1**: In T1:$F_{T1} >= M_{T1}$. In the control group, women may perform better or the same than men due to the learning topic being in the female domain.

2. **Hypothesis 2**: $F_{T2} < M_{T2}$. Given that men have higher adoption rates and proficiency in AI, we expect them to perform better than women in the AI-assisted treatment. Women may be less likely to take up the treatment and we will analyze this in the appendix.

3. **Hypothesis 3**: $F_{T3} == M_{T3}$ by providing guidance we expect that both female and male learning is enhanced so that the gender gap in exam score in T2 is closed by women achieving a higher learning potential with AI.

4. **Hypothesis 4**: The performance of students with high GPA is higher than the performance of those with lower GPAs, but the gap is smaller in the AI guided treatment.

5. **Hypothesis 5**: female students with high GPA are the ones who benefit the most from the AI guided treatment.

6 Potential factors underlying treatments effects

We propose four factors that could be behind any treatment effects or lack of differences in test scores across treatments. In the post-study survey, we ask several questions that measures attitudes towards the tool. By comparing how students respond to these different attitude questions across treatments, we can identify causal mechanisms behind the gap. One example is: if students in the AI guided treatment report higher levels of engagement and self assessed learning, it suggests that potentially higher scores in the test are driven by higher engagement and motivation. Another example is: if women in the AI guided treatment report higher levels of confidence than in the AI assisted treatment, we know that this is due to the AI guide. Below are the four categories of explanations which we think may be driving these gaps along with the survey questions we use to measure these potential mechanisms.

1. Complement vs. Substitute

- I think the tools helped me a lot in understanding the material.
- I used the tools mostly to get the answer to the practice questions.
- I feel like I learned a lot working through the questions.

2. Confidence/Comfortable

- I was not comfortable using the tools.
- I feel I was prepared enough for the test.
- I found the test to be quite difficult for the level of the lecture and practice questions.

3. Cheating Perceptions

- I believe using the tools as aids to directly solve practice questions is equivalent to cheating.
- I believe using the tools as a learning aid while studying Esperanto is equivalent to cheating.

4. Engagement/Motivation

- Using the tools did not make me feel engaged with the material.
- I felt afraid of becoming over-reliant on the tool to solve the questions.
- I put minimal effort into solving practice questions.
- I was highly motivated to solve the questions.
- I was highly motivated to solve the test.

In the main text, we will present summary indices of these four categories as outcomes following the econometric specifications above. We plan to present the results for each question separately in the appendix, following the same specifications. Each of the questions have four categories: Disagree, somewhat disagree, somewhat agree and agree. The choices in the negative statements will be reversed and indicator variables indicating agreement with the statement will be generated. The summary indices will be a simple average of the items in each category. We also provide a bar-graph comparing averages in each category across treatments, so it's 3 x 4 bars with confidence intervals.

7 Robustness checks, attrition, and take-up of ChatGPT

The two main robustness checks involve adding day and time fixed effects and clustering the standard errors at the session level. In case these additions substantially affect the results without any controls, we will present them in the main text instead of the versions with no controls.

In our sample, attrition may occur when students do not follow the instruction that they must not leave the survey page while taking the test. This instruction prevents students from accessing any learning aids during the test, just as exams in university courses usually take place. The notes they take are also removed from them before they begin the test. If students leave the survey page, they are not allowed to finish the test. We will then lose this observation as we do not observe the main outcome. We hope to have a small number of observations in this situation. However, if it is more than 10% of the sample, we will present some bounding exercises and present analyses of the characteristics of the attritors.

Finally, students allowed to use ChatGPT in the AI-assisted and AI-guided treatments may decide to not make use of this tool and use other online resources (although not other AI chatbots because these are blocked). We expect this the case in particular for women as previous work has documented lower adoption rates by women Carvajal, Franco and Isaksson (2024); Humlum and Vestergaard (2024). If we see that take-up is below 80%, we will report IV estimates in addition to the ITT estimates discussed so far.

To establish take-up, we will use the survey question: I used the tools intensively to solve practice questions. Additionally, we can use the observed use of ChatGPT among those assigned to the AI treatments. In the IV analysis, the first stage will be given by a regression on ChatGPT use on treatment assignment. The second stage will regress the outcomes mention above on the ChatGPT variable instrumented with the treatment assignment. The results from this analysis with give the local average treatment effect of ChatGPT use on the outcomes of interest. The ITT, on the other hand, will capture the effect of being expose to ChatGPT, without necessarily using it.

8 Exploratory analyses

We will have a unique source of data from the prompts that students write on ChatGPT and the search history on the web. We plan to analyze data from the prompts using ChatGPT. Specifically we will use the prompts as input, and upload them on ChatGPT, and ask GPT to assign scores on the quality of prompts for each participant. Due to the fact that GPT always answers differently for each prompt, we will create 50 ratings per subject and take the average rating as the score for each subject. This will give us

approximately 400 prompt ratings which we will compare across treatments, gender and GPA. We will rate prompts on clarity, specificity, relevance, level of detail, and conciseness. Our hypothesis is that prompts will be rated higher in AI-guided than AI-assisted. We also intend to use text analysis tools to provide additional assessments of the prompts.

Given the treatment variation between AI assisted and AI guided, we can also assess the differences between how students use ChatGPT in either case. We will do so through text analysis of the prompts and using the following survey questions:

- I was confident in formulating clear and effective prompts
- ChatGPT did not provide useful examples to learn from
- More encouragement to use AI in school would make me more likely to use it
- I am worried that relying on AI tools might make me lazy or less capable
- I feel like ChatGPT hindered my learning of Esperanto
- Would you have preferred to also have access to Google Search and Google Translate during this session?
- Would clearer guidelines on how to use the tools you were allowed to use improve your comfort and learning outcomes?
- If given a choice, which tool would you prefer to use for learning in the future? Options: AI tools like ChatGPT, Google Search, Other.

The last two are asked to all students so we can also use them to compare with the control group.

References

Carvajal, Daniel, Catalina Franco, and Siri Isaksson, "Will Artificial Intelligence Get in the Way of Achieving Gender Equality?," *NHH Dept. of Economics Discussion Paper*, 2024, (03).

Humlum, Anders and Emilie Vestergaard, "The Adoption of ChatGPT," *University of Chicago, Becker Friedman Institute for Economics Working Paper*, 2024, (2024-50).