esade

EsadeEcPol - Center
for Economic Policy

When GenAI increases inequality: evidence from a university debating competition

Working Paper

AUTHOR

Toni Roldán-Monés
Director, EsadeEcPol

September, 2024

When GenAI increases inequality: evidence from a university debating competition *

Antonio Roldán-Monés†

August 26, 2024

Abstract

This paper evaluates the impact of Generative Artificial Intelligence (GenAI) on productivity and work inequality. I run a Randomized Controlled Trial in a university debating competition, in which I randomly assign GenAI support to students to prepare a series of one-on-one debates. This novel setting allows me to measure productivity improvements in a task involving unpredictable verbal interactions and high cognitive and social skills. Contrary to most early findings in the GenAI literature, I find that high ability students benefit significantly more from GenAI than their lower ability counterparts. Analysis of mechanisms suggests that high ability students are more effective at extracting and using the information provided by GenAI. They also experience larger improvements in their perception of time needed to prepare debates when allowed to use GenAI. I suggest a possible explanation to reconcile these results with previous findings: when tasks require higher-order skills and unpredictable interactions, and answers cannot be copy-pasted from the AI, high ability workers are likely to benefit more from GenAI.

Keywords: Generative AI, Artificial Intelligence, Productivity, Technology, Complementarity, ChatGPT, Labor Market Inequality, Randomized Controlled Trial, Higher-Order Skills

JEL Classification: D80, J24, M15, M51, O33

*The experiment reported in this paper is registered at The American Economic Association's registry for randomized controlled trials, AEA RCT Registry ID: AEARCTR-0011113. I am grateful to the attendants at the seminar at the LSE Department of Social Policy and to Berkay Ozcan and Luis Garicano for their invaluable support throughout the process. I would also like to give special thanks to David Autor for his detailed comments on a previous draft. I would like to thank Miguel Almunia, Michael Becher, Antonio Cabrales, Jorge Galindo, Claudia Hupkau, Ignacio Jurado, Mathew Kraft, Kiko Llaneras, Antoni-Ítalo de Moragas, Monica Martinez-Bravo, Angel Martínez, Javier Martínez, Luis Miller, José Montalbán, Adam Oliver, Pedro Rey-Biel and Andrés Velasco for useful comments on previous stages of this work. I am also grateful to Teresa Raigada, Jorge Galindo, Nuria Aparicio, Sergio Salas, Ángel Martínez and Javier Martínez, the Esade Decision Lab and the Esade Communications team for their support with organization of the debates. Special thanks go to Ignacio Rigau, Miguel Almunia and Jose Ignacio Conde-Ruiz for allowing me to implement the study with their students and to Gemma Lligadas and Ignacio Rigau for introducing me to the exciting world of student debating competitions.

†Department of Social Policy, London School of Economics (LSE) and Universitat Ramon Llull, EsadeEcPol Email: a.roldan-mones@lse.ac.uk

1 Introduction

Business leaders, governments and researchers across the globe are expecting Generative Artificial Intelligence (GenAI) systems to have a profound impact on work (Chui et al., 2023). Previous waves of automation had a negative impact on "routine", blue-collar jobs, while boosting productivity of high-skilled workers, leading to rising inequalities (Autor et al., 2003; Acemoglu and Autor, 2011; Acemoglu and Restrepo, 2017; Felten et al., 2019). Yet, it remains an open question who will benefit from GenAI and on what specific tasks (Wilmers, 2024). In some cases, GenAI might be productivity-compressing and, in other cases, it might be inequality-expanding (Autor, 2024). Early findings in the literature on GenAI point to two main effects of GenAI on work inequalities. First, because of the expanded technical capacities of Large Language Models (LLMs), previously shielded high-skill professions are expected to be deeply affected (Eloundou et al., 2023; Felten et al., 2023). Second, by boosting the productivity of low performers more than that of high performers, GenAI has been found to compress the productivity distribution in a variety of tasks, including mid-level professional tasks, knowledge-intensive consultancy tasks, consumer services or coding (Noy and Zhang, 2023; Doshi and Hauser, 2023; Brynjolfsson et al., 2023; Choi and Schwarcz, 2023; Peng et al., 2023; Dell'Acqua et al., 2023).

In this paper, I analyze the effect of GenAI on productivity and test the "productivity compression hypothesis" in a unique setting allowing me to measure improvements in a task involving unpredictable verbal interactions and "higher-order" cognitive skills. Deming (2022) uses the term "higher-order skills" to refer to a series of "soft" and cognitive skills at the top of the cognitive skills pyramid such as social perceptiveness or critical thinking that have been proven to be of crucial importance for high-earning workers of all kinds (Börner et al., 2018; Deming, 2017; Weinberger, 2014).Debating - similarly to negotiating, managing, or leading – requires higher-order skills, including rhetoric ability, social intuition, argumentative agility, and persuasion capacity (Adams, 1810; Corbett and Connors, 1999; Aristotle, 1960). While most previous studies analyze the effects of GenAI on productivity in written tasks, mine is one of the few existing papers analyzing in-person realistic spoken interactions. Against most previous results, I find that individuals with higher ability, including those on a merit scholarship and those showing high initial debating performance, benefit significantly more from GenAI than low ability students.

I run a randomized controlled trial (RCT) to evaluate the effects of ChatGPT (the most widely used GenAI system) on debating performance in a university debating competition, involving 142 undergraduate students. One of the clearly relevant applications of ChatGPT is that it provides very fast summaries of complex ideas and concepts, arguments in favor or against any topic and endless examples and metaphors. For these reasons, ChatGPT can be a powerful ally for improving debating skills. I test two main hypotheses: (1) whether ChatGPT improves overall debating performance and (2) whether it contributes to reduce inequality in debating results.[1] The main outcomes I measure are (a) debating points (per student) and (b) probability of winning the debate.

The debate contest followed a simplified version of the "British Parliament" style competition, consisting of three to four rounds of short, one-on-one debates. After the first round of debates (the baseline), half of the students - the treatment group - were randomly assigned a 20-minute intense training of ChatGPT and were allowed to use it as support throughout the contest. The control group could only use conventional resources on the internet. Students were also randomly assigned to debating positions and debating partners. Each debate was audio recorded and sent to three different independent expert judges that did not have any information about the experiment. The rubric for the evaluation followed the usual metrics in international debating competitions. Ten prizes of 100€ in Amazon Vouchers were also offered to the winners to incentive maximum effort among students.

I find that ChatGPT has a positive but not significant effect on overall debating performance. Treatment debaters are 9.2% more likely to win than control debaters and score on average 2.2% (equivalent to an increase by 0.15 standard deviations) higher than control individuals, but the differences are not statistically different from zero.

However, this result masks substantial heterogeneity. I find that ChatGPT helps significantly more those students in the top of the skill distribution. Using a measure of innate student ability - whether they are on a merit scholarship or not - I find that high ability students experience significantly larger improvements from using ChatGPT than non-high ability students. The coefficient estimate shows an improvement of 12% among treatment students who are on a scholarship, and a 1.7% (insignificant) effect of ChatGPT among those without a merit scholarship.

[1]This study was preregistered at the AEA registry for randomized controlled trials, with register number AEARCTR-0011113. See https://www.socialscienceregistry.org/trials/11113.

Using an additional measure of ability, I find that among those in the top 50% in debating points in the baseline round (before the treatment was implemented), treatment students have on average 5.2% higher points than control students - the equivalent of a 0.47 standard deviation increase compared to control students-. Among those with lower baseline debating points (bottom 50%), treatment individuals do not benefit at all from ChatGPT. These findings suggest that for tasks such as debating that require higher-order skills, GenAI is complementary to ability and may increase inequalities in productivity.

The analysis of the judges' debate evaluation by subcategory (see Annex A.5 for details) points to potential mechanisms behind my main result. First, the impact of ChatGPT seems to depend on how effectively the information is extracted from the AI and used. For high-skilled students, ChatGPT significantly boosts scores in four out of five debating indicators, such as the *credibility* and *superiority* of their arguments and their *refutation* and *rhetoric* capacity. However, for low ability students - seemingly less capable to effectively prompt the machine - access to GenAI has no significant effects in four out of five indicators. Secondly, the impact of ChatGPT for high and low ability students varies depending on the debating task for which it is being used. For instance, while ChatGPT is clearly useful for low-ability students to improve the *clarity* of their presentation (a measure associated with the structure of the information presented, rather than the quality), it does not help them in other tasks associated with the capacity to discriminate good from bad quality content. This underscores the importance of understanding who benefits from GenAI and on what specific tasks.

The students' end-line survey responses provide some further evidence on the possible mechanisms that might be at work. High-ability individuals experience positive, significant and large effects of ChatGPT on self-reported perceptions regarding "having had sufficient time to prepare the debates". When given access to ChatGPT, their perception of time sufficiency increased to well above that of low-ability students.

I suggest a possible theoretical explanation to reconcile these results with the earlier findings supporting the "productivity-compression" hypothesis of GenAI. In predictable written tasks, GenAI and Machine Learning (ML) models learn patterns of behavior of best and worst performers (Brynjolfsson et al., 2023). This allows AI systems to reproduce the content of best performers, helping poor performers improve (more than good performers) through simple prompting, little reflection, and copy-pasting answers.

In such environments, poor performers can be expected to be at least as competent as the AI system they can access. However, in social environments with repeated knowledge-intensive interactions, closer to those found in day to day real managerial jobs of all kinds, high-skilled individuals - because of previous deeper knowledge, their superior persuasive abilities, or, more generally, their higher-order skills - will be better able to extract and use the information provided by GenAI to their advantage than low-skilled individuals.

Most studies analyzing the effect of GenAI on productivity focus on testing writing tasks at which LLMs are especially effective. These include programming (Peng et al., 2023), professional writing, such as writing emails or press releases (Noy and Zhang, 2023), law examinations (Choi and Schwarcz, 2023), and creative writing (Doshi and Hauser, 2023). Only few studies have tested the impact of GenAI in realistic work environments or involving social interactions. Dell'Acqua et al. (2023), for instance, use a large sample of consultants to study the productivity effects of ChatGPT at 18 written tasks, such as writing a 500-word memo for the CEO or coming up with ideas for good marketing slogans. Brynjolfsson et al. (2023) test the effects of a specifically trained Machine Learning model for customer support with real customer service agents. A writing bot helped resolving (highly predictable) written questions, the answers to which were more than 80% of the time automatically copy-pasted by the agents in a chat. These studies find that GenAI generally boosts productivity of all workers, while compressing the initial inequality in task performance because of larger improvements of worst performers. Closer to my research, although in a very different open-ended entrepreneurial decision-making environment, Otis et al. (2023) run a field experiment over several months to assess the impact of AI-generated advice on revenues and profits of small businesses in Kenya. In such context, also requiring complex problem-solving skills, they find that high initial performers benefited more than low performers from AI assistance.

This paper makes four contributions. First, this is one of the few studies analyzing the productivity impact of GenAI in a verbal task requiring a complex set of higher-order skills. Second, contrary to most findings in the early literature, my results suggest that in such settings, GenAI might increase the productivity gap between low and high ability individuals. Third, I show that the benefits of GenAI among low and high ability participants vary significantly depending on the specific task they perform. Finally, I suggest a

possible theoretical explanation: when answers cannot be directly extracted from the AI and copy-pasted, high-skill workers are likely to benefit more of the advantages of GenAI.

Understanding what types of workers might benefit from GenAI and on what tasks is a question of utmost importance. My findings complement the existing literature by presenting a case where productivity inequality increases as a result of GenAI. Given the proven importance of social and critical thinking skills for high-earning workers of all kinds Deming (2017), (Green, 2024), if my findings generalize, they might have broader implications for understanding the economic and social impacts of GenAI.

The rest of the paper is organized as follows: In Section 2, I describe the design of the debate competition. In Section 3, I discuss the experimental design and implementation. In Section 4, I explain the empirical strategy and the hypotheses tested. Section 5 shows the results and mechanisms in more detail. Section 6 concludes.

2 Context of the intervention

The intervention took place at ESADE Business School and at CUNEF University over three different debating days. The two first sessions, with 38 and 50 students registered respectively, where organised at ESADE on the 20th and 24th of March 2023. The third session, with 58 participants, was organized on the 19th of April 2023. The target population were undergraduate university students. Most participants were from three courses: two Debating courses at Esade and an Economic Policy class at CUNEF. Most students were enrolled in either law, business or economics degrees. The first session was run in English, the other two sessions in Spanish.

The experiment was presented to participants as a debating competition with the aim of testing the impact of different technological tools in debates. Students signed an informed consent before the start of the competition (see Annex A.1) and were asked to fill out a registration form and baseline survey (see Annex A.2) by their professors ahead of the scheduled debate competition days.

The challenge consisted in three to four rounds of short one-on-one debates on public policy topics over a three-hour session. At the beginning of the competition students were told that all debates were going to be recorded and sent for evaluation by independent judges. Participants with the highest number of points according to the judges' criteria would be the winners.

At the moment of the intervention, ChatGPT3.5 had been out for about four months.

Less than half of participating students declared having used ChatGPT before.

2.1 Debating rules and tools

I designed the technical aspects of the debating competition with the support of professional debating teachers at ESADE. The rules of the debate were sent in advance to participants (see Annex A.3 for details). The design of the debates followed the British Parliament (BP) style. BP is used, for instance, in the world's largest international official debating tournament, the World Universities Debating Championship (WUDC).

In my experiment, for evaluation purposes, I chose to do individual debates rather than group debates, as commonly done in BP debates. Also, given limited class-time slots offered by professors, I shortened the length of debates to 3+2 minutes per debater: each student had three minutes for an opening statement and two minutes for refutation and conclusion. Preparation time for each round of debates was 15 minutes. Students were asked to bring their computers and cellphones to the competition and did not receive any materials in advance to prepare the debates.

2.2 Spaces, monitoring and audio-recording

All participants debated at the same time in a large room in groups of two. Treatment and control groups were separated in different rooms to prepare the debates and did not interact among them except for the competition time. A team of six people was monitoring the whole time to make sure there were no interactions between treatment and control participants and avoid cheating or contamination. Debates were audio-recorded by students with their cellphones and uploaded to a folder.

At the end of the debating competition, each student had done three or four rounds of debates. In total, each student had about two hours of debating time, involving preparation and actual debates. After the debates finalized, the recordings were sent to the judges for evaluation. In total, there were 230 debates recorded, each of ten minutes, which is equivalent to 2300 minutes of recordings (about 39 hours).

2.3 Policy topics debated

Students were randomly assigned to debating positions in each of the three different debating days. The final eight debating topics were selected from a set of twenty topics

previously circulated with an informal group of 10 academic economists. Debates addressed a variety of topics, such as taxes, education or trade. All debate topics had two opposing positions: one "supported by strong evidence" in economics and the other "supported by weak or no evidence".

Examples of topics debated were "Rent Controls: Should the state set housing prices?" or "Retirement age and young employment: Would lowering the retirement age help young people to find work?". Annex A.4 shows the full list of topics. Debate topics were deliberately selected so the "supported by weak or no evidence" policy position was half the time coming from policies typically associated with the left and half the time from the right. The reason why I chose eight different policy topics was that I wanted to avoid potential information leakages among students among different debate days.

2.4 Judges and evaluation criteria

Expert judges were chosen according to two criteria: they had to be either former debating champions or debating teachers at some university. Nine judges ended up participating in the experiment. The reason why I chose to have nine judges was time restrictions on the judges' side. Judges were randomly assigned a set of debates to evaluate but were not given any details about the study. They also received a rubric for the evaluation involving five different criteria. They had to give ten points to each participant in five standard categories: (1) clarity and validity of the defended position; (2) credibility of the evidence used; (3) formal quality and rhetoric; (4) ability to refute the rival's arguments and (5) superiority of the arguments used (see Annex A.5 for more details). Judges were asked to provide two sets of final outcomes: total debating points of each participant and winner of the debate. Every debate was evaluated by three different judges in order to reduce the probability of results being explained by potential judge biases. To calculate the final scores, I computed the average of the three evaluations for each debate. All judges were paid to do the evaluation.

2.5 Incentives for participants

All participants were offered a certified diploma just for participating. Ten prizes of 100€ in Amazon Vouchers were also offered. Given that participants were randomly assigned different tools for debating, prizes were given according to the tools used: five prizes to

top performers in the treatment group and five in the control group. The reason why I introduced economic incentives was to make sure students made the maximum possible effort in their debates. There were no other academic rewards involved for participating.

3 Randomization, implementation and data

Figure A1 shows the timeline of the intervention.

3.1 Randomization

Randomization was done *in situ*, using a STATA command, at the individual level on each of the three debating days. The randomization took three steps: First, participants were randomly assigned to treatment and control groups. Second, participants were randomly assigned to debating partners in each of the four rounds. Third, participants were randomly assigned debating positions in each round. Each of the four rounds of debates corresponded to a different policy topic. As a result of this randomisation strategy, treatment students could be at any time debating with other treatment or control students, and also repeat debating partners (which happened very rarely in practice).

Table 1 shows balancing in baseline characteristics between treatment and control group individuals. Randomization was balanced with respect to the information collected at baseline, which included basic socio-demographic characteristics, academic achievements, previous experience in debating competitions or courses, preferred language for debates and previous experience in using ChatGPT, among other variables.

3.2 Implementation

Upon arrival, students were given an ID number and were put together in a large room. Over the first 20 minutes, there were two introductory interventions to explain the rules of the debating competition and how to record and upload the debates. Over that time, the IDs were used to randomly assign students to treatment and control, as well as to partners and positions for the three to four rounds of debates. In the first round of debates ChatGPT was prohibited for everyone.

After the first round, treatment students were sent to a separate room and received a 20-minute training on ChatGPT by a ChatGPT heavy user. Control students were given an extra 20-minute break. Then, the debate matching and topics for the following rounds were projected on big screens. Treatment and control students remained separated in two

rooms and were given 45 minutes to prepare the three following debates. AI tools were explicitly prohibited for control students. Treatment students were explicitly told not to tell about ChatGPT. When the four rounds of debates were finalized, students were asked to fill an endline survey.

3.3 Data

In this section I describe the data collection process, the kind of information I collected at base- and endline and the outcome measures I constructed.

Baseline information Ahead of the experiment I received the ethics approval from ESADE's Ethics Committee. All students were automatically pre-registered and asked to fill a baseline survey as a condition to participate in the competition. The baseline survey included a consent form agreed with the ESADE Decision Lab (see Annex A.1 and A.2). The survey asked for relevant student characteristics, socio-economic background information and political views of the students. It included questions on age, gender, type of studies, parent's education level, scholarships, past academic results, mother language, preferred language for debating, previous debating experience and previous knowledge of the topic. I also asked whether students felt comfortable speaking in public and if they had used AI tools in the past.

Endline survey At the end of the debating session, students were asked to fill a short end-line survey asking for their views about the debating contest, as well as the use of their time and the debating tools they had access to. The survey included relevant questions for productivity, such as whether they felt they had enough time to prepare the debates and views on the policy topics debated. I have used some of these questions to explore potential mechanisms in the mechanisms subsection.

Outcome variables There are two main outcome variables: individual debating points and winning the debate. Every student did three to four debates, so I have three to four observations per individual. For each student, each observation is a five minute (3+2) recording of his two interventions in each round of debates. The recorded debates were sent for evaluation to three different judges (see Section 2.4) . Judges were asked to give up to 50 points to each debater following a rubric five indicators typically used in

debating competitions (see Annex A.5). In order to construct my main outcome variable, individual debating points, I compute the average points given by the three judges to each participant, separately for each of the three to four rounds of debates. I classify an individual as winner of the debate if their average score (across three judges) in a debate was higher than that of their their rival.

4 Empirical strategy

In this section I explain the estimation strategy for the two main hypotheses tested.

First, I test whether access to ChatGPT improves individual debating points or the probability of winning a debate.

To test this, I compare average results of students with and without ChatGPT access, running OLS regressions of the following form:

$$Y_{ird} = \alpha_d + \alpha_r + \beta Treat_i + \lambda X_i' + \epsilon_{itd} \tag{1}$$

Where Y_{ird} refers to individual debating points or a dummy indicating having won the debate for individual i in debating round r on debating day d. The α_d's represent debate day fixed effect, as randomization took place separately each day; the α_r's represent debate round fixed effects to account for learning over the course of the debate competition and to control for potential variation in debate difficulty. The coefficient of interest is β, which measures the causal effect of having been assigned ChatGPT. The set of controls, denoted by the vector X_i, include the outcome in round 1 (baseline debate points), to increase the power of the experimental design. In specifications with controls, I additionally control for age, gender, parental education, whether studies subject related to economics, prior debating experience, whether the students has prior experience using ChatGPT, whether the student is recipient of a scholarship, whether the student feels comfortable in the debating language, and ability measured by high school diploma grades. Standard errors in this specification are clustered at the individual level, to take into account the fact that I observe each individual several times and outcomes are likely to be correlated within individuals across rounds.

Secondly, I test whether participants with lower ability benefit more from ChatGPT than those with higher ability. If this were the case, ChatGPT could be regarded as a productivity compressor, closing the gap in productivity between low and high ability individuals.

The specification has the flavor of a difference-in-difference design, as I estimate the difference in the performance in later rounds of those assigned to and those not assigned to using ChatGPT, across low and high ability individuals:

$$Y_{ird} = \alpha_d + \alpha_r + \beta Treat_i + \gamma HighAbility_i + \delta Treat_i \times HighAbility_i + \lambda X_i + \epsilon_{ird} \quad (2)$$

HighAbility can refer to two variables: In the first case, it is a binary variable that takes the value one if the participant is a scholarship (based on academic excellence) recipient, and zero otherwise. In the second case, it is a dummy that is equal to one if the student ranked in the top 50% of the distribution of debate points in the first round of debates, and zero if they ranked in the bottom 50%. The coefficient of interest is δ, which quantifies the interaction effect of assignment to using ChatGPT for individuals with high ability. A positive coefficient would indicate that ChatGPT has a greater positive effect among participants who have higher ability. Again, standard errors are clustered at the individual level.

5 Results

In this section I present the results of the preregistered experiment. Table 2 shows the estimated effects of the treatment on debating points and the probability of winning the debate. In Column 1, which estimates equation 1 and includes debate day and round fixed effects, and controls for baseline performance, treatment group individuals score 0.518 points higher than control individuals. When including additional controls (Column 2), the coefficient increases to 0.672, but remains imprecisely estimated and is not significant at conventional levels. When looking at "winning the debate" (Columns 3 and 4), ChatGPT increases the probability of winning by between 3.9 and 4.6 percentage points, but neither of these estimates is significant at conventional levels. Treatment debaters score on average 2.2% higher (equivalent to an increase by 0.15 standard deviations) than control individuals, and they are 9.2% more likely to win, but neither of the results is significant. The fact that I find no overall significant effects does not necessarily mean that ChatGPT has no impact on debating performance. Given my sample size of 142 students, I estimate the minimum detectable effect size in 0.4SD at 90% confidence levels.

The nature of the debate might be altered by the debating partners' performance,

which could be affected by the latter's treatment status. Even though my main specifications control for rival's treatment status, I also analyze the interaction between own treatment status and rival's treatment status to see whether the effect of treatment is different when competing against another treated student during a debate. This is particularly relevant for winning the debate (if both treated students perform better, probability of winning might not rise for either of them). The results are presented in Table A5. Column 1 shows the effect on total debating points. Individuals using ChatGPT score on average 0.605 points more than control individuals (not significant), and this is not affected by whether they debated against a student in treatment (also using ChatGPT) or control group. Column 2 shows the effect on the probability of winning a debate for treatment individuals. The probability of winning a debate is 10.5 percentage points higher (not significant) when treatment individuals debate with a control group person. When treatment individuals compete against a treated student, their probability of winning is the same as that of control individuals (interaction effect=-0.125).

Heterogeneity These overall results hide interesting heterogeneity. I use a measure of student ability - whether a student is recipient of a merit scholarship (based on academic excellence) - to test whether the effect of ChatGPT varies by student ability. Table 3 shows the effect the of treatment by whether the student is a scholarship recipient, estimated using equation 2. I find that students on such scholarships experience significantly larger improvements from ChatGPT than those without a scholarship. They improve their total debating points by 3.1 points compared to control students, while those without a scholarship just show a small positive but insignificant improvement. The result is summarized in Figure 1. The coefficient estimate represents a significant improvement by 11.8% for students using ChatGPT among those on a scholarship, and a 1.5% (insignificant) effect of ChatGPT among those without a scholarship. High ability treated students are also significantly more likely to win the debate (even though the coefficient becomes insignificant in the specification with controls, column 4), whereas their low-ability counterparts do not benefit from the treatment in this respect.

I complement the analysis by checking whether the impact of ChatGPT varies by baseline debating performance, an alternative measure of student ability. Table A2 shows the heterogeneous effects by baseline score, estimated using Equation 2. I find that those with high scores at baseline (top 50% of the debate performance distribution) benefit

from ChatGPT - they improve their total debating points by 2.36 points (equivalent to a 0.8 standard deviation increase) compared to control students-, while those with lower baseline debate points (bottom 50%) do not benefit at all. [2]

Taken together, these findings suggest that for tasks that require higher-order skills, GenAI is complementary to ability and may increase inequalities in productivity.

Mechanisms In order to gain a better understanding for what types of tasks and for what types of users GenAI is useful, I next exploit detailed information contained in the judges' evaluations. The rubric requires judges to evaluate individual debaters on five dimensions: (1) "clarity and validity of the defended position"; (2) "the evidence is credible"; (3) "formal quality of the participant rhetoric"; (4) "the ability to refute the rival's position" and (5) "the arguments are superior to those of the rival".

When separately estimating the points in the five rubric categories for low and high ability students measured by whether they were recipients of a scholarship, I find that in most categories, it is high-ability students who benefit. As Table 4 shows, ChatGPT significantly boosts high-ability participants' productivity in four out of five dimensions, including *credibility* and *superiority* of the arguments and *refutation* and *rhetoric* capacity (when pooled together, the overall effect in credibility and rhetoric is neutralized by the low-ability students).

These results reinforce the idea that the impact of ChatGPT depends on how effectively the information is extracted from the AI and used, which, in turn, is a function of ability. To illustrate this point further its interesting to look at the "credibility of evidence" indicator, which reflects the students capacity to discriminate between relevant, misleading or false evidence. The indicator shows a very strong effect of ChatGPT for high-ability students and a negative (but insignificant) effect for low-ability students.Given that previous research has pointed to problems of limited reliability of LLMs and invented references (Bommasani et al., 2021; Weidinger et al., 2021), being able to discern good from bad information seems important to extract the benefits of ChatGPT. One way of reading these results is that ChatGPT does provide relevant evidence, but extracting it effectively from the AI crucially depends on one's ability. Having more information but little capacity to discriminate, might even lead to reduce produc-

[2]The result is summarized in Figure A2..Tables A6 and A7 show descriptive statistics comparing scholarship recipients with non-recipients, and top with bottom baseline performers, respectively.

tivity.

Another relevant insight can be extracted from Table 4: ChatGPT affects different debating tasks differently. For instance, low-skill seem to significantly benefit from ChatGPT in one specific task: improving the clarity of their defended position. This indicator can be interpreted as helping to improve the structure (rather than the content) of their presentation. These findings highlight that a good analysis of the impact of GenAI on inequality will need to look at how GenAI interacts with specific tasks and users.

Finally, I exploit the information of the end-line survey to analyze the effects of the treatment on two questions: (1) Did you think we provided enough time to prepare for the debates? - with 100 being plenty of time - and; (2) How useful were the AI tools / materials we gave you to prepare the debates? - with 100 being very useful. As reported in Table A3, I find a positive and significant effect of ChatGPT on self-reported perception of time needed to prepare the debates, but only for high-ability individuals: they self-rate the sufficiency of time 31 points (or 1.06 standard deviations) higher than control individuals. I also find a positive effect of the treatment on students' perceptions of the usefulness of the tools used for the debate: among those not on a scholarship, treatment individuals value the usefulness of materials 36 points higher than students having only internet access (or 1.07 standard deviations), and the effect was similar in size among those on a scholarship (albeit not significant).

Since I randomly assigned participants to debating positions, I can also study whether the impact of ChatGPT varies depending on the position defended in the debates. As explained in Section 2.3, all debates had a policy position "supported by strong evidence" in economics and a policy position "supported by weak or no evidence". I find that ChatGPT clearly favors positions "supported by strong evidence". Table A4 shows the results of estimating equation 1, separately for debates where individuals were assigned to defend the "weak evidence" and the "strong evidence" position. When individuals had to defend the strong evidence position, treatment had a positive effect and increased total points by 1.7 on average (0.38 SD). However, when individuals had to defend the weak evidence position, being assigned to using ChatGPT had no positive effect in performance. This result has implications for the ongoing debate as regards to the potential risks of inadequate or misleading information that have been raised in the past relation to LLMs.

6 Conclusion

A novelty of GenAI systems is that they acquire knowledge through observation rather than rules, which allows them to perform sophisticated tasks, traditionally reserved for highly skilled professionals (Autor, 2024). Early experiments testing the impact of ChatGPT on work productivity in a variety of written tasks show a common pattern: GenAI systems help low performers more than high performers, thus compressing the productivity distribution. This study complements previous findings by exploring the effects of GenAI on a novel task, a debating contest, which provides an ideal setting to test whether GenAI can increase productivity when different higher-order skills are essential for productivity.

Contrary to initial results in the GenAI literature, I find that high-skilled individuals benefit more from the interaction with ChatGPT than low-skilled individuals. I also suggest a possible explanation to reconcile these results with existing literature: in written, predictable interactions, low performers will do at least as good as the GenAI system to which they have access; but when higher-order skills are required in realistic, unpredictable social contexts, high-skill workers are likely to enjoy stronger complementarities with AI. If these findings replicate in other contexts involving higher-order skills, such as in-person negotiating or selling, for instance, they would have relevant implications to understand the impact of GenAI on work inequalities.

References

Acemoglu, Daron and David Autor, "Chapter 12 - Skills, Tasks and Technologies: Implications for Employment and Earnings," in David Card and Orley Ashenfelter, eds., *Handbook of Labor Economics*, Vol. 4, Elsevier, 2011, pp. 1043–1171.

_ **and Pascual Restrepo**, "Robots and Jobs: Evidence from US Labor Markets," Working Paper 23285, National Bureau of Economic Research March 2017.

Adams, John Quincy, *Lectures on Rhetoric and Oratory: Delivered to the Classes of Senior and Junior Sophisters in Harvard University*, Vol. 1, Cambridge: Hilliard and Metcalf, 1810.

Aristotle, *The Rhetoric of Aristotle*, New York: Appleton-Century-Crofts, Inc, 1960.

Autor, David, "Applying AI to Rebuild Middle Class Jobs," Working Paper 32140, National Bureau of Economic Research February 2024.

Autor, David H., Frank Levy, and Richard J. Murnane, "The Skill Content of Recent Technological Change: An Empirical Exploration," *The Quarterly Journal of Economics*, 11 2003, *118* (4), 1279–1333.

Bommasani, Rishi, Drew A. Hudson, Ehsan Adeli, Russ Altman, and Simran Arora, "On the Opportunities and Risks of Foundation Models," *ArXiv*, 2021.

Brynjolfsson, Erik, Danielle Li, and Lindsey R Raymond, "Generative AI at Work," Working Paper 31161, National Bureau of Economic Research April 2023.

Börner, Katy, Olga Scrivner, Mike Gallant, Shutian Ma, Xiaozhong Liu, Keith Chewning, Lingfei Wu, and James A. Evans, "Skill discrepancies between research, education, and jobs reveal the critical need to supply soft skills for the data economy," *Proceedings of the National Academy of Sciences*, 2018, *115* (50), 12630–12637.

Choi, Jonathan H. and Daniel Schwarcz, "AI Assistance in Legal Analysis: An Empirical Study," *Minnesota Legal Studies Research Paper No. 23-22*, 2023.

Chui, Michael, Eric Hazan, Roger Roberts, Alex Singla, Kate Smaje, Alex Sukharevsky, Lareina Yee, and Rodney Zemmel, "The Economic Potential of Generative AI: The Next Productivity Frontier," June 2023. Accessed: 2024-05-22.

Corbett, Edward P.J. and Robert J. Connors, *Classical Rhetoric for the Modern Student*, 4th ed., New York: Oxford University Press, 1999.

Dell'Acqua, Fabrizio, Edward McFowland, Ethan R. Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R. Lakhani, "Navigating the Jagged Technological Frontier: Field Experimental Evidence of the Effects of AI on Knowledge Worker Productivity and Quality," *Harvard Business School Technology & Operations Mgt*, 2023, *Unit Working Paper No. 24-013.*

Deming, David J., "The Growing Importance of Social Skills in the Labor Market," *The Quarterly Journal of Economics*, 06 2017, *132* (4), 1593–1640.

_ , "Four Facts about Human Capital," *Journal of Economic Perspectives*, August 2022, *36* (3), 75–102.

Doshi, Anil Rajnikant and Oliver Hauser, "Generative Artificial Intelligence Enhances Creativity but Reduces the Diversity of Novel Content," *SSRN*, 2023.

Eloundou, Tyna, Sam Manning, Pamela Mishkin, and Daniel Rock, "GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models," March 2023, (2303.10130).

Felten, Edward, Manav Raj, and Robert Channing Seamans, "The Effect of Artificial Intelligence on Human Labor: An Ability-Based Approach," *Academy of Management Proceedings*, 2019, *2019* (1), 15784.

Felten, Edward W, Manav Raj, and Robert Seamans, "Occupational heterogeneity in exposure to generative ai," *Available at SSRN 4414065*, 2023.

Green, Andrew, "Artificial intelligence and the changing demand for skills in the labour market," *OECD, Artificial Intelligence Papers*, 2024, (14).

Noy, Shakked and Whitney Zhang, "Experimental Evidence on the Productivity Effects of Generative Artificial Intelligence," 2023.

Otis, Nicholas G., Rowan Philip Clarke, Solene Delecourt, David Holtz, and Rembrand Koning, "The Uneven Impact of Generative AI on Entrepreneurial Performance," OSF Preprints hdjpk, Center for Open Science December 2023.

Peng, Sida, Eirini Kalliamvakou, Peter Cihon, and Mert Demirer, "The Impact of AI on Developer Productivity: Evidence from GitHub Copilot," 2023.

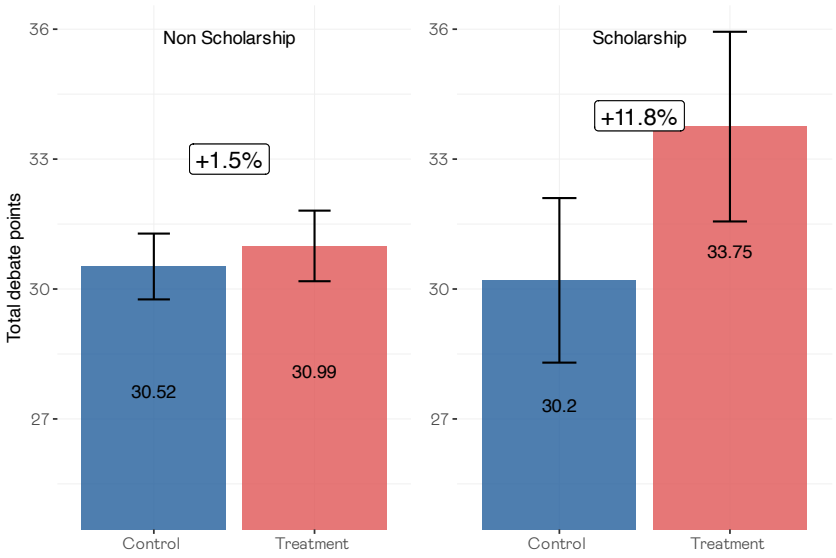
Weidinger, Laura, John Mellor, Maribeth Rauh, Conor Griffin, Jonathan Uesato, Po-Sen Huang, Myra Cheng, Mia Glaese, Borja Balle, Atoosa Kasirzadeh, Zac Kenton, Sasha Brown, Will Hawkins, Tom Stepleton, Courtney Biles, Abeba Birhane, Julia Haas, Laura Rimell, Lisa Anne Hendricks, William Isaac, Sean Legassick, Geoffrey Irving, and Iason Gabriel, "Ethical and social risks of harm from Language Models," *arXiv*, 2021, *2112.04359.*

Weinberger, Catherine J., "The Increasing Complementarity between Cognitive and Social Skills," *The Review of Economics and Statistics*, 12 2014, *96* (5), 849–861.

Wilmers, Nathan, "Generative AI and the Future of Inequality," *An MIT Exploration of Generative AI*, mar 27 2024. https://mit-genai.pubpub.org/pub/24gsgdjx.

Figures

Figure 1: Heterogeneity in treatment effects of ChatGPT by student ability



Notes: This figure shows the average debate points of control and treatment individuals for those with and without a merit scholarship, conditional on control variables. The spikes represent 90% confidence intervals (predictive margins of the treatment indicator ($Treat$) in equation 2 by baseline performance), and above the bars I show the percent difference in the outcome between treatment and control group.

Tables

Table 1: Balancing between treatment and control group

	(1) Treat	(2) Control	(3) Difference (1)-(2)	(4) p-value Col. (3)
Age	20.31	20.33	-0.01	0.96
Female	0.45	0.48	-0.03	0.74
Father holds Master or higher degree	0.44	0.44	0.00	1.00
Economics background	0.48	0.51	-0.03	0.74
Scholarship	0.13	0.17	-0.04	0.48
Has prior debating experience	0.44	0.52	-0.08	0.32
Has won a debate prize before	0.24	0.18	0.06	0.41
Knows ChatGPT	0.39	0.41	-0.01	0.87
Has used ChatGPT before	0.41	0.38	0.03	0.73
Baseline debate points (0-50)	30.18	28.67	1.51	0.08
Enjoyment (0-100) of speking in public	65.65	70.29	-4.65	0.33
Feels comfortable in debating language	0.59	0.58	0.01	0.87
Political position (1=left, 10=right)	6.81	6.16	0.65	0.07
Polarised (0-100)	53.62	54.02	-0.40	0.91
N	71	71		

Notes: The table shows balancing between the treatment and control group for the sample of students who registered to participate in the debating competitions. Column 1 reports the mean in the treatment group and Column 2 reports the mean in the control group. Column 3 reports the difference in the mean across the two groups, and Column 4 reports the p-value of a t-test of the equality in means across the two groups.

Table 2: Effect of ChatGPT on the debate performance

	Debate points		Win	
	(1)	(2)	(3)	(4)
Treat	0.518	0.672	0.039	0.046
	(0.545)	(0.498)	(0.054)	(0.051)
Constant	27.337***	13.865***	-0.242	-0.813
	(1.527)	(4.928)	(0.166)	(0.511)
Mean dep. var.	28.67	28.67	0.48	0.49
SD dep. var.	4.53	4.53	0.50	0.50
R^2	0.13	0.21	0.05	0.11
Obs.	364	364	364	364
Baseline score	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes

Notes: Significance levels are indicated by * $<$ .1, ** $<$ .05, *** $<$ .01. This table shows results from regressions of equation 1, where the outcome variable is debate points (Columns 1 and 2) or a dummy equal to one if person i won in debate round r (Columns 3 and 4). Specifications with controls (Columns 2 and 4) include the following control variables: age, gender, parental education, whether studies subject related to economics, prior debating experience, whether the students has prior experience using ChatGPT, whether the student is recipient of a scholarship, whether the student feels comfortable in the debating language, and ability measured by high school diploma grades. Standard errors are clustered at the individual level because each individual is observed between 2 and 3 times, depending on the number of debates they completed. The total number of individuals included in each regression is 141 out of 142 randomized individuals. One individual did not complete the baseline survey.

Table 3: Effect of ChatGPT on total points by whether is recipient of merit scholarship

	Debate points		Win	
	(1)	(2)	(3)	(4)
Treat	0.233	0.473	0.027	0.052
	(0.589)	(0.558)	(0.062)	(0.061)
Scholarship	-1.217	-0.322	-0.203**	-0.113
	(1.043)	(1.056)	(0.094)	(0.097)
Treat x Scholarship	3.815***	3.082*	0.321**	0.211
	(1.411)	(1.688)	(0.125)	(0.162)
Constant	32.965***	22.451***	0.477***	0.163
	(0.706)	(4.838)	(0.074)	(0.540)
Mean dep. var.	28.67	28.67	28.67	28.67
SD dep. var.	4.53	4.53	4.53	4.53
R^2	0.12	0.20	0.02	0.08
Obs.	364	364	364	364
Baseline score	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes

Significance levels are indicated by * $<$ .1, ** $<$ .05, *** $<$ .01. Columns 1 and 2 show results from regressions of equation 2, where the outcome variable in Columns 1 and 2 is the total debating points of individual i in debate round r (calculated as the average across different judges evaluating the same debate of individual), and in Columns 3 and 4, a dummy variable equal to one if the individual won the debate. Standard errors are clustered at the individual level because each individual is observed between 2 and 3 times, depending on the number of debates they completed. The total number of individuals included in each regression is 141 out of 142 randomized individuals. One individual did not complete the end-line survey.

Table 4: Effect of ChatGPT on total debating points by sub-category

	Clarity		Credibility		Rethoric		Refutation		Superiority of arguments	
	(1) Low	(2) High	(3) Low	(4) High	(5) Low	(6) High	(7) Low	(8) High	(9) Low	(10) High
Treat	0.238***	-0.197	-0.134	1.756***	0.081	1.736***	0.126	1.858***	0.090	1.430***
	(0.083)	(0.306)	(0.204)	(0.491)	(0.093)	(0.259)	(0.106)	(0.199)	(0.094)	(0.230)
Constant	4.781***	8.684***	0.932	4.333***	3.777***	9.521***	1.225	2.881***	3.803***	6.657***
	(1.019)	(0.379)	(2.486)	(0.765)	(1.107)	(0.444)	(1.370)	(0.354)	(1.054)	(0.320)
Mean dep. var.	6.12	6.02	5.12	5.28	6.41	6.22	5.82	5.50	5.94	5.67
SD dep. var.	0.73	1.07	1.73	1.62	0.81	0.84	0.93	1.14	0.81	1.04
R^2	0.30	0.61	0.20	0.62	0.31	0.60	0.23	0.64	0.22	0.61
Obs.	304	60	304	60	304	60	304	60	304	60

Significance levels are indicated by * $< .1$, ** $< .05$, *** $< .01$. The table shows coefficients from regressions of equation 1, where the outcome variable is the total debating points of individual i in debate round r for sub-category of the rubric (calculated as the average across different judges evaluating the same debate of individual), separately for low ability (non merit scholarship recipients) and high ability (merit scholarship recipients) students. All specifications include the same controls as those reported in the notes to Table 2. Standard errors are clustered at the individual level because each individual is observed between 2 and 3 times, depending on the number of debates they completed. The total number of individuals included in each regression is 141 out of 142 randomized individuals because one individual did not complete the end-line survey.

A Online Appendix

A.1 Consent Form

The ESADE Debate Challenge is a debate competition in which participants are provided with various tools and resources to prepare for debates. Ten prizes in the form of Amazon vouchers worth €100 will be awarded per category, depending on the tools with which the participants compete. In the assessments, students compete only against students in their own category.

The aim of this study is to test the impact of different technological tools in democratic debates. To this end, the information (audios) collected during this event will be analyzed anonymously and scientifically by ESADE researchers. Therefore, by agreeing to participate in the ESADE Debate Challenge, you also agree to participate in a scientific study. The characteristics of the study are explained in more detail below.

This study is led by Antonio Roldán Monés at the ESADE Campus St. Cugat (Creapolis) and the Decision Lab. This project has been approved by the ESADE Research Ethics Committee (CUHSR protocol number: 011/2023).

You must be at least 18 years of age to participate in this study or provide informed consent from your parent/guardian.

If you agree to participate in this study, you will be asked to do the following:

- Complete an online questionnaire before the in-person debate.
- Participate in a debate competition that includes four debates and complete three very brief surveys.
- The participants themselves will record the debate on their cell phones and send the recording to an ESADE phone number where all recordings will be centralized. The files will be identified only by a code and will not be used or made available for any purpose other than the research project. The files will be destroyed at the end of the study.
- Participation in this study will take a total of 180 minutes of your time, but may be slightly extended due to logistical difficulties.
- There are no known risks beyond everyday life associated with your participation in this study
- All participants will receive a certificate from EsadeEcPol for their participation. Ten prizes of 100 euros will be awarded to the winners (5 in each of the two categories).
- Participation in this study is voluntary and you can withdraw from it at any time. You will not receive any direct benefit from the study.

- If you have any further questions or wish to report any problems related to the study, please contact the principal investigator, Antonio Roldán Monés, by e-mail at antonio.roldan@esade.edu.
- If you have questions about your rights and welfare as a volunteer participant in the study, please contact the Esade Research Ethics Committee at ethics@esade.edu.
- The confidentiality of your research records will be strictly maintained by ensuring that all data will be kept secure and that only the principal investigator and the research team will have access to these data. This means that no one else will have access to your data at any time during or after the study.

Basic information about the processing of your personal data

- Data controller: The data controller is the ESADE Foundation.
- Contact: lopd@esade.es
- Purpose: Consent to participate in research studies
- Legal basis: Consent, legitimate interest in the research studies and compliance with legal obligations.
- Addressee: Esade EcPol, Esade Decision Lab - Research Office, Principal Investigator(s)
- Data management: Data will be deleted when it is no longer necessary to fullfill the purpose for which it was collected. The most relevant information will be retained permanently. The criteria for retention or deletion will be based on public records regulations or result from the performance of public duties.
- Rights: you have the right to request from the controller information about your personal data and its rectification or erasure, or to restrict processing, or to object to processing, as well as the right to data portability, the right to withdraw your consent at any time, and the right to lodge a complaint with a supervisory authority.

By checking the box below, you agree to participate in the study and acknowledge that you have read, understand, accept and will comply with the above instructions and conditions.

- I agree

A.2 Baseline Survey

1. Full name
2. Gender (Male, Female)
3. Date of birth (Drop down)
4. Do you feel comfortable talking about complex topics in English?
   - ☐ Yes, I have no issues
   - ☐ Yes, quite comfortable
   - ☐ No, but I can hold my own
   - ☐ No, I find it very difficult
5. Do you feel comfortable talking about complex topics in Spanish?
   - ☐ Yes, I have no issues
   - ☐ Yes, quite comfortable
   - ☐ No, but I can hold my own
   - ☐ No, I find it very difficult
6. Degree you are currently pursuing:
   - ☐ Law and International Relations
   - ☐ Law
   - ☐ Economics
   - ☐ Business Administration
   - ☐ Other
7. University where you are pursuing these studies:
   - ☐ ESADE
   - ☐ Other
8. In your last high school year, on a scale of 0 to 10, where 0 is the lowest grade and 10 is the highest grade, what was approximately your average grade at the end of the year?
   - ☐ (0-2)
   - ☐ (2-4)
   - ☐ (4-6)
   - ☐ (6-8)
   - ☐ (8-10)
9. Are you currently receiving any type of scholarship?
   - ☐ Yes, an academic excellence scholarship
   - ☐ Yes, other income-related scholarship

☐ No

10. What is the highest level of education completed by your father?

    ☐ Primary school
    ☐ Secondary education
    ☐ Professional training
    ☐ University degree
    ☐ Master's degree
    ☐ PhD

11. What is the highest level of education completed by your mother?

    ☐ Primary school
    ☐ Secondary education
    ☐ Professional training
    ☐ University degree
    ☐ Master's degree
    ☐ PhD

12. On a scale from 0 to 100, with 100 being "a lot" and 0 being "not at all", how much do you enjoy speaking in public?

13. Have you participated in a debate competition before?

    ☐ Yes, several times
    ☐ Yes, once
    ☐ No, never

14. Have you ever received a prize in a debate competition?

    ☐ Yes, several times
    ☐ Yes, once
    ☐ No, never

15. On a typical day, approximately how much time do you spend watching, reading or listening to news about politics and current affairs? Please answer in hours and minutes. For example, if you spend one hour and twenty minutes, you would enter 01 under "HOURS" and 20 under "MINUTES".

16. In politics, we sometimes refer to the "left" and "right". Where would you place yourself on this scale? 0 means "left" and 10 means "right".

17. On a scale where 0 means you have very unfavourable feelings and 100 means you have very favourable feelings towards people who hold political views that are opposite to yours, where do you position yourself? A value of 50 means that your feelings are neither favourable nor unfavourable.

18. Which of the following software do you know?

    ☐ Google Drive
    ☐ Tableau
    ☐ ChatGPT
    ☐ Overleaf
    ☐ Jasper
    ☐ Grammarly

19. Which of the following software have you used recently?

    ☐ Google Drive
    ☐ Tableau
    ☐ ChatGPT
    ☐ Overleaf
    ☐ Jasper
    ☐ Grammarly

20. Price controls: Please indicate your level of agreement with the following statements: [0=Strongly disagree 100=Strongly agree].

    (a) A rent price control system should be implemented in all medium and large cities (0-100).
    (b) Price control systems are a bad idea (0-100)
    (c) Implementing a rent price control system in all medium and large cities would have positive consequences (0-100)
    (d) Do you agree or disagree with the implementation of a rent control system in all medium and large cities and neighbourhoods? [0=Strongly disagree 100=Strongly agree].
    (e) If there was a referendum tomorrow on implementing a system of rent price controls in all medium and large neighbourhoods and cities, how likely is it that you will vote in favour? [0=would NOT vote in favour with 100% certainty; 100 =would vote in favour with 100% certainty].

21. Central Bank Political Control: Please indicate your level of agreement with the following statements: [0=Strongly disagree 100=Strongly agree].

    (a) EU governments should regain political control of the ECB in order to be able to finance themselves on more advantageous terms (0-100)
    (b) The ECB's monetary policy should be subordinated to the fiscal needs of the member states (0-100)
    (c) In a situation where an EU member state is forced to make cuts, it is always better for the ECB to offer an unconditional bailout to that state (0-100)
    (d) Do you agree or disagree with EU member state governments regaining political control over the European Central Bank and thus, over monetary policy? [0=Strongly disagree 100=Strongly agree]

(e) If there was a referendum tomorrow on regaining political control over the direction of the European Central Bank and its monetary policy, how likely is it that you will vote in favor? [0=would NOT vote in favor with 100% certainty; 100 =would vote in favor with 100% certainty].

22. Retirement age and youth employment: Please indicate your level of agreement with the following statements: [0=Strongly disagree 100=Strongly agree].

    (a) Lowering the retirement age is a good measure to increase youth employment (0-100)

    (b) If there are 100 jobs in society, it is the government's responsibility to ensure that older people retire earlier in order to free up jobs for younger workers (0-100)

    (c) Lowering the retirement age from 67 to 60 would significantly reduce youth unemployment without negative effects (0-100)

    (d) Do you support or oppose the government's intervention to distribute jobs more equitably among different generations in society? [0=Strongly disagree 100=Strongly agree]

    (e) If there was a referendum tomorrow to lower the retirement age in order to increase youth employment, how likely is it that you will vote in favour? [0=would NOT vote in favour with 100% certainty; 100 =would vote in favour with 100% certainty].

23. Job guarantee: Please indicate your level of agreement with the following statements: [0=Strongly disagree 100=Strongly agree]

    (a) If a country has unemployed workers, the state should offer them a job through a job guarantee program (0-100)

    (b) A job guarantee program would lower the unemployment rate without negatively affecting other economic indicators (0-100)

    (c) Guaranteeing employment for all people of working age should be a recognised right, regardless of the public expenditure involved (0-100)

    (d) Are you in favour or against the government passing a law guaranteeing public employment for all unemployed people? [0=Strongly disagree 100=Strongly agree]

    (e) If there were a referendum on such a law tomorrow, how likely is it that you will vote in favour? [0=would NOT vote in favour with 100% certainty; 100 =would vote in favour with 100% certainty].

24. Taxes and tax collection: Please indicate your level of agreement with the following statements: [0=Strongly disagree 100=Strongly agree]

    (a) Reductions in taxes on labor, such as the personal income tax, cause people to work more and ultimately increase tax revenues. (0-100)

    (b) A tax increase leads to a taxpayer response in the form of lower consumption and employment, which ultimately reduces tax revenues. (0-100)

(c) When there are tax cuts in a country/region, workers from other regions move to where taxes are lower, which helps to increase final revenues (0-100)

(d) Are you in favor or against the government cutting taxes such as personal income tax or VAT? [0=Strongly disagree 100=Strongly agree]

(e) If there were a referendum tomorrow on lowering the general VAT rate from 21% to 10% and cutting income tax in half, how likely is it that you will vote in favor? [0=would NOT vote in favor with 100% certainty; 100 =would vote in favor with 100% certainty].

25. Determinants of social mobility: Please indicate your level of agreement with the following statements: [0=Strongly disagree 100=Strongly agree]

(a) The economic situation of the parents in childhood is not relevant for the future of a person. (0-100)

(b) A person's effort is the main determinant of his or her success in life. (0-100)

(c) The effort of a person from a poor background is rewarded in the same way as those of a person from a wealthy family. (0-100)

(d) Are you for or against the government redistributing income from the rich to the poor in order to reduce inequality of opportunity? [0=Strongly disagree 100=Strongly agree]

(e) If there were a referendum tomorrow to eliminate all wealth and inheritance taxes, how likely is it that you will vote in favour? [0=would NOT vote in favour with 100% certainty; 100 =would vote in favour with 100% certainty].

26. School repetition: Please indicate your level of agreement with the following statements: [0=Strongly disagree 100=Strongly agree]

(a) Repeating a year is an effective measure to improve the level of learning in the vast majority of cases where it is applied (0-100)

(b) Repeating a year does not increase the probability that a student will drop out of the educational system early (0-100)

(c) In terms of cost-benefit, repeating a year is a superior educational policy to tutoring in small groups or other reinforcement programs (0-100)

(d) Do you agree or disagree with the government passing a law that severely limits the cases in which a student can be required to repeat a year? [0=Strongly disagree 100=Strongly agree]

(e) If there was a referendum tomorrow to limit the cases in which a student can be required to repeat a year, how likely is it that you would vote in favor of it? [0=would NOT vote in favor with 100% certainty; 100 =would vote in favor with 100% certainty].

27. Trade policy: Please indicate your level of agreement with the following statements: [0=Strongly disagree 100=Strongly agree]

(a) The interests of national industry should be considered before opening trade with any country (0-100)

(b) Even if a product is much cheaper abroad, it should not be imported if it would result in job losses in national industry (0-100)

(c) It is preferable, from an economic efficiency standpoint, for consumers to pay higher prices so that national industry does not lose jobs (0-100)

(d) Do you agree or disagree with the government passing a law establishing a tariff on products imported from other countries to make domestic products more attractive? [0=Strongly disagree 100=Strongly agree]

(e) If there was a referendum tomorrow to limit imports from third countries, how likely is it that you would vote in favor of it? [0=would NOT vote in favor with 100% certainty; 100 =would vote in favor with 100% certainty].

A.3 Debate challenge regulation (sent to students)

The following document establishes the rules for the ESADE Debate Challenge, acceptance of which is a prerequisite for registration and participation in the competition. In addition, all participants must complete an Initial Questionnaire which includes a declaration of consent that the data collected during the event will be used for academic purposes.

A.3.1 Article 1. Eligibility

The competition is open to students from all colleges and universities and to participants in the Pre-University Debate League. Contestants may be asked to show proof of enrolment in university or in the Pre-University Debate League. In addition, minors under the age of 18 must provide a signed consent form from their parent or guardian.

A.3.2 Article 2. Required materials

All students must bring their laptop and a cell phone capable of audio recording, as well as the appropriate chargers for both devices, as they are essential for participation in the competition.

A.3.3 Article 3. Topics to be debated

There are eight different topics, each with its corresponding debate question. Each participant will have to defend the position assigned to him or her at random, either for or against, regardless of his or her personal opinions on the subject. The selected topics deal with current issues in the field of public policy. On the day of the debate, four of these eight topics will be randomly selected and discussed by the participants.

A.3.4 Article 4. Date, language and venue

The confrontation rounds will be held on March 20 and 24 in the auditorium of ESADE Creapolis in Sant Cugat del Vallès. On March 20, the debates will be held in English, while on March 24 they will be held in Spanish. The schedule for Monday, March 20, is from 3:00 p.m. to 6:30 p.m., and for Friday, March 24, from 3:30 p.m. to 7:45 p.m. All participants should arrive at the Creapolis Building auditorium 15 minutes before the designated time to register and begin on time.

A.3.5 Article 5. Structure and sequence of the individual discussion rounds

Each participant competes individually. The debates take place in the auditorium and are organised as direct confrontation between the speakers. After being briefed on the topic to be debated and the assigned position, 20 minutes are allowed for preparation. The debate lasts 10 minutes and proceeds as follows:

- First, the speaker with the position in favour begins the debate with a 3-minute presentation.
- Next, the speaker representing the opposing side gives their first 3-minute presentation.
- Then, the speaker in favour has 2 minutes to respond to their opponent's arguments.
- To conclude the debate, the speaker of the opposing side has another 2 minutes to address the counter-arguments of his opponent.

A.3.6 Article 6. Preparation tools and resources

This tournament is part of a research study in which participants will be randomly divided into two different groups. Each group will have access to different sources and tools to prepare for their respective debates, which may initially lead to some inequities. However, all participants will be evaluated based on the resources assigned to their respective groups, and prizes will be awarded based on individual scores, taking into account the tools available to each participant.

Speakers will have the option of reading their speech directly from a piece of paper or a screen if they deem it appropriate. They may also use brief notes to remember the main points of their argument.

A.3.7 Article 7. Communication between groups

From the moment the participants have been divided into two groups, communication between the members of the two groups is prohibited, except during the debate. Failure to comply with this rule may result in the participant's exclusion from the debate tournament.

A.3.8 Article 8. Recording of the debates

The debates will be recorded using the cell phone of one of the participants for later analysis. Below are the instructions for recording and sending the files:

- Before you begin recording, enable airplane mode on the mobile device being used for recording. This will avoid interruptions from incoming calls.
- Use an audio recording application pre-installed on the phone, such as "Voice Memos" on iPhone or "Recorder" on Android.
- Before the start of the debate, participants should state their identification number (assigned to them upon arrival), the question to be debated, and the position defended to facilitate file identification.

- The cell phone will be passed between the debaters as if it were a microphone to achieve better sound quality, considering that there will be more people in the room.
- Speak in a moderate tone of voice to ensure proper recording.
- It is recommended that participants bring a charged cell phone battery and charger if needed.
- At the end of the discussion and before leaving the table, the audio file will be sent via WhatsApp to +34 645 155 884.
- Before leaving the table, please notify someone from the organisation to verify that the audio file was received correctly.
- Once the recording has been sent via WhatsApp, the file will be uploaded to a OneDrive folder within 24 hours. The link to the corresponding folder will be provided at the end of the event.

A.3.9 Article 9. Evaluation of the debates

The recordings of the debates will be judged by a panel of experts in the field of debates. Each debate will be evaluated by three different judges. The evaluation will focus primarily on the substantive aspects of the debate, although the rhetorical and oral skills of the participants will be also considered. In this way, the overall performance of each speaker will be evaluated. The jury will decide which of the speakers are the winners. The evaluation will take place over a period of three to four weeks after the debate.

A.3.10 Article 10. Evaluation Criteria

Participants will be evaluated in each of the following categories:

1. Ability to respond to the question posed.
2. Coherence between thesis and argument.
3. Clarity of presentation: the arguments are easily recognisable.
4. Correctness of argument: it is not limited to the presentation of evidence.
5. Credible and adequately presented evidence.
6. Variety of evidence.
7. Appropriate and persuasive use of language.
8. Ability of the arguments to respond to the opponent.
9. Superiority of arguments compared to those of the opponent.
10. Ability of arguments to integrate counterarguments.

A.3.11 Article 11. Prizes and recognition

Cash prizes in the form of Amazon gift cards will be awarded to the participants who achieve the ten highest scores in the competition. Since the debaters have different resources to prepare their arguments, the prizes will be awarded considering the participants in equal categories. A total of 10 prizes of 100 euros each will be awarded. In addition, the remaining participants will receive a participation diploma awarded by EsadeEcPol, ESADE's Center for Economic Policy, in recognition of their efforts and commitment to the competition.

A.3.12 Article 12. Changes in the rules and regulations

The Organizing Committee reserves the right to make changes to the rules and regulations at any time and without prior notice.

A.4 Topics debated

The following topics were debated during the debate competitions:

1. Rent Controls: ¿Should the state set housing prices?
2. Job Guarantee: Should the State guarantee the full employment of the working-age population by directly providing jobs to the unemployed?
3. Central Bank Political Control: Should EU governments regain political control of the ECB in order to finance themselves on more advantageous terms and avoid austerity?
4. Retirement age and young employment: Would lowering the retirement age help young people to find work?
5. Taxes and tax collection: Does lowering taxes help improve tax collection?
6. Determinants of social mobility: Is the family socio-economic background a strong determinant of people's job opportunities in life?
7. School repetition: Is repeating a course an effective way to improve the level of learning?
8. Trade policy: Should the government punish or even forbid products from third countries that may pose a threat to the national industry?

A.5 Rubric to score debates

Please, assign 0 to 10 points per category to each participant. You may conclude that both debaters deserve 10 or 0 points in the same category, at your discretion. Remember: your mission is to assess the validity of the argument construction mainly, not so much the form that it acquires. Focus on examining the substance and content being debated. "Total score" should reflect the sum of the previous five scores and should be higher for the winner of the debate (there cannot be a tie).

Table A1: Rubric for debate scoring

Dimension	**Evaluation**	Points given to position **"In favor"**	**"Against"**
1.	Clarity, correctness and validity of the defended position. His thesis answers the question.		
2.	Credible and well-presented evidence.		
3.	Formal quality and rhetoric of the participant (not the content). Convincing use of language.		
4.	Ability to refute the rival's position		
5.	Superiority of the arguments to those of the rival.		
	Total score:		

A.6 Additional figures

Figure A1: Timeline of the experiment

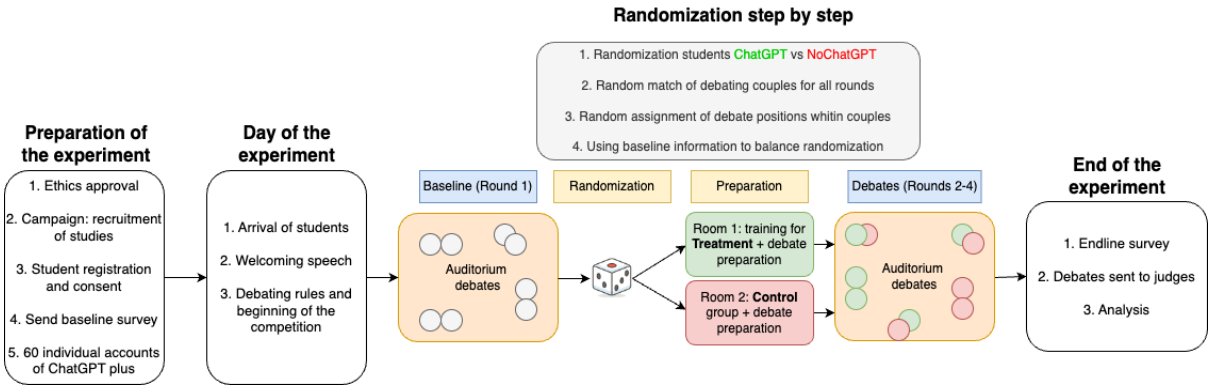
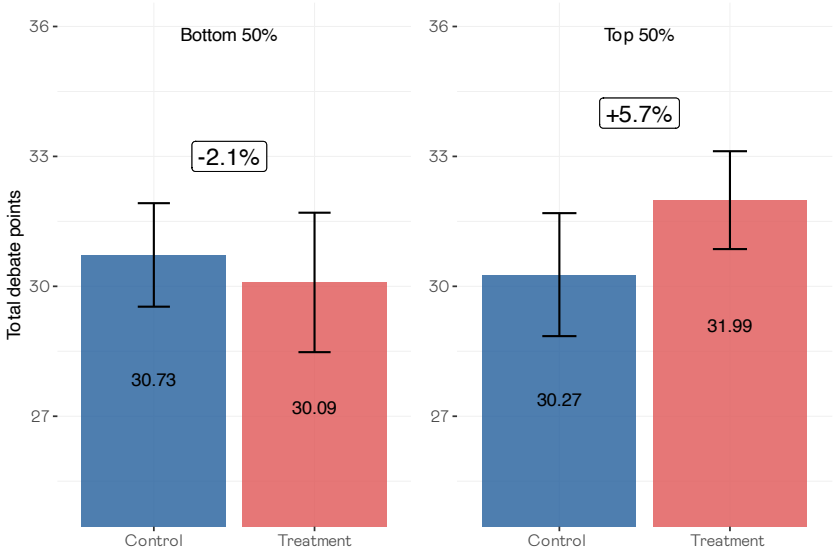
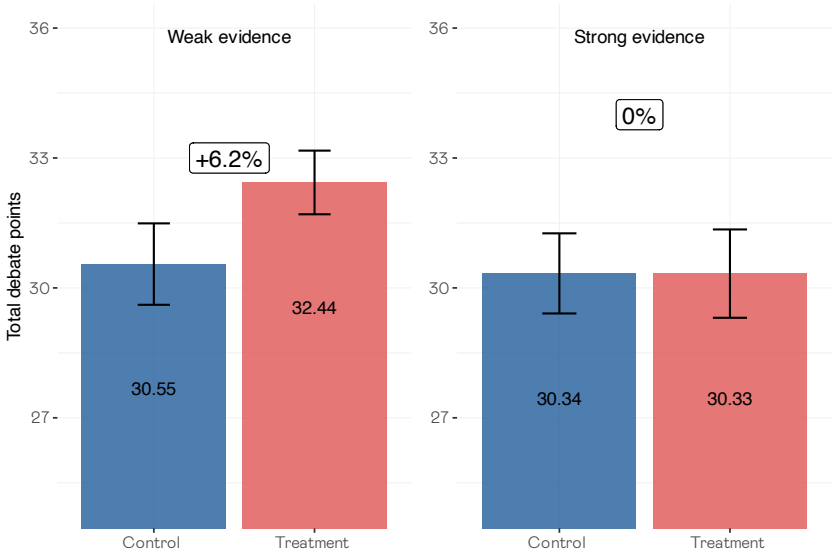


Figure A2: Heterogeneity by baseline debate performance



Notes: This figure shows the average debate points of control and treatment individuals for those scoring in the bottom and those in the top 50% of baseline debate points, conditional on control variables. The spikes represent 90% confidence intervals (predictive margins of the treatment indicator ($Treat$) in equation 2 by baseline performance), and above the bars I show the percent difference in the outcome between treatment and control group.

Figure A3: Treatment effects of ChatGPT by debate position defended



Notes: This figure shows the average debate points of control and treatment individuals for those defending the bad and those defending the evidence-based policy, conditional on baseline performance and control variables with 90% confidence intervals (predictive margins of the treatment indicator ($Treat$) in equation **??** by debate position), and the percent difference in the outcome between treatment and control group.

A.7 Additional tables

Table A2: Effect of ChatGPT by baseline debate performance

	Debate points		Win	
	(1)	(2)	(3)	(4)
Treat	-0.966	-0.636	-0.027	0.001
	(0.762)	(0.803)	(0.073)	(0.078)
Top 50%	-1.055	-0.455	0.088	0.147
	(1.051)	(1.116)	(0.108)	(0.108)
Treat x Top 50%	2.865***	2.356*	0.103	0.053
	(1.093)	(1.256)	(0.109)	(0.117)
Constant	28.317***	17.499***	0.043	-0.383
	(2.281)	(5.415)	(0.230)	(0.571)
Mean dep. var.	28.67	28.67	28.67	28.67
SD dep. var.	4.53	4.53	4.53	4.53
R^2	0.15	0.23	0.06	0.12
Obs.	364	364	364	364
Baseline score	Yes	Yes	Yes	Yes
Controls	No	Yes	No	Yes

Significance levels are indicated by * $<$ .1, ** $<$ .05, *** $<$ .01. Columns 1 and 2 show results from regressions of equation 2, where the outcome variable is the total debating points of individual i in debate round r (calculated as the average across different judges evaluating the same debate of individual). In Columns 3 and 4, the outcome is a dummy variable equal to one if the individual won the debate. Standard errors are clustered at the individual level because each individual is observed between 2 and 3 times, depending on the number of debates they completed. The total number of individuals included in each regression is 141 out of 142 randomized individuals. One individual did not complete the end-line survey.

Table A3: Effect of ChatGPT on use of time and tools

	Time sufficient (0-100)		Useful tools (0-100)	
	(1)	(2)	(3)	(4)
	Low	High	Low	High
Treat	5.171	31.040**	35.839***	24.790
	(4.778)	(11.195)	(5.786)	(16.196)
Constant	52.877***	29.703***	47.201***	62.774***
	(5.106)	(9.193)	(6.468)	(10.362)
Mean dep. var.	64.25	38.75	46.60	56.55
SD dep. var.	26.78	29.24	33.47	34.84
R^2	0.11	0.40	0.32	0.13
Obs.	117	20	107	19

Significance levels are indicated by * < .1, ** < .05, *** < .01. The table shows results from regressions of a variable measuring self-reported valuations of whether time given to prepare debates was sufficient (0-100, with 0 being "not at all" and 100 being "plenty of time") and whether the tools given to prepare the debates were useful (0-100, with 0 being "not useful at all" and 100 "very useful") on a treatment dummy and debate day fixed effects, separately for low ability (non merit scholarship recipients) and high ability (merit scholarship recipients) students. The total number of individuals included in each regression differs between the columns because not all individuals answered all the questions at endline.

Table A4: Effect of ChatGPT on total points by debating position

	(1)	(2)
	Weak evidence	Strong evidence
Treat	-0.095	1.671***
	(0.744)	(0.629)
Constant	32.666***	32.675***
	(0.873)	(0.868)
Mean dep. var.	29.37	28.12
SD dep. var.	4.53	4.44
R^2	0.10	0.14
Obs.	183	183

Significance levels are indicated by * < .1, ** < .05, *** < .01. Columns 1 and 2 show results from regressions of equation **??**, where the outcome variable is the total debating points of individual i in debate round r (calculated as the average across different judges evaluating the same debate of individual). Column 3 shows the differences-in-difference specification, and column 4 shows the specification using individual fixed effects.

Table A5: Effect of ChatGPT on debate performance by own and rival's treatment status

	(1) Total points	(2) Winning debate
Treat	0.605	0.105
	(0.664)	(0.072)
Rival treated	0.026	-0.026
	(0.642)	(0.078)
Treat x Rival treated	0.140	-0.125
	(0.961)	(0.110)
Constant	13.908***	-0.851*
	(4.943)	(0.510)
Mean dep. var.	28.67	0.48
SD dep. var.	4.53	0.50
R^2	0.21	0.11
Obs.	364	364

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. The table shows results of regressing outcomes on treatment status and an interaction between own treatment status and that of ones rival. Both specifications include baseline debate points and the full set of controls as described in the notes to Table 2.

Table A6: Summary statistics by scholarship status

	(1) Scholarship	(2) No scholarship	(3) Difference Difference (1)-(2)	(4) p-value p-value Col. (3)
Age	20.00	20.38	-0.38	0.30
Female	0.57	0.45	0.13	0.29
Father holds Master or higher degree	0.24	0.47	-0.23	0.05
Economics background	0.24	0.54	-0.30	0.01
Scholarship	1.00	0.00	1.00	.
Has prior debating experience	0.52	0.47	0.05	0.66
Has won a debate prize before	0.24	0.21	0.03	0.75
Knows ChatGPT	0.38	0.40	-0.02	0.84
Has used ChatGPT before	0.38	0.40	-0.02	0.89
Baseline debate points (0-50)	29.38	29.42	-0.03	0.98
Enjoyment (0-100) of speking in public	79.79	65.95	13.84	0.04
Feels comfortable in debating language	0.95	0.52	0.43	0.00
Political position (1=left, 10=right)	4.57	6.83	-2.26	0.00
Polarised (0-100)	57.20	53.21	3.99	0.40
N	21	121		

Notes: The table shows summary statistics for the scholarship and non-scholarship recipients for the sample of students who registered to participate in the debating competitions. Column 1 reports the mean among those with a scholarship and Column 2 reports the mean among those with no scholarship. Column 3 reports the difference in the mean across the two groups, and Column 4 reports the p-value of a t-test of the equality in means across the two groups.

Table A7: Summary statistics by baseline performance

	(1) Top 50%	(2) Bottom 50%	(3) Difference (1)-(2)	(4) p-value Col. (3)
Age	20.19	20.45	-0.26	0.32
Female	0.44	0.49	-0.04	0.60
Father holds Master or higher degree	0.50	0.38	0.12	0.14
Economics background	0.50	0.50	0.00	1.00
Scholarship	0.13	0.15	-0.02	0.73
Has prior debating experience	0.35	0.60	-0.24	0.00
Has won a debate prize before	0.32	0.11	0.21	0.00
Knows ChatGPT	0.47	0.32	0.15	0.07
Has used ChatGPT before	0.49	0.32	0.17	0.05
Baseline debate points (0-50)	33.74	25.33	8.40	0.00
Enjoyment (0-100) of speking in public	71.80	64.52	7.29	0.13
Feels comfortable in debating language	0.66	0.50	0.16	0.05
Political position (1=left, 10=right)	6.32	6.65	-0.32	0.38
Polarised (0-100)	55.97	51.97	4.00	0.24
N	68	72		

Notes: The table shows summary statistics for the bottom and top 50% performers at baseline for the sample of students who registered to participate in the debating competitions. Column 1 reports the mean in the top 50% and Column 2 reports the mean in the bottom 50%. Column 3 reports the difference in the mean across the two groups, and Column 4 reports the p-value of a t-test of the equality in means across the two groups.