Three experiments on technology, inequality and ideology

A thesis submitted to
London School of Economics

in fulfilment for the award of the degree of

Doctor of Philosophy
in Social Policy

By
Antonio Roldán-Monés

Department of Social Policy, London School of Economics
2nd Floor, Old Building
Houghton Street, London WC2A 2AE, UK

September 2024

Contents

Declaration

I certify that the thesis I have presented for examination for the MPhil/PhD degree of the London School of Economics and Political Science is solely my own work other than where I have clearly indicated that it is the work of others (in which case the extent of any work carried out jointly by me and any other person is clearly identified in it).

I declare that my thesis consists of 76032 words.

Statement of co-authored work

I confirm that Chapter 2 was co-authored with Claudia Hupkau and Lucas Gortázar. I am the corresponding author of the paper.

I confirm that Chapter 3 was co-authored with Jorge García-Hombrados, Marcel Jansen, Berkay Ozcan, Angel Martínez and Pedro Rey Biel. The authors are cited alphabetically.

Statement of inclusion of previous work

I can confirm that chapter 2 was the result of an experiment launched in Spring 2021, when I had not yet been accepted to the PhD program at LSE. However, the bulk of the work, all the data analysis and the writing were done during my time at the LSE.

I can confirm that the original idea of Chapter 3 was presented for funding to La Caixa Social Research Fund before starting my doctorate, although the implementation of the experiment, the data analysis and the writing was done after I had started my doctorate.

Statement of use of third party for editorial help

I confirm that my thesis was not edited by third parties.

Abstract

This dissertation consists of three Randomized Controlled Trials analyzing the relationship between technology and inequality as well as the ideological constraints to evidence-based policy implementation.

In the first chapter I investigate the impact of a Generative Artificial Intelligence system (GenAI) on productivity and inequality, examining a novel set of skills in a unique setting: a university debating competition. Early experimental research in the area highlights two key findings: GenAI significantly boosts productivity in a variety of tasks, and it disproportionately benefits lower-performing workers, potentially reducing work inequalities. However, most studies in this literature are focused on written tasks. Contrary to previous studies, my findings reveal that GenAI's effects vary considerably when higher-order skills like persuasion or social perceptiveness are needed. My results show that GenAI exhibits stronger complementarities with top performers than with low performers at debating, suggesting that GenAI could reinforce inequalities depending on the context and the skills required.

The second chapter explores an innovative pedagogic design to address educational inequalities through online learning in a post-pandemic context. While previous research has shown the effectiveness of in-person tutoring, evidence for 100% online environments was very limited. Our study demonstrates that a cost-effective 2-students-to-1-mentor online tutoring model significantly improves maths learning outcomes for children in vulnerable settings. The intervention significantly increased standardised test scores (+0.26 SD) and end-of-year maths grades (+0.49 SD), while reducing the probability of repeating the school year. The intervention also raised aspirations, as well as self-reported effort at school. The two-on-one design allows us to significantly reduce costs and improve scalability, while showing similar results as one-on-one tutoring programs.

The third chapter studies political constraints to scientific communication. Utilising a nationwide experiment, the paper examines the role of political alignment between knowledge brokers (such as think tanks and media outlets) and real policymakers in the adoption of an evidence-based policy. Results show that policy adoption increases by over 65% when the informing institution aligns

ideologically with policymakers, while opposite ideologies do not affect adoption. Both think tanks and newspapers are equally effective when aligned ideologically. Finally, a proposed three-stage framework reveals that ideological alignment influences belief updating but not the decision to get exposed to new evidence among politicians.

Introduction

I am part of a generation of economists that was trained in an era of extraordinary optimism and trust in human progress. Economists of the so-called *Great Moderation* period thought they had found a formula for economic stability (Galí and Gambetti, 2009); the implementation of effective institutions, such as central bank independence, had led to a drastic reduction in the volatility of business cycles (Bernanke, 2004). In economic development, the recipes of the Washington Consensus – fiscal responsibility, orthodox monetary policy and structural reforms – seemed infallible to guarantee economic success (Williamson, 1990). In politics, the success of liberal democratic regimes, expressed through the improvements in the standards of living and the extension of individual freedoms in advanced democracies seemed an unavoidable path for all countries (Fukuyama, 1989).

At the core of that view, there was strong a faith in the possibilities of technology. Information and Communication Technologies (ICTs) would bring sustained productivity increases by allowing further economic integration and specialisation. Internet would lead to the democratisation of information, eliminating economic hierarchies and empowering individuals independently of their geography or social context. This optimistic view of the possibilities of technology in the early 2000s was well exemplified by the words of Marc Andreessen, Netscape co-founder, in 2005: "Today, the most profound thing to me is the fact that a 14-year-old in Romania or Bangalore or the [former] Soviet Union or Vietnam has all the information, all the tools, all the software easily available to apply knowledge however they want" (Autor, 2024). By allowing for information and ideas to travel instantly without constraints, ICTs would not only expand economic growth and opportunities, but also they would make it impossible for autocracies to thrive. Ronald Reagan concluded, at a conference in the UK, in 1989: "the Goliath of totalitarianism will be brought down by the David of the microchip" (New York Times, 1989).

The wave of optimism also extended to policy studies. In its opening chapter, the 2008 edition of the Oxford Handbook of Public Policy – published just ahead of the global financial crisis - referred to the "high modernist" approach to public policy. The "high modernist" view saw policy

problems in the same way as engineers see mechanical problems: as a matter of technical expertise (Goodin et al., 2008). Consequently, solutions to most social and economic challenges could be found through the systematic application of technical ability and effective policy designs. The increasing availability of data as well as improvements in econometric techniques – which characterised the so-called empirical revolution in economics - allowed for the continued expansion of the evidence-based policy agenda (Banerjee et al., 2016).

Such "technocratic hubris, married to a sense of mission to make a better world" (Goodin et al., 2008) was what brought me into politics in the first place. When I was elected Member of the Spanish Parliament of Ciudadanos, a centrist platform, in 2015, I had the conviction that the methods and evidence-based policies I had learned at some of the most preeminent economics departments and policy schools were going to be instrumental for our political success. It was just a matter of time, I thought, that voters would realise how good - socially sensible, rigorous, and based on the best available evidence - our policy programs were and, as a result, political support for our platform would inevitably grow. Even *The Economist*, the British weekly magazine, called us "The OECD party" (The Economist, 2018). What could go wrong?

My predictions seemed correct for a while. My party - of which I was the head of policy and economics spokesperson - became one of the most successful political start-ups in Europe: in less than three years the party grew as a major political force in Spain, gathering more than four million votes, 59 out of 350 seats at the national parliament and widespread recognition. But it did not last long. Quickly, Ciudadanos became an irrelevant political force. It is not the purpose of this work to analyze the reasons of Ciudadanos' political failure. But the fall of my party was not unrelated to the deep crisis of the liberal economic order that started unfolding.

Contrary to the optimistic predictions of the previous decade, economic instability and political backlash became the new global norm. The rise of populist and extremist parties across the globe was accompanied by a revival of identity politics and rising political polarisation (Fukuyama, 2018). This translated into increasing distrust in experts and science (van der Linden, 2023).

Because of its impact on work and democracy, accelerated technological change came to be seen as a central element explaining the rising anxiety (Brynjolfsson and McAfee, 2014). The whole intellectual construction of my formative years, based on an almost blind trust in the possibilities of technology and human rationality was falling apart. When I left politics, I suspected that part of the problem of the old intellectual paradigm was that politics had largely been left out of the picture. Economists and political scientists had underestimated the political implications of the deep economic shocks and technological disruptions we were going through. Moreover, after my experience as a policy-maker, I became convinced that individuals, when dealing with policy, had

very little to do with the rational, calculating and depoliticised view of the human brain I had learned in my studies. To achieve effective policy change, economists and policy-makers needed a more realistic understanding of human behaviour in polical environments.

Trying to better understand the relationship between technology and inequality, as well as the political and behavioral limitations to implement effective evidence-based policies to reduce those inequalities was what brought me (back) to study a PhD in social policy.

In what follows, I explore the overarching ideas of my work. The three chapters of the thesis and my contribution to each of them are summarised in detail in the following section.

On the relationship between technology, skills and inequality

In his 1930 essay, "Economic Possibilities for our Grandchildren", John Maynard Keynes predicted that technological advancements would render work largely obsolete. Keynes imagined a future where technology driven productivity gains would be so large that the fundamental problem of scarcity would be resolved.

History proved him wrong. Over the last one hundred years, successive technological innovations reduced the relative cost of capital in relation to labor and automation displaced many jobs, particularly in agriculture and manufacturing. However, despite the recurring fears of generalised unemployment, technological progress also created new industries and new economic opportunities, leading, on aggregate, to higher employment rates in developed economies (Autor, 2015).

Depending on how different technological advancements affected existing work, the return to certain skills changed, influencing wage structures, and the overall distribution of income in the economy.

The effects of technology on work are typically determined by whether technologies complement or substitute human skills. Innovations are complementary to skills when they enhance (or augment) workers' productivity. Conversely, job automation or substitution takes place when two phenomena occur at once. First, when a task formerly performed by a human can be programmed to be performed by an available technology, such as a software or a robot. Second, when the relative cost of using that technology becomes clearly inferior to that of a worker (Autor et al., 2003a). Since the first industrial revolution, technological development has continuously expanded the perimeter of tasks that can be automated. In recent years, improvements in computational speed, data storage and data retrieval, among other innovations, have dramatically reduced the cost of doing predictive tasks (Agrawal et al., 2018).

For a comprehensive understanding of the impact of technology on work, it is essential to examine the relationship between tasks and skills. A task refers to a specific work activity that results in

the production of goods and services. A skill is a worker's set of capabilities for executing various tasks. Workers utilize their skills to complete tasks in exchange for wages and adjust the tasks they perform in response to changes in technology. In a seminal work, Autor et al. (2003b) classified tasks into three broad groups: analytical or cognitive, interpersonal, and manual or physical. Within each group, they differentiated between routine or mechanical tasks and those requiring higher skills. They then quantified the task content of various occupations.

By reducing the time needed to do certain important tasks, such as accessing and structuring new information, the so-called first wave of automation in the 1970s - driven by advances in computerisation – allowed workers with higher levels of education to focus on more complex, value-added activities, such as interpreting and applying that information. This contributed to liberate time for decision-making, augmenting "the accuracy, productivity and thoroughness of professional judgement, thus magnifying its value" (Autor, 2024). This improved the productivity of high-skill workers relative to low-skill workers, leading to a rising divide driven by a *skill-biased technological change* (Card and DiNardo, 2002).

In a pioneering work in this literature, Krueger (1993) used survey data of the 1980s to study the effect of computers on wages. He estimated that workers using computers on their job earned 10 to 15 percent higher wages. Since those jobs were typically performed by workers with relatively high levels of education, Krueger predicted that the expansion in computer use also explained a very substantial part of the return to education. In a posterior work, David Autor and co-authors studied the effects of the introduction of computers on the demand of qualified employment (Autor et al., 1998). Between 1940 and 1996, the introduction of computers increased the relative demand of college workers because of their stronger complementarities with computers. The authors conclude that the rapid decline in technological capital costs increased the relative demand for skilled workers - who had a comparative advantage in performing non-routine tasks – and pushed their demand and wages upwards, thus leading to rising inequalities.

This was accompanied by a process of job polarisation (Autor et al., 2006). Many jobs in the middle of the income distribution became obsolete. These jobs included administrative, clerical or production jobs in large factories – such as automobiles - that were mostly based on routine tasks. New and cheaper algorithms and robots could substitute humans in substantial parts of the production process, especially at those more simple, repetitive or predictable tasks. Autor et al. (2003b) analyzed U.S. data over several decades and confirmed that sectors initially intensive in routine tasks experienced greater automation and employment decline over time, while employment in occupations with fewer routine tasks grew. Likewise, Acemoglu and Restrepo (2017a), exploiting data from the US between 1990 and 2007, show that those regions in the US more heavily exposed to

robotization, especially in urban industrial areas, experienced negative effects on overall employment and wages. Cortes et al. (2017) studied individual job transitions from routine jobs in the US from 1977 to 2005. They found that less qualified workers on routine jobs tended to move to manual occupations requiring lower cognitive abilities. Conversely, more qualified routine workers migrated towards occupations demanding more intense cognitive skills. They also found that workers that remained in routine occupations saw their relative wages grow at a much lower rate than those that moved to any other non-routine occupation.

Job polarisation also meant that not all low-skilled jobs were affected in the same way. Many low-skilled occupations abundant in the services sector, such as waiters or hairdressers, requiring social or interpersonal skills, typically performed in open-ended environments and involving a variety of non-repetitive actions, remained largely shielded from automation. The reason was simple: those jobs were based on tasks that could not be reduced to a series of simple commands to be performed by a robot. Autor et al. (2013) show that between 1980 and 2005, occupations in the second skill quartile of the distribution fell as a share of employment, while those in the lowest skill quartile expanded sharply. Meanwhile, the demand for high-skill and highly paid jobs requiring "tacit knowledge", or high cognitive abilities, such managers, rapidly increased. This led to an (irregular) U-shaped effect of technology on employment opportunities.

For more than 30 years, skill driven inequalities continued to rise in many advanced economies (Acemoglu and Restrepo, 2017b). According to Goldin and Katz (2007) for the case of the US, from 1973 to 2005, family incomes for the lowest 20 percent of earners remained nearly unchanged, while incomes for the top 5 percent increased at more than three times the rate of those in the middle income quintile.

"Tacit knowledge" is an important concept in the literature - especially useful in the study of the latest generation of Artificial Intelligence (AI) technologies. It is usually contrasted with explicit knowledge and associated with a series of inherently human tasks intuitively done by humans but that seemed, for many years, shielded from automation (Polanyi, 1966). This includes essential tasks typically performed by high earning professional workers of all kinds, such as summarising or analysing text, writing a persuasive email, or giving a speech. These tasks involve higher cognitive and social abilities that require a complex understanding of the environment in which the task is being performed. Now, new AI and Machine Learning (ML) technologies are starting to trespass the "tacit knowledge" frontier.

The rise of Generative AI

The first wave of automation drove changes in large industries, where significant parts of the production process could be reduced to a series of programmable tasks. The most recent wave of automation, characterised by AI and ML technologies, operates fundamentally different from traditional computer programs. ML algorithms do not rely on explicit instructions to function. Instead, they infer instructions from examples. For instance, given a dataset of professional emails, ML systems can learn to identify specific patterns of words that are likely to come together and create new "original" texts without having explicit codification instructions.

According to Brynjolfsson et al. (2023), this capacity highlights a critical advantage of ML systems vis-à-vis previous systems: their ability to learn and perform tasks requiring tacit knowledge, previously only attainable through human experience. The transformative ability of acquiring knowledge through observation rather than rules, allows these technologies to improvise and make expert judgements, traditionally reserved to highly skilled professionals (Autor, 2024).

The beginning of 2023 is considered to be a "tipping point to Artificial Intelligence" history (Mollick, 2022). ChatGPT, a large language model (LLM) built to imitate human conversations – based on AI and ML technologies - reached 100 million monthly active users in January 2023, just two months after its launch, a faster growth than any other internet service, social network or consumer application in history (UBS, 2023). The potential applications of these new GenAI systems were so large and their development so fast that some scholars thought we were getting to an "AI's Jurassic Park moment" because of the risks posed to humanity (Marcus, 2022).

Researchers from various fields are exploring the specific impacts of these technologies in different domains, including the effects on work, productivity and inequality. A first consensus from the early economics literature is that because of their ability to create expert judgement and translate it into usable content, these technologies will affect a much broader set of jobs, including those in high-earning professions of all kinds (Felten et al., 2023). A McKinsey report predicts large displacement of jobs in sectors like finance, healthcare or legal services, where GenAI can handle tasks such as drafting documents, diagnosing medical conditions or analysing financial data (Chui et al., 2023). Eloundou et al. (2023) expect that "around 80% of the U.S. workforce could have at least 10% of their work tasks affected by the introduction of LLMs, while approximately 19% of workers may see at least 50% of their tasks impacted".

Despite these gloomy predictions, researchers are still far from having a coherent understanding as regards to when and how those technologies will complement or substitute different workers, and who will benefit from GenAI on what tasks. (Mollick, 2023). For instance, while the latest versions of those technologies are not particularly good at solving simple math problems, they can

be extraordinarily effective at creative writing, programming or writing legal expert memos when accompanied by humans. Likewise, when left on their own, they typically provide rather superficial knowledge, make mistakes, invent references and hallucinate (Bommasani et al., 2021).

A growing body of experimental evidence shows that, in a variety of complex writing tasks, GenAI substantially raises average work productivity. These tasks include, among others, mid-level professional writing tasks, customer support and variety of consultancy tasks (Doshi and Hauser, 2023; Brynjolfsson et al., 2023; Choi and Schwarcz, 2023). A second early finding in the literature is that productivity improvements in most of those areas are not only large but also tend to be unevenly distributed: GenAI seems to work as a productivity compressor, reducing performance inequality by helping more those at the bottom of the skill distribution (Dell'Acqua et al., 2023).

Given the unprecedented expansion and potential impact of those technologies, understanding how GenAI will affect work inequalities is among the top concerns for governments, researchers and business leaders worldwide (Alliance, 2024). My first paper provides experimental evidence on the effects of GenAI on productivity and inequality, exploiting a novel setting – a university debating competition - that allows me to test the impact of GenAI on a set of "higher order skills", typically hard to test in controlled environments. The results I find go against most of the findings in the recent GenAI literature: best performers and those on a merit scholarship benefit significantly more from the interaction with GenAI than low-performing students. I hypothesise that in environments requiring unpredictable verbal interactions – where answers cannot be copy-pasted - high skill individuals are likely to benefit more, not less, from GenAI, thus increasing work inequalities.

My work complements the work of Deming (2017) and others on the importance of social and "higher-order skills" in modern labor-markets. Deming argues that social skills —which include coordination, negotiation, persuasion, social perception, and empathy— help lower transaction and coordination costs in the workplace. The demand for workers who possess strong social skills, particularly in combination with analytical cognitive abilities, has risen over the last decades (Weinberger, 2014). Those skills are central for high-earning workers of all kinds. Debating is a good proxy for those skills, as it requires high cognitive and social abilities in a highly unpredictable environment.

It will be the task of future research to provide a more comprehensive theory as regards to how GenAI relates to different human abilities. But if these findings replicate, they could have relevant implications in the study of technology and inequality.

Adapting to a permanent revolution

In their most recent book, Acemoglu and Johnson (2023) argue that the negative effects of technology on inequality are not unavoidable: they are a function of the institutional arrangements that are in

place in the economy. Successive waves of automation have continuously expanded the tasks that can be performed by machines, demanding new skills and competencies to adapt to those changes. That "race" between the need for skilled labor and the education system's ability to provide it is the focus of a classic book by Claudia Goldin and Lawrence Katz (Goldin and Katz, 2008). The authors argue that during periods when education systems successfully expanded and adapted to new technological requirements, wage inequality tended to decrease. Conversely, when technological advancements outpaced educational attainment, wage inequality increased.

The GenAI revolution will lead to the automation of some tasks, but it will also augment human capabilities, creating new opportunities for those who can adapt. Adapting to GenAI will, in turn, require a profound rethinking of educational systems and large efforts in up-skilling and re-skilling, to help workers develop complementary skills that AI cannot easily replicate, such as complex problem-solving, critical thinking, creativity or socio-emotional skills (Autor, 2024).

Exploring how to use online tutoring and technology to boost key skills of disadvantaged students is the focus of my second chapter. A vast amount of research has shown that socioeconomic status is a strong predictor of student achievement (see, for instance, Blanden et al. (2023), for a recent review). The most recent PISA results show that students from the wealthiest 25% families outperform their peers from the poorest 25% families by more than a year of schooling in many countries (OECD, 2023). In Spain, for example, students from a family in the lowest income quintile are four times more likely to repeat the school year than those from a family in the top quintile of the population (Cabrales and Roldán, 2020). Those educational inequalities translate into poor labor-market outcomes later in life. Repeating the school year often leads to early drop out, which, in turn, is strongly correlated with poor labor-market outcomes, low earnings and high unemployment (Jackson and Holzman, 2020).

These educational gaps are also deeply affected by technological divides. One such example is the expansion of shadow education, or private supplementary tutoring, facilitated by the boom of highly effective online learning technologies in recent years. Shadow education has grown exponentially driven by a variety of cultural, demographic and economic factors, becoming a global phenomenon. Since access to private tutoring is restricted to wealthier families, shadow education is having very significant implications for educational inequalities (Gortázar and Moreno, 2022).

Another example is the educational gap caused by the COVID-19 pandemic. Generalised social distancing restrictions led to the closure of schools in over 150 countries - affecting 1.6 billion learners globally (Azevedo et al., 2022). In person teaching had to be suspended for several months and entire educational systems moved to remote learning. However, students from lower-income households often lacked access to the necessary platforms and internet connectivity to participate effectively

in online education. According to UNICEF, at least one-third of the world's schoolchildren were unable to access remote learning during school closures (Mascheroni et al., 2021). This digital divide disproportionately affected students from disadvantaged backgrounds, rural areas, and low-income countries, leading to rising educational inequalities (Betthäuser et al., 2023).

Governments announced several initiatives to respond to learning losses. For instance, the UK launched a National Tutoring Program in September 2022, initially funded with £1 billion, offering both face-to-face and online tuition. While strong evidence existed on the effectiveness of in person one-on-one and small group tutoring programs, there was very little experimental evidence on the effectiveness of 100% online tutoring programs (Nickow et al., 2020).

Online tutoring provides some clear advantages vis-à-vis in person tutoring. First, it has the benefits of accessing a larger pool of tutors and eliminates commuting expenses for both students and tutors. Furthermore, remote after-school tutoring, as opposed to in-person tutoring during school hours, minimises logistical difficulties for schools and teachers regarding the coordination of time and space for tutoring sessions (Kraft et al., 2022).

But online learning has several limitations as well. David Deming and coauthors show that effectiveness of online education is limited because of high levels of student attrition, explained by lower levels of engagement and interaction compared to traditional settings (Deming, 2017). Furthermore, experimental work has identified problems of self-discipline and motivation affecting more negatively students with lower prior academic achievement (Escueta et al., 2020).

Given the mobility restrictions caused by the pandemic, we designed an innovative 100% online program to respond to rising learning losses among vulnerable children. The experimental program was targeted at students from poor neighbourhoods in Madrid and Barcelona and was designed in collaboration with a team of pedagogic experts of Empieza por Educar, the Spanish branch of Teach for All, an NGO specialized in training teachers for difficult learning environments.

For us it was critical to design an effective pedagogic plan to overcome the mentioned learning limitations in online environments. The program had a strong socio-emotional focus and followed the "No Excuses" approach, which has been shown to be effective in other settings (Angrist et al., 2013). Moreover, we introduced an innovative two-students-to-one-tutor design with students from the same class. We hypothesised that the student dynamics would help to reduce attrition. We also established a semi-automated system of alerts when students were not showing up to the tutoring sessions and invested resources in monitoring and personalised contact with families.

The results of the program, which are summarised in the following section, were very positive. A very short online intervention focused on maths but also on socio-emotional skills improved student outcomes, drastically reduced school repetition and boosted students' aspirations. At the same time,

we managed to get very good results in terms of attendance rates – the median number of completed sessions in the treatment group was 20 out of a target of 24. This indicates that our two-on-one design might have helped to mitigate some of the shortcomings found in the literature in online education.

Given the global rise of educational technologies, innovative online applications, personalized learning algorithms, asynchronous interactions with tutors through chats or other AI bots designed to support tutors and students, further research will be needed to understand how to offer fulfilling, inclusive and effective educational experiences. Some of these rising technologies already offer a big potential. Access to EdTech platforms such as MOOCs, Khan Academy or Coursera is helping to democratize education, broadening the access to good quality content previously inaccessible, regardless of geographical location. The development of algorithms used to tailor educational content to students could help to significantly improve educational outcomes. Innovations involving gamification might also help to improve engagement and retention in online environments. Furthermore, the use of technology in class and blended learning approaches might contribute to improve teachers' efficiency (Molina et al., 2024).

In terms of broader policy improvements, the use of data analytics in education will allow educators and policymakers to track progress in real-time, identifying trends, and making data-driven decisions to improve teaching strategies or education policies with potentially very relevant implications. A key empirical question will be to test what combination of human work and technology works best to design public policies that help improve educational outcomes, while bridging digital divides.

How to successfully implement evidence-based policies is the concern of my third chapter.

From policy design to policy implementation

A fundamental difference between policy studies and research on political science or economics is that policymaking is mostly a matter of action, and action requires persuasion (Goodin et al., 2008). To change reality and contribute to the betterment of life, policymakers need to carry people with them. This implies that, to be relevant, policy researchers not only need to be able to use the relevant tools to design effective policies, they also need to have a deep understanding of how human cognition works. After all, when leaving the immaculate world of theoretical conjectures and econometrics, policy programs become constructed narratives that navigate in the real world of power, interests, and human psychology.

One way of thinking about the policy-making process is to divide it into three parts: context evaluation, policy design and policy implementation. My two first papers focus on the two first parts

of the policy process. My last paper focuses on the third: policy implementation.

Although first-best policy designs tend to look good on paper, they often face a variety of limitations that constrain their implementation. Those constraints can be of different forms. A first group of challenges when implementing policies is related to political economy considerations, including interests and power-relations involved in a reform process (Tommasi and Velasco, 1996; Rodrik, 1996; Khemani, 2017) or the political equilibrium that results from a policy intervention (Acemoglu and Robinson, 2013). Policy adoption can also be limited by other types of constraints, such as administrative or bureaucratic (Aucoin, 1990). In many cases, even when the right information gets to policymakers and other political constraints are surmounted, a policy might not be implemented because it is simply too costly to implement (Tommasi and Velasco, 1996). The focus of my last chapter is on policy implementation and the role that political ideology plays in the policy communication process for the adoption of evidence-based policies.

Related literature has been in expansion over recent years to try to offer responses to a variety of pressing societal puzzles. First, despite overwhelming evidence supporting scientific facts, such as climate change or the benefits of vaccines, large segments of the population seem to be impervious to those facts (Pinker, 2021). Second, in the information age, despite having more access than ever to reliable sources, fake news and misinformation represent a growing threat to democratic systems (van der Linden, 2023). Third, rising political polarization seems to be intensifying everywhere, reducing the space to find common ground for policy advancement (Klein, 2020).

To try to find answers to some of these questions, some researchers are proposing a departure from the "bounded rationality" paradigm – which identified some systematic deviations of the *homo economicus* rationality assumptions – towards a somewhat "motivated" view of the human mind (van der Linden, 2023). According to this "motivated reasoning" view, biases are not seen as individual deviations or bugs of the human rational brain, but rather a feature of it. In economics, Bénabou and Tirole (2016) summarize this shift in the following way: "the pendulum has started to swing again toward some form of adaptiveness, or at least implicit purposefulness, in human cognition." This new perspective implies that beliefs often fulfill important psychological and functional needs of the individual.

Rather than having a dispassionate view of evidence, individuals search for answers that make them feel good with themselves, that confirm their pre-conceived view of the world or that help them be accepted and liked by their peers (Pinker, 2021). Such underlying (political) motivation in many circumstances "leads us to (willingly) distort our perception of the evidence to fit our worldview" (van der Linden et al., 2017). Humans can be seen thus as "rationalising" (rather than rational) agents, using their brain to justify some pre-established priors (Haidt, 2012).

Political ideology, like religion, is an essential feature of our social identity. Almost by definition, social identities create in-group and out-group dynamics that are deeply entrenched in the human brain (Tversky and Kahneman, 1974). When dealing with political issues, researchers have found that individuals' predefined partisan or ideological beliefs strongly condition the way they approach new information (Kahan, 2016). This leads to patterns of information acquisition that deviate from standard Bayesian models of learning (Alonso and Padró i Miquel, 2023). In fact, in politics, humans tend to show "endogenous directionality" as they search for comforting and confirming beliefs to avoid cognitive dissonances (Bénabou and Tirole, 2016). More specially, ideological biases have been shown to condition the way individuals analyse policies (Dan M. Kahan and Braman, 2011). We tend to scrutinise information more when it comes from distrusted sources and we also uncritically accept evidence coming from reliable (or partisan) sources (Kahan, 2013). These systematic ideological deviations are also present in politicians and policymakers of different kinds (Banuri et al., 2019).

There are different ways in which motivated reasoning affects policy communication. A growing literature has been dedicated to analysing "messenger effects" (Fielding et al., 2020; Maclean et al., 2019). Depending on the case, the "messenger" of the policy – a government institution, a think tank, a scientific body or a media outlet - might be considered reliable or politically biased and this might influence beliefs about the evidence provided. For instance, in the literature on science communication in the context of climate change, a good amount of research has shown that the perceived credibility of the messenger is very important for people to trust scientific information (see, for instance, Diamond and Zhou (2022)).

In a recent paper by Nobel Prize winner Abhijit Banergee and co-authors, they randomise the way in which relevant scientific information is communicated by altering the person communicating the message and the features of the message itself. They send SMS messages to 25 million people in West Bengal using Banerjee (a widely known and respected economist from the region) as the treatment messenger of the scientific information on social distancing and hygiene measures related to Covid-19. They find that take up of positive health measures increases significantly among those people that received the video against a control group using conventional government resources (Banerjee et al., 2020).

Similarly, in the psychology literature, research has demonstrated the importance of reference groups, such as political parties, to generate trust. Cohen (2003), for instance, has shown that people's attitudes toward a social policy depend less on the evidence provided than on the stated position of one's political party. Also, research in economics and political science has demonstrated that the perceived political bias of a media outlet changes the way we approach information (Gentzkow and Shapiro, 2006).

A related literature analyses the importance of "the message". Some experiments have tested how different ways of presenting scientific information – the format - might be more persuasive, such as the use of Policy Briefs or systematic reviews (Masset et al., 2013). Second, the evidence itself might simply not be considered good enough, contradictory with other existing evidence, or considered irrelevant for a determined context. For instance, Nakajima (2021) find that policymakers have preferences for larger studies and studies conducted in similar contexts as their own jurisdiction. However, they do not differentiate between experimental and observational studies (Nakajima, 2021).

The unique design of the paper exploits a simple, non-partisan, cost-less policy to analyse the effects of the messenger and the format of the message on the implementation of an evidence-based policy in a nation-wide experiment with more than 5000 local politicians. We use real-world authoritative political institutions, such as think tanks and media outlets, to introduce ideological variation in the messenger. We also randomise the format in which the evidence is communicated. Results indicate that when the ideology of the institution disseminating the evidence and the local politician receiving the evidence are aligned, policy adoption increases by over 65%. However, when information comes from institutions with opposing ideologies, the effect is negligible. The format in which the information is presented (policy brief vs. newspaper article) showed no significant difference in effectiveness.

Given the importance of effective evidence-based policy dissemination for human progress, future research should continue exploring the different factors that might constrain scientific communication and influence policy implementation in much more depth. For instance, investigating the roles played by governmental agencies, different levels of government or independent scientific bodies could yield valuable insights.

Having been a politician myself, my goal with this PhD was not only to get a deeper understanding of the drivers of inequality and effective policy designs to fight those inequalities, but also to get a deeper understanding of how politics might constrain the adoption of those policies. In the following section, the three chapters of the thesis are summarised in detail.

Bibliography

Acemoglu, Daron and James A. Robinson, "Economics versus Politics: Pitfalls of Policy Advice," *Journal of Economic Perspectives*, May 2013, *27* (2), 173–92.

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

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

_ **and Simon Johnson**, *Power and Progress: Our Thousand-Year Struggle Over Technology and Prosperity*, New York: PublicAffairs, 2023.

Agrawal, Ajay, Joshua Gans, and Avi Goldfarb, *Prediction Machines: The Simple Economics of Artificial Intelligence*, Boston: Harvard Business Review Press, 2018.

Alliance, AI Governance, *Generative AI Governance: Shaping a Collective Global Future*, Geneva: World Economic Forum, 2024.

Alonso, Ricardo and Gerard Padró i Miquel, "Competitive Capture of Public Opinion," NBER Working Papers 31414, National Bureau of Economic Research, Inc 2023.

Angrist, Joshua D., Parag A. Pathak, and Christopher R. Walters, "Explaining Charter School Effectiveness," *American Economic Journal: Applied Economics*, October 2013, *5* (4), 1–27.

Aucoin, Peter, "Administrative reform in public management: paradigms, principles, paradoxes and pendulums," *Governance: an international journal of policy and administration*, 1990, *3* (2), 115–137.

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

Autor, David H., "Why Are There Still So Many Jobs? The History and Future of Workplace Automation," *Journal of Economic Perspectives*, September 2015, *29* (3), 3–30.

_ **, David Dorn, and Gordon H. Hanson**, "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review*, October 2013, *103* (6), 2121–68.

_ **, 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.

_ **,** _ **, and** _ , "The Skill Content of Recent Technological Change: An Empirical Exploration*," *The Quarterly Journal of Economics*, 11 2003, *118* (4), 1279–1333.

_ **, Lawrence F. Katz, and Alan B. Krueger**, "Computing Inequality: Have Computers Changed the Labor Market?," *The Quarterly Journal of Economics*, 1998, *113* (4), 1169–1213.

_ **,** _ **, and Melissa S. Kearney**, "The Polarization of the U.S. Labor Market," *American Economic Review*, 2006, *96* (2), 189–194.

Azevedo, Joao Pedro Wagner De, Maryam Akmal, Marie-Helene Cloutier, F. Halsey Rogers, and Yi Ning Wong, "Learning Losses during COVID-19 : Global Estimates of an Invisible and Unequal Crisis," Policy Research Working Paper Series 10218, The World Bank October 2022.

Banerjee, Abhijit, Marcella Alsan, Emily Breza, Arun G Chandrasekhar, Abhijit Chowdhury, Esther Duflo, Paul Goldsmith-Pinkham, and Benjamin A Olken, "Messages on COVID-19 Prevention in India Increased Symptoms Reporting and Adherence to Preventive Behaviors Among 25 Million Recipients with Similar Effects on Non-recipient Members of Their Communities," Working Paper 27496, National Bureau of Economic Research July 2020.

Banerjee, Abhijit Vinayak, Esther Duflo, and Michael Kremer, "The influence of randomized controlled trials on development economics research and on development policy," *The state of Economics, the state of the world*, 2016, pp. 482–488.

Banuri, Sheheryar, Stefan Dercon, and Varun Gauri, "Biased Policy Professionals," *World Bank Economic Review*, 2019, *33* (2), 310–327.

Bernanke, B. S., "The Great Moderation," 2004. Remarks at the meetings of the Eastern Economic Association, Washington, DC. Available at Federal Reserve.

Betthäuser, Bastian, Anders Bach-Mortensen, and Per Engzell, "A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemicDid students learn less during the COVID-19 pandemic? Reading and mathematics competencies before and after the first pandemic wave," *Nature Human Behaviour*, 2023, pp. 1–11.

Blanden, Jo, Matthias Doepke, and Jan Stuhler, "Educational inequality," in "Handbook of the Economics of Education," Vol. 6, Elsevier, 2023, pp. 405–497.

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

Brynjolfsson, Erik and Andrew McAfee, *The Second Machine Age: Work, Progress, and Prosperity in a Time of Brilliant Technologies*, New York: W. W. Norton & Company, 2014.

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

Bénabou, Roland and Jean Tirole, "Mindful Economics: The Production, Consumption, and Value of Beliefs," *Journal of Economic Perspectives*, September 2016, *30* (3), 141–64.

Cabrales, Antonio and Antonio Roldán, "Dos acuerdos educativos para la legislatura: una propuesta transversal," March 2020. Policy Brief #1, EsadeEcPol.

Card, David and John E. DiNardo, "Skill-Biased Technological Change and Rising Wage Inequality: Some Problems and Puzzles," *Journal of Labor Economics*, 2002, *20* (4), 733–783.

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.

Cohen, Geoffrey, "Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs," *Journal of personality and social psychology*, 12 2003, *85*, 808–22.

Cortes, Guido Matias, Nir Jaimovich, and Henry E. Siu, "Disappearing routine jobs: Who, how, and why?," *Journal of Monetary Economics*, 2017, *91*, 69–87. The Swiss National Bank/Study Center Gerzensee Special Issue: "Modern Macroeconomics: Study Center Gerzensee Conference in Honor of Robert G. King" Sponsored by the Swiss National Bank and the Study Center Gerzensee.

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.

Diamond, Emily and Jack Zhou, "Whose policy is it anyway? Public support for clean energy policy depends on the message and the messenger," *Environmental Politics*, 2022, *31* (6), 991–1015.

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).

Escueta, Maya, Andre Joshua Nickow, Philip Oreopoulos, and Vincent Quan, "Upgrading Education with Technology: Insights from Experimental Research," *Journal of Economic Literature*, December 2020, *58* (4), 897–996.

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

Fielding, Kelly S, Matthew J Hornsey, Ha Anh Thai, and Li Li Toh, "Using ingroup messengers and ingroup values to promote climate change policy," *Climatic Change*, 2020, *158*, 181–199.

Fukuyama, Francis, "The End of History?," *The National Interest*, 1989, (16), 3–18.

_ , *Identity: The Demand for Dignity and the Politics of Resentment*, New York: Farrar, Straus and Giroux, 2018.

Galí, Jordi and Luca Gambetti, "On the Sources of the Great Moderation," *American Economic Journal: Macroeconomics*, January 2009, *1* (1), 26–57.

Gentzkow, Matthew and Jesse M Shapiro, "What Drives Media Slant? Evidence from U.S. Daily Newspapers," Working Paper 12707, National Bureau of Economic Research November 2006.

Goldin, Claudia and Lawrence F Katz, "Long-run changes in the US wage structure: Narrowing, widening, polarizing," 2007.

_ and Lawrence F. Katz, *The Race between Education and Technology*, Harvard University Press, 2008.

Goodin, Robert, Michael Moran, and Martin Rein, *The Oxford Handbook of Public Policy*, Oxford University Press, 06 2008.

Gortázar, Lucas and Juan Moreno, "Shadow Education in Spain: how private tutoring is becoming an essential good," 2022.

Haidt, Jonathan, *The Righteous Mind: Why Good People Are Divided by Politics and Religion*, New York: Pantheon Books, 2012.

Jackson, Michelle and Brian Holzman, "A century of educational inequality in the United States," *Proceedings of the National Academy of Sciences*, 2020, *117* (32), 19108–19115.

Kahan, Dan M., "Ideology, motivated reasoning, and cognitive reflection," *Judgment and Decision Making*, 2013, *8* (4), 407–424.

_ , *The Politically Motivated Reasoning Paradigm, Part 1: What Politically Motivated Reasoning Is and How to Measure It*, John Wiley & Sons, Ltd,

Kahan, Hank Jenkins-Smith Dan M. and Donald Braman, "Cultural cognition of scientific consensus," *Journal of Risk Research*, 2011, *14* (2), 147–174.

Khemani, Stuti, "Political economy of reform," Policy Research Working Paper Series 8224, The World Bank October 2017.

Klein, Ezra, *Why We're Polarized*, New York: Avid Reader Press / Simon Schuster, 2020.

Kraft, Matthew A, John A List, Jeffrey A Livingston, and Sally Sadoff, "Online tutoring by college volunteers: Experimental evidence from a pilot program," in "AEA Papers and Proceedings," Vol. 112 2022, pp. 614–18.

Krueger, Alan, "How Computers Have Changed the Wage Structure: Evidence from Microdata, 1984–1989," *The Quarterly Journal of Economics*, 1993, *108* (1), 33–60.

Maclean, Johanna Catherine, John Buckell, and Joachim Marti, "Information Source and Cigarettes: Experimental Evidence on the Messenger Effect," Working Paper 25632, National Bureau of Economic Research March 2019.

Marcus, Gary, "AI's Jurassic Park moment," https://garymarcus.substack.com/p/ais-jurassic-park-moment 2022.

Mascheroni, Giovanna, Marium Saeed, Marco Valenza, Davide Cino, Thomas Dreesen, Lorenzo Gisuseppe Zaffaroni, and Daniel Kardefelt-Winther, "Learning at a Distance Children's remote learning experiences in Italy during the COVID-19 pandemic," Technical Report, UNICEF Office of Research - Innocenti 2021.

Masset, Edoardo, Marie Gaarder, Penelope Beynon, and Christelle Chapoy, "What is the impact of a policy brief? Results of an experiment in research dissemination," *Journal of Development Effectiveness*, 2013, *5* (1), 50–63.

Molina, Ezequiel, Cristobal Cobo, Jasmine Pineda, and Helena Rovner, "The AI Revolution in Education: What You Need to Know," 2024. Digital Innovations in Education; Brief No.1.

Mollick, Ethan, "ChatGPT is a Tipping Point for AI," *Harvard Business Review*, 2022.

_ , "The future of education in a world of AI," https://www.oneusefulthing.org/p/the-future-of-education-in-a-world 2023.

Nakajima, Nozomi, "Evidence-based decisions and education policymakers," *Unpublished Paper*, 2021.

New York Times, "Reagan Gets a Red Carpet From British," 1989.

Nickow, Andre, Philip Oreopoulos, and Vincent Quan, "The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence," Working Paper 27476, National Bureau of Economic Research July 2020.

OECD, *PISA 2022 Results (Volume I): What Students Know and Can Do*, Paris: OECD Publishing, 2023.

Pinker, Steven, *Rationality: What It Is, Why It Seems Scarce, Why It Matters*, New York: Viking, 2021.

Polanyi, Michael, "The logic of tacit inference," *Philosophy*, 1966, *41* (155), 1–18.

Rodrik, Dani, "Understanding Economic Policy Reform," *Journal of Economic Literature*, 1996, *34* (1), 9–41.

The Economist, "Spain's centrist Ciudadanos are on the march," *The Economist*, 2018. Accessed: 2023-06-27.

Tommasi, Mariano and Andres Velasco, "Where Are We in the Political Economy of Reform?," Working Papers 11, Universidad de San Andres, Departamento de Economia 1996.

Tversky, Amos and Daniel Kahneman, "Judgment under Uncertainty: Heuristics and Biases: Biases in judgments reveal some heuristics of thinking under uncertainty.," *science*, 1974, *185* (4157), 1124–1131.

UBS, "Let's chat about ChatGPT," https://www.ubs.com/global/en/wealth-management/our-approach/marketnews/article.1585717.html 2023. Accessed: 2023-04-26.

van der Linden, Sander, *Foolproof: Why Misinformation Infects Our Minds and How to Build Immunity*, New York: W. W. Norton & Company, 2023.

_ **, Anthony Leiserowitz, Seth Rosenthal, and Edward Maibach**, "Inoculating the Public against Misinformation about Climate Change," *Global Challenges*, 2017, *1* (2), 1600008.

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

Williamson, John, "The Washington consensus," *Washington, DC*, 1990.

Summary of the three chapters

Chapter 1

Title: "When GenAI increases inequality: evidence from a university debating competition"

Motivation and Research Questions

Historical trends in automation have shown adverse effects on routine, blue-collar jobs, while enhancing the productivity of high-skilled workers, thereby exacerbating inequalities. In contrast, early studies on GenAI suggest it may have different implications. Due to the advanced capabilities of Large Language Models (LLMs), high-skill professions such as programmers, designers or video-producers previously thought to be shielded from automation, are expected to be negatively affected by GenAI. Additionally, a growing body of experimental evidence shows that GenAI enhances the productivity of lower performing individuals more than that of high performers in a variety of written tasks, suggesting GenAI might act as a productivity compressor.

This study examines the "GenAI productivity compressing hypothesis" by evaluating productivity improvements in debating, a task requiring verbal interactions and higher-order social and cognitive skills.

Speaking, presenting ideas and arguing convincingly are key human skills, vital for a wide range of activities, and common to high-earning workers of all kinds. One of the clearly relevant applications of ChatGPT, the most widely used GenAI, is that it provides very fast summaries of complex ideas and concepts, well-organized arguments in favor or against any topic and endless examples and metaphors.

While most existing evidence on the "GenAI productivity compressing hypothesis" focuses on written tasks and online environments, this study analyzes GenAI's impact on productivity in a setting that allows to measure a set of skills largely overlooked by previous literature. Contrary to early findings in the literature, the study reveals that individuals with high-performance initially gain more from GenAI than those with lower initial performance. Furthermore, those with a stronger academic background benefit significantly more from GenAI.

Literature and Contribution

Research on the effects of technology on inequality has boomed over the last two decades (Autor et al., 2003; Acemoglu and Autor, 2011; Acemoglu and Restrepo, 2017; Felten et al., 2019). More recent studies have focused on the impact of AI on work. My paper is part of a very recent literature studying the impact of new models of GenAI on productivity and inequality in a variety of different tasks and in different contexts.These tasks include coding (Peng et al., 2023), professional writing tasks, such as writing emails or press releases (Noy and Zhang, 2023), law exams (Choi and Schwarcz, 2023), creative writing (Doshi and Hauser, 2023), online written customer service (Brynjolfsson et al., 2023) or a variety of management consultancy tasks (Dell'Acqua et al., 2023). All these studies test different *written skills* at which Large Language Models (LLMs) tend to excel and find that GenAI compresses the distribution of skills by helping more low performing individuals. However, few studies have explored the impact of GenAI on productivity in tasks involving higher-order social interactions.

My paper contributes in three important ways to the literature. First, it is one of the few existing 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 literature, my results suggest that in such settings GenAI might lead to increasing inequalities. Finally, I suggest a possible explanation to reconcile my results with previous literature: when tasks require higher-order skills, and answers cannot be directly extracted from the AI and copy-pasted, high-skilled workers are likely to benefit more of the advantages of GenAI.

Data and Methods

I run a randomized controlled trial in a real debating competition involving 141 undergraduate students from two different universities in Spain. The contest was designed following the standards of official international university debating competitions. The intervention took place at Esade Business School in Barcelona and at CUNEF University in Madrid over three different debating days in late March and April 2023. About half of the students - the treatment group - were randomly assigned to a short 20-minute training of ChatGPT3.5 and were allowed to use it as support throughout the contest. The other half of students – the control group - were only allowed to use conventional internet access to prepare the debates.

All registered students were asked to fill a baseline survey as a condition to participate in the competition. The survey asked for relevant student characteristics, socio-economic background information and political views of the students. Each student debated in three to four rounds of one-on-one debates.

In order to measure the impact of ChatGPT on productivity and inequality at debating I test two

main hypotheses: (H1) whether ChatGPT improves debating points of participants; (H2) whether it contributes to reduce inequality in debating skills.

My two main outcome variables were individual debating points and win/lose the debate. To build those variables, all debates were recorded and sent to three different judges. I computed the average points given by the three judges to each participant for each round of debates. All judges were chosen among debating teachers or former debating champions and did not know anything about the experiment. Students were also asked to fill out an end-line survey at the end of the competition day. In that survey I asked students on their use of the technology and whether they felt comfortable with the time given (among other related questions).

Randomisation was done at the individual level in three blocks, corresponding to the three different debating days. The treatment was introduced after the first round of debates to capture the effect of ChatGPT not only between participants but also within subjects. The randomisation was done *in situ* on each of the three debating days, once all participants had arrived at the competition using a Stata command.

I also randomised participants in three steps. First, participants were randomly assigned to treatment and control groups. Second, participants were randomly assigned to debating partners in each round. Third, participants were randomly assigned debating positions (either in favor or against the debate proposition) for each round. The randomization was balanced with the information collected at baseline.

Candidate's contribution

This is a single-authored paper.

Chapter 2

Title: "Online tutoring works: Experimental evidence from a program with vulnerable children" (with Lucas Gortazar and Claudia Hupkau)

Motivation and Results

Evidence shows that lockdowns and school closures resulting from the pandemic had a profound negative impact on the learning and socio-emotional development of pupils around the world and that students from disadvantaged families were hit harder than those from more comfortable backgrounds (Maldonado and De Witte, 2021; Schult and Lindner, 2021; Engzell et al., 2021; Tomasik et al., 2021). Researchers estimated that learning losses in developed countries could amount to a drop in future GDP of about 1.5% per year in the long-run (Hanushek and Woessmann, 2020), and deepen inequalities (Elliot Major et al., 2020).

In response, we launched an innovative tutoring program with "Empieza por Educar", the Spanish

branch of Teach for All, an NGO, from April to June 2021. The program offered free, 100% on-line, after-school tutoring targeted at pupils aged 12 to 15 from disadvantaged backgrounds. Eighteen state and grant-assisted schools in low-income districts of Madrid and Catalonia regions participated in the program. The tutoring sessions focused on math and socio-emotional support (motivation, well-being, work routines), lasted eight weeks and were delivered in groups of two students.

The intervention significantly increased standardised test scores (+0.26 SD) and end-of-year maths grades (+0.48 SD), while reducing the probability of repeating the school year. The intervention also raised aspirations, as well as self-reported effort at school.

Literature and Contribution

Evidence gathered by over 100 experimental studies before the pandemic showed that one-on-one and small group face-to-face tutoring programs have a very positive impact in a wide array of contexts and program designs, with programs improving learning outcomes by 37% of a standard deviation on average (Nickow et al., 2020). However, very little evidence existed as regards to the effectiveness of online tutoring programs with real teachers (not educational software) before or during the pandemic.

The closest to our research is Carlana and La Ferrara (2021). The authors ran a one-on-one online tutoring program in Italy, with tutoring provided by volunteer university students during the harshest lockdown period in April to June 2020, when all kids were at home. Kraft et al. (2022) also implement an online tutoring program for middle school students with college volunteers. They find positive but insignificant effects on math and reading. Our project departs from these studies in several ways: we worked with paid professional teachers, while they worked with volunteers; our lessons were in groups of two students, while the prior studies offered one-on-one tutoring, and our program took place when the schools were fully re-opened in Spain, while the other programs took place amid the pandemic.

Governments and international institutions have committed large amounts of money to respond to the pandemic learning losses. Our paper informs on an innovative and cost-effective policy design immediately relevant for post-pandemic education policies across the world.

Data and Methods

Recruitment of participants for the RCT was done through schools, who were asked to advertise the program among the students most in need for math tutoring. The experimental strategy relied on over-subscription. No compensation for students not assigned to the treatment was offered. Randomisation was done in various steps. First, we assigned students who enrolled in the program randomly into treatment and control groups. In a second step, we randomly assigned treatment students within the same classrooms into groups of two. In the last step, we randomly assigned

tutors to the different groups.

We collected a rich array of family and household characteristics at the stage of online registration for the program, where parents had to fill out a detailed survey. After the completion of the registration period and before randomisation, we ran a baseline student survey that included a maths test and a questionnaire on prior attainment, wellbeing, and other socio-emotional outcomes. During implementation, we collected data for each tutoring session to monitor attendance and dropout in real time.

When tutoring sessions finalised, we administered an endline survey. The endline survey again included a math test as well as the questions regarding well-being, locus of control and self-rated ability. We also ran an online and phone survey to parents of treated and control group students in early July, once the school year was over. Qualitative surveys were also administered to tutors, school principals, and math teachers of treated students to get a better understanding of implementation and process aspects. We also run another survey with parents one year after finalising the program.

Candidate's contribution

I am the corresponding author and lead researcher in this project. I developed the idea, put together the team for launching the program and managed to get the financing. I worked on the RCT design with two other co-authors that also contributed to the program design, the data analysis, and the writing up. The paper was published at the Journal of Public Economics in January 2024.

Chapter 3

Title: "Ideological Alignment and Evidence-Based Policy Adoption" (with Jorge Garcia-Hombrados, Marcel Jansen, Berkay Ozcan, Angel Martínez and Pedro Rey).

Motivation and Research Questions

Understanding how to improve the dissemination of scientific knowledge to policymakers is crucial for economic and social progress. However, despite persistent efforts to promote evidence-based policymaking, a significant gap persists between available evidence and the policies ultimately implemented (European Commission, 2022; OECD, 2020). In this paper we run a field experiment with more than 5000 local politicians in Spain to understand which constraints might be at play to evidence based policy adoption. Our main intervention involves sending out information presenting the results of a study that shows the large benefits on the local economy of a costless and politically neutral policy. The intervention consists in improving the content of the Wikipedia page of the municipality, which has been demonstrated in (Hinnosaar et al., 2021) to increase the number of tourists, while generating large net benefits for local economies.

The information was sent to a random draw of local politicians, introducing variations in the

format (policy brief vs newspaper article) and the political affiliation of the messenger. The level of policy adoption – the main outcome variable – is determined by changes in the Wikipedia page.

Our unique setting allows us to investigate three research questions: First, whether lack of awareness about research prevents the adoption of an almost zero-cost, politically neutral, evidence-based policy. Second, whether information-framing matters for policy implementation. Third, whether political alignment with the messenger of the policy – in this case research disseminating institutions - influences policy implementation. To answer these questions, we conducted a country-wide field experiment collaborating with prominent and authoritative institutions with opposing ideologies who disseminated research findings to a large sample of local policymakers.

Literature and Contribution

This project aims to contribute to the very scarce literature studying the reasons why evidence-based policy is so seldom implemented by politicians. Research in this area has focused on analyzing the effectiveness of different tools for research dissemination. For instance, Masset et al. (2013) run an experiment to study the effectiveness of policy briefs as a mechanism for changing policy beliefs. Another strand in the literature has analysed the role of prior beliefs in conditioning the way policymakers react to new available information. Using survey experiments, Christensen and Moynihan (2020), find that politicians tend to engage in motivated reasoning and to rely more on prior political attitudes and less on policy information than the general public. Similarly, (Banuri et al., 2019) conduct experiments with policy professionals and show that these are subject to decision making traps, including confirmation bias correlated with ideological priors.

Closer to our research, (Hjort et al., 2021) explore if research findings alter policy-makers beliefs and lead to actual policy change. They find that informing mayors in Brazil about research increases the probability that their municipality implements the policy by 10 percentage points.

Our results indicate that merely providing information increases policy adoption by 38% relative to the uninformed control group, although this increase is marginally above conventional significance thresholds (p-value=0.13). However, when the ideologies of policymakers and informing institutions align, we find that the probability of policy adoption increases by more than 65% compared to the control group (p-value=0.03). Conversely, when information comes from institutions with opposite ideologies, the coefficient is small and statistically non-significant. The effect size of receiving a policy brief from an ideologically nonsalient prestigious institution is nearly half that of a policy brief from an institution with an aligned ideology. Finally, results show that both formats (policy brief and newspaper article) are similarly effective in influencing policy adoption.

Our main contribution is that we study, for the first time, how ideological alignment affects *policy implementation* using a country-wide sample of policymakers and real and authoritative ideological

institutions to inform them. Second, we add to the literature on motivated reasoning, polarization, and partisan bias, which has shown, that when research evidence aligns with a particular ideology, it affects the general public's belief updating and compliance with policies (Druckman et al., 2021; Druckman and McGrath, 2019; Guilbeault et al., 2018; Butler and Broockman, 2011). Our use of a non-ideological policy allows us to isolate the effect of the informant's ideology along the policy adoption process.

Data and Methods

Our sample consists of all Spanish municipalities considered touristic according to a list of objective criteria. We randomly divided our sample of 5,678 touristic municipalities into six groups of similar size to have five treatment arms and one control group (each group included about 950 municipalities). The randomization was stratified according to three criteria: the political party ruling the municipality, the municipality's population, and the number of touristic accommodations available in the municipality. The analytical sample of local politicians is randomly split in treatment and control groups.

To measure the degree of adoption of the policy recommended in the brief -the improvement of the Wikipedia page- we look at Wikipedia pages over time using various techniques. Additional information on the ruling political party and characteristics of the municipality was gathered from the publicly available databases of the Spanish Ministry of Interior and from the Padrón Continuo de Habitantes (Spanish Population registry).

Our RCT design allows us to compare email opening rates, access to policy briefs and newspaper articles once policymakers learn the ideology of the informing institution, and changes in Wikipedia across groups to investigate three questions. First, does providing information to policymakers increase policy adoption? Second, does the ideological alignment between the informing and the policymaker affect policy adoption? Third, does the instrument used to describe the summary evidence (newspaper vs policy brief) affect policy adoption?

We use linear probability models to estimate the effect of the different treatment arms on the probability of implementing a recommended change in the Wikipedia page of a municipality.

Candidate's contribution

I worked on this paper with five co-authors.

Bibliography

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.

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.

Banuri, Sheheryar, Stefan Dercon, and Varun Gauri, "Biased Policy Professionals," *World Bank Economic Review*, 2019, *33* (2), 310–327.

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

Butler, Daniel M. and David E. Broockman, "Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators," *American Journal of Political Science*, 2011, *55* (3), 463–477.

Carlana, Michela and Eliana La Ferrara, "Apart but Connected: Online Tutoring and Student Outcomes during the COVID-19 Pandemic," Discussion Paper IZA DP No.14094, IZA Institute of Labor Economics 2021.

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

Christensen, Julian and Donald P. Moynihan, "Motivated reasoning and policy information: politicians are more resistant to debiasing interventions than the general public," *Behavioural Public Policy*, 2020, p. 1–22.

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.*

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

Druckman, James N and Mary C McGrath, "The evidence for motivated reasoning in climate change preference formation," *Nature Climate Change*, 2019, *9* (2), 111–119.

Druckman, James N., Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan, "How Affective Polarization Shapes Americans' Political Beliefs: A Study of Response to the COVID-19 Pandemic," *Journal of Experimental Political Science*, 2021, *8* (3), 223–234.

Elliot Major, Lee, Stephen Machin, and Andrew Eyles, "Generation COVID: Emerging work and education inequalities," *CEP Covid-19 analysis No. 011*, 2020.

Engzell, Per, Arun Frey, and Mark D. Verhagen, "Learning loss due to school closures during the COVID-19 pandemic," *Proceedings of the National Academy of Sciences*, 2021, *118* (17).

European Commission, "Staff Working Document - Supporting and connecting policymaking in the Member States with scientific research," 2022.

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.

Guilbeault, Douglas, Joshua Becker, and Damon Centola, "Social learning and partisan bias in the interpretation of climate trends," *Proceedings of the National Academy of Sciences*, 2018, *115* (39), 9714–9719.

Hanushek, Eric A. and Ludger Woessmann, "The economic impacts of learning losses," 2020, (225).

Hinnosaar, Marit, Toomas Hinnosaar, Michael E. Kummer, and Olga Slivko, "Wikipedia Matters," *Journal of Economics & Management Strategy*, 2021, pp. 1–13.

Hjort, Jonas, Diana Moreira, Gautam Rao, and Juan Francisco Santini, "How Research Affects Policy: Experimental Evidence from 2,150 Brazilian Municipalities," *American Economic Review*, May 2021, *111* (5), 1442–80.

Kraft, Matthew A, John A List, Jeffrey A Livingston, and Sally Sadoff, "Online tutoring by college volunteers: Experimental evidence from a pilot program," in "AEA Papers and Proceedings," Vol. 112 2022, pp. 614–18.

Maldonado, Joana Elisa and Kristof De Witte, "The effect of school closures on standardised student test outcomes," *British Educational Research Journal*, 2021.

Masset, Edoardo, Marie Gaarder, Penelope Beynon, and Christelle Chapoy, "What is the impact of a policy brief? Results of an experiment in research dissemination," *Journal of Development Effectiveness*, 2013, *5* (1), 50–63.

Nickow, Andre, Philip Oreopoulos, and Vincent Quan, "The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence," Working Paper 27476, National Bureau of Economic Research July 2020.

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

OECD, "Public Governance Reviews," *Building Capacity for Evidence-Informed Policy-Making. https://www.oecd.org/gov/building-capacity-for-evidence-informed-policy-making-86331250-en.htm*, 2020, p. 80.

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

Schult, Johannes and Marlit Annalena Lindner, "Did students learn less during the COVID-19 pandemic? Reading and mathematics competencies before and after the first pandemic wave," 2021.

Tomasik, Martin J., Laura A. Helbling, and Urs Moser, "Educational gains of in-person vs. distance learning in primary and secondary schools: A natural experiment during the COVID-19 pandemic school closures in Switzerland," *International Journal of Psychology*, 2021, *56* (4), 566–576.

Chapter 1

When GenAI increases inequality: evidence from a university debating competition

Antonio Roldán-Monés[1]

Abstract

This paper evaluates the impact of Generative Artificial Intelligence (GenAI) on productivity and work inequality. While existing literature has focused on the impact of GenAI on writing tasks, I exploit a unique setting to analyze productivity improvements on a task involving verbal interactions and requiring high cognitive and social skills. I run a Randomized Controlled Trial with undergraduate students in a debating competition. Participants were randomly assigned to either GenAI support for debate preparation or to conventional internet resources. The small overall impact of GenAI I find hides significant heterogeneity between students. In particular, and contrary to previous findings, high-performing students and those with a stronger academic background benefit significantly more from GenAI than their lower-performing counterparts. Moreover, high-skilled individuals with access to GenAI experience large improvements in their perception of time needed to prepare debates, while low-ability students do not. I suggest a possible explanation to reconcile these results with previous literature: when tasks require higher-order skills, and answers cannot be directly extracted from the AI and copy-pasted, high-skilled workers are likely to benefit more of the advantages of GenAI. These findings could have important implications for the broader understanding of potential economic and social impacts of these technologies.

Acknowledgements

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 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,

[1]Department of Scoail Policy, LSE, and Esade Business School, Universidad Ramon Llull.

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.

1.1 Introduction

Business leaders, governments and researchers across the globe are expecting Generative Artificial Intelligence (GenAI) systems to have a deep impact on work (Chui et al., 2023). Yet, it remains an open question how these new technologies will affect work inequalities (Wilmers, 2024). Previous waves of automation proved to have 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). However, early findings in the literature on GenAI suggest its effects might go in a different direction. First, because of the expanded technical capacities of Large Language Models (LLMs), previously shielded high-skill professions are expected to be profoundly affected (Eloundou et al., 2023; Felten et al., 2023). Second, by boosting the productivity of low performers more than that of high performers within specific jobs, GenAI has been found to contribute to compress the productivity distribution, making GenAI a skill leveler (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 study, I test the validity of the "GenAI skill leveler hypothesis" in a unique setting allowing me to measure productivity improvements in a task involving realistic human 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, teamwork or critical thinking that have been proven to be of critical importance for high-earning workers of all kinds (Börner et al., 2018; Deming, 2017; Weinberger, 2014).[2] Debating - similarly to negotiating, managing, teaching, or leading – involves highly unpredictable social interactions and requires higher-order skills, including rhetoric ability, critical thinking, social intuition, argumentative agility, and persuasion capacity [3] (Weston, 2020). While most evidence on the "GenAI skill leveler hypothesis" is in written tasks and online settings, this is one of the very few studies to analyze the impact of GenAI on productivity in a setting involving social interactions. Contrary to the early results in the literature, I find that individuals who initially showed high performance benefit significantly more from GenAI than low initial performers. I also find that those with better

[2] Deming (2022) follows Bloom's (1956) taxonomy of educational objectives, which delineates a hierarchical framework with factual knowledge at the bottom. The hierarchy progresses through stages including pattern recognition and classification, advancing to more complex objectives such as the application of knowledge to novel situations, experimentation, and the integration of new ideas. At the pinnacle, the taxonomy encompasses evaluation and decision-making as well as the design and creation of new concepts.

[3] O*NET, the most widely used data base for skills and occupational information in the US, classifies "persuasion", "negotiation" and "social perceptiveness" within a broader set of "social skills" and other "Cross-Functional Skills", defined as "developed capacities that facilitate performance of activities across jobs". "Cross-Functional Skills" also includes "complex problem-solving skills" defined as "Developed capacities used to solve novel, ill-defined problems in complex, real-world settings"(O*NET OnLine, n.d.). Effective debating requires all these skills, see for instance Weston (2020)

academic background profit more than the rest from GenAI.

I run a randomized controlled trial (RCT) to assess 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 speaking, presenting and debating skills. I test two main hypotheses[4]: (1) whether ChatGPT improves overall debating performance and (2) whether it contributes to reduce inequality in debating performance. 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, 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 (either debating professors at university or former debating champions) 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 important heterogeneity. I find that ChatGPT helps significantly more those students in the top of the skill distribution. Among those in the top 50% in debating points in the first baseline round of debates (before the treatment was implemented), treatment students have on average 5.2% higher points than control students - the equivalent to 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.

Using an additional measure of innate student ability, whether they are on an merit scholarship, 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% for students

[4]AEARCTR-0011113. See https://www.socialscienceregistry.org/trials/11113.

using ChatGPT having received a scholarship, and a 1.7% (insignificant) effect of ChatGPT among those also in the treatment group but without a scholarship. 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 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". Interestingly, students on a merit scholarship in the control group were very pessimistic about having had sufficient time to prepare the debates, while non-scholarship students in the control group seemed much more (over-)confident. However, the effect of the treatment was very different in the two groups. When merit scholarship students were given access to ChatGPT, their perception of time sufficiency increased to well above that of non-scholarship recipients. High ability students seem to be more effective at leveraging the benefits of AI to improve their productivity in this specific task.

I suggest a possible theoretical channel to reconcile these results with existing literature. 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 are using. However, in social environments with repeated human interactions, high-skilled individuals - because of previous deeper knowledge[5], their superior persuasive abilities, or, more generally, their higher-order skills - will be better able to use the information provided by ChatGPT to their advantage than low-skilled individuals.[6]

Most studies analyzing the GenAI skill leveler hypothesis have focused on testing a narrow set of writing skills at which LLMs are especially effective. These include code programming (Peng

[5]The distinction between superficial and deep learning is well-documented in cognitive and educational psychology. Classic theories by Piaget (1952) and Vygotsky (1978) emphasize that deep learning occurs through active engagement and construction of knowledge, where learners build upon their existing knowledge base to be able to engage in higher order thinking. Deep learning, according to this literature, is essential for tasks requiring higher cognitive engagement, such as teaching or complex problem-solving.

[6]Another way of looking at such complementarities in the debating context is by using a simplification of Aristotle's classic model on rhetoric, in which debate performance (DP) would be a function of two variables: Logos (L) and Ethos (Et). Aristotle wrote three books on the Art of Rhetoric. According to his writing, speech can produce persuasion either through the character (*êthos*) of the speaker, the emotional state (*pathos*) of the listener, or the argument (*logos*) itself. For simplicity I will not consider the listener here. Logos would be the logical skill associated with having a higher rational or cognitive capacity. Ethos would be the rhetorical skill related to the character, eloquence, or "higher-order" skill of the speaker, needed to effectively adapt the debating strategy to the right social context to maximize persuasion. In this simple model, ChatGPT can be seen as a technological development that provides a faster flow of good arguments that directly boost L for good and bad debaters. However, by liberating time to sharpen their persuasion strategy, only good debaters (with both high L and Et) will harness the effects of ChatGPT.

et al., 2023), professional writing skills, 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.

To be a good debater it is not enough to have access to a longer list of better arguments. One needs to (in the jargon of debating coaches) "own" those arguments – to internalize them—and present them in a compelling way, sounding self-confident, reacting quickly to the rivals' comments, picking the right evidence, showing empathy and emotion, understanding the audience, and finding the right metaphor at the right time (Adams, 1810; Corbett and Connors, 1999; Aristotle, 1960). For these reasons, , although they are becoming increasingly persuasive (Salvi et al., 2024), powerful AI-driven autonomous debating systems have struggled systematically to perform better than humans (Slonim et al., 2021).

This paper contributes in three important ways to the literature. First, this is one of the few existing 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 literature, my results suggest that in such settings GenAI might lead to increasing inequality. Third, I show that ChatGPT has a very large impact on top students (but not in the rest) in a measure of self-perceived productivity: whether they consider they had sufficient time to prepare the debates. Finally I suggest a possible explanation to reconcile my results with previous literature: when tasks require higher-order skills, and answers cannot be directly extracted from the AI and copy-pasted, high-skilled workers are likely to benefit more of the advantages of GenAI.

These findings are important, especially, in light of the demonstrated growing relevance of social and higher-order cognitive skills for high-earning professions in the data economy (Börner et al.,

2018). For instance, the OECD finds that in occupations with the highest exposure to AI, social and emotional skills such as management, teamwork, negotiation, or persuasion see the largest increase in demand (Green, 2024). Likewise, Deming (2017) shows that the economic return of social skills in the United States more than doubled for a cohort of young people entering the labor market in the 2000s compared to those who entered in the 1980s.

If my results generalize, this research might have broader implications for understanding the economic and social impacts of GenAI, especially for high-earning occupations characterized by complex human interactions.

The rest of the paper is organized as follows: In Section 1.2, I describe the design of the debate competition. In Section 1.3, I discuss the experimental design and implementation. In Section 1.4, I explain the empirical strategy and the hypotheses tested. Section 1.5 shows the results and mechanisms in more detail. Section 1.6 concludes.

1.2 Context of the intervention

The intervention took place at ESADE Business School in Barcelona and at CUNEF University in Madrid 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 in Madrid on the 19th of April 2023. The target population were undergraduate university students. I established collaborations with specific professors to get their students to participate. Thus, most participants were from three courses: two Debating courses at Esade and an Economic Policy class at CUNEF.[7] Most of them were enrolled in either law, business or economics degrees. The first session was run in English, the other two sessions were done 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.

[7]The debating course in English, called "Global Debate Skills 1" is an optional course for first year students in the double degree of law and international relations at the ESADE Law School. The debating course in Spanish, called "Debate League: Techniques" is also a first-year optional course open to law, business and international relations students. Students from CUNEF were third year students of "Economic Policy", a mandatory subject in the double degree of law and business.

The debate competition took place during usual class time. Although there were some exceptions, students in each of the three courses competed with students from their same course. 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.

1.2.1 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 standard format used in international short debating competitions, 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. BP debates typically involve one intervention of 5 or 7 minutes per member of the debating team. Each student had three minutes for an opening statement and two minutes for refutation and conclusion. Preparation time for each round of debates was 20 minutes. In total, each debate round took about 30 minutes, including preparation time. Students were asked to bring their computers and cellphones to the competition and did not receive any materials in advance to prepare the debates.

1.2.2 Spaces, monitoring and audio-recording

All participants debated at the same time in a large room or auditorium in groups of two. However, after the first round of debates, 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 on each of the three debating days to make sure there were no interactions between treatment and control participants and avoid cheating or contamination. There was a risk that, if students in the control group found out that their rivals were using ChatGPT, this could affect their performance in the debates. Debates were audio-recorded by students with their own cellphones. At the end of the debates students were asked to upload the recording to a folder using their individual IDs specifically given for the competition.

At the end of the debating competition, each student had done three or four rounds of debates, depending on the day of the competition. At CUNEF University, we had less class time to run the competition and could only do three rounds. For the treatment group this meant one debate before

and two to three debates after the ChatGPT training. 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 independent expert judges for evaluation. Every student had done either three or four five-minute (3+2) interventions. In total, there were 230 debates recorded, each of ten minutes, which is equivalent to 2300 minutes of recordings (about 39 hours).

1.2.3 Policy topics debated

Each debate topic was announced on a big screen just before the preparation time. 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, inequality, monetary policy, trade or labor-markets. All debate topics had an "evidence-based policy" and a "bad policy" side. That is: one side of the debate is supported by ample evidence in economics or social sciences and the other side is not, according to the criteria of those economists.

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 "bad policy" idea was half the time coming from policies typically associated with the left and half the time from policies typically associated with the right, so arguing for the bad policy stance was not associated with one particular political ideology. The reason why I chose eight different policy topics was that I wanted to avoid potential information leakages among students in the two first rounds of debates at ESADE. In the third round at CUNEF (in Madrid) I randomly selected four out of the eight topics that had been used in previous debates. Students were randomly assigned to debating positions and partners in each of the three different debating days, after arriving to the competition.

1.2.4 Judges and evaluation criteria

Expert judges were chosen according to two criteria: they had to be either former debating champions in official competitions 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 categories. The categories were established following standard debating rubrics: (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 ten euros per hour to do the evaluation.

1.2.5 Incentives for participants

All participants were offered a certified diploma from EsadeEcPol (ESADE), an economic policy think tank in Spain, or CUNEF, just for participating. Ten prizes of 100€ in Amazon Vouchers were also offered. Given that participants were randomly assigned different tools for debating that could potentially drive their results, prizes were given according to the tools used. That is: five prizes to the top performers in the treatment group and five prizes to the top performers in the control group. The incentives were announced at registration and the prizes were given a few weeks after the contest ended. The reason why I introduced economic incentives was to make sure students made the maximum possible effort in their debates. For most students, assistance to the debating competition was mandatory, but there were no other academic rewards involved for participating.

1.3 Randomization, implementation and data

Figure 1.1: Timeline of the experiment

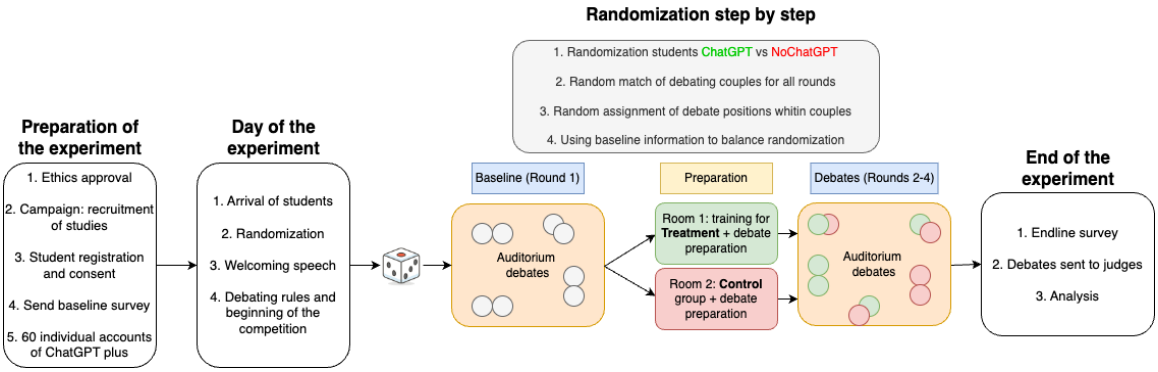


Figure 1.1 shows the timeline of the intervention. Planning and design of the experiment took place between January and March 2023. Ahead of the experiment I received the ethics approval from ESADE's Ethics Committee. All the students were pre-registered with the collaboration of the ESADE Decision Lab. Five days ahead of the competition, participants received a baseline survey including a consent form. The day before the competition, the experiment was pre-registered. Participation closed 15 minutes after the indicated arrival time. In one of the sessions, there was an uneven number of participants, so I asked a randomly selected person to leave. Those students that

had not filled the base-line survey did so at arrival.

1.3.1 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 (defending either the evidence-based side or the bad policy side) 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 1.1 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.

1.3.2 Implementation

Upon arrival, students were given an ID number and were put together in a large room or auditorium. Over the first 15-20 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. When the introduction concluded, treatment and control students were sent to different rooms to prepare the debates. However, in the first round of debates ChatGPT prohibited for everyone. When the 15 minutes for debate preparation ended, all students were asked to go back to a large room. Then debates and recordings started. After the debate ended, students were asked to upload the recorded material to an online folder.

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 and policy expert. Control students were given an extra 20-minute break. Then, the matching and debate 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 debates. AI tools were explicitly prohibited for control students. Treatment students were explicitly told not to tell about ChatGPT. Communication between treatment and control groups was supervised for the whole time and limited to the actual

debating time to avoid contamination. When the four rounds of debates were finalized, students were asked to fill an endline survey. All debates were sent via WeTransfer to the expert judges to correct. A few weeks later, a communication was sent to the winners of the debating contest.

1.3.3 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 All students of the relevant courses that participated in the study were automatically 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. Finally, students were asked about their interest in politics, political preferences, a measure of polarisation, and ex-ante views on eight policy topics.

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. Finally the survey included a series of questions regarding their opinions 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 1.2.4) . 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.

1.4 Empirical strategy

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

H1: ChatGPT improves public speaking and debating skills, measured by individual debating points or probability of winning a debate

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

$$Y_{ird} = \alpha_d + \alpha_r + \beta Treat_i + \lambda X_i' + \epsilon_{itd} \tag{1.1}$$

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.

H2: ChatGPT reduces inequality in debating skills

Here I test whether participants with lower baseline debating points benefit more from ChatGPT than those with higher baseline debating points. I use debate points from the first round of debates to divide participants into top (50%) and bottom (50%) performers. Then I compare average debating points in rounds two, three and four for top and bottom performers among individuals who use ChatGPT and among individuals who do not use ChatGPT. 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 initial performers:

$$Y_{ird} = \alpha_d + \alpha_r + \beta Treat_i + \gamma Top50 + \delta Treat_i \times Top50 + \lambda X_i + \epsilon_{ird} \tag{1.2}$$

$Top50$ is a binary variable that takes the value one if the participant ranked in the top 50% of the distribution of debate points in the first round and zero if they ranked in the bottom 50%. The coefficient γ indicates how initial debating skills correlate with later debating outcomes for the top 50% performers in the control group. The coefficient δ quantifies the interaction effect of assignment to using ChatGPT for individuals who ranked in the top 50% of the debating skill distribution. A positive coefficient would indicate that ChatGPT has a greater positive effect among participants who had higher baseline debating skills. Again, standard errors are clustered at the individual level.

1.5 Results

In this section I present the results of the preregistered experiment.

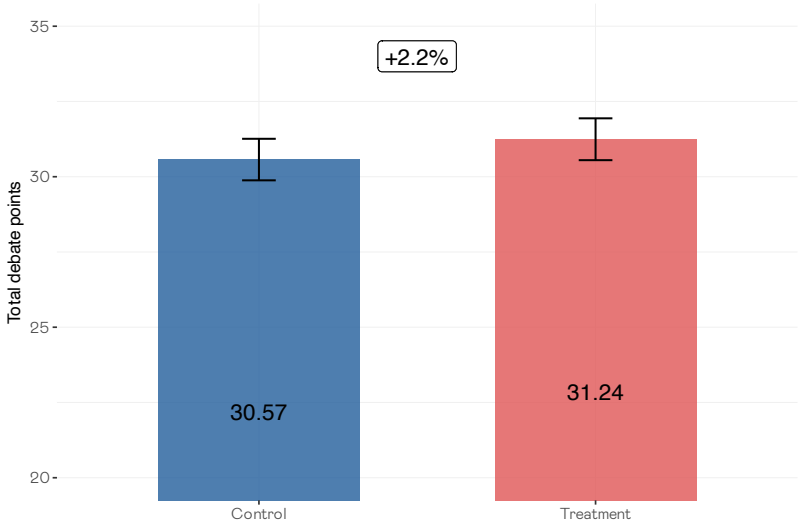
Result 1: Positive but not significant overall effect of ChatGPT on debate performance

Table 1.2 shows the estimated effects of the treatment on debating points and the probability of winning the debate. In Column 1, which estimates equation 1.1 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 0.676, 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 4.2 percentage points, but neither of these estimates is significant at conventional levels. Figures 1.2b and 1.2a summarize these results. They show the average debating points among the control and treatment group and the share of individuals winning the debate, conditional on control variables and debate day and round fixed effects. 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.

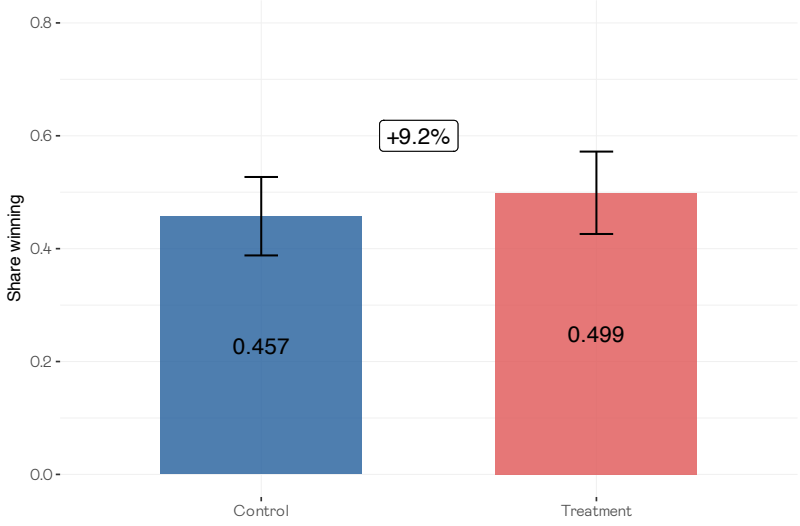
I next exploit the information contained in the judges' evaluations, by looking at points given in the five dimensions of the rubric for debate scoring separately. These dimensions are: (1) clarity and validity of the defended position; (2) the evidence is credible; (3) formal quality of the participant rhetoric; (4) the ability to refute the rival's position and (5) the arguments are superior to those of the rival. In Table 1.3, I show the results of regressing total points in each of these categories on the treatment dummy, debate day and round fixed effects, and controls for demographics (equation 1.1). I find positive, significant and large coefficient estimates for the effect of the treatment in three out

Figure 1.2: Treatment effects of ChatGPT

(a) Total debate points



(b) Probability of winning the debate



Notes: This figure shows the average of the outcome variables for control and treatment individuals, conditional on baseline performance and control variables. These correspond to the predictive margins of the treatment indicator ($Treat$) in equation 1.1. The spikes correspond to the 90% confidence intervals. Above each graph, I show the percent difference in the outcome between treatment and control group.

of five categories: clarity, correctness and validity of the defended position"; "ability to refute the rival's position" and "superiority of the arguments to those of the rival". These three categories have a direct connection with an effective use of ChatGPT to quickly provide and structure arguments. In contrast, the category "Formal quality and rhetoric of the participant (not the content). Convincing use of language", which is more indirectly related to ChatGPT, is positive but not significant. The main exception is "The evidence presented is credible", where the coefficient I estimate is zero. This latter result could be interpreted in line with previous research on LLMs pointing to problems of limited reliability of the information or invented references (Bommasani et al., 2021; Weidinger et al., 2021).

The nature of the debate might be altered by the debating partners' performance, which could be affected by the latter's treatment status. I analyze this by including a control for one's rivals treatment status and the interaction between own and rival's treatment status. The results are presented in Table A2 and are consistent with the findings presented in this section. Column 1 shows the effect on the probability of winning a debate for treatment individuals, conditional on their rival being assigned to treatment or not. The probability of winning a debate is 10.5 percentage points higher for treatment individuals that debated with a control group person. However, the effect is not significant. Furthermore, treatment individuals assigned to debate against another treatment individual have the same probability of winning as control individuals (interaction effect=-0.125). When it comes to 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.

Result 2: ChatGPT increases inequality in debate performance

I now check whether the impact of ChatGPT depends on baseline debating performance. Table 1.4 shows the heterogeneous effects by baseline score, estimated using equation 1.2. Contrary to the ample evidence found in the literature that LLMs work as productivity levelers for a variety of professional tasks (Dell'Acqua et al., 2023; Noy and Zhang, 2023; Doshi and Hauser, 2023), 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 2.15 points (equivalent to a 0.47 standard deviation increase) compared to control students-, while those with lower baseline debate points (bottom 50%) do not benefit at all. The result is summarized in Figure 1.3a. The coefficient estimate represents a performance improvement by 5.2% of using ChatGPT among top 50% performers, and a negative (albeit insignificant) impact of -1.9% of using ChatGPT on bottom 50% performers.

I also use an additional measure of student ability - whether a student is recipient of a merit

scholarship (a scholarship granted based on academic excellence) - and 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 effect. The result is summarized in Figure 1.3b. The coefficient estimate represents a significant improvement by 12% for students using ChatGPT among those on a scholarship, and a 1.5% (insignificant) effect of ChatGPT among those without a scholarship. Taken together, these findings suggest that for tasks that require higher-order skills, GenAI is complementary to ability and may increase inequalities in productivity.[8]

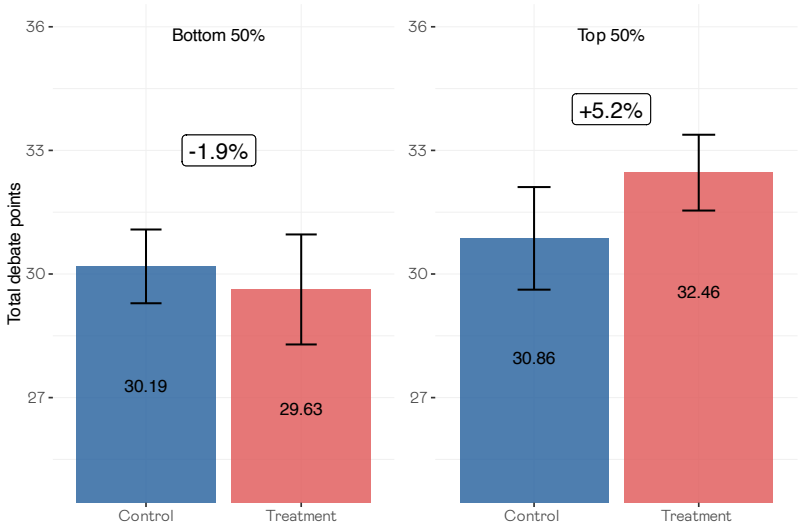
Finally, I exploit the information of the end-line survey on potential mechanisms that might be explaining the results. First, I analyze the overall effects of the treatment on two questions: (1) Did you think we provided enough time to prepare for the debates? - with 0 being equal to totally insufficient time and 100, plenty of time - and; (2) How useful were the AI tools / materials we gave you to prepare the debates? - with 0 being not useful at all and 100 being very useful. As reported in Figure 1.4a and Table 1.6, I find a positive and significant overall effect of ChatGPT on self-reported perception of time needed to prepare the debates: treated individuals self-rate the sufficiency of time 10 points (or 15.6%) higher than control individuals. I find an even stronger effect of the treatment on students' perceptions of the usefulness of the tools used for the debate, shown in Figure 1.4b: treatment individuals value the usefulness of materials 30 points higher (out of 100) than students having only internet access, equivalent to a 69.2% difference between treatment and control group.

To get a better understanding of why only high ability individuals seem to gain in productivity from the treatment, I also study heterogeneity in the above mechanisms. When looking at time sufficiency for debate preparation, shown in Figure 1.5b, the results point to a double effect: students on a merit scholarship in the control group are concerned about not having had enough time (44/100), while non-scholarship students in the control group seem much more (over-)confident (63.19/100). However, the effect of the treatment changes completely the picture: the perception moves to above 76/100 for treatment students with a scholarship (+72.3%), while it only increases to 68/100 for treatment students with no scholarship (non-significant). These results might hint to a relevant mechanism at work: good students self-report higher productivity gains from using the Chat, as

[8]Tables A3 and A4 show descriptive statistics comparing top with bottom baseline performers, and scholarship recipients with non-recipients, respectively. Top performers are similar to bottom performers in age, gender and scholarship status. However, top performers are less likely to have prior debating experience, but are more likely to have won a debating prize before. They are also more likely to have known and used ChatGPT before the experiment, and enjoy speaking in public more than bottom performers. Scholarship recipients are more likely to be female, are less likely to have a father with a Master or higher degree, enjoy speaking in public more and feel more comfortable in the debating language.

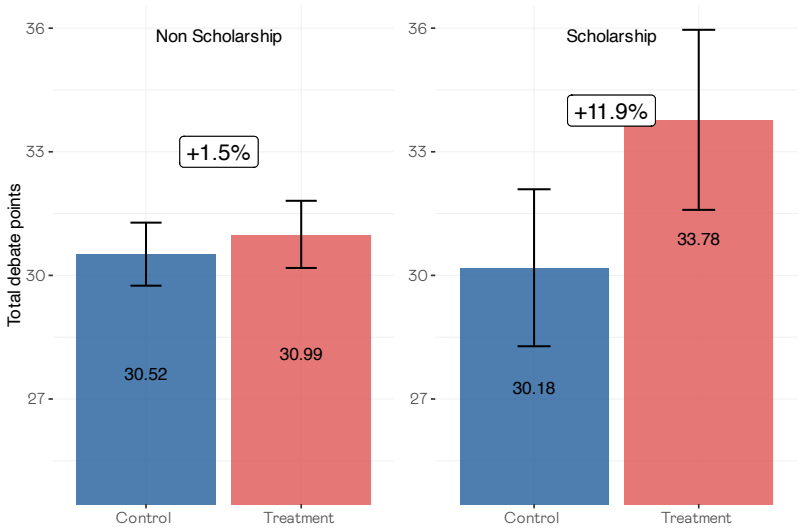
Figure 1.3: Heterogeneity in treatment effects of ChatGPT on total debate points

(a) 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 1.2 by baseline performance), and above the bars I show the percent difference in the outcome between treatment and control group.

(b) By merit scholarship



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 1.2 by baseline performance), and above the bars I show the percent difference in the outcome between treatment and control group.

measured by perceived time needed to prepare debates. High ability individuals seem to be more effective at exploiting Chat GPT to liberate time to maximize their persuasion capacity.

When it comes to participants' perceived usefulness of the tools they were given for preparation, I find that gains are large and similar for high and low ability students (see Figure 1.6): all treatment students feel AI support was very helpful. Taken together, these results indicate that the large gains found for high ability students are not driven by the fact that they find it more useful, but that they are more capable of using AI efficiently. Again, this supports the idea that higher-order skills are complementary to the input provided by ChatGPT.

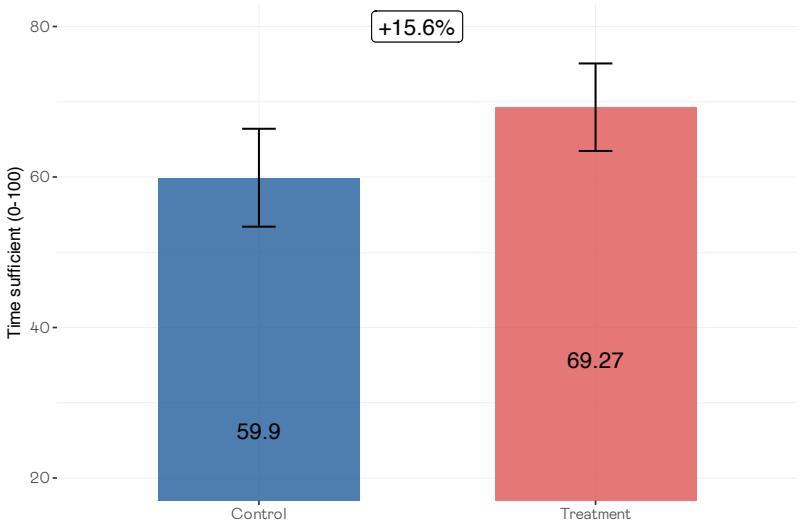
1.6 Conclusion

Since their public appearance in late 2022, Generative AI technologies such as ChatGPT have gone through an unprecedented expansion (UBS, 2023). Scholars and business analysts predict that these systems will disrupt entire industries and transform labor markets. A novelty of those systems, based on LLMs, is that they can perform sophisticated tasks that were unthinkable just a few years ago, such as writing creative texts, providing business ideas, or creating realistic video clips through simple human prompting. Unlike in previous waves of automation or robotization, many of these tasks are common to knowledge workers in a large variety of high-earning occupations (Mollick, 2022). Early experiments testing the impact of ChatGPT on work productivity on a variety of written tasks have shown a common pattern: GenAI systems seem to help low performers more than high performers, thus compressing the productivity distribution.

To my knowledge, this is one of the very few studies testing the impact of GenAI on high-skill tasks in a demanding interactive human context. The beauty of the debating contest is that it provides an ideal setting to test whether GenAI can increase productivity when different higher-order skills, such as rhetoric persuasion or critical thinking, are essential for productivity. I find, contrary to most results in the existing literature, that high-skilled individuals benefit more from the interaction with ChatGPT than low-skilled individuals. I find some evidence that high-skilled individuals benefit more from using the chat and suggest a mechanism 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-skilled 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 important implications for the broader understanding of the potential economic and social impacts of those fascinating technologies. Further research should also explore whether these effects persist with subjects different from students and when using newer versions of

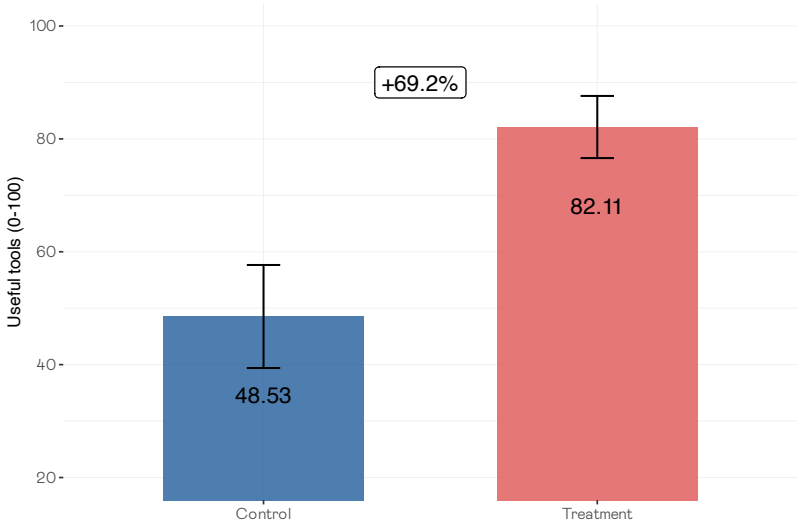
Figure 1.4: Impact of ChatGPT on use of time and tools

(a) Time sufficient



Notes: This figure shows the average valuations of whether time given to prepare debates was sufficient (0-100, with 0 being "not at all" and 100 being "plenty of time") for control and treatment individuals, conditional on debate day fixed effects. The spikes represent 90% confidence intervals (predictive margins of the treatment indicator ($Treat$)), and above the bars I show the percent difference in the outcome between the treatment and control group.

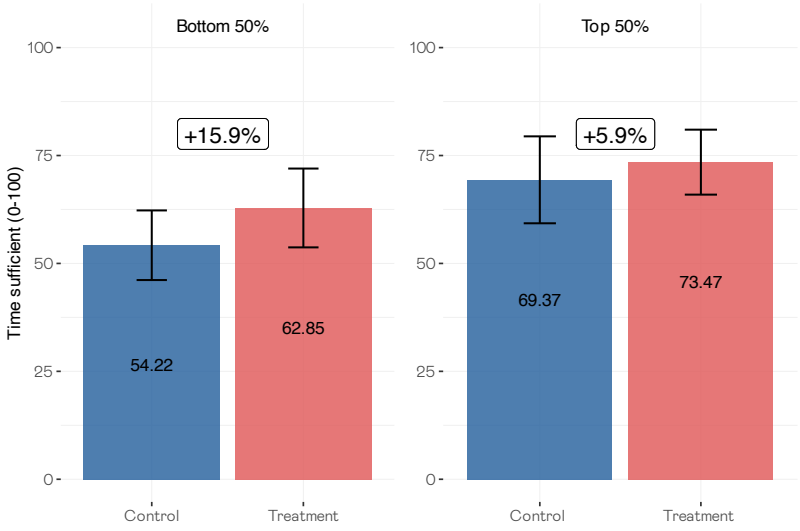
(b) Useful tools



Notes: This figure shows the average valuations of whether the tools given to prepare the debates were useful (0-100, with 0 being "not useful at all" and 100 "very useful") for control and treatment individuals, conditional on debate day fixed effects. The spikes represent 90% confidence intervals (predictive margins of the treatment indicator ($Treat$)), and above the bars I show the percent difference in the outcome between the treatment and control group.

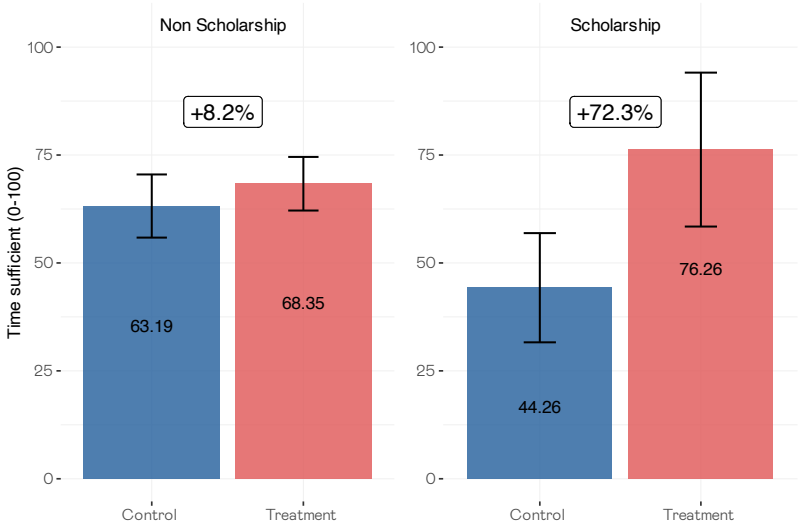
Figure 1.5: Mechanism: Time Sufficient

(a) By baseline debate performance



Notes: This figure shows the average valuations of whether time given to prepare debates was sufficient (0-100, with 0 being "not at all" and 100 being "plenty of time") for control and treatment individuals and by whether the individual scored in the bottom or top 50% at baseline, conditional on debate day fixed effects. The spikes represent 90% confidence intervals (predictive margins of the treatment indicator ($Treat$)), and above the bars I show the percent difference in the outcome between the treatment and control group.

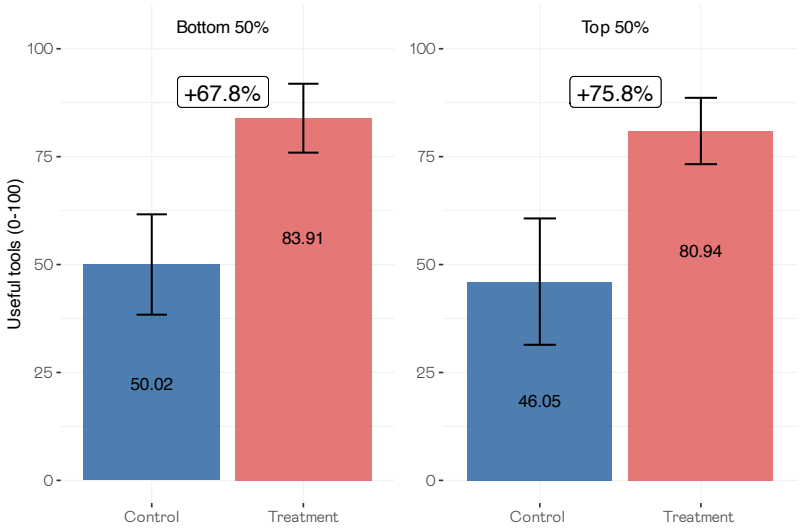
(b) By merit scholarship



Notes: This figure shows the average valuations of whether time given to prepare debates was sufficient (0-100, with 0 being "not at all" and 100 being "plenty of time") for control and treatment individuals and by whether the individual is recipient of a merit scholarship, conditional on debate day fixed effects. The spikes represent 90% confidence intervals (predictive margins of the treatment indicator ($Treat$)), and above the bars I show the percent difference in the outcome between the treatment and control group.

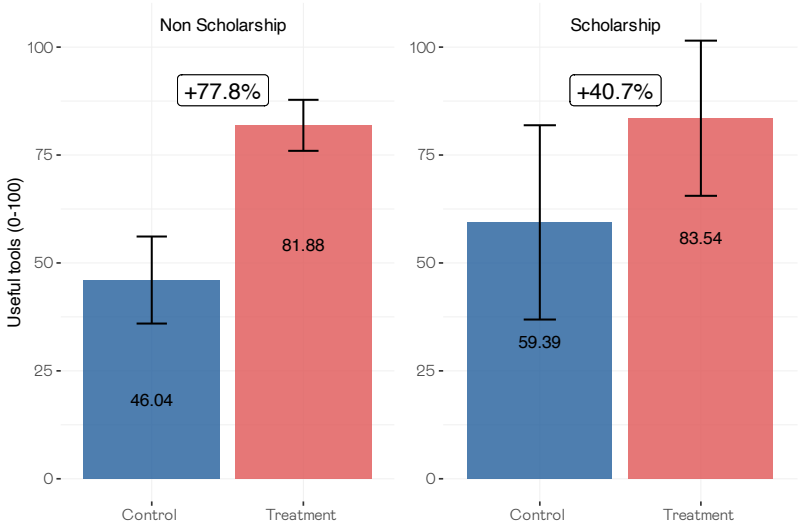
Figure 1.6: Mechanism Useful tools

(a) By baseline debate performance



Notes: This figure shows the average valuations of whether the tools given to prepare the debates were useful (0-100, with 0 being "not useful at all" and 100 "very useful") for control and treatment individuals and by whether the individual scored in the bottom or top 50% at baseline, conditional on debate day fixed effects. The spikes represent 90% confidence intervals (predictive margins of the treatment indicator ($Treat$)), and above the bars I show the percent difference in the outcome between the treatment and control group.

(b) By merit scholarship



Notes: This figure shows the average valuations of whether the tools given to prepare the debates were useful (0-100, with 0 being "not useful at all" and 100 "very useful") for control and treatment individuals and by whether the individual is recipient of a merit scholarship, conditional on debate day fixed effects. The spikes represent 90% confidence intervals (predictive margins of the treatment indicator ($Treat$)), and above the bars I show the percent difference in the outcome between the treatment and control group.

GenAI.

Bibliography

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 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).

Mollick, Ethan, "ChatGPT is a Tipping Point for AI," *Harvard Business Review*, 2022.

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

O*NET OnLine, "O*NET OnLine, National Center for O*NET Development," https://www.onetonline.org/. Accessed: 2024-05-29.

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.

Piaget, Jean, *The Origins of Intelligence in Children*, New York: International Universities Press, 1952.

Salvi, Francesco, Manoel Horta Ribeiro, Riccardo Gallotti, and Robert West, "On the conversational persuasiveness of large language models: A randomized controlled trial," *arXiv preprint arXiv:2403.14380*, 2024.

Slonim, Noam, Yonatan Bilu, Carlos Alzate et al., "An autonomous debating system," *Nature*, 2021, *591*, 379–384.

UBS, "Let's chat about ChatGPT," https://www.ubs.com/global/en/wealth-management/our-approach/marketnews/article.1585717.html 2023. Accessed: 2023-04-26.

Vygotsky, L. S., *Mind in Society: Development of Higher Psychological Processes*, Harvard University Press, 1978.

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.

Weston, Anthony, *A Rulebook for Arguments*, 5th ed., Indianapolis: Hackett Publishing Company, Inc., 2020.

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.

Tables

Table 1.1: Balancing table

	(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 1.2: Effect of ChatGPT on the debate performance

	Debate points		Win	
	(1)	(2)	(3)	(4)
Treat	0.518	0.676	0.039	0.042
	(0.545)	(0.501)	(0.054)	(0.052)
Constant	27.337***	13.860***	-0.242	-0.809
	(1.527)	(4.915)	(0.166)	(0.521)
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	0.10
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 1.1, 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 1.3: Effect of ChatGPT on total debating points by sub-category

	(1)	(2)	(3)	(4)	(5)
	Clarity	Credibility	Rethoric	Refutation	Superiority of arguments
Treat	0.312***	-0.009	0.129	0.201**	0.166*
	(0.080)	(0.187)	(0.091)	(0.101)	(0.089)
Constant	5.213***	0.920	4.063***	1.052	3.783***
	(0.994)	(1.849)	(0.989)	(1.226)	(0.990)
Mean dep. var.	6.10	5.15	6.38	5.77	5.90
SD dep. var.	0.80	1.72	0.82	0.96	0.86
R^2	0.31	0.20	0.29	0.24	0.24
Obs.	364	364	364	364	364

Significance levels are indicated by * $<$ .1, ** $<$ .05, *** $<$ .01. The table shows coefficients from regressions of equation 1.1, 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). All specifications include the same controls as those reported in the notes to Table 1.2. 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. One individual did not complete the end-line survey.

Table 1.4: Effect of ChatGPT by baseline debate performance

	Debate points		Win	
	(1)	(2)	(3)	(4)
Treat	-0.907	-0.558	-0.023	0.003
	(0.753)	(0.799)	(0.073)	(0.078)
Top 50%	0.260	0.679	0.179**	0.214**
	(0.792)	(0.815)	(0.083)	(0.090)
Treat x Top 50%	2.685**	2.151*	0.091	0.036
	(1.112)	(1.262)	(0.110)	(0.118)
Constant	32.330***	21.520***	0.320***	-0.165
	(0.670)	(4.961)	(0.070)	(0.499)
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	0.22	0.05	0.11
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 1.2, 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 1.5: Effect of ChatGPT on total points by whether is recipient of merit scholarship

	Debate points		Win	
	(1)	(2)	(3)	(4)
Treat	0.233	0.477	0.027	0.050
	(0.589)	(0.560)	(0.062)	(0.061)
Scholarship	-1.217	-0.332	-0.203**	-0.108
	(1.043)	(1.058)	(0.094)	(0.097)
Treat x Scholarship	3.815***	3.113*	0.321**	0.194
	(1.411)	(1.687)	(0.125)	(0.160)
Constant	32.965***	22.510***	0.477***	0.132
	(0.706)	(4.794)	(0.074)	(0.545)
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	0.07
Obs.	364	364	366	366
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 1.2, 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 1.6: Effect of ChatGPT on use of time and tools

	Time sufficient (0-100)			Useful tools (0-100)		
	(1)	(2)	(3)	(4)	(5)	(6)
Treat	9.370**	8.636	5.162	33.574***	33.891***	35.841***
	(4.416)	(6.166)	(4.798)	(5.415)	(7.173)	(5.811)
Top 50%		15.155**			-3.962	
		(6.542)			(9.349)	
Treat x Top 50%		-4.540			0.999	
		(8.892)			(10.994)	
Scholarship			-18.929**			13.358
			(7.513)			(12.556)
Treat x Scholarship			26.841**			-11.696
			(11.677)			(15.513)
Constant	47.825***	40.129***	52.002***	50.889***	52.912***	47.694***
	(4.844)	(5.487)	(4.875)	(5.489)	(6.642)	(6.326)
Mean dep. var.	59.81	59.81	59.81	48.48	48.48	48.48
SD dep. var.	0.50	0.50	0.50	0.50	0.50	0.50
R^2	0.15	0.20	0.19	0.28	0.28	0.29
Obs.	137	137	137	126	126	126

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. Columns 2 and 5 additionally control for a dummy equal to one if the individual scored in the top 50% at baseline and its interaction with the treatment dummy. Columns 3 and 6 additionally control for a dummy equal to one if the individual is recipient of a merit scholarship and its interaction with the treatment dummy. The total number of individuals included in each regression differs between the columns because not all individuals answered the questions at endline.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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.

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"	Points given to position "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 tables

Table A2: 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 1.2.

Table A3: 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.

Table A4: 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.

Chapter 2

Online tutoring works: Experimental evidence from a program with vulnerable children

Lucas Gortazar[1] Claudia Hupkau[2] Antonio Roldán-Monés[3]

Abstract

We provide evidence from a randomized controlled trial on the effectiveness of a novel, 100-percent online math tutoring program, targeted at secondary school students from highly disadvantaged neighborhoods. The intensive, eight-week-long program was delivered in groups of two students during after-school hours, mostly by qualified math teachers. The intervention significantly increased standardized test scores (+0.26 SD) and end-of-year math grades (+0.49 SD), while reducing the probability of repeating the school year. The intervention also raised aspirations, as well as self-reported effort at school. The two-on-one design allows us to significantly reduce costs and improve scalability, while showing similar results as one-on-one tutoring programs.

Acknowledgements

We thank seminar participants at LSE-CEP, Universidad Carlos III, Paris School of Economics, the World Bank, the Inter American Development Bank, Universidad Pablo Olavide, and Universidad Autonoma de Madrid. Pablo García-Guzmán, Natalia Collado, Ana Herrero, and Angel Martinez provided excellent research assistance. We thank Antonio Cabrales, Jonathan de Quidt, Luis Garicano, Nagore Iriberri, Matthew A. Kraft, Mónica Martinez-Bravo, Pedro Rey-Biel, and Jenifer Ruiz-Valenzuela for their comments. We are grateful to Empieza por Educar for helping with the implementation of the program, including Miriam Arriola, Zaida Pillado, Lara Crespo, Oscar Ferri, Carla Carmona, Paula Royo, Miguel Costa, Carles Lopez, Luis Oliver, Justo Quintanar, and Beatriz Morilla. We are also grateful to Iria Mata, Dirk Rosquillas,

[1]ESADE Center for Economic Policy and World Bank. Email: lucas.gortazar@esade.edu

[2]Department of Economics, CUNEF Universidad; and Centre for Economic Performance, London School of Economics. Email: claudia.hupkau@cunef.edu.

[3]Department of Social Policy, London School of Economics and Esade Center for Economic Policy. Email: antonio.roldan@esade.edu

Miguel Ujeda, and Gonzalo Romero for their support in various layers of implementation. We are grateful to the tutors and the schools that participated in the intervention. We acknowledge financial support from Esade, Fundación COTEC, Fundación CEOE, United Way Spain, Banco Santander, UNIR, Huawei, Fundación Atlantic Copper, Edvolution and Google. We also want to thank Fátima Báñez and Cristina Garmendia for their essential support in the early stages of the project. This project underwent ethics review by the Research Ethics Committee of the Universitat Ramon Llull and obtained IRB approval from the same university. The experiment reported in this paper is registered at The American Economic Association's registry for randomized controlled trials, AEA RCT Registry ID: AEARCTR - 0007138. This chapter has been published in the Journal of Public Economics. Available online here: https://www.sciencedirect.com/science/article/pii/S0047272724000185.

2.1 Introduction

Intensive, in-person tutoring in one-on-one and small group settings has been shown to have substantial positive effects on learning at moderate cost (Nickow et al., 2020). The Covid-19 pandemic and associated lockdowns, which disrupted education in over 150 countries (Azevedo et al., 2021) and disproportionately affected disadvantaged children (Betthäuser et al., 2023), has brought tutoring programs center stage as a cost effective policy to close educational gaps that have widened during the pandemic.[4]

Most of these programs were and are delivered online. On the one hand, social distancing rules that were in place throughout the pandemic made this necessary. On the other hand, technologies and new habits adopted during lockdowns have made online tutoring more accessible to families from all backgrounds. Yet, very little evidence exists as regards to its effectiveness. Online tutoring has the advantage that it can draw on a larger pool of potential tutors, not limited to local labor markets, and it reduces costs associated with commuting for both tutors and students (Kraft et al., 2022). Compared to in-person tutoring conducted during school hours, where students are typically pulled out of their regular classes, remote after-school tutoring also imposes fewer logistical challenges on schools and teachers in terms of co-ordination of time and space for sessions.

In this paper, we study the effectiveness of an intensive, eight-week math tutoring program on academic and socio-emotional outcomes of secondary school children in Spain. It offered free, 100-percent online after-school tutoring to pupils aged 12 to 15 from very disadvantaged backgrounds. The program, called *Menπores*, has four key features.[5] First, the whole organization of the program and the tutoring sessions were implemented online. Second, the large majority of tutors delivering the program were paid-for, qualified math teachers. Third, the tutoring sessions were done in groups of two students per tutor. Fourth, the program focused on math and social-emotional support (motivation, well-being, and work routines). This focus was chosen because our target population was teenage children aged 12 to 15, and evidence suggests that tutoring in mathematics tends to be more effective for students in higher grades, while literacy interventions have been shown to be more effective in pre-school and primary school settings (Nickow et al., 2020). Further, the focus on socio-emotional support was introduced to mitigate the detrimental effects of the pandemic and associated school closures on children's mental health (Newlove-Delgado et al., 2021), and because of the growing evidence as to the importance of socio-emotional skills in educational attainment and

[4]For instance, in July 2022, President Biden launched the National Partnership for Student Success (NPSS) in the US, a three-year $122 billion federal program to provide high-quality tutoring, summer learning, and after-school programs. In September 2022, the UK launched a new edition of the National Tutoring Program (originally funded with £1bn), offering both face-to-face or online tuition.

[5]The program is called *Menπores*, with the Greek letter π used as a reference to mathematics. However, it is pronounced as "mentores", the Spanish word for mentors.

future labor market outcomes (Heckman et al., 2006; Kosse et al., 2020; Kosse and Tincani, 2020; Eisner et al., 2020).

We implemented the program in partnership with *Empieza por Educar* (ExE), the Spanish branch of Teach for All, an NGO specialized in training young teachers working in schools attended by vulnerable and low-income students. The recruitment of program participants was done in two steps. First, we identified a number of schools that showed interest in the program. Second, we asked principals and teachers in participating schools to identify students most in need for support in math and disseminate the program among them and their families. Among all students who signed up, we randomly assigned slightly more than half to the program. Randomization was blocked by classrooms to increase the power of our experimental design. This also ensured that students who ended up in the same group knew each other. Within blocks, treatment students were randomly divided into groups of two, and were subsequently randomly assigned to a tutor.

We collected a rich array of child and family characteristics, such as prior attainment, family size, household income, and immigration background, at the stage of online registration. We ran base- and endline surveys of pupils, which included a standardized math test and questions on socio-emotional well-being, aspirations, and past performance. To minimize attrition, these surveys were run during regular math class among all pupils in classrooms with participating students. At the end of the program, we also ran a parent survey to collect information on academic results, such as the final math grade, whether the subject was passed, and whether the school year had to be repeated. We also collected very rich real-time data throughout the duration of the program capturing participation, connection time, and quality of the connection.

Our first set of results is based on the in-class test and student survey. Using our standardized math test, which was graded externally, we find an intention-to-treat (ITT) effect of the intervention of 0.26 SD, which is significant at the 10 percent level (p-value: 0.077). To put these numbers into context, Guryan et al. (2023)'s evaluation of high-dosage, two-on-one math tutoring (60 mins/day during the entire school year) for 9th and 10th graders in Chicago high schools finds ITT effects between 0.09 and 0.14 SD for math test scores and reductions in the likelihood of failing the course by between 15 and 24 percent.

In terms of non-cognitive outcomes, we find that the program raised students' aspirations: students in the treatment group were 13.5 (p-value: 0.022) percentage points more likely to state that they would like to go onto the academic track after compulsory schooling (i.e. *Bachillerato*), equivalent to a 31 percent increase compared to the control group mean. This result is important because aspirations have been shown to positively affect future educational achievement (Khattab, 2015), and because attending the academic track at upper secondary school is linked to higher earnings

later in life thus potentially increasing social mobility (Matthewes and Ventura, 2022). We do not find a positive impact on stated intentions to go to university, possibly because the decision to go to university lies too far away in the future for the students in the intervention (the average age of participants was 13).

The training of tutors had a particular focus on student motivation, which tutors were meant to foster using the growth mindset approach developed by Dweck (1986). This approach is based on the idea that when effort is valued over success and teacher feedback is specific, describing the praised behavior rather than simply affirming a correct answer or giving feedback about the person's ability (Dweck, 1999), this will positively affect student effort, motivation, perseverance, and ultimately academic achievement (Chalk and Bizo, 2004). We find that students assigned to treatment were 11.4 percentage points (p-value: 0.064) more likely to state that they exerted high effort always or most of the time at school, which corresponds to an increase by 18 percent when compared to the control group mean but is not robust to multiple hypothesis testing. However, we do not find an impact on student's motivation for school. We neither find an effect on perseverance measured by the grit scale developed by Duckworth and Quinn (2009). It is likely that our program was too short to be able to change this outcome. In fact, recent research suggests that grit is a highly heritable personality trait with limited malleability (Rimfeld et al., 2016).

One of the objectives of the focus on motivation and the growth mindset was to foster in students the belief that ones conduct and actions influence the result obtained (also called internal locus of control), as opposed to feeling a lack of control over the environment and circumstances, making any effort useless because ones own actions cannot change the situation or outcome (external locus of control). Contrary to our hypothesis, we find that students assigned to treatment show a more external locus of control than control students (p-value: 0.081), but this result is not robust to multiple hypothesis testing. This effect is driven by an increase in the probability to agree that, when bad things happen in their lives, it tends to be the fault of others, among students assigned to treatment. A possible interpretation of this result is that the program reduced self-blame among individuals in the treatment group - students that may have believed until then that the fact that they are low achieving is entirely their own fault.

Since our program had a focus on math and was targeted at very low performing students in this subject, we expected the intervention to have a positive impact on self-perceived math competencies or the likelihood of stating they like mathematics. However, we find no such effect. These results are surprising in light of the positive impact of the intervention on actual achievement (both externally graded math tests and teacher-assessed outcomes). Given the positive relationship between perceived ability and outcomes (Spinath et al., 2006), the failure to raise students' self-image in mathematics

may have limited the potential longer-term effects of our program. To check whether there were spillover effect on other subjects, we also asked whether students liked more and felt more confident in Spanish. It is possible, for instance, that as a student becomes better in math, they gain a comparative advantage in that subject and lose interest in other subjects. It is also possible that improved results in math could motivate students overall and make them more motivated for other subjects. However, our results show no such spillover effects.

Given the post-pandemic context of the intervention, with students likely still being affected negatively in terms of mental health (Newlove-Delgado et al., 2021), we hypothesized that the intervention might have a positive impact on student well-being due to the positive group dynamics and the presence of an adult reference as a tutor and mentor (Kosse et al., 2020). However, we do not find an impact on overall well-being. Yet, when looking at one of the questions included in the well-being index separately - satisfaction with school - we do find a relatively large coefficient estimate equivalent to a 0.22 SD increase (p-value: 0.077), which is however not robust to multiple hypothesis testing. In sum, while the program was not successful at raising well-being overall, it seems to have increased satisfaction in the dimension most closely linked to the context of the intervention: school.

The second set of results is based on the parent-survey. We will address concerns about selective attrition for outcomes from this survey further below. We find a positive and significant ITT-effect of the program on end-of-year parent-reported math grades of 0.85 points (on a 1 to 10 numerical scale), equivalent to a 0.49 SD increase. Further, we find a significant increase of about 30 percent with respect to the control group mean in the likelihood of passing the math course (also parent-reported). Further, we find a large and significant effect on grade retention: the program decreased the likelihood of repeating the school year by 8.9 percentage points, equivalent to a 74 percent decrease with respect to the control group, which had a repetition rate of 12 percent. We also provide suggestive evidence that the positive effects of the program are persistent one year after the end of the program.

While attrition for the in-class survey was low - the response rate was 88 percent - and equal for the treatment and control group, it was more pronounced and 13 percentage points higher for the control group in the parent-survey, on which we rely to measure end-of-year academic outcomes. This raises concerns about the internal validity of our estimates and could bias our results. For instance, experimenter demand effects might cause parents of treated children to report more positive results. We address this concern in different ways and show that results on parent-reported outcomes are robust to using inverse probability weights and provide bounds to our estimates using Lee (2009)'s and Behaghel et al. (2009)'s approach. We find that bounds only include positive impacts on the final math grade (ITT-estimate=0.852, bounds=[0.304,1.317]), whether the student passed the subject

(ITT-estimate=0.205, bounds=[0.143,0.357]) and only negative impacts on whether the student had to repeat the school year (ITT-estimate=-0.089, bounds=[-0.270,-0.060]). This gives reassurance that when taking into account sample attrition, the main conclusions from our analysis continue to hold.

Analysis of mechanisms suggests that the program was more effective for higher achieving students at baseline. This is consistent with results in Guryan et al. (2023), who find that their program had positive treatment effects on math test scores for all but the bottom quartile in baseline achievement. We find that when the tutor and the student were of the same gender, the impact on standardized test scores is slightly higher (although imprecisely estimated), possibly because students felt more similar to their tutor (as has been shown for instance by Dee (2005)). Results on parent-reported academic outcomes also tend to be slightly higher when students were matched to someone of their own gender in the group. Contrary to existing evidence (e.g. in Duflo et al. (2011)), we do not find that the ability match in the group mattered for the impact of the program. While our lack of statistical power does not allow us to draw strong conclusions from this analysis, it provides suggestive evidence that should be investigated further in future research.

Our study contributes to the understanding of whether online tutoring can work as an effective tool for closing learning gaps for disadvantaged students. The closest to our research is the online tutoring program implemented in Italy in Spring 2020 by Carlana and La Ferrara (2021). They find large positive effects on student achievement (+0.26 SD) and positive effects on socio-emotional skills, aspirations, and psychological well-being. Kraft et al. (2022) also implement an online tutoring program for middle school students with college volunteers. They find positive but insignificant effects on math and reading.[6]

Our program departs from these studies in three fundamental ways. First, they were delivered by volunteer university students, while *Menπores* used mostly paid-for, qualified secondary school teachers. Second, our tutoring was implemented in groups of two students, instead of one-on-one. Third, and more importantly, these programs were implemented in exceptional circumstances. For the case of Carlana and La Ferrara (2021), the program took place during the harshest lockdown period in Italy from April to June 2020 (when all kids were at home and schools were closed).[7] In the case of Kraft et al. (2022), in early 2021 in the US, when schooling was still highly disrupted. Our

[6]Before the pandemic, a sizable amount of research was dedicated to understanding the effectiveness of educational software tools and online learning for university students (Escueta et al., 2020). During the pandemic, some authors explored the effectiveness of different remote learning methods, such as online peer mentoring to support university students (Hardt et al., 2022; Kofoed et al., 2021) or parental educational support through phone calls and text messages (Angrist et al., 2022). However, none of these studies analyzes the effects of online real-time tutoring between teachers and secondary school students.

[7]The Italian Statistical Institute estimates that around 3 million Italian students aged 6-17 may not have been reached by remote learning during the lockdown (Instituto Nazionale di Statistica, 2020).

program, instead, was implemented one year after the onset of the pandemic, several months after schools were fully re-opened in Spain. In that sense, we believe our results show the effectiveness of online tutoring in normal times, when tutoring can be considered a complement rather than a substitute for regular schooling.

Our contribution is relevant both in terms of policy and for further academic research. Governments are investing large amounts of money in tutoring programs (both in face-to-face and online formats). Our evidence suggests that this money is well spent. The intervention costs approximately €300 per student, and has a positive impact of 0.26 SD on our standardized math test, translating into a 0.087 SD increase per €100 spent.[8] This compares favorably with summer schools analyzed in Cooper et al. (2000), with a cost-effectiveness of 0.066 SD per €100 spent (based on an impact of 0.23 SD and a cost of €350 per student). It also compares favorably with increasing instruction time by one hour per day, which according to Higgins et al. (2012) costs €1,020 for an increase of 0.24 SD in test scores, resulting in a cost-effectiveness rate of 0.0235 SD per €100 spent.

Regarding potential future scaling up, we would expect our results to be replicable at a larger scale, provided students have devices and internet connections, which is more likely in developed countries. The main limitation to reproduce such good results at scale is likely to be the availability of high quality tutors.

In terms of costs, programs with paid-for professionals are more expensive than programs with volunteers. However, at large scale, volunteer programs are likely to face more practical and political economy limitations than programs with paid tutors. First, availability of large amounts of volunteers is likely to be a significant limitation in normal times. Second, large government-supported tutoring programs with unpaid workers are likely to encounter resistance from teacher unions, at least in advanced economies. Third, paid work is likely to generate higher engagement and lower tutor turnover. Indeed, our monitoring data shows that volunteer tutors delivered on average three fewer sessions and 200 minutes less of tutoring than our professional, paid-for tutors. Our innovative two-on-one online design offers additional cost savings in relation to in-person programs and one-on-one online programs, while achieving very similar results.

The rest of the paper is organized as follows. Section 2.2 describes the context of the intervention. In Section 2.3, we present the study design and in Section 2.4 we describe the data. The empirical strategy is presented in Section 2.5, and results and robustness checks are shown in Sections 2.6 and 2.7, respectively. Section 2.8 concludes.

[8]The cost of €300 per student is based on the following calculations derived from the project implementation: Every group of two students received up to 24 hours of tutoring, hence the direct cost per student in terms of tutor wages is the compensation for 12 hours of tutoring time per student. Tutors were paid at 19 euros per hour, including social security cost, resulting in wage costs of €228 per tutored student. The cost of training (including an online course and two live webinars), administrative and supervision costs amounted to approximately €70 per tutored student.

2.2 Context of the intervention

Our intervention took place in two large regions of Spain, Madrid and Catalonia. In both regions, schools were largely back to normal after the pandemic at the time our intervention took place: On March 9th 2021, just before the start of our intervention, only 0.5 percent of classes in Spain were operating remotely due to quarantines.

In primary school and the first two grades of lower secondary school (Grades 1 to 8, ages 6 to 13), the relevant years for our study, classes had been operating under a face-to-face model since September 2020. In order to guarantee social distancing, class sizes were slightly reduced. To avoid additional physical contact between students, break times, lunch times and extra-curricular activities were minimized or eliminated. This meant that some of the students in our study potentially had up to two hours more time outside school in the afternoons compared to the pre-pandemic scenario. The number of hours of instruction, however, remained the same as in any other regular year.

To cope with the various learning models and anticipate potential future school closures, the Ministry of Education and Vocational Training and regional ministries made large efforts to provide schools with tablets and computers for the school year 2020/21. The Autonomous Community of Madrid, for instance, invested more than €6.1 million (or $6.9 million) in 36,100 tablets for their schools (Comunidad de Madrid, 2020). Because schools lent these devices to students who did not have access to a computer or tablet, only a very small share (6 percent) of students who enrolled in our program did not have the technology at home to attend online tutoring sessions. We supplied these students with tablets that were later donated to their schools.

2.3 Study design

In this section we describe the intervention design, recruitment of participants and tutors and the timeline of implementation.

2.3.1 The program *Menπores*

Our online tutoring program, called *Menπores*, was an intensive intervention consisting of three 50-minute sessions per week over a period of eight weeks. The target population were students in Grades 7 and 8 (grades 1 and 2 of secondary school, students aged 12 to 15), attending schools in highly disadvantaged neighborhoods. We chose this target for two reasons. First, disadvantaged students were disproportionately affected by learning loss during the pandemic (Haelermans et al., 2021; Blainey and Hannay, 2021) and most likely to benefit from the intervention. The need to invest and experiment with remedial programs which could facilitate catch up for the learning loss of these students was and still is a priority in education policy in many countries (World Bank,

2021a,b). Second, evidence suggests that tutoring in mathematics tends to be more effective for students in higher grades (Nickow et al., 2020), and budget, logistical and time constraints meant that we could deliver tutoring only in one subject area and only in secondary schools.

Tutoring sessions were delivered online mostly by qualified math teachers in groups of two students per tutor. We decided to concentrate hiring efforts on qualified math teachers for several reasons: First, existing evidence on face-to-face tutoring shows that they are significantly more effective than non-professionals or volunteer tutors (Nickow et al., 2020). Second, while we had initially planned a second treatment arm with tutoring delivered by volunteer university students as in Carlana and La Ferrara (2021), we were neither able to recruit sufficient participating students nor sufficient volunteer tutors in the short time-frame we were operating in.[9] The timing of our intervention (towards the end of the academic year, when university students tend to be more busy because of final examinations) and the fact that life in Spain had largely gone back to normal by March 2021 (students were no longer locked inside their homes as they had been between March 2020 to May 2020) are possible explanations for the low response to our call.

The group composition was fixed throughout the program, with the same students attending meetings with the same tutor in each session. The students in each tutoring group of two were from the same class or grade from the same school. This was done in order to increase the power of our experimental design as well as to guarantee that students knew each other and would find it easier to connect and accommodate. We decided to go for a two-on-one student-tutor ratio for three reasons. First, the pedagogic team in charge of implementation suggested that being in a group with another child had the potential to generate mutual motivation and peer pressure not to abandon the program. Second, existing evidence for face-to-face programs in Nickow et al. (2020) shows that two-on-one tutoring is nearly as effective as one-on-one tutoring. Moreover, evidence from a two-on-one math tutoring program in Chicago (Guryan et al., 2023) shows that this design can be highly effective even for older (secondary school) children. Third, this design is relevant from a scalability perspective, as it significantly reduces cost per student.[10]

A key element of the program was its online nature. The fact that face-to-face interactions outside the classroom were severely constrained by social distancing rules (avoiding breaks, lunch at school or extra-curricular activities) made this the only viable option. Additionally, the demand and interest in online tutoring has surged rapidly since 2020, while to date very limited evidence on its effectiveness exists.

[9] Like Carlana and La Ferrara (2021), we launched a call searching for volunteer math tutors at five large public and private universities in Barcelona and Madrid, but received less than 50 applications.

[10] We decided not to go for a three-to-one ratio as we thought it would have been exceedingly challenging from a logistic point of view to coordinate four people to be available at the same time three times a week.

2.3.2 Content and methodology of the tutoring sessions

We designed the academic and pedagogic content of the intervention together with *Empieza por Educar* (ExE), the Spanish partner of the US based network *Teach for All*. ExE is an NGO specialized in training young teachers working in schools attended by highly vulnerable and low-income students in the regions of Madrid and Catalonia.[11] Its core activity is based on a highly selective model of teacher training, identifying teacher candidates with top academic and socio-emotional skills that are relevant for the teaching profession, as well as an interest in the profession and in social change.

The academic content of the tutoring program was based on the national mathematics curriculum and covered the expected knowledge from 1st and 2nd graders in secondary schools in Spain. Additionally, the program aimed at providing psycho-social and socio-emotional support to students. This was done for several reasons. First, to potentially mitigate the detrimental effects of the pandemic and associated school closures on children's mental health (Newlove-Delgado et al., 2021). Second, there is growing evidence as to the importance of socio-emotional skills in educational attainment and future labor market outcomes (Heckman et al., 2006; Kosse et al., 2020; Kosse and Tincani, 2020; Eisner et al., 2020). There was therefore an explicit mandate for tutors to spend time in the sessions providing such support and reserve at least ten out of the 50 minutes to discuss any issues, fears or concerns the children might be facing at home or at school. The pedagogical approach of the sessions was inspired by the *No Excuses* methodology, which has been shown to be effective in raising academic and non-cognitive outcomes in the context of urban US charter schools for vulnerable children (Dobbie and Fryer, 2013). This methodology emphasizes high expectations, increased instructional time, individualized support, continuous feedback and intensive data collection on student progress to guide instruction. We provide details on how tutors were trained in these aspects in Section 2.3.4.

The tutoring program was also aimed at improving student motivation through the growth mindset approach developed by Dweck (1986). In this approach, effort is valued more than success and teacher feedback is aimed at describing the praised behavior rather than simply affirming a correct answer or giving feedback about the person's ability (Dweck, 1999). This in turn is meant to positively affect student effort, motivation, perseverance, and academic achievement (Chalk and Bizo, 2004).

[11] Every year, ExE selects around 80 candidates out of an applicant pool of between 2000 and 3000 to receive training and support during the two years they work in such schools in pedagogy, classroom management, school and community transformation and leadership skills.

2.3.3 Recruitment of schools and participants

The recruitment of program participants was done in two steps. First, we identified a number of schools that showed interest in the program. Second, we asked schools who had agreed to participate to identify potential beneficiaries from their pool of students and disseminate the program among them.

For recruitment of participant schools, we leveraged ExE's large network of teachers and schools in the regions of Catalonia and Madrid. School principals were initially contacted by ExE and informed about the program and its characteristics, its target population (disadvantaged students in the 1st and 2nd grade of secondary school and lagging behind in mathematics), and the fact that the program was to be evaluated scientifically through a randomized controlled trial.

Emails were sent and calls were made to gauge interest to around 32 schools. We had calculated that in order to reach our target sample size of 400 enrollments, we would need to get about 20 schools on board.[12] After the initial emails and calls, recruitment efforts were intensified among those schools that showed an interest (i.e., those that replied to emails or answered calls and consulted with the governing bodies of their schools to see whether there was support for participation). Recruitment ended when the target number of schools had been reached, as time, budgetary and operational constraints meant that we could not deliver tutoring to more than about 200 students. Eventually, 18 schools signed participation agreements.[13]

For the selection of potential participants, there were no strict eligibility rules. Instead, we relied on the knowledge of teachers and principals to identify four to six students per classroom that were most in need for math tutoring.

In the second step, parents of children identified by the school as in need were directed to an online registration form. It is notoriously hard to reach lagging behind, disadvantaged students and their parents for opt-in programs (Robinson et al., 2022). We therefore asked both schools as well as coordinators from ExE to actively help parents fill out the registration form to ensure we reached our target population. The online registration form included an information sheet for parents and children, informing them of the fact that the program was to be evaluated and that not all students that registered would eventually be selected. Parents were also asked to give consent for their children's participation and the usage of data for research purposes. In the registration process we collected detailed data on household and student characteristics and whether the student that

[12] On average, there are two classrooms per grade level, the intervention was targeted at two grade levels, 7th and 8th grade, and we expected enrollment of between 4 and 6 students per classroom at most, which makes an expected enrollment of 2 $grades \times$ 2 $classes/grade \times$ 5 $students \times 20 = 400$.

[13] Principals signed an agreement detailing the school's role in the study, including: (i) the identification of a group of students that would benefit most from the program; (ii) dissemination of the application material among these students and their families; (iii) ensuring the administration of baseline and endline surveys during school hours; and (iv) participating in a final survey themselves.

was being registered had access to a tablet or other device to participate in the online sessions.

2.3.4 Selection and training of tutors

Our implementation partner ExE designed and implemented the selection and training for tutors based on their longstanding experience with teacher selection. A key criterion for selection was to hold a post-graduate (Master's) degree in Teacher Training in a scientific specialization (math, physics, chemistry or biology), which is a formal requirement to teach mathematics in secondary education in Spain. While holding a Master's degree in teacher training was a desired characteristic, it was not binding. Other skills, such as motivation for the program, having taught in low-income schools, and prior teaching experience, were also considered. Advertisement of the positions was done through various channels, including online hiring portals, ExE's own network of current teachers and alumni, and other teachers whom they work with. A total of 199 applicants which met the minimum pre-requisites were sent a formal application form, and applied. Out of these, 110 candidates were sent a link for an online interview. Out of the 110 candidates interviewed we hired 37 professional tutors. In parallel, we recruited a small number of university students as volunteer tutors and ended up including eight such tutors in the program.[14]

Before the start of the program, tutors received between 15 to 20 hours of online training through ExE's teacher training platform. Training included two remote training modules and two online webinars with expert teachers. Training focused on the following key areas: how to establish strong ties with students, student motivation, lesson planning, learning verification and formative assessments, math academic content knowledge and tutoring methodology.

2.3.5 Timeline

Figure 3.1 shows the timeline of the intervention. Planning and design took place between January and March 2021, and the registration period for parents and children started in early March 2021, lasting for about two weeks. A total of 375 complete registrations with valid consents were received during this time window. After registrations were closed, baseline tests and surveys were administered in all participating classrooms, that is, in all classes where at least one student had registered

[14] As mentioned previously, we had initially planned a third treatment arm with volunteer mentors only. Although we advertised the program at five large public and private universities, we only received 50 applications. We attribute the small number of applications to the fact that the program required a high level of time commitment and coincided with end of term examinations at university. This was also the reason why most of the applicants finally decided to drop out of the process before the start of tutoring sessions. After initial screening and interviews, we were able to include only eight volunteer tutors, who completed the entire application process. We decided to keep these tutors in our pool and included them in the randomization. This allowed us to fulfill the initial commitment to schools to provide tutoring to around 200 students. We include students tutored by volunteers in all the results presented. Results are very similar when students that were taught by volunteer tutors are excluded. We discuss these in Section 2.7.3.

for the program.

Students were randomly assigned to treatment and control group, and in case they had been selected to be in the treatment group, to a partner and tutor, during the Easter break (early April 2021). The tutoring sessions started in the second week of April 2021 and ended in early June, at the time where the final grade evaluation takes place.

Endline tests and questionnaires to students were administered after the end of the intervention and before the end of the academic year (second week of June 2021). We also asked tutors, principals and math teachers to complete brief online surveys at the end of the program. Finally, we administered an online and phone survey to parents during the month of July 2021.

2.3.6 Experimental design and randomization

The experimental strategy relied on over-subscription. No compensation for students not assigned to the treatment was offered, as at the time of the randomization we did not have funds available that could have covered the cost of a second round of the program at a later stage.

Randomization was done in various steps. First, we assigned the initially 375 students who enrolled in the program randomly into treatment (205 students) and control group (170 students). Randomization was at the person level in blocks, where a block consisted of all students of a class at a school that had signed up for the program. When the number of students from the same class who enrolled was two or less, we combined classrooms of the same grade level within the same school into one block. We did this in order to get blocks of sufficient sizes to assign an even number of students within each block to the treatment group. The total number of blocks was 68, distributed across 18 schools. In a second step, we randomly ordered treatment students within each block and assigned them sequentially into groups of two. For instance, if a given block had four treatment students, students one and two in the random order were assigned to the same group, and students three and four to another group.

In the last step, we randomly assigned tutors to groups of two students. In general, all tutors were assigned to three tutoring groups, hence providing support to six students. Volunteer tutors were assigned only one group. Randomization of tutors was stratified by geographic area, where those tutors based in Catalonia who indicated they spoke Catalan were assigned to students based in Catalonia, and those based in Madrid to those who were based in the region of Madrid.

2.3.7 Implementation

Students and tutors were able to organize their own schedule and agree on weekly meeting times.[15] Each student and mentor received personal and unique credentials for accessing a specifically created domain within an online platform from a large, US-based technology firm, consisting of a tool to organize emails, calendars, files and most importantly, hold online meetings. Tutors had to hold sessions through the platform and could only communicate with students through this channel.[16] Students who registered and stated they did not have access to a computer or tablet and/or internet were provided with a tablet with internet access for the duration of the program. In total, 13 students were given tablets, which were donated to their schools at the end of the program.

A key advantage of the online format was that student attendance could be monitored in real time. Throughout the program we collected data for each tutoring session via a management and monitoring dashboard that was fed with data from the technological platform where the virtual sessions were taking place. This data allowed us to immediately identify issues with the connection and quality of video calls and pupils who did not attend their sessions. With this information we could draw up plans of action with tutors, families, and schools to help get them back into the program.

Figure 2.2 shows the distribution of total minutes and total number of tutoring sessions attended by students. Only seven students (3.4 percent of those assigned to the treatment group) actively dropped out of the program before it began. Among students assigned to treatment, the median number of minutes of tutoring received was 952, representing 80 percent of the target number of minutes (1200). The median number of sessions attended was 20, corresponding to 83 percent of the envisaged number of sessions (24).[17]

2.4 Data

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

[15] Students and tutors were asked and had to confirm at the registration and application stage, respectively, that they were available at least three days a week between 4 p.m. and 7 p.m.

[16] This was done both for organizational as well as for legal reasons of child protection: all communication through these channels could be monitored by us and the implementation team.

[17] Medians are calculated on the entire sample assigned to treatment, including zeros for the seven students that dropped out before the start of the intervention. Among those that did start the program, there are six students for whom we could not match the online meeting data and consequently have no information on the number of sessions attended or minutes of tutoring received.

2.4.1 Baseline information

We collected a rich array of family and household characteristics at the stage of online registration for the program, where parents had to fill out a detailed survey. This survey included questions on household composition, civil status of the respondent (the mother or father), education level and household income, as well as the origin of the respondent and the child that was being registered for the program, and the language typically spoken at home. We also asked whether the child was receiving tutoring support of any kind and whether the child had a device (computer or tablet) and internet connection available at home with which to connect to the sessions.[18]

After the completion of the registration period and before randomization, we ran a baseline student survey that included a math test and a questionnaire on prior attainment, well-being, and other socio-emotional outcomes. The baseline test was completed by all students in classrooms where there was at least one student registered for the program, thereby avoiding stigmatization or association of the test with the program. The tests were paper-based and administered by the children's math teachers or their main classroom teachers (also called *tutor* in Spanish) during a regular math class or the weekly lesson reserved for general matters.[19] Because of the timing of the baseline test - right before the Easter holiday and after grading for the first term had finished - students were not missing regular math content to do the test. We explicitly instructed teachers not to mention the program *Menπores* while they ran the tests, so that students who had registered would not associate this assessment to their likelihood of being selected.

Because there is no official standardized test for the age groups included in the program (grade 7 and 8), we created our own assessment. Together with ExE experts with experience as secondary math teachers, we designed two math tests based on the national curriculum for the respective grade levels. Sample questions are shown in Appendix A.1. The test for 7th graders included seven questions, while the test for 8th graders included six questions.

The second part of the survey, which covered well-being, socio-emotional skills, and prior attainment, was identical for both grade levels. Well-being questions were based on the well-being module in the age 14 survey of the Millennium Cohort Study (University College London et al., 2020). Students were asked six questions on how they felt about different aspects of their lives, which they had to rate on a scale of 1 to 7, where '1' meant not at all happy and '7' meant completely happy. The exact questions can be found in Appendix A.2. We calculate the average scores across the six items to create a Likert-type well-being scale.

[18] As noted in Section 3.3, having a device and internet connection was not a pre-requisite for participation, as we provided internet-enabled devices to students who did not have one at home.

[19] In Spanish secondary schools, all classes have one hour per week reserved for a class with their *tutor* in which they discuss general matters.

The second set of questions comprised three items from the CARALOC Pupil Questionnaire (University College London et al., 2021), which assesses locus of control, and are answered with 'yes' or 'no', which we assign value zero and one, respectively. We calculate the average across the answers to these three questions, where a number closer to one indicates a more internal locus of control. An internal locus of control indicates that students believe they are more in control of the results of their actions in their daily lives. A more internal locus of control has been found to be associated with better academic outcomes (Shepherd et al., 2006). Additionally, we asked students to self-assess their ability in Spanish language and math. Finally, we asked students whether and how often they had attended online classes during the school closures from mid-March to June 2020, to be able to control for potential learning losses experienced during the onset of the pandemic.

2.4.2 Outcome measures

During the second week of June, when tutoring sessions had finalized, we administered an endline survey. The endline survey again contained a standardized math test and also included the questions regarding well-being, locus of control and self-rated ability discussed in Section 2.4.1.

We added new questions on socio-emotional skills and aspirations. Socio-emotional skills were captured in several dimensions. First, as in Carlana and La Ferrara (2021), we measured grit using the Short Grit Scale developed by Duckworth and Quinn (2009). This includes eight questions with a 5-point scale, which are then aggregated into an overall Likert-scale by averaging the valuations across all questions. The exact questions can be found in Appendix A.2. Second, we measured school motivation using three items from the school motivation grid of the sixth wave (age 14 survey) of the Millennium Cohort Study (University College London et al., 2020). These covered the frequency with which students (1) exerted high effort at school, (2) thought school was interesting, and (3) they found school a waste of time, with answers ranging from 1 (never) to 4 (always). We look at these outcomes individually and also aggregate them into a school motivation index by adding the values for each question and dividing it by the maximum sum (12). We also ask questions on interest in language and math. To measure aspirations, we ask about the plans students have after completing compulsory schooling at age 16 (vocational track, academic track or dropping out of school), as well as their intentions to go to college.

We also conducted an online and phone survey of parents of study participants in early July, when the school year was over. Parents were asked two key questions about their child's academic outcomes: The final math grade, measured on a 1 to 10 numerical scale, obtained by their child at the end of the year, and whether the child would have to repeat the school year.[20] We also asked

[20]We could not obtain administrative data from the schools for these outcome measures for legal reasons.

about whether their children had received any other remedial education support program (besides Menπtores in case of the treatment group).

2.4.3 Sample and balancing

Our randomization sample consisted of 375 students whose parents had registered through the online form and provided consents and background characteristics on the child and the parents.

Balancing between treatment and control group characteristics of the randomization sample is shown in Table 2.1. The table shows the control group mean and standard deviation (column 2), and the treatment-control difference estimated from a regression of the characteristic as the dependent variable on a treatment dummy and block fixed effects (column 3). Finally, it shows the normalized difference between treatment and control group, equal to the coefficient reported in column 3 divided by the standard deviation of the control group (column 4). Note that balance variables were not pre-registered. However, they represent all pre-determined characteristics that were collected during the registration process and the baseline child survey.

There are no significant baseline differences between the treatment and control group, except in the number of children aged 18 or below in the household. The p-value for an F-test of the null hypothesis that baseline characteristics are jointly the same for treatment and control group is equal to 0.818.

2.5 Empirical strategy

The estimation of the effect of the intervention is done using two empirical specifications. For outcomes that are measured both at base- and at endline, we estimate the following difference-in-difference regressions by OLS:

$$Y_{ibt} = \alpha_b + \beta Treat_i + \gamma Post_t + \delta Treat_i \times Post_t + \lambda X_i + \epsilon_{ib} \tag{2.1}$$

where Y denotes the outcome for student i, in block b at time $t \in \{pre, post\}$, i.e. either before or after the intervention. The α_b's are block fixed-effect (indicating the classroom of the student). $Treat_i$ is a dummy variable equal to one if the individual is in the treatment group, and $Post_t$ is a dummy variable equal to one for the period after the end of the program. The vector X represents a set of pre-determined student and parent characteristics that include student age, grade, gender, region fixed-effects, a dummy indicating school meal eligibility, a set of dummy variables indicating baseline math grade categories (fail, pass, good), a set of dummy variables indicating the frequency of online lessons during school closures in April and May 2020, a dummy indicating whether the student had a tablet or computer at home before the program, a dummy indicating whether the

student was receiving other tutoring before the program, categorical variables indicating the number of people below age 18 at home, the language spoken at home, parental education, household income, an indicator for whether the responding parent is a single parent, and a dummy variable indicating whether the parent is of Spanish origin. Finally, ϵ_{ib} is an error term. The coefficient of interest, δ, corresponds to the intention-to-treat (ITT) estimate, which measures the effect of being assigned to participate in *Menπores*. Standard errors in this specification are clustered at the individual level.[21]

For outcomes measured only at endline, we estimate the impact of being assigned to the program with the following OLS regression:

$$Y_{ib} = \alpha_b + \delta Treat_i + \lambda X_i + \epsilon_{ib} \tag{2.2}$$

where the variables are defined in the exact same way as in the difference-in-difference specification above. Again, the coefficient of interest measuring the ITT-effect is captured by δ. We report heteroskedasticity-robust standard errors for this specification.

The results presented here deviate in two ways from what had been pre-registered. First, we had not initially registered some of the academic outcomes (math grade, whether the course was passed and whether the school year was repeated). The reason for this was that even though we included a clause for consent to linking the student data to administrative records, there was a high chance that this data would not be made available. This is due to the fact that in Spain, the process for access to administrative data is not yet well established, and often depends on factors that are outside researchers' control.[22] Indeed, the data owners - the ministries of education (*Consejeria de Educacion*) in Madrid and Catalonia - did not give us access to this data after the end of the program, stating legal reasons. Because we believed these outcomes were extremely important and we had funding available at the end of the program, we decided to solicit the information on academic outcomes through a survey to parents. Additionally, our pre-registration did not include results on self-perceived affinity and ability in math and language. The reason was that we included these outcomes in the base- and endline surveys after pre-registration, but we considered it relevant to look at them as we thought they could potentially constitute mechanisms which might explain persistent effects of the program.

Second, we had initially planned to analyze whether the effectiveness of the program depended on whether students in a tutoring group were matched with someone they considered a friend. Due to time and logistic restrictions, we were not able to collect data on friendships between group members

[21] While individual level assignment to the treatment implies standard errors should be clustered at the individual level in the difference-in-difference setting (see for instance the discussion in Abadie et al. (2017)), we also report results when clustering standard errors at the block level in column 5 of Tables 2.9-2.12.

[22] See for instance this recent initiative by one of the authors of this paper and signed by all co-authors to streamline this process and make it more transparent: https://www.esade.edu/ecpol/wp-content/uploads/2023/05/AAFF_ESP_EsadeEcPol_Brief38_UsoDatos_v6.pdf.

and were not able to perform this analysis.

We look at many outcome variables, which raises the risk for false positives. To adjust for the fact that we are testing multiple hypotheses and may incorrectly reject null hypothesis of no effects, we calculate Romano-Wolf step-down adjusted p-values, which control for the family-wise error rate and allow for dependence among p-values. To do so, we group our outcomes into five families: (1) academic achievement; (2) self-perceived ability and affinity; (3) aspirations; (4) school attitudes and motivation and (5) socio-emotional outcomes.

2.6 Results

In this section we discuss the implications of selective attrition and present our main results.

2.6.1 Selective attrition

Before discussing our main results, we check for selective attrition between base- and endline for the various outcomes we study.

Most of the outcomes of interest described above - the score on the standardized test, socio-emotional outcomes and aspirations - were collected through endline surveys administered during math classes at school. Despite the fact that this meant that attrition was very low - 328 out of 375 students (87.5 percent) participated in the endline survey and math test - our estimates could still be biased if attrition was different between the treatment and control group.

To check whether there is selective attrition at endline, Table 2.2 shows balancing conditional on having an endline observation in the in-class math test and questionnaire. None one of the characteristics, except for the number of children aged 18 and below in the household, are significantly different between treatment and control group. The p-value for an F-test of the null hypothesis that baseline characteristics are jointly the same for treatment and control group is equal to 0.765. Additionally, we find no significant difference in missingness between treatment and control group for the outcomes measured through the in-class test and student questionnaire (see Table A1 in the online appendix).

The second set of academic achievement variables - end-of-year math grades, whether the math course was passed and whether the child had to repeat the grade - were collected via an online and phone survey to parents at the end of the school year. Attrition for this survey was higher, with 62 percent of parents responding at endline overall (233 out of 375). Missingness for the outcomes measured through the parent survey was significantly higher for control group students, whose parents were 13 percentage points less likely to respond at endline (see Table A1 in the online appendix). In Table 2.3, we show balancing conditional on having an endline observation in the parent-reported

outcomes. There are no statistically significant differences in most baseline characteristics between treatment and control group. However, we reject the null hypothesis that baseline characteristics are jointly equal between treatment and control group due to some characteristics showing significant differences. For instance, we find that treatment students whose parents responded at endline were more likely to be recipients of free school meals, and their parents were more likely to hold only compulsory schooling or below (versus high school diploma or above), suggesting slightly negative selection into responding at endline among the treatment group. While these analyses suggest differential attrition at endline between treatment and control group, it is reassuring that we observe no differences in terms of baseline academic achievement between the two groups. In robustness checks (Section 2.7.1), we will use inverse probability weights and provide Lee (2009) and Behaghel et al. (2009) bounds to treatment effects to account for non-random sample selection in these outcomes.

2.6.2 Academic outcomes

Table 2.4 summarizes the results for the impact of the intervention on academic outcomes. The dependent variable in column 1 is the score on the math test, standardized by grade level.[23] The difference-in-difference estimate indicates that treatment students improved their score by 0.26 SD more than control students, which is significant at the 10 percent level.[24]

Columns 2 to 4 show treatment effect estimates for teacher-assessed outcomes reported by parents at the end of the school year. We find that treatment group students have a 0.85 points higher end-of-year math grade than control students, corresponding to an increase by about 0.49 SD compared to the control group. Treatment students are also 20.5 percentage points more likely to have passed the subject (math), corresponding to a 30 percent increase in the likelihood of passing compared to the control group mean. Further, we find a large, negative and significant effect on grade retention. Treatment students were 8.9 percentage points less likely to have to repeat the school year, corresponding to a 74 percent drop in the repetition probability compared to the control group. After accounting for multiple hypothesis testing, results become just insignificant at the 10 percent level for the standardized test score and repeating the school year (the Romano-Wolf step-down adjusted p-values for these outcomes are 0.102), and remain significant at conventional levels for the final math grade and passing the subject.

In online appendix Table A2, we show results from parent-reported academic outcomes collected from a survey we implemented one year later (in autumn 2022). The response rate for the follow

[23] The test score is standardized at the grade level (for grade 7 and 8, respectively) and using the mean and standard deviation of the control group at baseline.

[24] This result is based on standardizing the test score at the year group level among participating students only. The effect size is 0.23 SD and remains significant at the 10% level when standardizing the test score at the grade level among all students who took the test, including those that did not participate in the study.

up questionnaire was only 45 percent (168 out of 375), out of which 120 had also replied in the first questionnaire, and results are mostly insignificant. However, the magnitude and direction of coefficients is very similar to those estimated based on the survey results right after the end of the program: Final grades of students who were assigned to treatment were 0.48 points higher (+0.32 SD). Treatment students were also 12 percentage points less likely to have repeated the school year that started after the end of the program. While the small sample size means that we cannot estimate these effects precisely, they are indicative of potential positive long-run effects of the program.

2.6.3 Self-perceived affinity and ability

In Table 2.5 we present results on outcomes measuring self-perceived ability and affinity towards math. Columns 1 and 3 show, respectively, that the program did not increase the likelihood of pupils stating that they thought they were good at math or that they liked math. These results are surprising in light of the positive impact of the intervention on actual achievement (both externally graded math tests and teacher-assessed outcomes). Given the positive relationship between perceived ability and outcomes (Spinath et al., 2006), the failure to raise students' self-image in mathematics may have limited the potential longer-term effects of our program. For comparison, we also asked the same questions about Spanish language, to check whether there was some sort of crowding out (i.e., students shifting preferences toward math or away from math in favor of other subjects). It is also possible that improved results in math could motivate students overall and make them more motivated for other subjects. However, we do not find evidence that this happened (see columns 2 and 4).

2.6.4 Aspirations, perseverance, effort and motivation

We now look at the impact of the program on aspirations, perseverance, and motivation. Column 1 of Table 2.6 shows that treatment group students were 13.5 percentage points more likely than control group students to state that they would like to go on to complete a *bachillerato* (academic high school track), the pre-requisite for entering university in Spain, after completing compulsory education. This corresponds to a 31 percent higher probability than the control group, and the result is robust to adjusting for multiple hypothesis testing. We believe the impact of the program on raising aspirations to choose the academic track are important for several reasons: First, because aspirations have been shown to positively affect future educational achievement (Khattab, 2015). Second, attending the academic track at upper secondary school is linked to higher earnings later in life and may thus increase social mobility (Matthewes and Ventura, 2022). We do not find an increase in the likelihood of stating that students plan to go on to higher education (college/university) after-

school (column 2). While choosing the academic track at upper secondary school tends to be highly correlated with planning to go to university, the fact that we do not find an impact here might be because this is a decision that lies very far in the future for the students in our intervention, who were on average just 13 years old.

In the one-year follow up (online appendix Table A2), we find a zero effect on parent's beliefs about their children's plans to attend the academic upper secondary route (doing the *bachillerato*), and a 4.4 percentage points increase (not significant) in the likelihood of parents stating that they believe their children will go to college after school. Note that these results are not comparable to those immediately after the end of the program, because aspirations at endline were asked to students, while in the one-year follow-up they were solicited from parents.

Given the programs specific focus on student motivation using the growth mindset approach developed by Dweck (1986), we tested whether the program positively affected student motivation, effort, and perseverance, which would be potential mechanisms driving also the increase in academic achievement. In column 3 of Table 2.6, we assess whether program assignment had any impact on grit, a measure of perseverance and conscientiousness. We do not find evidence that this was the case. It is likely that our program was too short to be able to change this outcome. In fact, recent research suggests that grit is a highly heritable personality trait with limited malleability (Rimfeld et al., 2016).

Column 4 shows the impact of program assignment on self-perceived effort at school. Students in the treatment group were 11.4 percentage points more likely than the control group to state that they exerted high effort at school always or most of the time, corresponding to a 18 percent higher probability than in the control group. This result is however not robust to taking into account multiple hypothesis testing. Column 5 shows the effect of program assignment on our school motivation index. We do not find an impact on this outcome. It thus seems that while our program was able to raise students' self-perceived effort, it was potentially too short or not specific enough in order to raise student motivation or perseverance.

2.6.5 Well-being and socio-emotional outcomes

In the aftermath of the Covid-19 pandemic, there were considerable concerns about the longer-run effects of the lockdown and school closures on children's mental health (Newlove-Delgado et al., 2021). We expected the intervention to have a positive impact on socio-emotional well-being due to the positive group dynamics and the presence of an adult reference as a tutor and mentor (Kosse et al., 2020). Table **??** shows the ITT-estimates of the impact of the program on measures of well-being and locus of control. We find no impact of the program on overall subjective well-being measured

by the well-being index (column 1). However, when looking at one of the questions included in the well-being index separately - satisfaction with school - we find a relatively large coefficient estimate equivalent to a 0.22 SD increase (column 2), which is however not robust to multiple hypothesis testing (Romano-Wolf p-value: 0.11). While the program was not successful at raising overall well-being, this provides suggestive evidence that it did raise satisfaction in the dimension most closely linked to the context of the intervention: school.

Column 3 shows that the intervention had a significant negative impact on our measure of locus of control, meaning that treatment students were less likely to believe they can influence what happens in their lives. This result does not remain significant after taking into account multiple hypothesis testing (Romano-Wolf p-value: 0.11). While this result is counter-intuitive - we would have expected the intervention, if anything, to increase internal locus of control - we interpret this as evidence for a reduction in self-blame among treatment students. When looking at one of the variables composing the locus of control index separately - whether students agreed with the statement that when something bad happened to them it tended to be the fault of others - we find that this increases significantly more for treatment students than for control students. Possibly, the intervention made treatment students believe that their low achievements might not be their own fault, but due to a lack of external support (e.g., from teachers or parents). These results should be taken with caution and more research is needed to understand exactly how our tutoring intervention might affect locus of control.

2.6.6 Tutor, teacher and parent feedback

At endline, we collected feedback from parents, tutors and schools, asking them to evaluate their experience with the program. While these evaluations are of a purely subjective character and have no causal interpretation, we nevertheless believe they are important to analyze potential obstacles and lessons for a potential scale up of the program in the future.

In the final survey of the families of the pupils participating in the program, we found a general satisfaction with *Menπores*. More than 80 percent of the families agreed or strongly agreed with the statement 'My mentored child is more confident in the subject of mathematics'. Some 80 percent of families agreed or strongly agreed with the statement: 'Tutoring has improved my child's results in mathematics at school'. Finally, 85 percent agreed with the statement: 'The mathematics reinforcement program has been useful for my child'.

Mathematics teachers and headteachers of the participating schools rated the impact of the program positively. More than 70 percent of teachers and 57 percent of headteachers surveyed agreed or strongly agreed that the program had been useful for their pupils. Some 69 percent

of teachers believed that the program was a good support for their teaching. Finally, 71 percent thought that the program should continue, which is also shared by 100 percent of headteachers surveyed. More than 40 percent of surveyed mathematics teachers believed that the fact that pupils participated in the program helped them to work better, and another 42 percent believed that the coordination meetings with the mentors were useful. Some teachers said that they were overwhelmed by the additional workload during the program (due to coordination with tutors and administering base- and endline tests). In the open-ended responses, several teachers and headteachers suggested to start the program before April and make it longer.

We also analyze what tutors perceived to be the main obstacles to students attending the sessions. Figure 2.3 shows a summary of the results from this analysis. Around 40 percent of tutors mentioned clashes with other extracurricular activities as a common cause for non-attendance. Poor internet connections and feeling tired or unwell were also important. Almost 20 percent of tutors mentioned a lack of quiet study space and interruptions by parents or siblings a problem for effectively delivering tutoring sessions.

In terms of lessons for a potential scale-up, in order for the program to be positively regarded by teachers it should be ensured that it does not mean additional workload for them. It would also be valuable to test whether a different timing (more towards the beginning or middle of the academic year) and a longer duration would make the program even more effective. Finally, scheduling the sessions in groups of two so they do not clash with other activities is both challenging and important, and tutors and students should be given enough flexibility in order to accommodate other commitments. At a larger scale, students could be matched with a group mate depending on their availability in order to facilitate scheduling.

2.6.7 Heterogeneous effects and mechanisms

In this section we present some insights into what might have worked best in our intervention. We look at whether there are differential treatment effects depending on the tutor-student gender match, and the gender and ability composition of the group. For instance, if students were of the same gender, they might have been less embarrassed to interact, and this might have improved classroom dynamics and tutoring effectiveness. If the tutor was of the same gender as the student, teaching might have been more effective as students felt more similar to their tutor (as has been shown, for instance, by Dee (2005)). With respect to group ability composition, the tutor might have been able to teach at the right level when students were more similar in their baseline ability (Duflo et al., 2011; Banerjee et al., 2016). We also study whether the tutoring program was more beneficial for students who had higher or lower initial test scores.

Figure 2.4 shows the interaction effects of the different characteristics with the treatment variable (Table 2.8 shows the full set of coefficients on the treatment dummy and the interactions terms). While interaction terms tend not to be significant, which would be expected due to our small sample size, the signs and magnitudes of the coefficients provide several interesting insights. For the standardized test, the program seems to have been less effective for those in the bottom 50% of the baseline ability distribution. This is consistent with results in Guryan et al. (2023), who find that their program had positive treatment effects on math test scores for all but the bottom quartile in baseline achievement. The program seems slightly more effective in terms of math test scores if the student and the tutor were of the same gender, although this interaction is very imprecisely estimated. The gender match among the students in the group or whether students were of similar baseline ability does not seem to have mattered for impacts on math test scores. This latter finding is in contrast with those in Duflo et al. (2011), who find positive effects of ability tracking.

For final math grades and whether the subject was passed, program effects do not seem to differ by baseline performance, but seem higher when the students in the group were of the same gender. Again, the group ability match does not affect program effectiveness for these outcomes. Further research at a larger scale is needed to get a better understanding of these aspects. Our evidence suggests that gender matching among group members might be an important aspect, at least for students in this age group.

One of the findings from our study is that despite the fact that tutors had a clear mandate to address socio-emotional aspects, such as motivation and emotional well-being, we find little or no impact of our intervention on these outcomes. While the program might simply have been too short to have an effect on such outcomes, it raises the question of whether tutors were actually implementing socio-emotional support in their sessions. While we do not have monitoring data on this aspect, we have data on open-ended questions to tutors at endline regarding best practices that they would recommend for future editions of the program. About 32 percent of tutors mentioned the emphasis on motivation and socio-emotional aspects as important for the success of the intervention (see Figure 2.3). The fact that almost a third of tutors mentioned this explicitly, and that this factor seems to have been more important than, for instance, the use of technology (such as digital whiteboards) during sessions (mentioned by 20 percent of tutors) or being informed about what students were working on during their regular math classes (mentioned by 9 percent of tutors) suggests that indeed tutoring sessions were not merely focused on academic content. In future research it will be important to keep track and monitor more explicitly to what extent socio-emotional support is actually implemented, or randomly vary the degree of this support in order to understand its importance for program effectiveness.

2.7 Discussion of results and robustness checks

We now discuss several potential concerns around the internal and external validity of our main results and perform several checks to see whether our results remain robust to different specifications, and taking into account selective attrition.

2.7.1 Selective attrition

As discussed in Section 2.6.1, the differential response rate to the endline parent survey among treatment and control group parents raises concerns about whether our estimates might be biased due to selective attrition. To quantify how much this might matter for our results, column 4 in Table 2.9 shows the estimated impact of program assignment on academic outcomes using inverse-probability weights. We can see that the effects size on our standardized math test (Panel A) is virtually identical when using these weights compared to our main result (reported in column 3). Panels B to D are outcomes that rely on parental responses to the endline survey. Effect size estimates are virtually identical, if anything, slightly higher, for all three outcomes when using inverse-probability weights. This is in line with the potential (downward) bias in effect sizes predicted from our analysis of selective attrition.

While the above robustness check is reassuring, the substantial attrition for parent-reported outcomes, which are also the most precisely estimated, raises concerns about experimenter demand effects, with parents of treated children possibly being more likely to report positive outcomes. Given that we do not have administrative data on these outcomes, we try to address this concern by calculating bounds on our estimates on parent-reported outcomes using Lee (2009)'s and Behaghel et al. (2009)'s approach.[25] The results of this exercise, shown in online appendix Table A3, indicate that bounds only include positive impacts on the final math grade (ITT-estimate=0.852, bounds=[0.304,1.317]), whether the student passed the subject (ITT-estimate=0.205, bounds=[0.143,0.358]) and only negative impacts on whether the student had to repeat the school year (ITT-estimate=-0.089, bounds=[-0.270,-0.060]). These results give reassurance that when taking into account sample attrition, the main conclusions from our analysis continue to hold.

In column 4 of Tables 2.10 to 2.12 we show the effect size estimates using inverse-probability weights for all non-academic results. Again, column 3 reports our main results for comparison. For each set of outcomes, the point estimates are very similar to each other.

To conclude, our estimates are robust to accounting for attrition using inverse probability weights, and, if anything, are slightly downwardly biased in the case of academic outcomes reported by parents. Treatment-effect bounds for these outcomes indicate that our main conclusions continue to

[25] We use Behaghel et al. (2009)'s approach for binary outcomes as in this case it provides tighter bounds.

hold even in the presence of non-random attrition at endline.

2.7.2 Alternative specifications

The results shown thus far correspond to those using our preferred, full specification. In columns 1-3 of Tables 2.9 to 2.12, we additionally show regressions using alternative specifications. In each table, column 1 shows the most basic specification, including only the treatment dummies and block fixed effects. In column 2 we add demographic characteristics (age, gender, grade, autonomous community, baseline math grade categories, whether had online classes during lockdown, whether had a device to connect to tutoring sessions available, whether received some form of academic tutoring at baseline) and in column 3 we add variables relating to socio-economic status (whether eligible for school meal subsidy, whether speaks Spanish at home, dummies for household income intervals, number of household members below age 18, parental education, whether living in a single-parent household, and whether the parent is of Spanish origin), corresponding to our main specification shown so far.

When looking at academic outcomes in Table 2.9, results are very stable across specifications, with effect size estimates mostly increasing as we add more controls to take into account heterogeneity at baseline across treatment and control group. For non-academic outcomes reported in the remaining Tables 2.10 to 2.12 the same holds: estimates are remarkably stable across specifications and so are their statistical significance levels. We therefore conclude that our main results are robust to the inclusion or exclusion of specific control variables.

Our preferred specification for outcomes where we have repeated measures at base- and endline is a difference-in-difference model. In Table 2.13 we check how the estimates differ when we estimate these results using Equation (2.2) (post estimator with lagged dependent variables) instead of Equation (2.1) (DID). At the bottom of the table we present the difference in the ITT-effect estimates between the two strategies and the p-value of a test for equality of the coefficients of the $Treat \times Post$ and the $Treat$-dummy across the two models. The effects are not substantially different using either method and most coefficients are very similar in magnitude across the two specifications. For all but one of the outcomes - the well-being index - we cannot reject the null hypothesis that the coefficients are equal across the two models. For one of our main outcomes, the standardized test score, the estimated effect is around 0.09 SD lower in the lagged dependent variable specification and becomes insignificant at conventional levels. However, looking at the predictive margins of treatment over different values of the baseline score shown in Figure 2.5, estimated from a model that includes the interaction of treatment and baseline test score, we can see that the average effect masks substantial heterogeneity. For medium to high values (above 0.5) of the standardized baseline test score, the effect of treatment is large and significant at the ten percent level (and ranges up to half a standard

deviation).

2.7.3 Volunteer tutors

We had initially planned a third treatment arm, where we wanted to compare the effectiveness of volunteer tutors with that of our professional tutors. As explained in Section 2.3.1, although we did not achieve enough volunteer applications in order to fully implement this third treatment arm, we included them in the randomization and eventually 19 students were taught by such volunteers. Tables A4 to A7 in the online appendix show results when excluding volunteer tutors. Point estimates tend to be slightly higher for our main results on academic achievement, which is consistent with evidence showing that professional tutors are more effective than volunteers (Nickow et al., 2020). Apart from differences in experience and qualifications as a potential explanation for why volunteers may be less effective, we find that in our program volunteer tutors delivered on average three fewer sessions and 200 minutes less of tutoring than professional, paid-for tutors. This confirms one of the main concerns with volunteers: a possible lack of commitment, which is likely to be less of a problem for paid tutors. However, we cannot draw any strong conclusions from this evidence given the small number of students tutored by volunteers.

2.7.4 Contamination of control group

In response to the pandemic school closures, many governments launched additional support programs to close learning gaps that emerged during lockdowns. Such competing programs were also launched in Spain around the same time as ours, which constituted a risk of contamination of the control group. To check whether this was likely a problem, during the endline survey we asked parents whether students had received any other tutoring or academic support program in math or other subjects during the period while *Menπores* was implemented. Indeed, as Figure 2.6 shows, nearly 40 percent of control group students received some other tutoring or academic support in math, compared to only around 12 percent of the treatment group. The control group was also more likely to have received additional support in another subject, indicating that schools and/or parents might have compensated control group students with other offers. It is also possible that the process of initial identification of students in need for individualized support in math made parents more aware of the needs of their children and ended up prompting them to seek more support, especially if they were not selected for participation in *Menπores*. Overall, these findings suggest that our impact estimates could be interpreted as lower bounds.

2.7.5 External validity

Our program was specifically targeted at schools in disadvantaged areas, which means that our sample is not representative of the population of Spanish 7th and 8th grade students as a whole. To get a sense of how students at participating schools compare to our schools, online appendix Table A8 shows summary statistics of learning outcome indicators and socio-economic characteristics, separately for the entire population of schools in the region of Madrid and for those schools that participated in our experiment.[26] In column 1 we can see that the ESCS index, measuring student socio-economic status, of schools participating in the program is half a standard deviation below the regional average, approximately placed in the 25th percentile in the overall socio-economic distribution. Columns 2 to 5 show that students in participating schools were on average much lower performing, by between 73 (Spanish), 16 (Math), 90 (English) and 74 (Social subjects) percent of a standard deviation with respect to the regional average. Participating schools also have a lower overall share of children born to Spanish parents. Overall, students at our participating schools are on average lower performing and more disadvantaged than the average population of students in Madrid. This should be kept in mind when interpreting our results.

An additional external validity concern is related to the opt-in nature of the program. The effects we find apply to children whose parents are motivated enough to actively register them to after-school tutoring. However, it is well-documented that many remedial programs targeted at low-performing, marginalized children and youth do not tend to reach those who most need them (Robinson et al., 2022). While registration to our program was voluntary, our implementation partner went to great lengths to ensure parents registered the children that had previously been identified by their teachers or school principals as in need for additional support. This included information desks with computers in schools and hands-on help with online registration. While this might have encouraged participation among parents that would have otherwise not undertaken the effort to register their children, after-school, opt-in programs will unlikely reach the same population as during-school, pull-out tutoring.

2.8 Conclusion

Governments and international organizations around the world still struggle to find efficient and scalable interventions to close educational gaps. The pandemic crisis contributed to widening those gaps. But it also opened up the possibility to implement new online tutoring formats. While face-to-face tutoring has been widely evaluated, very little experimental evidence exists on the effectiveness

[26] We cannot do the same exercise for the schools in Catalonia as we do not have access to school level statistics for that region.

of online tutoring programs for secondary school students.

In this study, we show that in a normal schooling environment, our 100-percent online intensive tutoring program in small groups of two students improved academic outcomes and aspirations of socially disadvantaged students. The 8-week program significantly increased standardized test scores (+ 0.26 SD), end of year math grades (+0.49 SD) and the probability of passing the subject (by about 30 percent with respect to the control group mean), while reducing the probability of repeating the school year (by about 74 percent with respect to the control group). In terms of non-academic outcomes, the intervention significantly contributed to raising aspirations. Students assigned to treatment were 13.5 percentage points more likely to state that they would go to the academic track after compulsory schooling. Although not robust to adjusting for multiple hypothesis testing, we also find a 11.4 percentage points increase in the likelihood of stating that treatment students exerted high effort at school always or most of the time. They were also significantly more likely to say they were satisfied with school.

Our results are highly relevant to inform on how to design effective policy responses to reduce educational inequalities. Online tutoring programs have the advantage of reaching children at a lower cost and can be provided to any child with an internet connection, including those in remote places where traditional tutoring programs are harder to deliver. Moreover, our two-students-per-tutor format has the benefit of being more cost-effective than other alternatives with professional teachers, such as face-to-face small groups or one-on-one online programs. Beyond this, global private tutoring is projected to grow at an annual rate of 9 percent per year between 2022 and 2027, mostly due to the larger growth of its online segment. A policy strategy of publicly funded tutoring could contain and respond to this growing demand for more personalized services among middle-classes (Report Linker, 2022), and may be especially relevant for lower-income or lagging students in order to contain widening educational gaps.

In terms of implementation, a key advantage of the online format is that attendance can be tracked in real time and one can react with action plans immediately when students are starting to lag behind or be absent. In our experiment, this ability might have been one of the reasons explaining the high attendance rate of the program, in spite of its intensity (three sessions per week) and the fact that most participants came from highly disadvantaged backgrounds. When thinking about implementing such a program at a larger scale, these considerations are important, as data driven monitoring is a key advantage of online programs.

A potentially major challenge for the implementation of a program like ours at scale is reaching students from disadvantaged backgrounds in need of additional learning support. In the context of opt-in, on-demand tutoring it has been shown that take-up tends to be very low, but that it can be

improved substantially by targeted communications to parents and students (Robinson et al., 2022). In our study, principals and teachers at participating schools provided hands-on support for families to ensure they filled out the registration forms. We also had a highly motivated team communicating actively with the schools that showed interest in the program. At a larger scale, it is not obvious that such personalized support would be possible. Additionally, during the time our intervention took place, families and students were likely more receptive to additional support programs due to the dramatic impact of the pandemic on learning. It is not clear that in the present context the level of interest and motivation would be equally high. More research is needed to understand how take-up of opt-in educational resources, such as our after-school online tutoring, can be increased.

As regards to external validity, one of the main contributions of our study is that the program was implemented while schools were open, thus providing a complement to formal schooling, which is closer to a normal setting, and hence may depict what can be a promising avenue of intervention to support students in educational or social disadvantage.

A potential limitation of our design for large scale programs might be the secular shortage of qualified math teachers (Santiago, 2002). To what extent is it possible to select and train a large workforce of medium to highly qualified tutors? Although the online nature may help bridge the gaps between supply and demand in local labor markets, this policy will require creating professional pathways for tutors, assuming that tutoring will usually be a part-time job, that it will not be a lifetime career, and that it will require a social commitment towards vulnerable students in the system. The most likely candidates could be undergraduate and graduate students with interest in education and social change, recent graduates aiming for job opportunities or retired teachers aiming at contributing to their communities.

For future research, it will be relevant to explore in more detail the mechanisms driving our results: tutor characteristics and interactions with students, the type of training received or the number of students per tutor. It will also be important to explore whether the positive results of online tutoring shown here hold in different contexts: with primary school students, with variations in socio-emotional support or focusing on other subjects, such as reading. Also, it would be interesting to explore in more detail the potential benefits of small-group positive peer dynamics in online teaching. The remarkable academic effect of the program as well as the high attendance rates - the median number of completed sessions was 20 out of a target of 24 - indicate that our two-on-one design might have helped to mitigate some of the shortcomings found in the literature in online education, such as a lack of perseverance and motivation (Escueta et al., 2020). Likewise, it would be interesting to explore the effect of introducing complementary technologies, such as adaptive software with high quality content, asynchronous interactions with tutors through chats or even

more advanced AI bots, to support tutors in teaching and students in learning.

Bibliography

Abadie, Alberto, Susan Athey, Guido W. Imbens, and Jeffrey Wooldridge, "When Should You Adjust Standard Errors for Clustering?," Working Paper 24003, National Bureau of Economic Research November 2017.

Angrist, Noam, Peter Bergman, and Moitshepi Matsheng, "Experimental evidence on learning using low-tech when school is out," Technical Report 7 2022.

Azevedo, Joao Pedro, Halsey Rogers, Sanna Ellinore Ahlgren, Marie-Helene Cloutier, Borhene Chakroun, Gwang-Cho Changl, Suguru Mizunoya, Jean Nicolas Reuge, Matt Brossard, and Jessica Lynn., "The State of the Global Education Crisis : A Path to Recovery," Technical Report 2021.

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, Shobhini Mukherji, Marc Shotland, and Michael Walton, "Mainstreaming an Effective Intervention: Evidence from Randomized Evaluations of "Teaching at the Right Level" in India," Working Paper 22746, National Bureau of Economic Research October 2016.

Behaghel, Luc, Bruno Crépon, Marc Gurgand, and Thomas Le Barbanchon, "Sample attrition bias in randomized experiments: a tale of two surveys," Discussion Paper IZA DP No.4162, IZA Institute of Labor Economics 2009.

Betthäuser, Bastian, Anders Bach-Mortensen, and Per Engzell, "A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemicDid students learn less during the COVID-19 pandemic? Reading and mathematics competencies before and after the first pandemic wave," *Nature Human Behaviour*, 2023, pp. 1–11.

Blainey, Katie and Timo Hannay, "The impact of school closures on autumn 2020 attainment," White papers, RS Assessment from Hodder Education 2021.

Carlana, Michela and Eliana La Ferrara, "Apart but Connected: Online Tutoring and Student Outcomes during the COVID-19 Pandemic," Discussion Paper IZA DP No.14094, IZA Institute of Labor Economics 2021.

Chalk, Karen and Lewis A. Bizo, "Specific Praise Improves On-task Behaviour and Numeracy Enjoyment: A study of year four pupils engaged in the numeracy hour," *Educational psychology in practice*, 2004, *20* (4), 335–351.

Comunidad de Madrid, "Invertimos más de 6,1 M en adquirir 36.100 tablets para los centros educativos," 2020. Accessed on 27/11/2021 from https://www.comunidad.madrid/noticias/2020/11/18/invertimos-61-m-adquirir-36100-tablets-centros-educativos.

Cooper, Harris, Kelly Charlton, Jeff C Valentine, Laura Muhlenbruck, and Geoffrey D Borman, "Making the most of summer school: A meta-analytic and narrative review," *Monographs of the society for research in child development*, 2000, pp. i–127.

Dee, Thomas S., "A Teacher like Me: Does Race, Ethnicity, or Gender Matter?," *The American Economic Review*, 2005, *95* (2), 158–165.

Dobbie, Will and Jr. Fryer Roland G., "Getting beneath the Veil of Effective Schools: Evidence from New York City," *American Economic Journal: Applied Economics*, October 2013, *5* (4), 28–60.

Duckworth, A. L. and P. D. Quinn, "Development and validation of the Short Grit Scale (GRIT–S)," *Journal of Personality Assessment*, 2009, *91* (2), 166—-174.

Duflo, Esther, Pascaline Dupas, and Michael Kremer, "Peer Effects, Teacher Incentives, and the Impact of Tracking: Evidence from a Randomized Evaluation in Kenya," *American Economic Review*, August 2011, *101* (5), 1739–74.

Dweck, Carol S, "Motivational processes affecting learning.," *American psychologist*, 1986, *41* (10), 1040.

_ , *Self-theories: Their role in motivation, personality, and development*, Psychology press, 1999.

Eisner, Manuel, Denis Ribeaud, Giuseppe Sorrenti, and Ulf Zölitz, "The Causal Impact of Socio-Emotional Skills Training on Educational Success," *Mimeo*, 2020.

Escueta, Maya, Andre Joshua Nickow, Philip Oreopoulos, and Vincent Quan, "Upgrading Education with Technology: Insights from Experimental Research," *Journal of Economic Literature*, December 2020, *58* (4), 897–996.

Guryan, Jonathan, Jens Ludwig, Monica P. Bhatt, Philip J. Cook, Jonathan M. V. Davis, Kenneth Dodge, George Farkas, Jr. Fryer Roland G., Susan Mayer, Harold Pollack, Laurence Steinberg, and Greg Stoddard, "Not Too Late: Improving Academic Outcomes among Adolescents," *American Economic Review*, March 2023, *113* (3), 738–65.

Haelermans, Carla, Madelon Jacobs, Lynn van Vugt, Bas Aarts, Henry Abbink, Chayenne Smeets, Rolf van der Velden, and Sanne van Wetten, "A full year COVID-19 crisis with interrupted learning and two school closures: The effects on learning growth and inequality in primary education," *Maastricht University, Graduate School of Business and Economics. GSBE Research Memoranda No. 021*, 2021.

Hardt, David, Markus Nagler, and Johannes Rincke, "Tutoring in (Online) Higher Education: Experimental Evidence," *CESifo Working Paper No. 9555*, 2022.

Heckman, James J., Jora Stixrud, and Sergio Urzua, "The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior," *Journal of Labor Economics*, 2006, *24* (3), 411–482.

Higgins, Steve, Dimitra Kokotsaki, and Robert Coe, "The teaching and learning toolkit," Technical Report 2012.

Instituto Nazionale di Statistica, "Rapporto annuale 2020. La situazione del paese.," 2020. Accessed on 21/12/2020 from https://www.istat.it/storage/rapporto-annuale/2020/Sintesi2020.pdf.

Khattab, Nabil, "Students' aspirations, expectations and school achievement: what really matters?," *British Educational Research Journal*, 2015, *41* (5), 731–748.

Kofoed, Michael S, Lucas Gebhart, Dallas Gilmore, and Ryan Moschitto, "Zooming to Class?: Experimental Evidence on College Students' Online Learning during COVID-19," Technical Report, IZA Discussion Papers 2021.

Kosse, Fabian and Michela Tincani, "Prosociality Predicts Labor Market Success Around the World," *Nature Communications*, 10 2020, *11*, 5298.

_ , **Thomas Deckers, Pia Pinger, Hannah Schildberg-Hörisch, and Armin Falk**, "The Formation of Prosociality: Causal Evidence on the Role of Social Environment," *Journal of Political Economy*, 2020, *128* (2), 434–467.

Kraft, Matthew A, John A List, Jeffrey A Livingston, and Sally Sadoff, "Online tutoring by college volunteers: Experimental evidence from a pilot program," in "AEA Papers and Proceedings," Vol. 112 2022, pp. 614–18.

Lee, David S., "Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects," *The Review of Economic Studies*, 07 2009, *76* (3), 1071–1102.

Matthewes, Sönke Hendrik and Guglielmo Ventura, "On Track to Success? Returns to vocational education against different alternatives," *CVER Discussion Paper 038*, 2022.

Newlove-Delgado, T., S. McManus, K. Sadler, S. Thandi, T. Vizard, C. Cartwright, T. Ford, Mental Health of Children, and Young People group, "Child mental health in England before and during the COVID-19 lockdown," *Lancet Psychiatry*, May 2021, *8* (5), 353–354. © 2021. This manuscript version is made available under the CC-BY-NC-ND 4.0 license http://creativecommons.org/licenses/by-nc-nd/4.0/.

Nickow, Andre, Philip Oreopoulos, and Vincent Quan, "The Impressive Effects of Tutoring on PreK-12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence," Working Paper 27476, National Bureau of Economic Research July 2020.

Report Linker, "Private Tutoring Market: Global Industry Trends, Share, Size, Growth, Opportunity and Forecast 2022-2027," Technical Report 2022.

Rimfeld, K., Y. Kovas, P. S. Dale, and R. Plomin, "True Grit and Genetics: Predicting Academic Achievement From Personality," *Journal of Personality and Social Psychology*, 2016, *111*, 780–789.

Robinson, Carly D., Biraj Bisht, and Susanna Loeb, "The inequity of opt-in educational resources and an intervention to increase equitable access," *EdWorkingPaper Nr. 654*, October 2022, *Annenberg Institute at Brown University:* http://www.edworkingpapers.com/ai22-654.

Santiago, Paulo, "Teacher demand and supply: Improving teaching quality and addressing teacher shortages," *OECD Education Working Papers No. 1*, 2002, *OECD Publishing, Paris.*

Shepherd, Stephanie, Dean Owen, Trey J Fitch, and Jennifer L Marshall, "Locus of control and academic achievement in high school students," *Psychological reports*, 2006, *98* (2), 318–322.

Spinath, Birgit, Frank M. Spinath, Nicole Harlaar, and Robert Plomin, "Predicting school achievement from general cognitive ability, self-perceived ability, and intrinsic value," *Intelligence*, 2006, *34* (4), 363–374.

University College London, UCL Institute of Education, and Centre for Longitudinal Studies, *Millennium Cohort Study: Sixth Survey, 2015, SN-8156*, [data collection]. 7th Edition. UK Data Service, 2020.

_ , _ , **and** _ , *1970 British Cohort Study Response Dataset, 1970-2016, SN: 5641*, [data collection]. 4th Edition. UK Data Service, 2021.

World Bank, "Accelerate Learning Recovery," Technical Report 2021.

_ , "Remediating Learning Loss," Technical Report 2021.

Figures

Figure 2.1: Timeline of the *Menπores* program implementation

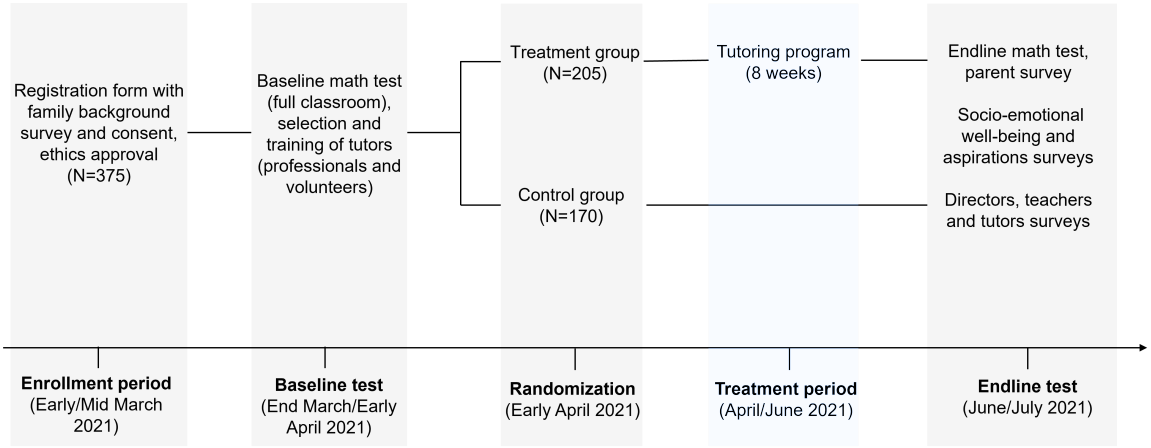
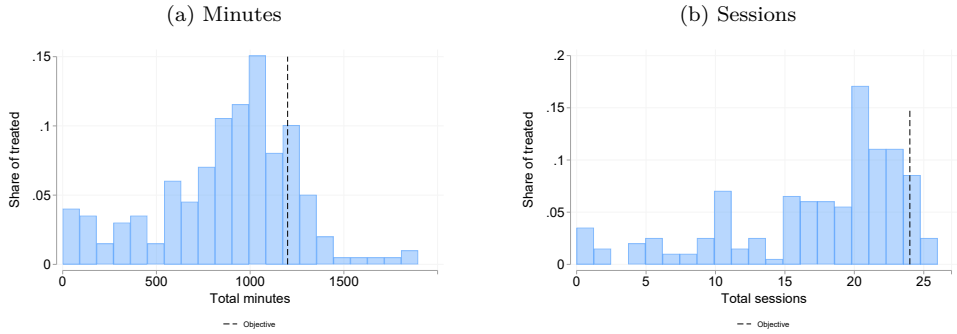
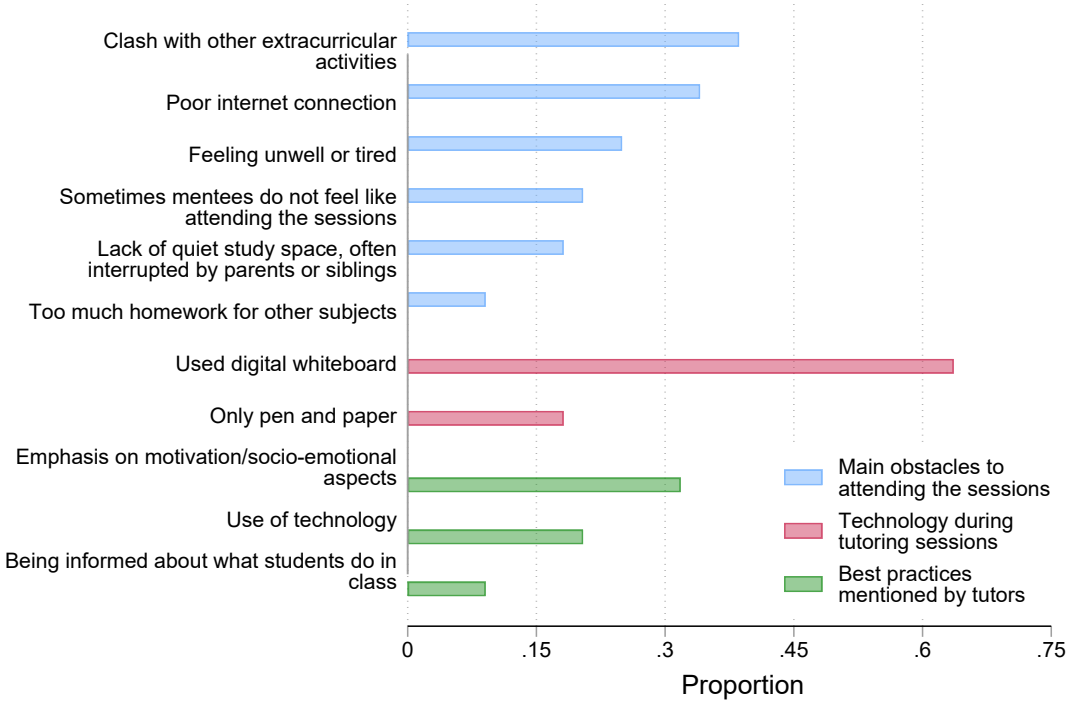


Figure 2.2: Distribution of total minutes and number of sessions of tutoring attended



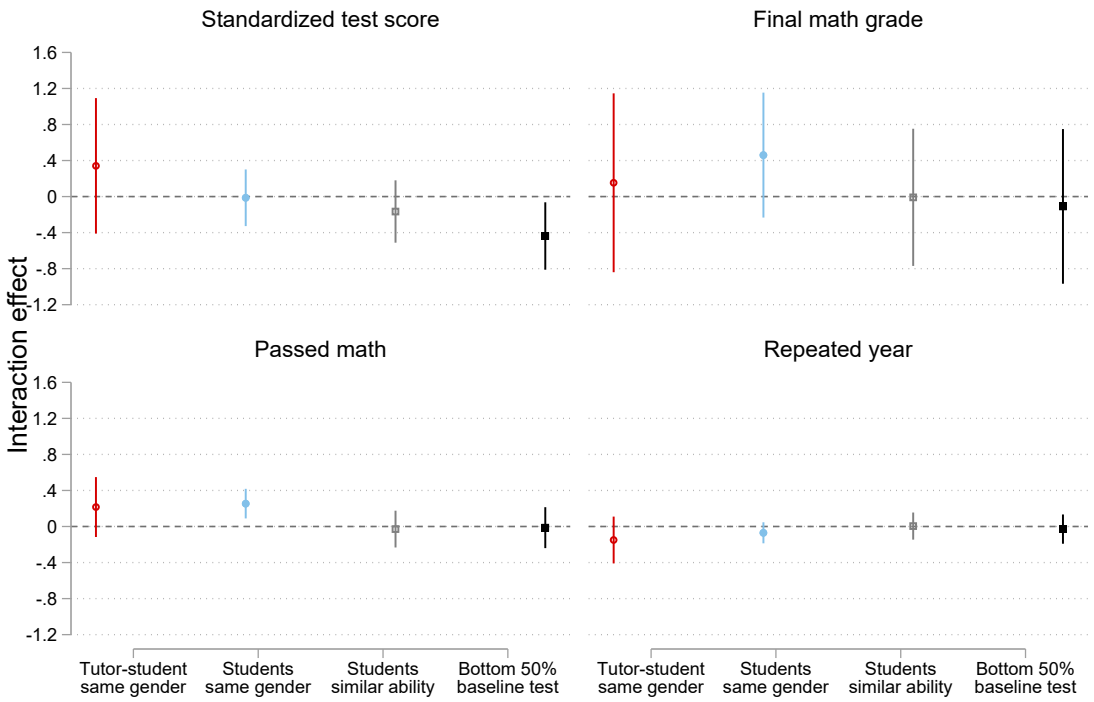
Note: This figure shows histograms of the total number of tutoring minutes attended (left panel) and the number of sessions attended (right panel). The data comes from electronic records of connection times to online meetings from 199 students out of 205 students in the treatment group. This includes seven students who dropped out of the program before it started (zero minutes and zero sessions), and excludes six students for whom we could not match meeting data.

Figure 2.3: Best practices mentioned by tutors at endline



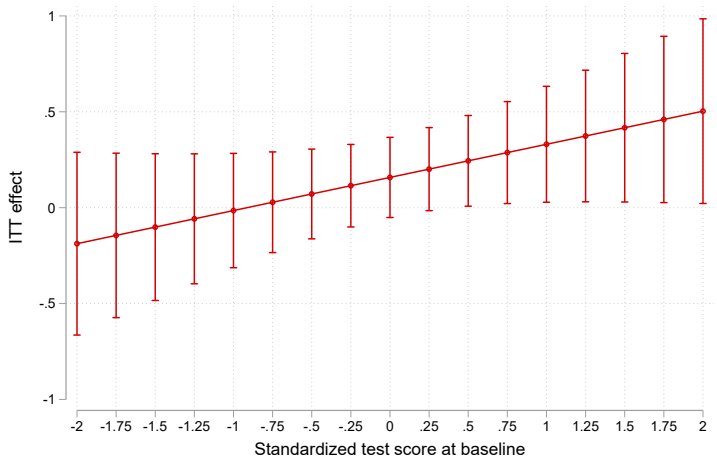
Note: This figure shows information collected through surveys to tutors at the end of the program ($N = 44$ out of a total of 45 tutors). They show the share of tutors who mentioned different types of best practices that they thought were important for the success of the program.

Figure 2.4: Heterogeneous effects by group composition and baseline performance



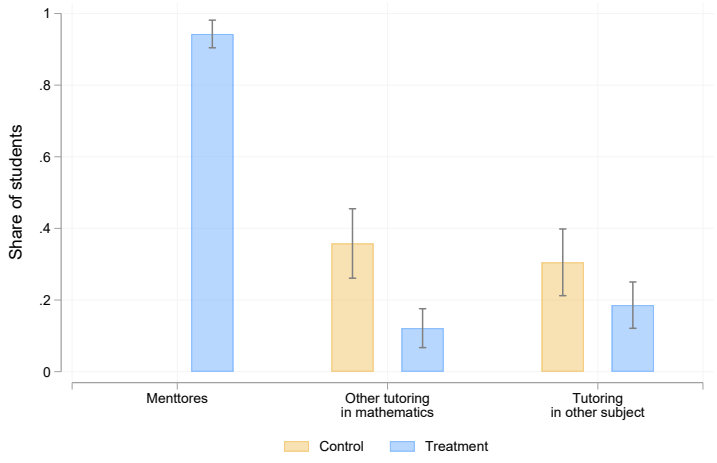
Note: This figure shows the coefficients from the interaction terms shown in Table 2.8. These are estimated by regressing the outcome on a treatment dummy and the interaction between the treatment dummy and a dummy indicating 1) whether the student and tutor were of the same gender; 2) whether the students in the group were of the same gender; 3) whether the students in the group were of similar ability at baseline and 4) whether the students were both in the bottom 50% of the baseline test score distribution. All regressions include the full set of controls as specified in the notes to Table 2.4 and block fixed effects. The plot shows 90% confidence intervals.

Figure 2.5: Predictive margins of treatment assignment by baseline test score



Note: This figure shows the predicted standardized test scores post-treatment for treatment and control group students by standardized test score at baseline. These are estimated by regressing standardized test scores on a treatment dummy, the lag of the dependent variable, and the interaction between the two variables, plus the controls specified in the notes to Table 2.4 and block fixed effects. The plot shows 90% confidence intervals.

Figure 2.6: Counterfactuals



Note: This figure shows the share or treatment and control group students who received i) the *Menπores* program, ii) another tutoring or academic support program in math, and iii) another tutoring or academic support program in another subject. Sample of students whose parents responded to the endline survey.

Tables

Table 2.1: Balancing between treatment and control group

	Control mean (SD)	N=375 Treatment/control difference (SE)	Normalized difference
Child characteristics			
Age	13.04 (0.835)	-0.106 (0.07)	-0.127
Girl	0.44 (0.497)	0.032 (0.05)	0.064
Born in Spain	0.83 (0.377)	0.015 (0.04)	0.039
Grade 8	0.48 (0.501)	0.000 (0.00)	0.000
Public school	0.25 (0.436)	-0.000 (0.00)	-0.000
Catalonia	0.31 (0.465)	0.000 (0.00)	0.000
School meal stipend	0.08 (0.267)	0.030 (0.03)	0.111
No laptop/tablet at home	0.06 (0.247)	0.035 (0.03)	0.140
Has access to internet	0.99 (0.108)	-0.003 (0.01)	-0.032
Receiving academic support (at baseline)	0.18 (0.387)	0.025 (0.04)	0.065
Baseline child survey outcome			
Partial completion, 1+ maths question (%)	0.91 (0.284)	-0.037 (0.03)	-0.129
Completed survey fully (%)	0.52 (0.501)	-0.028 (0.05)	-0.057
Baseline performance			
Has failed math before	0.38 (0.486)	0.035 (0.05)	0.073
Has failed at least one subject before	0.64 (0.483)	0.006 (0.05)	0.013
Repeated grade at least once	0.26 (0.442)	-0.036 (0.05)	-0.081
Pass (5-6.9) math in first term	0.37 (0.484)	-0.043 (0.05)	-0.088
Good (7-8.9) math in first term	0.06 (0.247)	0.019 (0.03)	0.076
Test score (%) at baseline	0.26 (0.180)	-0.017 (0.02)	-0.096
Parental/household characteristics			
Mother responded	0.76 (0.425)	0.004 (0.04)	0.009
Married/cohabiting	0.71 (0.454)	0.001 (0.05)	0.003
Spanish origin	0.52 (0.501)	0.010 (0.05)	0.021
Spanish/Catalan spoken at home	0.86 (0.343)	0.031 (0.03)	0.090
Compulsory schooling or below	0.52 (0.501)	0.015 (0.05)	0.031
Income < 1000 EUR	0.48 (0.501)	-0.024 (0.05)	-0.047
HH size	4.11 (1.088)	0.012 (0.12)	0.011
Nb. children age $\leq$ 18	1.88 (0.834)	0.193* (0.10)	0.231
Age of youngest child	9.96 (3.965)	-0.406 (0.41)	-0.102

Notes: The table shows balancing between treatment and control group observations for the sample of students who registered to participate in *Menπores*. For each variable, we report the control group mean and standard deviation in parenthesis in the second column. The third column shows the δ coefficients from specifications of the type $Y_{ib} = \alpha_b + \delta Treat_i + \epsilon_{ib}$, where Y_{ib} is the variable indicated in the first column, and the α_b's are block fixed effects. The third column shows the normalized difference between treatment and control group, derived by dividing the treatment/control difference by the standard deviation in the control group. Significance levels are indicated by * < .1, ** < .05, *** < .01. The p-value for an F-test of the null hypothesis that baseline characteristics are jointly the same for treatment and control group is equal to 0.818.

Table 2.2: Balancing table - endline respondents to in-class child questionnaire and test

	Control mean (SD)	N=328 Treatment/control difference (SE)	Normalized difference
Child characteristics			
Age	12.99 (0.790)	-0.104 (0.07)	-0.132
Girl	0.44 (0.498)	0.020 (0.06)	0.040
Born in Spain	0.83 (0.379)	0.030 (0.04)	0.079
Grade 8	0.50 (0.502)	0.000* (0.00)	0.000
Public school	0.25 (0.434)	0.000** (0.00)	0.000
Catalonia	0.32 (0.467)	0.000 (0.00)	0.000
School meal stipend	0.08 (0.266)	0.037 (0.03)	0.139
No laptop/tablet at home	0.06 (0.229)	0.031 (0.03)	0.134
Has access to internet	0.99 (0.117)	-0.008 (0.01)	-0.068
Receiving academic support (at baseline)	0.19 (0.391)	0.011 (0.05)	0.028
Baseline child survey outcome			
Partial completion, 1+ maths question (%)	0.93 (0.254)	-0.043 (0.03)	-0.170
Completed survey fully (%)	0.52 (0.501)	-0.024 (0.06)	-0.047
Baseline performance			
Has failed math before	0.39 (0.490)	0.015 (0.05)	0.031
Has failed at least one subject before	0.63 (0.485)	0.011 (0.06)	0.022
Repeated grade at least once	0.23 (0.425)	-0.021 (0.05)	-0.049
Pass (5-6.9) math in first term	0.34 (0.477)	-0.011 (0.05)	-0.022
Good (7-8.9) math in first term	0.08 (0.266)	0.010 (0.03)	0.039
Test score (%) at baseline	0.26 (0.184)	-0.013 (0.02)	-0.070
Parental/household characteristics			
Mother responded	0.77 (0.425)	0.005 (0.04)	0.012
Married/cohabiting	0.73 (0.445)	-0.000 (0.05)	-0.000
Spanish origin	0.53 (0.501)	0.007 (0.05)	0.013
Spanish/Catalan spoken at home	0.86 (0.353)	0.041 (0.04)	0.116
Compulsory schooling or below	0.52 (0.501)	-0.004 (0.06)	-0.008
Income < 1000 EUR	0.46 (0.500)	-0.028 (0.05)	-0.055
HH size	4.13 (1.095)	0.024 (0.13)	0.022
Nb. children age $\leq$ 18	1.90 (0.848)	0.225* (0.12)	0.265
Age of youngest child	10.05 (3.856)	-0.616 (0.44)	-0.160

Notes: The table shows balancing between treatment and control group observations for the sample of students who responded to the endline in-class child questionnaire and test. For each variable, we report the control group mean and standard deviation in parenthesis in the second column. The third column shows the δ coefficients from specifications of the type $Y_{ib} = \alpha_b + \delta Treat_i + \epsilon_{ib}$, where Y_{ib} is the variable indicated in the first column, and the α_b's are block fixed effects. The third column shows the normalized difference between treatment and control group, derived by dividing the treatment/control difference by the standard deviation in the control group. Significance levels are indicated by * < .1, ** < .05, *** < .01. The p-value for an F-test of the null hypothesis that baseline characteristics are jointly the same for treatment and control group is equal to 0.765.

Table 2.3: Balancing table - children of endline respondents to parent survey

	N=233		
	Control mean (SD)	Treatment/control difference (SE)	Normalized difference
Child characteristics			
Age	12.98 (0.751)	-0.034 (0.088)	-0.046
Girl	0.41 (0.494)	0.084 (0.068)	0.171
Born in Spain	0.83 (0.379)	0.065 (0.050)	0.170
Grade 8	0.47 (0.502)	0.000 (0.000)	0.000
Public school	0.23 (0.420)	0.000 (0.000)	0.000
Catalonia	0.28 (0.451)	0.000 (0.000)	0.000
School meal stipend	0.04 (0.204)	0.095** (0.040)	0.467
No online classes during lockdown	0.17 (0.379)	-0.022 (0.052)	-0.058
No laptop/tablet at home	0.11 (0.311)	0.003 (0.039)	0.009
Has access to internet	0.98 (0.146)	0.006 (0.016)	0.043
Receiving academic support (at baseline)	0.18 (0.389)	-0.042 (0.058)	-0.109
Baseline child survey outcome			
Partial completion, 1+ maths question (%)	0.92 (0.265)	-0.032 (0.04)	-0.121
Completed survey fully (%)	0.52 (0.502)	-0.034 (0.07)	-0.068
Baseline performance			
Has failed math before	0.37 (0.484)	0.090 (0.07)	0.185
Has failed at least one subject before	0.62 (0.487)	0.041 (0.07)	0.085
Repeated grade at least once	0.27 (0.446)	0.005 (0.06)	0.011
Pass (5-6.9) math in first term	0.35 (0.481)	-0.007 (0.07)	-0.014
Good (7-8.9) math in first term	0.08 (0.265)	-0.000 (0.04)	-0.001
Test score (%) at baseline	0.23 (0.154)	0.007 (0.02)	0.044
Parental/household characteristics			
Mother responded	0.78 (0.413)	-0.034 (0.06)	-0.082
Married/cohabiting	0.77 (0.420)	-0.007 (0.06)	-0.017
Spanish origin	0.49 (0.503)	0.106 (0.07)	0.210
Spanish/Catalan spoken at home	0.86 (0.349)	0.028 (0.04)	0.081
Compulsory schooling or below	0.48 (0.502)	0.139* (0.07)	0.276
Income < 1000 EUR	0.52 (0.502)	-0.088 (0.07)	-0.175
HH size	4.10 (1.043)	0.006 (0.17)	0.006
Nb. children age $\leq$ 18	1.82 (0.820)	0.240* (0.14)	0.293
Age of youngest child	10.22 (3.796)	-0.689 (0.59)	-0.181

Notes: The table shows balancing between treatment and control group observations for the sample of students who registered to participate in *Menπores*. For each variable, we report the control group mean and standard deviation in parenthesis in the second column. The third column shows the δ coefficients from specifications of the type $Y_{ib} = \alpha_b + \delta Treat_i + \epsilon_{ib}$, where Y_{ib} is the variable indicated in the first column, and the α_b's are block fixed effects. The third column shows the normalized difference between treatment and control group, derived by dividing the treatment/control difference by the standard deviation in the control group. Significance levels are indicated by * < .1, ** < .05, *** < .01. The p-value for an F-test of the null hypothesis that baseline characteristics are jointly the same for treatment and control group is equal to 0.001.

Table 2.4: Impact on academic outcomes

	In-class test	Parent-reported		
	(1) Standardized test score	(2) Final math grade	(3) Passed math	(4) Repeated year
Treat	-0.092	0.852***	0.205***	-0.089**
	(0.102)	(0.234)	(0.058)	(0.044)
RW p-value		[0.015]	[0.015]	[0.102]
Post	0.103			
	(0.112)			
Treat x Post	0.260*			
	(0.146)			
RW p-value	[0.102]			
Constant	-0.624	7.272***	0.669*	0.336
	(0.546)	(2.021)	(0.365)	(0.260)
Mean dep. var.	-0.01	5.10	0.68	0.12
SD dep. var.	1.00	1.75	0.47	0.32
R^2	0.30	0.64	0.63	0.58
Obs.	679	233	233	231
Unique ind.	367			

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at student level (column 1) and heteroskedasticity-robust SEs (columns 2-4) in parenthesis. Romano-Wolf step-down adjusted p-values for multiple hypothesis testing reported in brackets (based on 10,000 replications). The table shows the δ coefficients from Equation (2.1) in column 1 and from Equation (2.2) for columns 2-4, where outcomes are only measured at endline. The number of unique individuals included in the regressions is indicated at the bottom of the table for regressions using DID. All regressions include block fixed-effects (FE) and control for student age, grade, gender, region FE, a dummy indicating school meal eligibility, a set of dummy variables indicating baseline math grade categories (fail, pass, good) (including a category for missing baseline math grade), a set of dummy variables indicating the frequency of online lessons during school closures in April and May 2020, a dummy indicating whether the student had a tablet or computer at home before the program, a dummy indicating whether the student was receiving other tutoring before the program, categorical variables indicating the number of people below age 18 at home, the language spoken at home, parental education, household income, an indicator for whether the responding parent is a single parent, and a dummy variable indicating whether the parent is of Spanish origin.

Table 2.5: Impact on self-perceived ability and affinity

	In-class test			
	(1) Good at math	(2) Good at Spanish	(3) Likes math	(4) Likes Spanish
Treat	0.022	-0.000	-0.060	-0.080
	(0.045)	(0.055)	(0.060)	(0.063)
RW p-value			[0.599]	[0.545]
Post	0.059	0.006		
	(0.037)	(0.051)		
Treat x Post	-0.028	0.014		
	(0.049)	(0.066)		
RW p-value	[0.752]	[0.794]		
Constant	-0.547*	0.241	0.239	0.638
	(0.314)	(0.371)	(0.652)	(0.520)
Mean dep. var.	0.25	0.55	0.42	0.48
SD dep. var.	0.43	0.50	0.50	0.50
R^2	0.36	0.27	0.46	0.37
Obs.	659	660	321	323
Unique ind.	365	366		

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at student level (columns 1-2) and heteroskedasticity-robust SEs (columns 3-4) in parenthesis. Romano-Wolf step-down-adjusted p-values for multiple hypothesis testing reported in brackets (based on 10,000 replications). The table shows the δ coefficients from Equation (2.1) in columns 1-2 and from Equation (2.2) for columns 3-4. The number of unique individuals included in the regressions is indicated at the bottom of the table, and coincides with the number of observations for regressions that do not have a baseline measure of the dependent variable. All regressions include block FEs and the same controls as reported in the notes to Table 2.4.

Table 2.6: Impact on aspirations and motivation

	In-class test				
	(1) Bachi llerato	(2) College	(3) Grit	(4) High effort	(5) Motivation school
Treat	0.135**	0.023	0.075	0.114*	0.005
	(0.059)	(0.050)	(0.061)	(0.061)	(0.018)
RW p-value	[0.056]	[0.654]	[0.456]	[0.262]	[0.806]
Constant	0.292	1.105***	2.594***	0.272	0.716***
	(0.535)	(0.340)	(0.458)	(0.458)	(0.184)
Mean dep. var.	0.43	0.81	3.04	0.64	0.75
SD dep. var.	0.50	0.39	0.50	0.48	0.16
R^2	0.46	0.42	0.36	0.40	0.42
Obs.	318	315	327	321	317

Notes: Significance levels are indicated by * $<$ .1, ** $<$ .05, *** $<$ .01. Heteroskedasticity-robust SEs in parenthesis. Romano-Wolf step-down adjusted p-values for multiple hypothesis testing reported in brackets (based on 10,000 replications). The table shows the δ coefficients from Equation (2.2). All regressions include block FEs and the same controls as reported in the notes to Table 2.4.

Table 2.7: Impact on socio-emotional outcomes

	In-class test		
	(1) Wellbeing index	(2) School satisfaction	(3) Locus of control
Treat	0.151	-0.043	-0.010
	(0.105)	(0.162)	(0.032)
Post	-0.143	-0.133	0.030
	(0.098)	(0.129)	(0.027)
Treat x Post	0.002	0.292*	-0.063*
	(0.108)	(0.165)	(0.036)
RW p-value	[0.982]	[0.110]	[0.110]
Constant	5.818***	4.803***	0.216
	(0.688)	(1.058)	(0.187)
Mean dep. var.	6.23	5.47	0.60
SD dep. var.	1.50	1.35	0.30
R^2	0.64	0.29	0.29
Obs.	679	666	673
Unique ind.	367	367	367

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at student level in parenthesis. Romano-Wolf step-down adjusted p-values for multiple hypothesis testing reported in brackets (based on 10,000 replications). The table shows coefficients from Equation (2.1). All regressions include block FEs and the same controls as reported in the notes to Table 2.4.

Table 2.8: Heterogeneous effects

	(1) Standardized test score	(2) Final math grade	(3) Passed math	(4) Repeated year	(5) Bachi llerato
Panel A: Tutor-student gender match					
Treat	0.227	0.851***	0.204***	-0.088**	0.115*
	(0.147)	(0.235)	(0.058)	(0.044)	(0.060)
Treat x Tutor-student same gender	0.341	0.153	0.216	-0.149	0.296
	(0.456)	(0.599)	(0.201)	(0.157)	(0.207)
Constant	-0.613	7.242***	0.627*	0.365	0.317
	(0.549)	(2.040)	(0.369)	(0.265)	(0.530)
Panel B: Student gender composition					
Treat	0.266	0.637**	0.086	-0.057	0.030
	(0.183)	(0.289)	(0.076)	(0.052)	(0.077)
Treat x Students same gender	-0.013	0.460	0.254**	-0.070	0.225**
	(0.191)	(0.418)	(0.099)	(0.070)	(0.106)
Constant	-0.630	7.143***	0.598	0.355	0.139
	(0.556)	(2.017)	(0.380)	(0.263)	(0.558)
Panel C: Group ability composition					
Treat	0.312*	0.855***	0.215***	-0.091**	0.190**
	(0.160)	(0.278)	(0.073)	(0.044)	(0.076)
Treat x Similar ability	-0.166	-0.008	-0.028	0.005	-0.161
	(0.210)	(0.459)	(0.123)	(0.091)	(0.135)
Constant	-0.388	7.272***	0.667*	0.336	0.197
	(0.525)	(2.027)	(0.368)	(0.261)	(0.555)
Panel D: Bottom 50% baseline test					
Treat	0.420**	0.987***	0.229**	-0.087	0.116
	(0.167)	(0.333)	(0.093)	(0.062)	(0.083)
Treat x Bottom 50% ability	-0.438*	-0.109	-0.013	-0.029	0.059
	(0.227)	(0.517)	(0.137)	(0.098)	(0.134)
Constant	0.348	11.232***	1.105***	0.198	0.507
	(0.546)	(2.032)	(0.354)	(0.407)	(0.552)
Mean dep. var.	-0.01	5.05	0.67	0.13	0.43
SD dep. var.	1.00	1.76	0.47	0.33	0.50
Obs.	663	219	219	217	302

Notes: The table shows shows the coefficient on the treatment and the interaction between treatment assignment and different measures of heterogeneity in the composition of the group and baseline performance for students in the study sample, i.e. those who registered for *Menπores*. Estimates are from our ITT specification (Equation 2.1 for standardized test scores and Equation 2.2 for the remaining outcomes), including a dummy for treatment assignment interacted with indicators for each group with appropriate main effects added, including block fixed effects and our usual set of baseline covariates, as indicated in the notes to Table 2.4.

Table 2.9: Robustness - academic outcomes

	(1) +Block FEs	(2) +Demog	(3) +SES	(4) +IPW	(5) Block cl.
			Panel A: Standardized test score		
Post x Treat	0.229	0.253*	0.260*	0.269*	0.260*
	(0.140)	(0.142)	(0.146)	(0.160)	(0.151)
Constant	-0.278**	-0.769***	-0.624	-0.752	-0.624
	(0.127)	(0.265)	(0.546)	(0.589)	(0.502)
R^2	0.21	0.25	0.30	0.32	0.30
Obs.	679	679	679	679	679
			Panel B: Final math grade		
Treat	0.765***	0.836***	0.852***	0.879***	0.852***
	(0.228)	(0.220)	(0.234)	(0.243)	(0.280)
Constant	5.490***	4.331***	7.272***	7.525***	7.272***
	(0.625)	(0.913)	(2.021)	(1.908)	(1.871)
R^2	0.36	0.51	0.64	0.67	0.64
Obs.	233	233	233	233	233
			Panel C: Passed math		
Treat	0.161***	0.183***	0.205***	0.213***	0.205***
	(0.060)	(0.053)	(0.058)	(0.058)	(0.069)
Constant	0.893***	0.336**	0.669*	0.661*	0.669*
	(0.066)	(0.167)	(0.365)	(0.347)	(0.365)
R^2	0.35	0.56	0.63	0.66	0.63
Obs.	233	233	233	233	233
			Panel D: Repeated year		
Treat	-0.074**	-0.075**	-0.089**	-0.091**	-0.089
	(0.037)	(0.037)	(0.044)	(0.042)	(0.057)
Constant	0.049	0.205	0.336	0.280	0.336
	(0.035)	(0.136)	(0.260)	(0.271)	(0.252)
R^2	0.40	0.48	0.58	0.59	0.58
Obs.	231	231	231	231	231

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at the individual level (Panel A) or heteroskedasticity-robust SEs (Panels B-D) in parenthesis. Column 1 shows OLS regression of the dependent variable in the column heading on a treatment dummy and block fixed effects only. In column 2, the following additional controls are added: Student age, grade, gender, region FE, a dummy indicating school meal eligibility, a set of dummy variables indicating baseline math grade (including a category for missing baseline math grade), a set of dummy variables indicating the frequency of online lessons during school closures in April and May 2020, a dummy indicating whether the student had a tablet or computer at home before the program, a dummy indicating whether the student was receiving other tutoring before the program. In column 3 we further add controls relating to socio-economic status and parental characteristics: Categorical variables indicating the number of people below age 18 at home, the language spoken at home, parental education, household income, an indicator for whether the responding parent is a single parent, and a dummy variable indicating whether the parent is of Spanish origin. In column 4 we present the full specification as in column 3, but using inverse-probability weights to derive estimates. In column 5 we show estimates according the the specification in column 3, but clustering standard errors on the block level rather than at the individual level.

Table 2.10: Robustness - self-perceived ability and affinity

	(1) +Block FEs	(2) +Demog	(3) +SES	(4) +IPW	(5) Block cl.
			Panel A: Good at math		
Post x Treat	-0.012	-0.030	-0.028	-0.032	-0.028
	(0.049)	(0.049)	(0.049)	(0.045)	(0.052)
Constant	-0.018	-0.141	-0.547*	-0.532*	-0.547
	(0.026)	(0.122)	(0.314)	(0.316)	(0.348)
R^2	0.24	0.32	0.36	0.41	0.36
Obs.	659	659	659	659	659
			Panel B: Good at Spanish		
Post x Treat	0.001	0.012	0.014	0.019	0.014
	(0.064)	(0.064)	(0.066)	(0.072)	(0.071)
Constant	0.718***	0.758***	0.241	0.339	0.241
	(0.251)	(0.269)	(0.371)	(0.381)	(0.300)
R^2	0.17	0.22	0.27	0.29	0.27
Obs.	660	660	660	660	660
			Panel C: Likes math		
Treat	-0.052	-0.061	-0.060	-0.024	-0.060
	(0.052)	(0.058)	(0.060)	(0.058)	(0.063)
Constant	0.368	0.227	0.239	0.392	0.239
	(0.316)	(0.382)	(0.652)	(0.649)	(0.654)
R^2	0.32	0.38	0.46	0.50	0.46
Obs.	321	321	321	321	321
			Panel D: Likes Spanish		
Treat	-0.074	-0.067	-0.080	-0.098	-0.080
	(0.058)	(0.060)	(0.063)	(0.060)	(0.074)
Constant	0.382	0.572	0.638	0.663	0.638
	(0.319)	(0.349)	(0.520)	(0.537)	(0.391)
R^2	0.24	0.28	0.37	0.43	0.37
Obs.	323	323	323	323	323

Notes: Significance levels are indicated by * $<$ .1, ** $<$ .05, *** $<$ .01. SEs clustered at the individual level (Panels A-B) or heteroskedasticity-robust SE (Panels C-D) in parenthesis. Controls for specifications in columns 1-5 are as described in the notes to Table 2.9.

Table 2.11: Robustness - aspirations and motivation

	(1) +Block FEs	(2) +Demog	(3) +SES	(4) +IPW	(5) Block cl.
			Panel A: Bachillerato		
Treat	0.130**	0.128**	0.135**	0.127**	0.135**
	(0.058)	(0.057)	(0.059)	(0.060)	(0.067)
Constant	0.246	0.357	0.292	0.536	0.292
	(0.352)	(0.355)	(0.535)	(0.575)	(0.445)
R^2	0.26	0.36	0.46	0.48	0.46
Obs.	318	318	318	318	318
			Panel B: College		
Treat	0.002	0.008	0.023	0.022	0.023
	(0.048)	(0.048)	(0.050)	(0.052)	(0.051)
Constant	0.998***	0.749***	1.105***	1.188***	1.105***
	(0.032)	(0.167)	(0.340)	(0.362)	(0.402)
R^2	0.24	0.33	0.42	0.41	0.42
Obs.	315	315	315	315	315
			Panel C: Grit		
Treat	0.064	0.060	0.075	0.043	0.075
	(0.057)	(0.063)	(0.061)	(0.061)	(0.064)
Constant	3.422***	3.237***	2.594***	2.661***	2.594***
	(0.135)	(0.185)	(0.458)	(0.454)	(0.506)
R^2	0.21	0.27	0.36	0.45	0.36
Obs.	327	327	327	327	327
			Panel D: High effort		
Treat	0.098*	0.095	0.114*	0.105*	0.114*
	(0.054)	(0.058)	(0.061)	(0.063)	(0.067)
Constant	0.935***	0.922***	0.272	0.325	0.272
	(0.047)	(0.171)	(0.458)	(0.459)	(0.493)
R^2	0.25	0.30	0.40	0.43	0.40
Obs.	321	321	321	321	321
			Panel E: Motivation school		
Treat	-0.003	-0.003	0.005	0.008	0.005
	(0.017)	(0.018)	(0.018)	(0.020)	(0.021)
Constant	0.863***	0.875***	0.716***	0.789***	0.716***
	(0.068)	(0.095)	(0.184)	(0.188)	(0.161)
R^2	0.24	0.31	0.42	0.41	0.42
Obs.	317	317	317	317	317

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. Heteroskedasticity-robust SE in parenthesis. Controls for specifications in columns 1-5 are as described in the notes to Table 2.9.

Table 2.12: Robustness - socio-emotional outcomes

	(1) +Block FEs	(2) +Demog	(3) +SES	(4) +IPW	(5) Block cl.
			Panel A: Wellbeing index		
Post x Treat	0.019	-0.016	0.002	-0.058	0.002
	(0.114)	(0.107)	(0.108)	(0.105)	(0.113)
Constant	7.554***	7.332***	5.818***	6.077***	5.818***
	(0.355)	(0.529)	(0.688)	(0.713)	(0.517)
R^2	0.56	0.60	0.64	0.65	0.64
Obs.	679	679	679	679	679
			Panel B: School satisfaction		
Post x Treat	0.288*	0.320**	0.292*	0.297*	0.292*
	(0.157)	(0.159)	(0.165)	(0.171)	(0.167)
Constant	5.746***	5.827***	4.803***	4.709***	4.803***
	(0.571)	(0.744)	(1.058)	(1.087)	(0.894)
R^2	0.17	0.23	0.29	0.33	0.29
Obs.	666	666	666	666	666
			Panel C: Locus of control		
Post x Treat	-0.060*	-0.063*	-0.063*	-0.063*	-0.063*
	(0.034)	(0.035)	(0.036)	(0.037)	(0.038)
Constant	0.589***	0.507***	0.216	0.269	0.216
	(0.072)	(0.077)	(0.187)	(0.192)	(0.223)
R^2	0.18	0.24	0.29	0.34	0.29
Obs.	673	673	673	673	673

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at the individual level in parenthesis. Controls for specifications in columns 1-5 are as described in the notes to Table 2.9.

Table 2.13: Comparison of Specifications

	Standardized test score		Good at math		Good at Spanish		Wellbeing index		School satisfaction		Locus of control	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
	DID	Lagged dep. var	DID	Lagged dep. var	DID	Lagged dep. var	DID	Lagged dep. var	DID	Lagged dep. var	DID	Lagged dep. var
Treat	-0.092	0.166	0.022	-0.050	-0.000	-0.004	0.151	0.192**	-0.043	0.250	-0.010	-0.063*
	(0.102)	(0.128)	(0.045)	(0.047)	(0.055)	(0.055)	(0.105)	(0.095)	(0.162)	(0.154)	(0.032)	(0.034)
Post	0.103		0.059		0.006		-0.143		-0.133		0.030	
	(0.112)		(0.037)		(0.051)		(0.098)		(0.129)		(0.027)	
Treat x Post	0.260*		-0.028		0.014		0.002		0.292*		-0.063*	
	(0.146)		(0.049)		(0.066)		(0.108)		(0.165)		(0.036)	
Constant	-0.624	-1.415	-0.547*	0.118	0.241	0.337	5.818***	2.530***	4.803***	1.165	0.216	0.321
	(0.546)	(0.876)	(0.314)	(0.331)	(0.371)	(0.402)	(0.688)	(0.847)	(1.058)	(1.242)	(0.187)	(0.271)
Mean dep. var.	-0.01	-0.01	0.25	0.31	0.55	0.57	6.23	6.12	5.47	5.35	0.60	0.64
SD dep. var.	1.00	0.99	0.43	0.46	0.50	0.50	1.50	1.25	1.35	1.52	0.30	0.31
R^2	0.30	0.43	0.36	0.56	0.27	0.55	0.64	0.77	0.29	0.65	0.29	0.53
Obs.	679	328	659	318	660	319	679	328	666	319	673	325
Diff. coefficients (S.E.)	−0.093 (0.107)		−0.022 (0.033)		−0.018 (0.048)		0.190 (0.078)		−0.042 (0.116)		0.000 (0.026)	
P-value of difference	0.384		0.500		0.707		0.015		0.714		0.985	

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. Robust SEs in parenthesis for the lagged dependent variable specification and clustered SEs (at the individual level) for DID specifications. The table shows the coefficients from regressions of the form specified in Equations (2.1) and (2.2). All regressions include block FEs and control for student age, grade, gender, region, a dummy indicating school meal eligibility, dummies for math grade categories " fail", "pass" and "good"(self-reported by student) in the first term of the academic year, a set of dummy variables indicating the frequency of online lessons during school closures in April and May 2020, a dummy indicating whether the student had a tablet or computer at home before the program, a dummy indicating whether the student was receiving other tutoring before the program, categorical variables indicating the language spoken at home, parental education, household income, and household composition, an indicator for whether the responding parent is a single parent, and a dummy variable indicating whether the parent is of Spanish origin. Additionally, columns 2, 4, 6, 8, 10 and 12 control for the lag of the dependent variable (measured at baseline).

A Sample questions

A.1 Math test

Example for grade 7 Solve the following equation for x and simplify the solution if possible. You must write down the entire procedure.

- $3x + 5(x - 3) = 4x - 2(x - 5)$

  ☐ $x = \frac{1}{2}$
  ☐ $x = \frac{5}{6}$
  ☐ $x = \frac{6}{25}$
  ☐ $x = \frac{25}{6}$

Example for grade 8 Solve the following equation for x:.

- $x^2 + 2x - 15 = 0$

  ☐ $x = -3, x = 5$
  ☐ $x = 3, x = -5$
  ☐ x does not belong to the set of real numbers
  ☐ $x = 31, x = -33$

A.2 Questions on socio-emotional skills, well-being and aspirations

Grit Here are a number of statements that may or may not apply to you. There are no right or wrong answers, so please answer truthfully, considering how you compare to most people. Indicate one of "Very much like me", "Mostly like me", "Somewhat like me", "Not much like me", and "Not like me at all".

- New ideas and projects sometimes distract me from previous ones.
- Setbacks don't discourage me. I don't give up easily.
- I have been obsessed with a certain idea or project for a short time but later lost interest.
- I am a hard worker.
- I often set a goal but later choose to pursue a different one.
- I have difficulty maintaining my focus on projects that take more than a few months to complete.
- I finish whatever I begin.
- I am diligent. I never give up.

Locus of control For each of the following questions, mark "Yes" or "No":

- Do you usually feel that it's almost useless to try in school because most children are cleverer than you?
- When bad things happen to you, is it usually someone else's fault?
- Do you tend to get low grades, even when you study hard?

Well-being On a scale from 1 to 7, where 1 means "not happy at all" and 7 means "completely happy", how do you feel about the following parts of your life?

- Your school work
- The way you look
- The school you go to
- Your friends
- Your life as a whole
- Think about the period of lockdown during Covid-19. How did you feel during that period?

Aspirations What are your plans after you complete compulsory schooling?

- Select one option:
  - ☐ Vocational education
  - ☐ Continue studying (*Bachillerato*)
  - ☐ Find a job
  - ☐ I don't know
- Mark "Yes" or "No":
  - Would you like to go to college in the future?
  - If so, do you think it would be possible?

Motivation for school How often do you... (indicate one of "always", "most of the time", "sometimes", "never)

- ...put effort into school?
- ...find school interesting?
- ...feel that school is a waste of time?

Frequency of homework Thinking about last May, how much time did you devote to schoolwork per day on average? Select one option:

☐ Less than 15 minutes

☐ 15-30 minutes

☐ 30-60 minutes

☐ 1-1.5 hours

☐ 1.5-2 hours

☐ 2-2.5 hours

☐ More than 2.5 hours

Interest in math and reading How much do you like the following subjects? Select one option:

- Spanish/catalan language:
  - ☐ A lot
  - ☐ Quite a bit
  - ☐ I somewhat like it
  - ☐ A bit
  - ☐ I don't like it at all
- Math:
  - ☐ A lot
  - ☐ Quite a bit
  - ☐ I somewhat like it
  - ☐ A bit
  - ☐ I don't like it at all

B Online Appendix

Table A1: Missing outcomes data

Outcome	(1) Control mean	(2) Treatment/Control Difference
Academic		
Standardized math test	0.147	−0.035 (0.033)
Final math grade (parent-survey)	0.453	−0.128 (0.051)
Passed math (parent-survey)	0.453	−0.128 (0.051)
Repeated year (parent-survey)	0.453	−0.119 (0.051)
Self-perceived ability and affinity		
Liked math	0.171	−0.041 (0.035)
Liked Spanish	0.153	−0.021 (0.035)
Good at math	0.165	−0.015 (0.036)
Good at Spanish	0.171	−0.035 (0.036)
Aspirations and motivation		
Bachillerato	0.182	−0.054 (0.037)
College	0.182	−0.037 (0.037)
Grit score	0.153	−0.041 (0.034)
High effort	0.159	−0.022 (0.035)
Motivation school index	0.171	−0.022 (0.037)
Socio-emotional outcomes		
Wellbeing index	0.147	−0.035 (0.033)
School satisfaction	0.171	−0.033 (0.037)
Locus of control index	0.153	−0.031 (0.034)

Notes: The table shows the share of control students with missing outcome data (column 1) and the treatment-control differences in the share of missingness and their standard errors in parenthesis (column 2) derived from regressing a dummy indicating that the outcome is missing on a treatment dummy and block fixed effects. All outcomes come from the in-class student test and questionnaire, except where indicated otherwise.

Table A2: Impact on academic outcomes - One year later

	Parent-reported			
	(1) Final math grade 1 year later	(2) Repeated year 1 year later	(3) Bachillerato 1 year later	(4) College 1 year later
Treat	0.483	-0.121	-0.005	0.044
	(0.364)	(0.096)	(0.121)	(0.111)
Constant	4.164	-0.286	1.638**	0.220
	(3.233)	(0.495)	(0.807)	(0.686)
Mean dep. var.	5.26	0.17	0.50	0.78
SD dep. var.	1.52	0.38	0.50	0.42
R^2	0.61	0.62	0.68	0.66
Obs.	168	168	168	168

Notes: Significance levels are indicated by * $< .1$, ** $< .05$, *** $< .01$. SEs clustered at student level in parenthesis. The table shows the δ coefficients from Equation (2.2). All regressions include block FEs and the same controls as specified in the notes to Table 2.4. Outcomes measured 16 months after the end of the intervention.

Table A3: Bounds for estimates

Outcome	N	ITT estimate	Bound type	Lower bound	Upper bound	Confidence Interval
Parent-reported outcomes						
Final math grade	233	0.852 (0.234)	Lee (2009)	0.304	1.317	[−0.124, 1.733]
Passed math	233	0.205 (0.058)	Behaghel et al. (2009)	0.143	0.357	[0.052, 0.464]
Repeated year	231	−0.089 (0.044)	Behaghel et al. (2009)	−0.270	−0.060	[−0.387, −0.005]

Notes: This table shows Lee (2009) bounds for the impact of the program on all outcomes. For binary outcomes we compute bounds following Behaghel et al. (2009). This methodology allows to obtain tighter bounds when outcomes are binary and there is non-compliance.

Table A4: Impact on academic outcomes - professional tutors only

	In-class test	Parent-reported		
	(1) Standardized test score	(2) Final math grade	(3) Passed math	(4) Repeated year
Treat	-0.092	0.860***	0.220***	-0.093**
	(0.105)	(0.242)	(0.062)	(0.045)
Post	0.099			
	(0.112)			
Treat x Post	0.270*			
	(0.148)			
Constant	-0.333	8.468***	0.713	0.439
	(0.605)	(2.386)	(0.459)	(0.338)
Mean dep. var.	-0.01	5.10	0.68	0.12
SD dep. var.	1.00	1.75	0.47	0.32
R^2	0.31	0.63	0.64	0.58
Obs.	646	220	220	218
Unique ind.	348			

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at student level (column 1) and heteroskedasticity-robust SEs (columns 2-4) in parenthesis. Sample of students assigned to professional tutors only. The table shows the δ coefficients from Equation (2.1) in column 1 - where we have a baseline measure of the dependent variable - and from Equation (2.2) for columns 2-4, where outcomes are only measured at endline. The number of unique individuals included in the regressions is indicated at the bottom of the table for regressions using DID. All regressions include block fixed-effects (FE) and control for student age, grade, gender, region FE, a dummy indicating school meal eligibility, a set of dummy variables indicating baseline math grade categories (fail, pass, good) (including a category for missing baseline math grade), a set of dummy variables indicating the frequency of online lessons during school closures in April and May 2020, a dummy indicating whether the student had a tablet or computer at home before the program, a dummy indicating whether the student was receiving other tutoring before the program, categorical variables indicating the number of people below age 18 at home, the language spoken at home, parental education, household income, an indicator for whether the responding parent is a single parent, and a dummy variable indicating whether the parent is of Spanish origin.

Table A5: Impact on self-perceived ability and affinity - professional tutors only

	In-class test			
	(1) Good at math	(2) Good at Spanish	(3) Likes math	(4) Likes Spanish
Treat	-0.004	0.012	-0.074	-0.066
	(0.046)	(0.057)	(0.061)	(0.065)
Post	0.057	0.007		
	(0.037)	(0.051)		
Treat x Post	-0.032	0.005		
	(0.048)	(0.067)		
Constant	-0.383	0.278	0.381	0.632
	(0.325)	(0.396)	(0.667)	(0.521)
Mean dep. var.	0.25	0.55	0.42	0.48
SD dep. var.	0.43	0.50	0.50	0.50
R^2	0.38	0.26	0.48	0.37
Obs.	629	630	307	309
Unique ind.	347	347		

Notes: Significance levels are indicated by * $< .1$, ** $< .05$, *** $< .01$. SEs clustered at student level (columns 1-2) and heteroskedasticity-robust SEs (columns 3-4) in parenthesis. Sample of students assigned to professional tutors only. The table shows the δ coefficients from Equation (2.1) in columns 1-2 and from Equation (2.2) for columns 3-4. The number of unique individuals included in the regressions is indicated at the bottom of the table, and coincides with the number of observations for regressions that do not have a baseline measure of the dependent variable. All regressions include block FEs and the same controls as reported in the notes to Table A4.

Table A6: Impact on aspirations and motivation - professional tutors only

	In-class test				
	(1) Bachi llerato	(2) College	(3) Grit	(4) High effort	(5) Motivation school
Treat	0.134**	0.035	0.091	0.126**	0.012
	(0.061)	(0.050)	(0.063)	(0.062)	(0.019)
Constant	0.338	1.187***	2.494***	0.281	0.729***
	(0.549)	(0.336)	(0.513)	(0.477)	(0.193)
Mean dep. var.	0.43	0.81	3.04	0.64	0.75
SD dep. var.	0.50	0.39	0.50	0.48	0.16
R^2	0.46	0.43	0.40	0.42	0.43
Obs.	304	301	311	307	303

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. Heteroskedasticity-robust SEs in parenthesis. Sample of students assigned to professional tutors only. The table shows the δ coefficients from Equation (2.2). All regressions include block FEs and the same controls as reported in the notes to Table A4.

Table A7: Impact on socio-emotional outcomes - professional tutors only

	In-class test		
	(1) Wellbeing index	(2) School satisfaction	(3) Locus of control
Treat	0.176	0.041	-0.006
	(0.110)	(0.170)	(0.033)
Post	-0.137	-0.114	0.029
	(0.098)	(0.129)	(0.027)
Treat x Post	-0.006	0.275	-0.059
	(0.113)	(0.168)	(0.037)
Constant	5.845***	4.452***	0.240
	(0.724)	(1.068)	(0.193)
Mean dep. var.	6.23	5.47	0.60
SD dep. var.	1.50	1.35	0.30
R^2	0.66	0.30	0.29
Obs.	646	633	642
Unique ind.	348	348	348

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at student level in parenthesis. Sample of students assigned to professional tutors only. The table shows coefficients from Equation (2.1). All regressions include block FEs and the same controls as reported in the notes to Table A4.

Table A8: Socio-economic and scores distribution of schools in Madrid (all schools vs. *Menπores* schools)

	(1) ESCS	(2) Spanish - Score	(3) Maths - Score	(4) English - Score	(5) Social - Score	(6) % of native-born parents (both)
All schools (region of Madrid)						
Mean	0.00	0.00	0.00	0.00	0.00	71.2
SD	1.00	1.00	1.00	1.00	1.00	24.2
p10	-1.15	-1.37	-1.23	-1.45	-1.31	35.7
p25	-0.59	-0.59	-0.67	-0.68	-0.68	62.5
p50	-0.01	0.08	-0.03	0.04	0.00	78.8
p75	0.61	0.65	0.57	0.69	0.61	87.9
p90	1.26	1.19	1.27	1.38	1.31	92.9
Menπores schools						
Mean	-0.51	-0.73	-0.16	-0.90	-0.74	68.8
SD	0.54	0.60	0.71	0.59	0.41	14.1
p10	-1.07	-1.62	-1.01	-1.51	-1.23	50.0
p25	-0.90	-1.29	-0.50	-1.45	-1.10	59.8
p50	-0.58	-0.56	-0.28	-1.01	-0.75	69.1
p75	-0.10	-0.33	0.12	-0.30	-0.39	80.5
p90	0.09	-0.24	0.69	-0.23	-0.35	82.1

Notes: This table shows the socio-economic and scores distribution for all schools in the region of Madrid and the socio-economic and scores distribution for schools participating in the tutoring program. ESCS is an Index of Social, Cultural and Economic Status derived from a student background questionnaire with information from the household. The index as well as the score values are standardized to mean 0 and standard deviation one at the regional level. Source: LOMCE evaluations from academic year 2017/18: students participating. All statistics have been derived assuming equal weights for each school.

Chapter 3

Ideological Alignment and Evidence-Based Policy Adoption

Jorge Garcia-Hombrados[1] Marcel Jansen[2] Angel Martínez[3] Berkay Ozcan[4] Pedro Rey-Biel[5] Antonio Roldán-Monés[6]

Abstract

The implementation of evidence-based policies hinges on the dissemination of evidence to policymakers, a process influenced by the attributes of the sender. We conduct a country-wide RCT in which two ideologically opposite prominent think tanks, two major newspapers, and a research institution with nonsalient ideology communicate identical information about a low-cost, non-ideological, and effective policy based on published research findings to a large sample of Spanish local policymakers. We measure the impact of information directly on policy adoption and find heterogeneous effects. When the informing institution aligns ideologically with policymakers, communicating research results leads to a more than 65% increase in policy adoption compared to an uninformed control group, while informing from an opposite ideology does not lead to policy adoption. Our design also allows us to compare the impact of knowledge brokers, such as think tanks, and coverage in leading newspapers in adopting public policies. We find that, when ideologically aligned with policymakers, both are equally effective in increasing policy adoption. We propose a three-stage conceptual framework of policy adoption processes - selective exposure to information, belief updating, and policy implementation- and show that ideological alignment does not influence selective exposure to information. However, evidence from a post-intervention online experiment shows that ideological alignment affects belief updating regarding a recommended policy's effectiveness. Finally, we discuss the trade-offs between effectiveness and outreach when using ideologically aligned and nonsalient institutions to disseminate research evidence and comment on the economic impact of ideological alignment for policy implementation.

[1]Department of Economics, Universidad Autonoma Madrid and IZA.
[2]Department of Economics, Universidad Autonoma Madrid and IZA.
[3]ESADE Center for Economic Policy.
[4]Department of Sociology, London School of Economics.
[5]ESADE Business School and Universidad Ramon Llull.
[6]Department of Social Policy, London School of Economics and ESADE Center for Economic Policy.

Acknowledgements

The study was pre-registered with the American Economic Association RCT Registry (RCT ID: AEARCTR-0008967). We thank Javier Martínez for superb research assistance throughout the project. José María Cueto, Javier Lanza and Constanza Jeldres also provided excellent research assistance at various project stages. We are grateful to Miguel Almunia, Marina Agranov, Antonio Cabrales, Colin Camerer, Pol Campos-Mercade, Stefano DellaVigna, Nagore Iriberri, Stephen P. Jenkins, Joan Monrás, Adam Oliver, Gerard Padró-i-Miquel, Diego Puga, Simeon Schudy, Almudena Sevilla, Uri Simonsohn, Joel Sobel and seminar audiences at UCSD, USC, UCLA Anderson, LSE Social Policy, CEMFI, UPF, UB, IE, ESADE, U. Innsbruck, U. Pablo de Olavide, U. Konstanz, U. Lund, UCL QSS, the Coller Conference in Behavioral Economics and the 2nd Istanbul Applied Economics Workshop for their comments. We thank Federación de Municipios y Provincias, Fundación Alternativas, Fundación FAES, El Mundo, ElDiario.es and TIDES for their support of the experiment. This project has been funded by the Social Observatory of the "La Caixa" Foundation as part of the project LCF/PR/SR21/52560009. We also thank Wikipedia Foundation for initial funding. Jorge García-Hombrados acknowledges funding from grant PID2019-107916GB-I00 from Ministerio de Ciencia e Innovación. Pedro Rey-Biel acknowledges funding from Ministerio de Ciencia e Innovación (PID2022-142172NB-I00) and Universidad Ramón Llull.

3.1 Introduction

Understanding how to improve the dissemination of scientific knowledge to policymakers is crucial for economic and social progress. Despite worldwide efforts to disseminate research findings and promote evidence-based policymaking (OECD, 2020), a significant gap persists between available evidence and the policies ultimately implemented (European Commission, 2022). Knowledge brokers seek to bridge this gap between researchers and policymakers; however, information provision does not happen in a laboratory; it inevitably occurs in politicized contexts. While some knowledge brokers, like think tanks, are affiliated with specific ideologies, others are not.[7] This raises two pivotal questions: How does the ideological alignment between the institutions disseminating research and policymakers affect the adoption of evidence-based policies? And, are prestigious institutions with nonsalient ideologies the most effective in promoting evidence-based policies? To answer these questions, we conducted a country-wide field experiment collaborating with prominent and authoritative institutions with opposing ideologies who disseminated research findings to a large sample of local policymakers.

Our main experimental design keeps the information provided to policymakers constant and manipulates two key variables — the institution disseminating research evidence and the format of information delivery — to investigate their impact on policy adoption across three potential factors. First, similarly to Hjort et al. (2021), we analyze the influence of receiving information compared to an uninformed control group. Second, we introduce variation in the ideological leaning of the informing institution, exploring the causal effect of ideological alignment or misalignment between the policymaker and the institution. Moreover, we investigate whether prestigious ideologically nonsalient research institutions can be as effective as ideologically aligned institutions to foster the adoption of evidence-based policies. Last, our design also allows us to compare the effectiveness of knowledge brokers, such as think tanks, to that of media coverage in the adoption of public policies, following the recent literature on the importance of the medium of information delivery (e.g., Masset et al., 2013; Banerjee et al., 2020; Arnautu and Dagenais, 2021; Yian Yin and Wang, 2021). We compare the rate of policy adoption when policymakers receive information about the research evidence via policy briefs prepared by renowned political think tanks and when they access articles published in prominent news outlets and written by professional journalists.

To accurately assess the effect of ideological alignment between informing institutions and receiving policymakers, it is important to isolate other potential reasons that may affect policy adoption.

[7] For example, the Congressional Budget Office in the US, the What Works Network in the UK, or Ciencia en el Parlamento in Spain. Moreover, public engagement offices and media communication departments of many prominent universities act as non-ideological knowledge brokers.

The ideal research evidence and its associated policy recommendation need to be unequivocally effective, ideologically neutral, have a low implementation cost, be rigorous, prescriptive, timely, and within the decision remit of policymakers. Our policy recommendation, derived from the findings of Hinnosaar et al. (2021), satisfies these criteria. Their randomized controlled trial study showcases the efficacy of enhancing municipalities' Wikipedia pages to bolster tourism. This is particularly relevant for Spain, where the initial research and our study were conducted, since tourism is a pivotal sector of the Spanish economy, contributing around 12.4% of GDP (OECD, 2022). Moreover, our experiment occurred during Spain's recovery from the COVID-19 pandemic, a period wherein tourism emerged as a linchpin for economic revitalization (OECD, 2022). Finally, changing municipalities' Wikipedia page is non-ideological,[8] has a very low implementation cost, is within the responsibility of local policymakers, and its implementation is easily traceable.

The experiment was conducted in 5,678 municipalities (out of 8,131 municipalities in Spain) with a revealed interest in tourism and where we could identify local governments' ideology. Municipalities were randomly assigned to five treatment arms and a control group. The first three arms received the same information communicated either by an ideologically aligned think tank, a think tank on the opposite side of the ideological spectrum, or a researcher from a renowned foreign research institution with no salient ideology, which we call *nonsalient*. The other two treatment arms received links to an article describing the research published in the online version of ideologically aligned or opposite newspapers. Municipalities in the control group received no information. To measure policy adoption, we tracked changes in the municipal Wikipedia pages that were consistent with the recommended policy. Upon completing the experiment, we conducted an online survey, which included an experimental component, with a broader sample of municipal policymakers. This helped us better understand policymakers' attitudes toward evidence-based policymaking and how ideological alignment between policymakers and informing institutions shapes belief updates about the effectiveness of interventions. Finally, we use the figures on the impact of Wikipedia changes on touristic revenues by Hinnosaar et al. (2021) to estimate the cost of ideological misalignment between the institution communicating research results and the policymakers receiving it.

Our results indicate that merely providing information increases policy adoption by 38% relative to the uninformed control group, although this increase is marginally above conventional significance thresholds (p-value=0.13). However, the effect of information provision on policy adoption in our experiment conceals substantial heterogeneity. When the ideologies of policymakers and informing institutions align, we find that the probability of policy adoption increases by more than 65% compared to the control group (p-value=0.03). Conversely, when information comes from institutions

[8]Appendix A shows that this policy is equally supported between left- and right-wing policymakers using survey data.

with opposite ideologies, the coefficient is small and statistically non-significant, indicating that receiving evidence from an institution with an opposite ideology is similar to receiving no information. The effect size of receiving a policy brief from an ideologically nonsalient prestigious institution is nearly half that of a policy brief from an institution with an aligned ideology. However, the coefficient is not significantly different at conventional confidence levels from either the control group or the group of municipalities that get information from an aligned institution. Finally, comparing municipalities that received newspaper articles with those receiving policy briefs, the observed difference proves marginal and lacks statistical significance, implying that both formats are similarly effective in influencing policy adoption.

We then examine the different stages at which ideological alignment may interfere with translating research evidence to policy implemented policy. We propose a three-stage conceptual framework that reflects the behavioral barriers to evidence-based policy adoption in line with Linos (2023): (1) selective exposure to information, (2) belief updating, and (3) policy implementation. First, the literature on polarization has found evidence of partisan selective exposure, showing that individuals tend to avoid information that might contradict their ideological priors (Stroud, 2010) and select media outlets whose biases match their own preferences or prior beliefs (see Gentzkow et al., 2016 for a review of this literature). Second, ideological alignment between informing institutions and policymakers may affect policymakers' beliefs about research evidence. As Bénabou and Tirole (2016) have shown, beliefs often fulfill important psychological and functional needs of the individual, such as social or political identity protection, coherence with previous beliefs, or moral self-esteem. When dealing with politics, individuals often show politically motivated reasoning that makes them resistant to new evidence (Kunda, 1987; Taber and Lodge, 2006; Druckman et al., 2021; Dan M. Kahan and Braman, 2011). For instance, Gentzkow et al. (2018) show that even minor biases can lead individuals on both sides of the ideological spectrum to trust ideologically aligned but unreliable sources over factual and neutral ones and to change their beliefs about facts. Finally, ideological alignment might restrict policy implementation, even when policymakers are convinced about the effectiveness of the policy due to factors such as career concerns (Besley, 2005), the political economy of implementation (Cerna, 2013), the political economy of policy reform (Rodrik, 2018), or party cues (Cohen, 2003).

To examine the hypothesis at the core of the first stage in our framework, we measure the proportion of policymakers who chose to access the full information once they became aware of the informing institution's ideology.[9] The literature on strategic information acquisition, starting with

[9]The emails were sent from an account without ideological salience to maximize exposure. Once policymakers opened the email, they learned which institution was disseminating the research and could choose whether to access the information. See Section 3.3 for details of the design.

Crawford and Sobel (1982), discusses why, in certain instances, individuals may strategically prefer to be informed from aligned or opposite sources to gain more information (see, more recently, Alonso and Padró i Miquel, 2023.) Our findings reveal no differences in access to the full information across treatments, which is consistent with the hypothesis that policymakers did not appear swayed by the ideology of the informing institution when deciding to acquire further information.

To test the hypothesis guiding the second stage of our framework, we conducted a survey experiment with 1,600 policymakers from 1,196 different municipalities, including many from our main experiment. Participants in the survey were asked about their beliefs regarding the potential impacts of a purportedly beneficial policy, different from the one used in our main experiment. Subsequently, policymakers were randomly assigned to receive information from think tanks with either aligned, opposite, or nonsalient ideologies, presenting published research that highlighted the actual negative effects of the policy. We then inquired whether they would believe the study results and still advocate for implementing this detrimental policy. Such motivated political reasoning and ideological biases have also been widely documented among policymakers in the literature (Baekgaard et al., 2017; Butler et al., 2017; Banuri et al., 2019; Christensen and Moynihan, 2020; Vivalt and Coville, 2023). We show that those individuals who received information from an institution with an aligned or nonsalient ideology updated their beliefs much more than those receiving information from an ideologically opposite institution. This result implies that ideological alignment matters at the stage of updating beliefs about policy effectiveness.

Finally, we compare the outcomes of our main experiment on policy adoption and our survey experiment. In the survey experiment, those policymakers receiving information communicated by an ideologically nonsalient think tank updated their beliefs to the same extent as those receiving information from an ideologically aligned think tank. However, in the main experiment, the latter group was nearly twice as likely to adopt the policy than the former, although the difference between both treatments is statistically non-significant. Taken together, and assuming policymakers update their beliefs and act upon them similarly across both experiments, the results suggest that policy adoption may depend on more factors than just belief updating. Thus, comparing the results of both experiments suggests that ideological alignment affects policy adoption across both our framework's second and third stages.

Our paper builds upon Acemoglu and Robinson (2013)'s and, more recently, Dercon (2023)'s observation that politics have often been neglected when studying the effects of policy recommendations and advice. In response, a recently growing body of literature has focused on policymakers and their ideology and how it affects their beliefs and attitudes towards evidence using surveys, without measuring actual policy adoption (e.g., Banuri et al., 2019; Nakajima, 2021; Toma and Bell, 2022;

Lee, 2022; Vivalt and Coville, 2023). Closest to our research are recent studies analyzing the impact on actual policy adoption of policymakers' access to scientific evidence and methods by Hjort et al. (2021), and Mehmood et al. (2024). We bridge this literature to another related research that discusses the bottlenecks to randomized control trials' (RCT) policy adoption (Kremer et al., 2019; Wang and Yang, 2021; DellaVigna et al., 2022). In both of these strands of literature, ideology is not explicitly studied as a main factor. However, when analyzing how innovative policies are diffused across different governments, DellaVigna and Kim (2022) identifies the role of ideology as a prominent factor. Our main contribution is that we study, for the first time, how ideological alignment affects *policy implementation* using a country-wide sample of policymakers and real and authoritative ideological institutions to inform them. This provides a natural, unique, and controlled setting, mimicking how policymakers often inform themselves about evidence to assess how ideological alignment between the informing institution and the policymaker affects policy adoption in practice.

Second, we add to the literature on motivated reasoning, polarization, and partisan bias, which has shown, using ideologically charged examples such as climate change or COVID-19 vaccination, that when research evidence aligns with a particular ideology, it affects the general public's belief updating and compliance with policies (Druckman et al., 2021; Druckman and McGrath, 2019; Guilbeault et al., 2018; Butler and Broockman, 2011). When policies have an ideological component, it is hard to disentangle its effect from the informant's ideology. Our use of a non-ideological policy, together with our post-intervention experimental survey, allows us to isolate the effect of the informant's ideology along the policy adoption process.

Third, we contribute to the literature in political science that focuses on the importance of the messenger and the format of scientific communication. The *messenger effect* literature analyzes how the characteristics of the messenger, such as authority, credibility or likability, influence how information is received and acted upon (Afrouzi et al., 2023; Maclean et al., 2019; Favero et al., 2021; Diamond and Zhou, 2022; Banerjee et al., 2020). Afrouzi et al. (2023) also studies the role of ideology as a characteristic of messengers, along with those listed above, and like others, focuses on how the general public receives the message. In contrast, we contribute to this literature by estimating the effect of ideological alignment between the messenger and the policymakers on adopting evidence-based policies.

Fourth, a growing literature on policy communication emphasizes policy briefs as crucial and increasingly popular means of communicating evidence (see Masset et al., 2013, and a review in Arnautu and Dagenais, 2021). For example, Yian Yin and Wang (2021) show how policy briefs from think tanks have been vital for policymakers to obtain cutting-edge information about the

Covid-19 pandemic. On the other hand, the media also serves as a crucial conduit through which policymakers access evidence from scientific studies (Grossman, 2022). In a pilot survey conducted with Spanish mayors before our main experiment, we found that nearly 66% indicated media as an important source of information for learning about evidence. Our design allows us to compare the effectiveness of both instruments in fostering evidence-based policy adoption. We find that both can be equally effective in policy adoption, an additional novel contribution to this literature.

Finally, we contribute to the literature on belief formation among professionals such as central bankers (Malmendier et al., 2017), academics (DellaVigna and Pope, 2017), or policy analysts (Banuri et al., 2019).[10] For instance, Baekgaard et al. (2017) and Christensen and Moynihan (2020), using survey experiments, show that policymakers, more than the general public, tend to reject evidence contradicting their prior beliefs and resist de-biasing interventions. We contribute to this literature by focusing on local policymakers and expanding the analysis beyond belief formation to study actual policy adoption.

The next section describes the experimental design. In Section 3.3, we show how the experiment was implemented and how we constructed our data. The empirical strategy and the main results are presented in Sections 3.4 and 3.5. Section 3.6 introduces the conceptual framework and tests at which stage of the policy adoption process - information exposure, belief updating, or policy implementation - ideological matching affects evidence-based policy adoption. Finally, Section 3.7 discusses the implications of our findings and concludes. A detailed description and analysis of our online endline survey, a translation of all materials used in the experiment, the heterogeneous effects of the treatment arms, the treatment effects on other Wikipedia outcomes, calculations of the welfare effects of our intervention, and an explanation of the deviations with respect to the pre-analysis plan are reported in the Appendices.[11]

3.2 Experimental design

Our experiment examines whether informing local policymakers about peer-reviewed research evidence increases the adoption of a policy based on such research. We introduce experimental variation in (a) whether policymakers are informed or not, (b) the ideological alignment between the policymakers and the institution informing about the evidence, and (c) the communication format in which information is presented to the policymakers.

We present local policymakers with the results of Hinnosaar et al. (2021), which utilizes tourism

[10] There is a tangentially related political science literature which focuses on political biases and discriminatory behavior of policymakers towards their citizens when delivering services (Butler et al., 2017; Gaikwad and Nellis, 2021; Barceló and Vela Barón, 2023).

[11] The complete registered pre-analysis plan is publicly available at: https://www.socialscienceregistry.org/trials/8967.

as a case study to explore how online content can influence offline consumer behavior. The paper documents that Wikipedia can play a significant role in enhancing revenue from tourism in Spain. Tourists often research travel destinations on Wikipedia before they decide on their destination. Informative Wikipedia pages might thus positively influence the choice of touristic destinations and/or the length of the stay. To test this hypothesis, the authors relied on an RCT to assess the effects of changes in the Wikipedia pages on touristic outcomes for a selection of 60 Spanish municipalities. The edits were carried out by the authors and their editorial team without communicating with the municipalities.[12] The study results received limited media coverage, and therefore, local policymakers were arguably unaware of the study results at the time of our intervention.[13] Their experiment reveals that minor improvements, such as including photographs or completing the information about local festivities or tourist landmarks on the municipalities' Wikipedia pages, increased the number of overnight stays by 9%.

Our sample consists of all Spanish municipalities considered touristic according to a list of objective criteria explained in the next section. We randomly divided our sample of 5,678 touristic municipalities into six groups of similar size to have five treatment arms and one control group (each group included about 950 municipalities). The randomization was stratified according to three criteria: the political party ruling the municipality, the municipality's population, and the number of touristic accommodations available in the municipality.[14]

In the first treatment arm, the research evidence is provided in the form of a policy brief sent by a think tank with an ideology aligned with the municipal government: FAES if the municipality is classified as right-wing, and Fundación Alternativas if the municipality is left-wing. Both institutions are arguably the two most influential conservative and progressive think tanks in Spain, respectively.[15] In the second treatment arm, the same *policy brief* is endorsed by the think tank of the *opposite ideology*. In the third treatment arm, the same policy brief was sent using the letterhead of Berkay Özcan, a professor at the London School of Economics (and co-author of our study), which included LSE's logo and the university's description. This serves as our ideologically nonsalient treatment[16]

[12]In a follow-up paper, the authors find that the Wikipedia changes conducted in the experiment do not trigger more or less subsequent changes in the content of the Wikipedia page (Hinnosaar et al., 2022).

[13]Only the small online news portals La Informacion and Tourinews had published brief articles about the study results. See https://www.lainformacion.com/management/turismo-editar-pueblos-wikipedia/2815402/ and https://www.tourinews.es/resumen-de-prensa/curiosidades/wikipedia-clave-reactivar-turismo-rural_4461915_102.html.

[14]For stratification purposes we divided all municipalities in the sample into four groups according to the ideology of the party ruling the local government. The population of the municipality and the number of tourist accommodations were obtained from official registries and divided into tertiles for stratification purposes.

[15]Appendix B contains the description of both think tanks, as it was provided to policymakers. Such description was extracted from the webpages of both think tanks: https://fundacionfaes.org/ and https://fundacionalternativas.org/.

[16]In the landscape of Spanish municipal politics, LSE is not usually thought of as being part of Spanish politics.

The remaining two treatments use media outlets instead of think tanks as the informing ideological institutions. In the fourth treatment, the research evidence is presented in an article published on the webpage of a newspaper whose ideology is *aligned* with the municipal government: El Mundo, if the municipality is classified as right-wing, and Eldiario.es if the municipality is classified as left-wing. El Mundo and Eldiario.es are two of the main media outlets in Spain and are clearly identified on the right and left of the political spectrum, respectively (see, for instance, Majo-Vazquez, 2022).[17] The fifth treatment arm presents the research evidence in an article published in a newspaper with an *opposite ideology*.[18]

We commissioned policy experts to write policy briefs and professional journalists to write newspaper articles. Both included the same basic information summarizing the research evidence, its relevance for Spanish municipalities, and the type of changes recommended. The text of the three policy briefs was exactly the same across treatment arms. Similarly, the text published by both media outlets was identical. Municipalities in the control group received no information.

Figure 3.1 shows the experimental design and the sequence of our intervention. Once municipalities were selected and classified according to the ideology of their government, municipalities in all treatment arms received an initial email (and several reminders) from the same non-ideological research center specialized in tourism.[19]

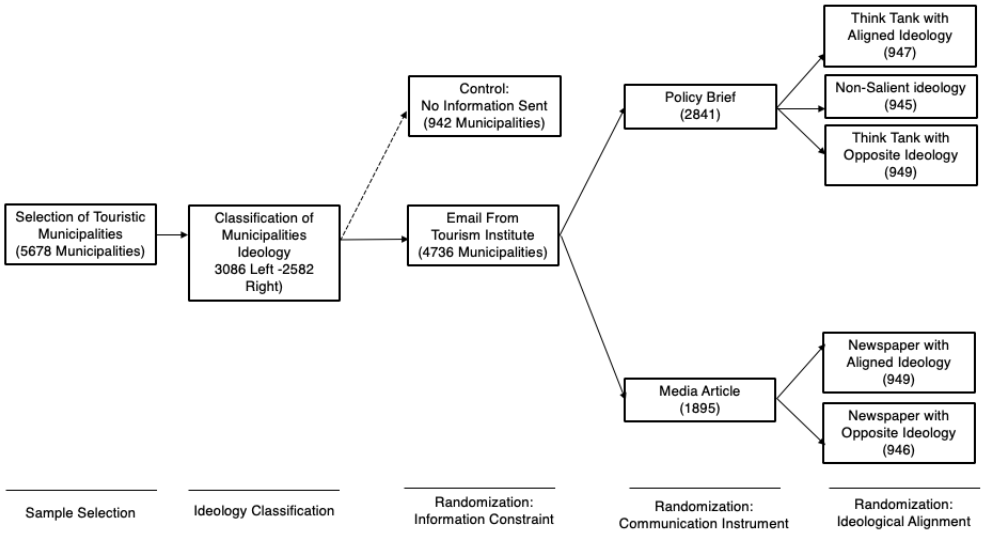
Thus, for our experimental design what is important is that LSE's ideology is perceived as being less salient than the think tanks and media outlets in our other treatments.

[17] Appendix B contains a translation of the description of both newspapers, as it was provided to policymakers. The published articles, both with the exact same text, can be found at https://www.elmundo.es/television/medios/2022/03/07/622663cffc6c832a3b8b45d5.html/ and https://www.eldiario.es/economia/estudio-demuestra-wikipedia-gran-aliado-recuperar-sector-turistico-espana_1_8888694.html/, respectively.

[18] There is no treatment arm using a non-ideologically aligned newspaper since we could not identify one in the Spanish political landscape. The effect of receiving evidence from a non-ideologically aligned institution is calculated by comparing policy briefs with respect to the control group.

[19] The email was sent from an institutional account of the Instituto Universitario de Turismo y Desarrollo Económico Sostenible (TIDES), which is a small research center specialized on tourism studies associated with the University of Las Palmas de Gran Canaria. https://tides.ulpgc.es/

Figure 3.1: Experimental Design



The email included a very short summary of the evidence and highlighted that the research paper presenting the evidence had been published in a prominent international academic journal. In addition, the email text recommended three changes to the municipalities' Wikipedia page to boost tourism: adding information on local festivities, references to touristic landmarks, and including new pictures of the municipality.[20] The message also emphasized the importance of having versions of the municipality's Wikipedia page translated into multiple languages, particularly English. The body of the email prominently displayed the institution that informed about the evidence, which varies by treatment, including the logo and an ideologically salient description of the institution. Finally, the email included two hyperlinks. The first one directs to the policy brief/newspaper assigned in the randomization and, therefore, varies by treatment arm. Both the policy briefs and the newspapers added details about the evidence and reinforced the importance of the changes recommended in the email to improve tourism in the municipality. The second link directed the reader to step-by-step instructions on how to change Wikipedia. The latter document was the same across treatment arms.

To sum up, the email's main content, the subject line, its sender, and the link to the instructions are the same across all treatment arms. On the other hand, the informing institution, policy briefs, and newspaper articles vary across treatment arms. Policy briefs, newspaper articles, the informing institution, and its ideology were only visible once the email was opened.

[20]Remarkably, these changes are doable even in very comprehensive Wikipedia pages of big cities, as it is always possible to add a new photograph of touristic landmarks or complete information about touristic landmarks and local festivities, which is particularly scarce even in the most comprehensive Wikipedia pages.

Our design allows us to compare email opening rates, access to policy briefs and newspaper articles once policymakers learn the ideology of the informing institution, and changes in Wikipedia across groups to investigate three questions. First, does providing information to policymakers increase policy adoption? Second, does the ideological alignment between the informing and the policymaker affect policy adoption? Third, does the instrument used to describe the summary evidence (newspaper vs policy brief) affect policy adoption? A detailed description of these questions and the empirical strategy used to test them is provided in Section 3.4.

3.3 Implementation of the intervention and data construction

We started by defining the universe of Spanish municipalities where tourism plays a relevant economic role. Spain is organized administratively in 17 autonomous communities, 50 provinces, and 8,131 municipalities. Municipalities have a mayor and a council responsible for managing local affairs, including urban planning, social services, waste management, and local taxation. Crucially, they also play an essential role in attracting tourism and investment to their local areas.

We included all touristic municipalities in Spain for which we could identify the ideology of the political party of the mayor, amounting to 5,678 municipalities. We define touristic municipalities as those meeting at least one of the following criteria: the municipality (a) has an officially registered touristic landmark, (b) has a tourism office, (c) has participated in FITUR, an annual International Tourism Fair in 2021 or 2022, (d) is located on the coast or near the Spanish border with France or Portugal, (e) is a member of SEGITUR, a state-owned agency which promotes innovation in tourism and/or (f) features in the database of credit card payments by tourists in 2021 provided by Caixabank, one of Spain's biggest commercial banks. We drop from the sample municipalities that, while fulfilling at least one of these criteria, meet all of the following exclusion criteria: (a) less than 500 inhabitants or more than 40% retired population, (b) do not have any touristic accommodations, and (c) do not show touristic expenditure in the pre-intervention period. While fulfilling one of the inclusion criteria (e.g., coastal, international border, etc.), the role of tourism in the local economy is likely to be small in municipalities that meet all the exclusion criteria. 73 municipalities were excluded when applying these exclusion criteria.[21]

We built a database of personal and active email addresses of all mayors and local councelors

[21] The original 60 municipalities used by Hinnosaar et al. (2021) were not informed that the original study took place, and the results of the study received very limited coverage in the Spanish media. Indeed, we explore heterogeneous effects of our treatment arms for these municipalities in Appendix C and find larger effects of ideological alignment on policy adoption for these municipalities.

members in charge of tourism.[22] Municipality-level population and touristic accommodations data were obtained from the Spanish National Institute of Statistics.[23] While the treatment arm and the ideology are assigned at the municipality level, we send the municipality's assigned treatment to the email addresses of all mayors and local councils in charge of tourism that we identified.[24]

The ideology of the municipality is a categorical variable with four groups: (1) Popular Party (Partido Popular, PP), which is the main conservative (right-wing) party; (2) Socialist Party (Partido Socialista Obrero Español, PSOE), which is the main progressive (left-wing) party, (3) other right-wing political parties and (4) other left-wing political parties. Party affiliation of local governments is publicly available from the central government.[25] In the vast majority of cases, the classification of the left-right ideology was straightforward. However, some municipalities in Spain are ruled by local parties whose ideology is unclear. In those cases, we proceeded as follows: first, we tracked mayors' party affiliation history. If the mayor's party was a spin-off of a previously existing party, we coded the ideology following the political views of that party. If none of these criteria clearly identified the ideology, the municipality was excluded from the sample. In total, we could not classify the ideology of 138 tourist municipalities.[26] Figure 3.2 below shows a map of Spain where all treated municipalities appear in dark color, control municipalities in lighter color, and municipalities not in our sample appear in white. The map shows that municipalities in our sample are located throughout Spain.

For the stratification of the sample, we relied on information on population, the number of tourist accommodations per capita, and the ideology of the municipality's mayor. We created three dummy variables for population and three dummy variables for tourist accommodations in the municipality, according to the tertiles of their respective distributions.

The experiment took place over seven months. The first email was sent in May 2022. During the following 7 months, we sent follow-up reminders on pre-established dates to maximize exposure (See Figure D.I in the appendix). The last reminder email was sent in early December 2022. To boost

[22] An initial rich database of the direct contact email addresses of local policymakers was facilitated by the Spanish Federation of Municipalities (Federación Española de Municipios y Provincias, FEMP). This database was further enriched vía webscrapping.

[23] This information can be found in the following links: https://www.ine.es/dyngs/INEbase/es/categoria.htm?c=Estadistica_P&cid=1254734710990PadrÃ§n and https://www.ine.es/experimental/viv_turistica/experimental_viv_turistica.htm

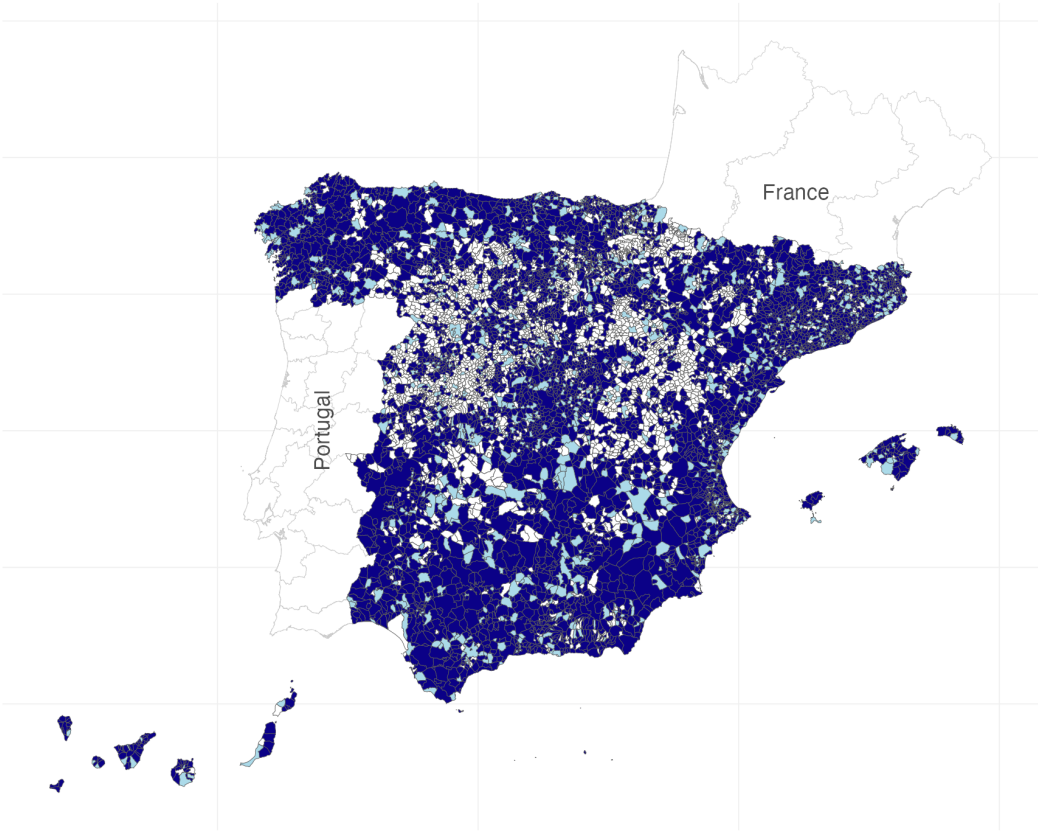
[24] The collection of email addresses was blind to the treatment assignment. Consistently, Table D.IV in Appendix D shows that the number of emails of mayors and local councils does not vary across treatment arms. An important assumption is that all the people targeted within the same municipality share the same ideology of the ruling political party. While unlikely, it is in principle possible in municipalities with coalition governments that a local policy maker contacted in the municipality does not share the ideology of the ruling political party. This would add measurement error to the treatment variable, biasing the estimates towards 0. In this case, our estimates should be interpreted as a lower bound for the true treatment effect.

[25] https://www.mptfp.gob.es/portal/politica-territorial/local/sistema_de_informacion_local_-SIL-/alcaldes_y_concejales.html

[26] This is less than 3% of the touristic municipalities. Measurement error in classifying the ideology would lead to underestimation of our treatment effects

email opening rates, we hired an independent company that phoned the municipal governments' headquarters of all treated municipalities. These calls only aimed to inform policymakers that they had been emailed about our intervention. The persons making the phone calls were given a script for this interaction and had no information about the purpose of the experiment, the treatment arms, or the content of the email. We measured email openings and click-through rates on the links to policy briefs/newspaper articles and instructions to change Wikipedia included in the emails using marketing software provided by MDIRECTOR.[27]

Figure 3.2: Spanish touristic municipalities



Note: Map from Spain with treated municipalities in a dark color, control municipalities in a light color, and municipalities not in the sample in white.

Our pre-registered main outcome variable measuring policy adoption is changes in the munici-

[27] We use this software in the first email and the following nine reminders. We identified during the intervention period that some emails sent with the MDIRECTOR software were not reaching the inbox of the targeted policymakers. To maximize the number of policymakers that received at least one email, we changed to Outlook from the tenth reminder on. Unfortunately, outlook does not allow to track email opening and click-through rates on the links included in the email, limiting our analysis to the first email and nine reminders sent.

palities' Wikipedia page along the recommended guidelines that occurred within the study period between May 25th and December 31st, 2022. Wikipedia's collaborative and open editing model allows us to trace all changes made on the Wikipedia page of the municipalities in the sample. This information includes the changes made, the time and date of the change, and the IP address from which it was done. We first web-scraped the history of the edits function in Wikipedia to identify all changes in the Wikipedia pages during the study period. Most of these changes were small edits arguably unrelated to our intervention. To identify which changes could be driven by our intervention, two coders independently reviewed all changes and, while remaining blind to treatment status, selected which changes were consistent with the changes explicitly recommended in the summary of evidence: adding information about local festivities, improving information about touristic landmarks and adding photographs of the municipality. This data creation exercise was conducted for the study and a placebo period in the same months of 2019, the last pre-Covid year before our intervention. Each coder performed the task separately, and differences in their judgments were resolved by a third coder who was also blind to treatment allocation. While some Wikipedia changes were undone by other Wikipedia users, we could observe all changes made within the study period, even if they were undone. The same process was carried out with the English Wikipedia pages of the municipalities.[28] While we do not observe the identity of the person who edited the Wikipedia page, and therefore, we cannot attribute each individual change with certainty to the action of the contacted policymaker (i.e., some of the edits might occur for other reasons), the existence of a control group allows to net out the effect of our treatment arms on the probability of conducting these changes from similar changes in Wikipedia that were not caused by our treatments during the study period. Reassuringly, Figure D.I in Appendix D shows that in the treatment group, most changes in Wikipedia were conducted soon after the email was sent, while reminders in the control group were uniformly spread throughout the period.

We also collected information on other Wikipedia-related outcomes, including the number of words and images on the municipalities' Spanish and English Wikipedia pages and the number of languages in which the municipality had a website in Wikipedia. Using web scraping techniques, we collected data on the latter variables before sending the first email and just after the end of the study period. We constructed variables to measure the variation in these outcomes from the beginning to the end of the experiment. These variables serve as alternative dependent variables to our main dependent variable. While they might be useful to understand how the information treatments shape Wikipedia outcomes, they are not ideal for measuring policy adoption in our experiment for two reasons: first, most of the changes registered in Wikipedia are minor edits unrelated to the changes

[28] Appendix B detailed the instructions given to the coders to identify the Wikpiedia changes that were compatible with the treatment recommendations.

we recommend, which could either increase or decrease the length of the text.[29] Even if these unrelated changes are minor, they would add measurement error, hindering the interpretation of the estimated treatment effect on this outcome as an indicator of policy change. Second, a significant portion of changes were undone within the period of interest, including approximately 30% of the changes consistent with recommended guidelines. While the reverted changes within the study period were identified for constructing the primary outcome variable (i.e., recommended changes at any point during the study period), they are not included in the Wikipedia outcomes described in this section. The latter set of outcomes simply registers the difference between the page's content at the beginning and the end of the study period. Given the abundance of irrelevant information and the understanding that more words do not always equate to greater clarity, it remains uncertain whether increasing the length of a Wikipedia page would necessarily improve it.

Finally, we conducted an online end-line survey targeting all Spanish municipalities, not only those we identified as touristic, between mid-April and early May 2023. These survey responses convey basic information about municipalities and local policymakers' attitudes toward evidence-based policymaking. Moreover, the survey featured an online experiment to assess political bias in belief updates. In total, invitations were extended to 17,044 local policymakers representing 7,576 municipalities. We achieved a response rate of nearly 10%, with 1,600 policymakers from 1,196 municipalities (15.8%) completing the survey[30]. Detailed information regarding this dataset and its outcomes can be found in Appendix A.

3.4 Empirical strategy

We investigate three main issues: (1) the pooled effect of information, (2) the role of ideological alignment between the policymakers and the informing institution, and (3) whether knowledge brokers are more effective than media outlets in promoting evidence-based policy adoption. To address them, we use linear probability models to estimate the effect of the different treatment arms on the probability of implementing a recommended change in the Wikipedia page of a municipality. Specifically, our pre-registered specification is the following:

$$Edits\ Wikipedia\ (0/1)_{is} = \sum_{m=1}^{5} \beta_m Treatment_{mis} + Strata\ FE_s + \mu_{is}, \tag{3.1}$$

where Edits Wikipedia (0/1) is a dummy variable equal to 1 if the Wikipedia page of municipality i from randomization strata s included a change in Wikipedia along the recommended guidelines

[29] These unrelated changes include small amendments to the history sections, minor edits in the text such as the removal of articles in sentences (not necessarily grammar errors), adding references, random images unrelated with tourism, and other minor changes unrelated to the changes recommended by the study.

[30] This is within the range of large-scale (2500+ obs) online surveys reported in Meng-Jia Wu and Fils-Aime (2022)

over the period studied, and 0 otherwise. $\sum_{m=1}^{5} Treatment_{mis}$ is a vector of dummy variables that indicate the treatment arm m to which municipality i from strata s is allocated. The omitted category is the control group, which received no information. *Strata FE* are randomization strata fixed effects, and μ is the error term. The coefficients of first interest are β_m, which yield the effect of receiving the treatment m relative to the control group. To test the three hypotheses described above, we not only rely on the estimation of the separate effects of the treatment arms, but we also estimate the combined effect of several treatment groups.[31]

We examine whether our stratified randomization successfully produced comparable treatments and control arms. To examine comparability, we estimate equation 3.1 using as dependent variables the characteristics of the municipalities' Wikipedia page and their socio-demographic characteristics, both measured before our intervention.

We start by examining balance across groups for the main outcome variable measured before the experiment, i.e., changes in the Wikipedia page. Specifically, we assess differences across groups regarding the probability of making a recommended change in Wikipedia during our placebo period, June 2019 to December 2019. These months correspond to 3 years before the start of the experiment, the last period before the COVID-19 outbreak. The results are reported in columns (1) and (2) of Table 3.1. The magnitude of the coefficients for the different treatment arms ranges between -0.5 and 0.3 percentage points, showing the expected small effects which are statistically indistinguishable from 0. The results of this placebo analysis show that the main outcome variable is balanced across treatment groups before the start of the treatment.

[31] For example, in our first result we calculate the pooled effect of all treatment arms combined to examine the pooled effect of information.

Table 3.1: Balancing checks across groups (characteristics measured at baseline)

	Recommended changes (0/1)		N words Sp		N words En		N images Sp		N image En		N languages		Tourist accom p/c		Population	
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)
Effect of treatment arms relative to control																
Aligned ideology - Policy brief	-0.0054	-0.0053	-1.33	5.47	-12.20	-10.59	0.04	0.10	-0.14	-0.14	-0.33	-0.31	-0.0041	-0.0042	3,768	3,864
	(0.0043)	(0.0043)	(34.39)	(34.85)	(10.94)	(10.69)	(0.36)	(0.36)	(0.19)	(0.19)	(0.25)	(0.25)	(0.0052)	(0.0051)	(3,049)	(3,071)
Opposite ideology - Policy brief	0.0030	0.0032	-75.14	-70.00	-16.43	-15.30	-0.17	-0.13	-0.02	-0.00	-0.34*	-0.33	-0.0064	-0.0059	1,714	1,777
	(0.0074)	(0.0074)	(53.56)	(54.00)	(11.15)	(11.31)	(0.35)	(0.35)	(0.15)	(0.15)	(0.20)	(0.20)	(0.0041)	(0.0042)	(1,936)	(1,929)
Nonsalient ideology - Policy brief	-0.0022	-0.0022	-22.70	-17.71	-14.32*	-13.24	-0.27	-0.23	-0.00	0.01	-0.09	-0.07	0.0058	0.0060	-70	41
	(0.0054)	(0.0055)	(39.48)	(39.21)	(8.55)	(8.67)	(0.34)	(0.34)	(0.17)	(0.17)	(0.15)	(0.16)	(0.0051)	(0.0052)	(638)	(649)
Aligned ideology - Newspaper	0.0009	0.0010	-0.25	1.86	-6.38	-5.83	0.04	0.06	0.04	0.05	-0.12	-0.12	-0.0022	-0.0020	276	339
	(0.0056)	(0.0056)	(58.54)	(58.59)	(11.27)	(11.23)	(0.38)	(0.37)	(0.19)	(0.18)	(0.26)	(0.26)	(0.0040)	(0.0040)	(1,039)	(1,056)
Opposite ideology - Newspaper	-0.0001	0.0001	-77.32**	-69.24*	-10.24	-8.29	-0.42	-0.36	0.08	0.10	-0.30	-0.27	-0.0061	-0.0058	1,286	1,425
	(0.0055)	(0.0055)	(38.03)	(38.67)	(8.21)	(8.06)	(0.30)	(0.30)	(0.20)	(0.20)	(0.26)	(0.26)	(0.0040)	(0.0038)	(1,286)	(1,280)
Mean dep var in control	0.0202	0.0202	1,066.37	1,066.37	214.91	214.91	27.14	27.14	16.67	16.67	35.57	35.57	0.06	0.06	7,020	7,020
Strata FE	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes	No	Yes
N	5,678	5,678	5,669	5,669	5,663	5,663	5,669	5,669	5,663	5,663	5,669	5,669	5,678	5,678	5,678	5,678
Pooled effects relative to control																
Any treatment	-0.0007	-0.0006	-35.35	-29.93	-11.91	-10.65	-0.16	-0.11	-0.01	0.00	-0.24	-0.22	-0.0026	-0.0024	1,395*	1,489*
	(0.0042)	(0.0043)	(36.79)	(36.83)	(7.39)	(7.29)	(0.31)	(0.31)	(0.13)	(0.13)	(0.20)	(0.20)	(0.0036)	(0.0035)	(767)	(780)
Aligned ideology	-0.0022	-0.0022	-0.79	3.66	-9.28	-8.20	0.04	0.08	-0.05	-0.04	-0.23	-0.21	-0.0031	-0.0031	2,020	2,099
	(0.0042)	(0.0042)	(41.18)	(41.34)	(9.86)	(9.72)	(0.34)	(0.34)	(0.17)	(0.16)	(0.24)	(0.24)	(0.0040)	(0.0039)	(1,438)	(1,456)
Opposite ideology	0.0015	0.0016	-76.23*	-69.62	-13.34*	-11.81	-0.30	-0.25	0.03	0.05	-0.32	-0.30	-0.0062*	-0.0059	1,500	1,601
	(0.0054)	(0.0055)	(43.57)	(43.86)	(7.96)	(7.85)	(0.30)	(0.29)	(0.15)	(0.15)	(0.21)	(0.21)	(0.0037)	(0.0037)	(1,186)	(1,182)
Policy brief	-0.0015	-0.0014	-33.11	-27.46	-14.32*	-13.05	-0.13	-0.09	-0.05	-0.04	-0.25	-0.23	-0.0016	-0.0014	1,805	1,895
	(0.0048)	(0.0049)	(35.29)	(35.33)	(8.07)	(8.06)	(0.33)	(0.32)	(0.12)	(0.12)	(0.18)	(0.18)	(0.0039)	(0.0039)	(1,128)	(1,140)
Newspaper	0.0104	0.0105	-38.72	-33.63	-8.31	-7.06	-0.19	-0.15	0.06	0.08	-0.21	-0.19	-0.0041	-0.0039	780	881
	(0.0067)	(0.0067)	(44.18)	(44.52)	(8.91)	(8.81)	(0.31)	(0.30)	(0.16)	(0.16)	(0.25)	(0.25)	(0.0036)	(0.0035)	(889)	(893)

Note: We examine balancing across groups by estimating the main specification for outcome variables measured before the start of the experiment. We first present differences between each treatment arm and the control group using the main specification estimated with and without strata fixed effects. Then, we calculate the pooled effects that will be computed in the main analysis. *Any treatment* yields the pooled effect of receiving the information across all treatment groups relative to not receiving any information. *Aligned ideology* yields the pooled effect of receiving the summary of study results endorsed by an institution (newspaper or think tank) with the same ideology relative to not receiving any information. *Opposite ideology* yields the pooled effect of receiving the summary of study results endorsed by an institution (newspaper or think tank) with the opposite ideology relative to not receiving any information. *Policy brief* yields the pooled effect of receiving the summary of study results through a policy brief relative to not receiving any information regardless of the ideology of the think tank. *Newspaper* yields the pooled effect of receiving the summary of study results through a newspaper article regardless of the ideology of the newspaper relative to not receiving any information. The table reports the balancing checks for Wikipedia outcomes measured before the start of the experiment: the probability of making a recommended change in the municipality's page in Spanish Wikipedia between May and December 2019, before the start of the experiment and the COVID pandemia, the number of words and images in the municipality's Spanish page in Wikipedia measured before the start of the intervention, the number of words and images in the municipality's English page in Wikipedia measured before the start of the intervention, the number of languages in which the municipality has a Wikipedia page, the number of touristic accommodations per inhabitant, the population of the municipality. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1

We also examine the balance across groups in terms of other Wikipedia outcomes and characteristics of the municipality. Specifically, we examine the comparability across treatments and control arms in terms of the number of words in the Spanish and English pages of Wikipedia, the number of images in the Spanish and English pages of Wikipedia, the number of languages in which the municipality has a Wikipedia page, the number of tourist accommodation per capita in the municipality, and the population of the municipality. Information on all these variables was collected on the last day before the start of the experiment. The results of the balancing checks are reported in Table 3.1. We also test for differences in the number of email addresses targeted in the treatment across groups in Table D.IV in Appendix D. While the results of these analyses show a few unsystematic statistically significant differences for some combinations of outcomes and treatment arms, they do not follow any pattern nor exceed the expected number of false positives.[32] Overall, the results of the balancing checks suggest that the randomization created balanced and comparable groups of municipalities across treatments and control arms.

One potential threat to the experiment is differential exposure to information, as measured by the share of recipients who opened emails across treatment groups. The email was sent from the same institutional email account of the same non-ideological research center regardless of the treatment arm. The ideologically aligned institution that disseminated the research evidence was only revealed once the email was opened. Thus, we should not expect differences in opening rates across treatment groups. We tested this proposition and found reassuring results, presented in Table D.V in Appendix D.

Another potential threat is that the main estimates of the effect may be confounded by spillover effects across municipalities, which may communicate the research evidence received to other municipalities. If available, spillovers would bias the estimated treatment effects downwards.[33] To investigate this concern, we included in the end-line survey a question asking policymakers whether they were aware of having received any information about the research evidence, either from our experiment or other sources. Only five policymakers assigned to the control group (out of a total of 236 respondents) reported having received any information, suggesting a very limited effect of spillovers. We investigate the existence of spillovers further through estimating the effect of distance to the nearest municipality in the sample of control municipalities for each treatment arm. The results, discussed in Section 3.5, rule out the existence of sizeable spillovers.

Goldsmith-Pinkham et al. (2022) identify potential biases in RCTs with mutually exclusive treat-

[32]Statistically, we expect 10 coefficients per 100 to be falsely significant at 10%, 5 coefficients per 100 to be falsely significant at 5%, and 1 coefficient per 100 to be falsely significant at 1%.

[33]If present, spillovers from treatment to control group would bias the main estimates downwards, resulting in our findings representing a lower bound for the true effects. This would not affect the main conclusions regarding the impact of political alignment on policy adoption.

ment arms when estimated with strata-fixed effects. To address these concerns, we also estimate the main regression with strata fixed effects using the procedure presented in Goldsmith-Pinkham et al. (2022) with nearly identical results.

Finally, we calculated the minimum detectable effect size (MDE) for different group comparisons. The MDE is the smallest effect we can identify with an 80% probability. The results are reported for dichotomous outcomes with different baseline probabilities in Table D.VI in Appendix D. The MDEs for hypotheses testing the effect of receiving any information, receiving information by an ideologically aligned institution, or comparing any group are approximately 2, 2.3 and 2.6 percentage points for outcomes with a baseline proportion of 3%.[34] While the calculated MDEs reported in the table represent non-negligible effects, they are likely larger than the true MDE because the pre-experimental power calculations could not take into account the stratification process and the subsequent use of strata fixed-effects in the regression, which would increase statistical power and reduce the MDE.[35]

We conducted a survey on the Social Science Prediction platform to examine other researchers' expectations about our hypotheses before making our results publicly available[36] We obtained responses from 84 researchers from 53 different universities around the World. We are reassured that most respondents agree with the statements that increasing tourism in our context is a welfare-enhancing policy (81%), and that changing municipalities' Wikipedia pages is an ideologically neutral policy (89%). Regarding the reach of the information provided in our experiment, respondents expect the opening rates of the emails that include the policy recommendation to be around 50%. Respondents also thought that click-through rates and the proportion of changes in Wikipedia would be higher when there is ideological alignment between informing institutions and policymakers. These rates were also expected to be higher in the non-salient ideology treatment than in the opposite ideology treatment (46.16% when aligned, 40.19% when ideologically nonsalient, and 30.2% when opposite ideology). Additionally, respondents expected a higher rate of changes in Wikipedia among those receiving the information from an ideologically opposite institution than among the control group (33.9% when aligned, 28.33% when ideologically nonsalient, 20.62% when opposite ideology and 13.85% in the control group). Finally, respondents were uncertain about the relative effectiveness of policy briefs compared to newspaper articles. While 28% of respondents thought they would be equally effective, the proportion of respondents who thought one would be more effective than the other was the same (36% each). In the next section, we discuss the results of our experiments and compare them with the expected results reported by the survey respondents.

[34] 3% is approximately the baseline probability of the main outcome in the control group.

[35] Table D.VII in Appendix D reports the power calculations for continuous outcomes, which correspond to secondary outcomes in the analysis such as the number of words and images in the municipalities' Wikipedia pages.

[36] The survey is accessible in the following link: https://socialscienceprediction.org/s/z486dd.

3.5 Results

In this section, we test our main hypotheses by focusing on our (pre-registered) primary variable measuring policy adoption: the probability of changing the municipality's page in Spanish Wikipedia according to the guidelines provided in the treatments. In Appendix F, we report the results using alternative outcomes such as the length of the Wikipedia page, the number of images on the page, the number of languages in which the municipality has a Wikipedia page, and the same outcomes in the English page of the municipality in Wikipedia. As discussed in the previous section, these alternative outcomes are less precise indicators of policy adoption than the main outcome examined in this section.

To examine our first hypothesis, we investigate the pooled effect of receiving information about study results on the probability of changing Wikipedia along the recommended guidelines during the study period relative to municipalities in the control group, which do not receive any information. The coefficients of the variable *Any treatment* in Columns (1) and (2) of Panel B in Table 3.2 show an increase of approximately 0.98 percentage points (equivalent to an increase by 38%) in the probability of changing the Wikipedia, although the effect is marginally above conventional significance thresholds (p-value=0.13).[37][38]

[37] Table D.I reports the p-values for the coefficients reported in Table 3.2 and other comparisons across treatment arms.

[38] The results reported in Columns (3) and (4) present the placebo exercise described in Section 3.4. They show the expected small and insignificant coefficients that reveal no differences in Wikipedia changes across treatment arms during the placebo period.

Table 3.2: Effects of the treatment arms on the probability of making a recommended change in Wikipedia

	Study period		Placebo period	
	(1)	(2)	(3)	(4)
Panel A: Effect of treatment arms relative to control				
Aligned ideology - Policy brief	0.0168**	0.0169**	-0.0054	-0.0053
	(0.0081)	(0.0082)	(0.0043)	(0.0043)
Opposite ideology - Policy brief	0.0019	0.0020	0.0030	0.0032
	(0.0078)	(0.0078)	(0.0074)	(0.0074)
Nonsalient ideology - Policy brief	0.0094	0.0097	-0.0022	-0.0022
	(0.0086)	(0.0087)	(0.0054)	(0.0055)
Aligned ideology - Newspaper	0.0167*	0.0167*	0.0009	0.0010
	(0.0091)	(0.0091)	(0.0056)	(0.0056)
Opposite ideology - Newspaper	0.0041	0.0043	-0.0001	0.0001
	(0.0075)	(0.0076)	(0.0055)	(0.0055)
Mean dep var in control	0.0255	0.0255	0.0202	0.0202
Strata FE	No	Yes	No	Yes
N	5,678	5,678	5,678	5,678
Panel B: Pooled effects relative to control				
Any treatment	0.0098	0.0099	-0.0007	-0.0006
	(0.0063)	(0.0064)	(0.0042)	(0.0043)
Aligned ideology	0.0167**	0.0168**	-0.0022	-0.0022
	(0.0075)	(0.0076)	(0.0042)	(0.0042)
Opposite ideology	0.0030	0.0032	0.0015	0.0016
	(0.0065)	(0.0065)	(0.0054)	(0.0055)
Policy brief	0.0094	0.0095	-0.0015	-0.0014
	(0.0067)	(0.0068)	(0.0048)	(0.0049)
Newspaper	0.0104	0.0105	0.0004	0.0005
	(0.0067)	(0.0067)	(0.0046)	(0.0047)

Note: Estimates in columns (1) and (2) examine the effect of the different arms on recommended changes between May and December 2022. These are the main results of the study. Estimates in columns (3) and (4) examine the effect of the different arms on recommended changes between May and December 2019, a placebo period before the start of the intervention. Regressions in columns (1) and (3) do not include strata fixed-effects and regressions in columns (2) and (4) are estimated with strata fixed-effects. *Any treatment* yields the pooled effect of receiving the information across all treatment groups relative to not receiving any information. *Aligned ideology* yields the pooled effect of receiving the summary of study results endorsed by an institution (newspaper or think tank) with the same ideology relative to not receiving any information. *Opposite ideology* yields the pooled effect of receiving the summary of study results endorsed by an institution (newspaper or think tank) with the opposite ideology relative to not receiving any information. *Policy brief* yields the pooled effect of receiving the summary of study results through a policy brief relative to not receiving any information regardless of the ideology of the think tank. *Newspaper* yields the pooled effect of receiving the summary of study results through a newspaper article regardless of the ideology of the newspaper relative to not receiving any information. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

The heterogeneity analyses reported in Appendix C show that the pooled effect of receiving any information is more relevant for left-wing municipalities, municipalities in the top tertile of the population size distribution, and municipalities above the median regarding the length of the Wikipedia page. These results are discussed in detail in Appendix C.

From the results of the analysis exploring the pooled effect of providing information on policy adoption for the full sample, we conclude:

Result 1: The pooled effect of all treatment arms of information provision on policy adoption is sizeable, but statistically not significant at conventional confidence levels.

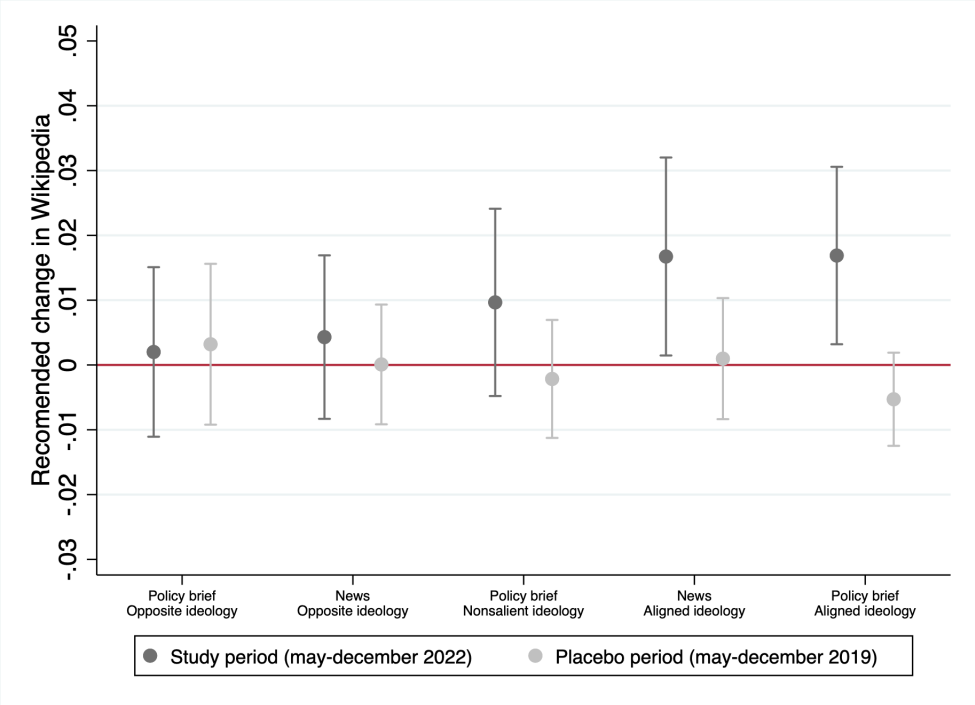
Next, we examine whether policy adoption is affected by the ideological alignment between the informing institution and policymakers. The results reported in columns (1) and (2) of Panel A in Table 3.2 show that dissemination by ideologically aligned institutions -think tanks or newspapers- increases the share of municipalities that implement the recommended changes in Wikipedia by 1.68 and 1.67 percentage points, respectively. In relative terms, this represents an increase of 66% and 65%, respectively, compared to the control group. The effects are statistically significant at the 5% significance level. The results are nearly identical when the estimation is conducted with and without strata-fixed effects. The difference is also significant at the 5% level when compared against the opposite-ideology treatment arms rather than against the control group (p-value=0.02).[39] The effects are evident when compared with the coefficients of the placebo exercise reported in Columns (3) and (4) and in Figure 3.3. Moreover, Figure D.II shows that the results are primarily driven by the inclusion of additional information on local festivities on the Wikipedia page, one of the changes we emphasized in both the email and the information provided.

The results of the analysis of heterogeneity reported in Table C.I in Appendix C show that the positive effect of receiving information from an ideologically aligned institution is larger for left-wing municipalities. However, this result should be treated with caution because, while the coefficient size is approximately three times larger, the difference between the effects is not statistically different at conventional confidence levels. Furthermore, the strength of the ideological alignment and misalignment with the partner institutions might be systematically different for right- and left-wing parties. While this is not a problem for the main estimations because ideology is a stratification factor in the randomization, the analysis of heterogeneity by policymaker's ideology should be interpreted cautiously.[40] The full results of the analyses of heterogeneity reported in Appendix C show that the effect of receiving information by an ideologically aligned institution on policy adoption seems to be larger for more populated municipalities and with lengthier Wikipedia pages. If population size and the length of Wikipedia pages indicate a municipality's capacity to implement the policy, these results suggest that alignment has a greater impact on municipalities with larger implementation capacity.

[39] Table D.I in the Appendix reports the p-values for the coefficients reported in Table 3.2 and other comparisons across treatment arms.

[40] This is discussed in detail in Appendix C.

Figure 3.3: Effects of the treatment arms on the probability of conducting recommended changes in the Spanish Wikipedia page of the municipality



Note: The figure displays the point estimates and 95% confidence intervals for the effect of the different treatment arms relative to the control group on the probability of conducting a recommended change in the municipality's page in the Spanish Wikipedia during the study period and the placebo period.

The coefficient estimate of *Nonsalient ideology - Policy brief* in Panel A, columns (1) and (2) of Table 3.2 is sizeable, indicating that information by an ideologically nonsalient institution increase policy adoption by 37%, although the p-value is above conventional significance levels (p-value=0.27) and the result should be interpreted with caution.[41] While the magnitude of the coefficient is approximately half as large as the effect of ideological alignment (0.94 versus 1.68 percentage points), both effects are not statistically significantly different from each other (p-value=0.34). The main estimates do not vary across specifications.[42]

The results of the heterogeneity analyses reported in Tables C.I in Appendix C show that the effect of receiving information by an ideologically nonsalient researcher has a strong effect in municipalities ruled by left-wing parties, while a negligible effect in municipalities ruled by right-wing parties. On the other hand, the results reported in Appendix C show no statistically differential effects of this treatment by the population of the municipality, length of its Wikipedia page, whether

[41] Table D.I in the Appendix reports the p-values for the coefficients reported in Table 3.2 and other comparisons across treatment arms.

[42] The results reported in Columns (3) and (4) present the placebo exercise described in Section 3.4. They show the expected small and non-significant coefficients that reveal no differences in Wikipedia changes across treatment arms during the placebo period.

the mayor belongs to one of the two main political parties (PP/PSOE) or a smaller one, and whether the mayor belongs to a party that promotes independence from Spain.

The coefficient estimate of *Opposite Ideology* in Panel B, columns (1) and (2) of Table 3.2 shows that receiving information communicated by an ideologically opposite institution does not affect the probability of changing the Wikipedia page along the recommended guidelines. The magnitude of the coefficients is negligible and not much higher than the coefficient for the same variable in the placebo analyses reported in Columns (3) and (4). The results suggest that receiving information from an ideologically opposite institution is not different from not receiving any information in terms of policy adoption. The results of the heterogeneity analyses reported in Tables C.II in Appendix C yield similar null effects of receiving information endorsed by an institution with an opposite ideology on the main policy adoption variable for right- and left-wing parties. The full results of the heterogeneity analyses reported in Appendix C also show equally null effects of receiving information by an ideologically opposite institution regardless of the population size of the municipality, length of its Wikipedia page, whether the mayor belongs to one of the two main political parties (PP/PSOE) or a smaller one, and whether the mayor belongs to a party that promotes independence from Spain.

Result 2: Information from aligned institutions promotes policy adoption, whereas information from ideologically opposite institutions does not increase policy adoption compared to an uninformed control group. The effect of receiving information from a prestigious ideologically nonsalient institution is nearly half as large as the effect of ideological alignment, although non statistically significant at conventional confidence levels.

Finally, we study whether the instrument used to communicate research evidence, i.e., policy brief vs. newspaper, differentially affects policy adoption. The coefficient estimates for *Policy Brief* and *Newspaper* reported in Panel B, columns (1) and (2) of Table 3.2 indicate that conditional on the ideological alignment of the municipality and the informing institution, the effects of policy briefs and newspaper articles on the probability of conducting a change in Wikipedia along the recommended guidelines are virtually the same. The difference in the coefficient estimates is small (0.1 percentage points) and largely statistically non-significant (p-value=0.82).[43]

From the results of the analysis comparing the effects of the newspaper and the policy briefs on policy adoption for the full sample, we conclude:

Result 3: Both policy reports and newspaper articles, when ideologically aligned with the receiving

[43]Table D.I in the Appendix reports the p-values for the coefficients reported in Table 3.2 and other comparisons across treatment arms.

policymaker, are equally effective in promoting policy adoption.

The results of the heterogeneity analysis reported in Appendix C show consistent null effects across different dimensions, including the mayor's ideology, the municipality's population, the length of its Wikipedia page, whether the mayor belongs to one of the two main political parties (PP/PSOE) or a smaller one, and whether the mayor belongs to a party that promotes independence from Spain.

The main results of the paper are unlikely to be driven by spillovers from information moving from treatment municipalities to control municipalities. First, the endline survey shows that, among the 236 respondents in control municipalities, only five of them reported having received any information. Secondly, the results of the spillover analysis reported in Table D.II show that for control municipalities, the distance to the nearest municipality in each treatment arm has no effect on our measure of policy adoption. Taken together, these results suggest very limited spillovers. Moreover, even if present, spillover from treatment to control group would bias the main estimates downwards, resulting in our findings representing a lower bound for the true effects. This would not affect the main conclusions regarding the impact of political alignment on policy adoption.

In Table D.III in Appendix, we replicate the main results of the analysis using the estimation method presented in Goldsmith-Pinkham et al. (2022). This method is used to account for contamination when estimating the effect of mutually exclusive treatments with control variables. The effects are nearly identical to those reported in Table 3.2.

How surprising are our main results? Comparing outside researchers' expectations in the Social Science Prediction Platform survey with actual results in our main experiment shows that respondents correctly anticipated the order of how effective our different treatments would be, although they overestimated them. Opening rates of the emails were the closest (49% predicted versus 38% opening rates at the email level; see Table D.V in the Appendix). Regarding click-through rates, once policymakers learned the ideology of the informing institution, survey respondents expected them to be much higher than the actual one (6.42%, see Table 3.3), in all treatments (46.16% in Aligned, 40.19% in Nonsalient and 30.2% in Opposite ideology, respectively). Similarly, as in DellaVigna et al. (2022), the rate of policy adoption was much lower than expected. Survey respondents expected high rates of policy adoption in all treatment arms: 33.9% with aligned ideologies, 28.3% with nonsalient ideology, 20.62% with opposite ideologies and 13.85% in the control treatment. Relative to control levels, the survey respondents predict that receiving information would increase the probability of treatment adoption by 144%, 103%, and 48% when informed by an institution with an aligned, nonsalient, or opposite ideology. The results of the experiment discussed above reveal much lower adoption rates and treatment effects in every treatment arm, highlighting the difficulties

of translating evidence into policy, even for the simplest policies.

Finally, we estimate the monetary cost of ideological misalignment between the policymaker and the informing institution to be 2,192 euros per municipality per year, in the context of the policy recommended using the estimated impact of Wikipedia changes on touristic revenues reported in Hinnosaar et al. (2021). Further details on this calculation are provided in Appendix G. Despite the large effect on adoption in relative terms, the low levels of adoption, even within the aligned institution treatment group, keep the cost moderate in the context of this policy recommendation. While the magnitude of the cost per municipality might seem small, it is important to highlight that the figure provided is not the cost of not implementing the policy (which we take from Hinnosaar et al., 2021) but the cost per municipality of informing policymakers using ideologically aligned versus opposite institutions.

3.6 At which stage of the policy adoption process does ideological alignment matter?

We propose a three-stage framework to understand how evidence-informed policies are ultimately implemented by policymakers: (1) information exposure, (2) belief updating, and (3) implementation.

Regarding the first stage, information acquisition, policymakers may choose to avoid exposure from sources with opposite ideologies (Stroud, 2010). To test whether ideological alignment could affect information exposure, we capitalize on the fact that in our experiment, all emails were sent from an email account of a small and non-ideological research center (TIDES), irrespective of the treatment arm. The identity of the endorsing institution is revealed prominently in the body of the email once policymakers have opened it. To access the full information on the policy recommendation or the instructions to change their municipalities' Wikipedia page, policy officials have to actively click on links that lead them to either a full-length policy report or a newspaper article and to step-by-step instructions on how to update Wikipedia. We cannot measure which sources of information policymakers seek or directly observe the amount of time or attention policymakers devote to the information in our emails. However, we can track whether email recipients open the email and whether, after learning which institution is providing the research information, they click on any of the mentioned links, the policy brief, the newspaper article, or the Wikipedia instructions, which arguably serve as a good proxy for policymakers' attention. Thus, our experimental design, separating email openings from click-through rates, allows us to differentiate the effect of information awareness from selective exposure to information provided by ideologically different sources. A

statistically significant difference in the click-through rates to the links across treatment arms could be interpreted as evidence supporting ideological alignment influencing policy adoption from the information exposure stage.

We previously showed in Section 3.4 that there were no statistically significant differences in the opening rates of emails across treatment groups. Now we test for differences across treatment groups regarding click-through rates to the links in the email. The results of these analyses are reported in Table 3.3 for the probability of making at least one click, and in Table D.VIII of Appendix D for the number of clicks in each of the links. Overall, we find no variations across treatment arms in the click-through rates to the policy brief/newspaper article or the Wikipedia instructions. The coefficients are small, and none is statistically significantly different from zero at the 5% significance level. Since data on click-through rates is available at both the email and municipality levels, we conducted analyses at both levels, consistently finding null effects. Taken together, these results suggest that selective exposure to research evidence does not explain the differences in *policy adoption* across treatment groups.

Ideological alignment may also affect how policymakers update their beliefs about the effectiveness of a policy in response to research evidence, which constitutes the second stage of our policy adoption framework. Specifically, policymakers may be more inclined to revise their beliefs when the informing institution is perceived as ideologically aligned or non-ideologically salient than when it is the opposite. Previous studies have shown that belief updates are more likely to take place when information is presented by trusted or in-group sources, such as by politicians who are from the same ideological positions or by non-ideological, neutral parties (Banuri et al., 2019; Afrouzi et al., 2023; Baekgaard et al., 2017; Christensen and Moynihan, 2020; Gentzkow et al., 2018; Cohen, 2003; Merkley and Stecula, 2021).

Table 3.3: Treatment effects in the probability of making at least one click

	At the municipality level		At the email level	
	(1)	(2)	(3)	(4)
Aligned ideology - Policy brief	0.0013	0.0017	0.0003	-0.0007
	(0.0173)	(0.0174)	(0.0099)	(0.0099)
Nonsalient ideology - Policy brief	0.0206	0.0207	0.0078	0.0064
	(0.0167)	(0.0168)	(0.0105)	(0.0106)
Aligned ideology - Newspaper	-0.0158	-0.0159	-0.0106	-0.0109
	(0.0157)	(0.0157)	(0.0107)	(0.0107)
Opposite ideology - Newspaper	-0.0250*	-0.0246*	-0.0120	-0.0119
	(0.0140)	(0.0141)	(0.0101)	(0.0099)
Reference group: Opposite ideology - Policy brief				
Mean dep variable	0.1227	0.1227	0.0642	0.0642
Strata FE	No	Yes	No	Yes
N	4,736	4,736	11,288	11,288

Note: The estimates presented in the table yield the effect of the different treatment arms on the probability of making at least one click in the links included in the email relative to the group of individuals that receive the summary of results endorsed by a think tank with an opposite ideology. The latter group is used as the committed category in the regressions since the control group did not receive the intervention email and we cannot measure clicks for them. Moreover, the number of municipalities included in the analysis is therefore smaller since the control group, which do not receive the email, are excluded from the analysis. The outcome variable is measured at the municipality level in columns (1) and (2). Because in some municipalities we had more than one email address, we estimate the effects in columns (3) and (4) with the outcome variable measured at the email address level. Estimates reported in columns (1) and (3) are estimated without strata fixed-effects and columns (2) and (4) are estimated with strata fixed-effects. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

To test the role of ideological alignment in belief updating, we conducted a survey with a large sample of local policymakers in Spain in late April/early May 2023.[44] The survey includes information on attitudes toward evidence-based policymaking and incorporates an online experiment to examine how ideological alignment affects the process of belief updating among policymakers. Participants were first asked their opinion on a non-ideological policy that differed from the policy used in our main experiment. Subsequently, they were presented with peer-reviewed evidence against this policy and asked whether they would implement it. Since the same think tanks provided information on the research evidence as in the main experiment, we can measure changes in beliefs once policymakers learn which ideological institution is presenting the evidence.[45]

The online survey experiment was divided into two parts. In the first part, we asked policymakers about their ex-ante beliefs about the effectiveness of displaying messages on road panels highlighting the number of casualties in road accidents to reduce the incidence of road accidents. Our prior was that most policymakers would believe such policy to be effective and, moreover, neutral from an ideological perspective. The vast majority of the policymakers surveyed expected either a beneficial

[44] A detailed description of the survey and the descriptive statistics are provided in Appendix A.

[45] The email inviting to complete the survey that includes the experiment was sent to all email addresses from an email account of a different non-ideologically aligned research institution, i.e., ESADE.

effect of the policy (i.e., a reduction in the number of casualties in road accidents) or no effect of the policy on this outcome. There were no ex-ante differences across policymakers of different ideologies (see Appendix A). Next, we presented policymakers with a summary of the counter-intuitive results in Hall and Madsen (2022): road panels announcing the number of casualties actually increase the number of accidents in their surroundings. The presented summary of the results was identical in all surveys, and we stressed that it had been published in a leading academic journal, i.e., *Science*. We randomized whether such information was communicated by a think tank with an aligned or opposite ideology or by an ideologically nonsalient institution. A representation of the experimental design and details about the randomization and sampling strategies are provided in Appendix A. Briefly, we rely on the same think tanks or ideologically nonsalient institutions as in our main experiment: FAES, Alternativas, and LSE. The randomization was conducted at the municipality level, and therefore, all the policymakers within the same municipality were in the same treatment arm. While uncommon, local governments in Spain may have formed coalitions of political parties with different ideologies. Because randomization was conducted at the municipality level, we excluded from the survey appointed policymakers who do not share the ideology of the mayor to avoid individuals within the same municipality receiving a different version of the online survey. For policymakers from municipalities included in the main experiment of policy adoption, we provided the same treatment (i.e., aligned, ideologically non-salient, or opposite). The new municipalities and those in the control group in the main experiment were randomized without stratification to receive the information from an ideologically aligned, nonsalient or opposite institution. In the survey experiment, there is no control group that does not receive any information.

After presenting the evidence, we asked policymakers whether they believed the study results and whether they would implement the road panels displaying the number of deaths. Table 3.4 shows the results for two samples. Panel A includes all individuals regardless of whether they believe the road messages are harmful. In Panel B, we restrict the analysis and explore belief updates among those individuals who, before presenting the evidence, did not believe that these messages increase road casualties. As expected, most individuals do not anticipate the negative effects of the panels documented in Hall and Madsen (2022), and the results are very similar across panels. However, the results reported in the table reveal that policymakers are more likely to believe the study results and to report that they would follow study recommendations when an ideologically aligned think tank communicates them. In other words, the results show that policymakers are more likely to believe a study published in the leading academic journal *Science* when it is communicated by a think tank with an aligned ideology than with an opposite ideology. Interestingly, we also find a large effect of receiving study results when an ideologically nonsalient institution communicated the

study, although the difference between both is not statistically significant.

Table 3.4: Effects of the treatment arms on update of beliefs about intervention effectiveness

	Believe study results (1)	Follow study recommendations (2)	Believe study results & follow recommendations (3)
Panel A: All individuals			
Aligned ideology	0.1816***	0.0931**	0.0799**
	(0.0357)	(0.0365)	(0.0327)
Nonsalient ideology	0.2075***	0.1425***	0.1437***
	(0.0398)	(0.0417)	(0.0373)
Reference group: Opposite ideology			
Mean dep var	0.4972	0.3408	0.2011
N	951	951	951
Panel B: Individuals with pre-treatment beliefs not aligned with study results			
Aligned ideology	0.1802***	0.0922**	0.0761**
	(0.0365)	(0.0369)	(0.0328)
Nonsalient ideology	0.2036***	0.1264***	0.1252***
	(0.0408)	(0.0425)	(0.0374)
Reference group: Opposite ideology			
Mean dep var	0.4942	0.3295	0.1936
N	916	916	916

Note: The table reports the effects of the on-line experiment. Panel A reports the effects of the different treatment arms on the probability of updating beliefs about the effectiveness of the intervention for all individuals who answered the survey. Panel B reports the effects of the different treatment arms on the probability of updating beliefs about the effectiveness of the intervention for individuals that belief the intervention has no negative impacts. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

The third stage of our framework encompasses everything that falls in between belief updating and policy implementation. This includes all the constraints typically associated with the actual implementation of the policy, such as political economy, budgetary or administrative limitations, as well as issues related to career incentives or party discipline largely explored in the literature (Besley, 2005; Cerna, 2013; Rodrik, 2018). For instance, a policymaker might be persuaded by a policy recommendation but might end up not implementing it because he or she does not have sufficient resources, stakeholders' support, or administrative capacity. On the other hand, a policymaker might not be persuaded by a policy recommendation from an ideologically aligned informer but still decide to implement it because of career concerns and/or electoral incentives to do so.

Comparing our main experiment with the online survey experiment allows us to study whether ideological alignment affects the second stage, namely belief updating, differently than the third

stage, namely policy implementation. While both experiments use the same think tanks and have comparable sub-samples, any comparison between the results should consider that both experiments' policies are different. For example, adopting a new policy, as in our main experiment, may be very different from removing an existing one, as shown by DellaVigna et al. (2022). However, both interventions are arguably ideologically nonsalient, with unequivocal rigorous evidence of their effectiveness. Comparisons would be informative if policymakers' decision processes were similar under both interventions. While the survey experiment (see results in Table 3.4) shows that policymakers update their beliefs similarly when the information comes from an ideologically aligned or an ideologically nonsalient institution, the main experiment shows that policymakers are, on average, nearly two times more likely to implement the policy when presented by an ideologically aligned institution than by an ideologically nonsalient institution, although these effects are not statistically different from each other. These findings suggest that there is something else, beyond policymakers' beliefs about policy effectiveness, driving actual policy implementation.

Our experimental design allows us to disregard two hypotheses that could explain this discrepancy. First, given that the proposed intervention is non-ideological and virtually costless, it is unlikely that policymakers encountered budgetary constraints for policy adoption. Second, we can investigate career concerns or reasons for party discipline. In the heterogeneity analysis reported in Table C.II in Appendix C, we find that the effect of receiving information from a think tank with aligned ideology is no different than receiving information from a think tank directly associated with the policymaker's political party.[46] This suggests that party discipline is unlikely driving the effect of ideological alignment on policy adoption during this stage. Further research is needed to understand how ideological alignment curbs policy adoption at this stage.

3.7 Discussion and Conclusion

For evidence-based policies to be adopted, research results must be communicated to policymakers. The extent to which policymakers are persuaded by institutions communicating evidence is crucial for scientific dissemination to have an impact in the real world. In this paper, we offer a unique experimental design that allows us to introduce variation in the ideology of the informing institution while keeping everything else constant. This allows us to test how much the ideological alignment between the informer - in this case, institutions with salient ideologies - and policymakers can influence policy adoption. We benefit from a unique collaboration with widely known think tanks and media outlets of opposite ideologies and a large database of direct contacts of local policymakers

[46] Fundación Alternativas was originally associated with the Socialist Party, and its president is a former member of parliament and a member of the socialist party, while Fundación FAES was originally associated with the Popular Party, and is presided over by a former conservative Spanish Prime Minister and a member of the party.

in Spain to isolate the effect of ideological alignment in a controlled environment.

Our results show that ideological alignment between informing institutions and policymakers substantially increases the adoption of an ideologically nonsalient policy based on scientific results, while information from ideologically opposite institutions is as ineffective as not receiving any information. Confirming this intuition, with actual estimates obtained in a controlled setting, is important because advocates of the adoption of evidence-based policies often focus on the creation and expansion of ideologically neutral institutions as the key to disseminating research and informing policymakers. However, institutions with salient ideologies, whether explicit or perceived, do exist and also play an active role in disseminating research results, whether directly, like think tanks, or indirectly, like media outlets. In fact, we estimate the effectiveness of ideological institutions in leading policymakers to implement evidence-based policies. However, they may fail with those policymakers whose ideologies are opposite. Understanding the trade-off between effectiveness and outreach is important for the debate regarding the role of institutions and which specific features make them better suited to act as knowledge brokers to disseminate research results. Additionally, our results contribute to the wider debate about designing institutions where the characteristics of who provides information matter. An interesting example is the judicial system: while in the United States courts, both sides get to pick their own expert witnesses, in Europe, it is more common to force both sides to agree on a single expert witness beforehand. Understanding the different incentives and trade-offs of both types of institutional designs in light of these messenger effects is crucial for designing effective state institutions.[47]

Our study also provides insights into the stages of policymakers' decision-making processes, from their access to research evidence to policy implementation. This is also crucial because we show that the ideology of the informing institution significantly affects the way policymakers update their beliefs and, ultimately, their adoption into policy, even in contexts in which the policy itself is not ideological.

Importantly, our paper may showcase a low threshold on the influence of ideology in evidence-based policy adoption since its effect may presumably be much larger when ideological discrepancies exist about the evidence itself and not only about the informing institution. Examples such as climate change or vaccine adoption easily come to mind. Further research is needed for a deeper understanding of the role of knowledge brokers in the adoption of evidence-based policies that are perceived as ideologically loaded.

Interesting avenues for further research include studying other aspects of the science communication process that could affect policy implementation beyond perceived ideology. For example,

[47] We are thankful to a previous anonymous referee for suggesting this example.

further research could focus on getting a more granular understanding of the role of other trusted or authoritative institutions, such as governmental agencies, policy evaluation institutions, different levels of government, other scientific institutions, or political parties, in disseminating scientific evidence. International institutions and governments invest resources and create new institutions to promote scientifically informed policy-making. Designing effective institutions that disseminate scientific evidence will be crucial to improving the adoption of research evidence in the context of increasing political polarization.

Bibliography

Acemoglu, Daron and James A. Robinson, "Economics versus Politics: Pitfalls of Policy Advice," *Journal of Economic Perspectives*, 2013, *27* (2), 173–192.

Afrouzi, Hassan, Carolina Arteaga, and Emily K Weisburst, "Is it the Message or the Messenger? Examining Movement in Immigration Beliefs," Technical Report, National Bureau of Economic Research 2023.

Alonso, Ricardo and Gerard Padró i Miquel, "Competitive Capture of Public Opinion," NBER Working Papers 31414, National Bureau of Economic Research, Inc 2023.

Arkhangelsky, Dmitry, Susan Athey, David A. Hirshberg, Guido W. Imbens, and Stefan Wager, "Synthetic Difference-in-Differences," *American Economic Review*, December 2021, *111* (12), 4088–4118.

Arnautu, Diana and Christian Dagenais, "Use and effectiveness of policy briefs as a knowledge transfer tool: a scoping review," *Humanities and Social Sciences Communications*, 2021, *8.*

Baekgaard, Martin, Julian Christensen, Casper Dahlmann, Asbjørn Mathiasen, and Niels Bjørn Grund Petersen, "The Role of Evidence in Politics: Motivated Reasoning and Persuasion among Politicians," *British Journal of Political Science*, 08 2017, *49*, 1–24.

Banerjee, Abhijit, Marcella Alsan, Emily Breza, Arun G Chandrasekhar, Abhijit Chowdhury, Esther Duflo, Paul Goldsmith-Pinkham, and Benjamin A Olken, "Messages on COVID-19 Prevention in India Increased Symptoms Reporting and Adherence to Preventive Behaviors Among 25 Million Recipients with Similar Effects on Non-recipient Members of Their Communities," Working Paper 27496, National Bureau of Economic Research July 2020.

Banuri, Sheheryar, Stefan Dercon, and Varun Gauri, "Biased Policy Professionals," *World Bank Economic Review*, 2019, *33* (2), 310–327.

Barceló, Joan and MAURICIO Vela Barón, "Political Responsiveness to Conflict Victims: Evidence from a Countrywide Audit Experiment in Colombia," *American Political Science Review*, 2023, p. 1–17.

Bénabou, Roland and Jean Tirole, "Mindful Economics: The Production, Consumption, and Value of Beliefs," *Journal of Economic Perspectives*, September 2016, *30* (3), 141–64.

Besley, Timothy, "Political Selection," *Journal of Economic Perspectives*, September 2005, *19* (3), 43–60.

Butler, Daniel M. and David E. Broockman, "Do Politicians Racially Discriminate Against Constituents? A Field Experiment on State Legislators," *American Journal of Political Science*, 2011, *55* (3), 463–477.

_ , Craig Volden, Adam M. Dynes, and Boris Shor, "Ideology, Learning, and Policy Diffusion: Experimental Evidence," *American Journal of Political Science*, 2017, *61* (1), 37–49.

Callaway, Brantly and Pedro H.C. Sant'Anna, "Difference-in-Differences with multiple time periods," *Journal of Econometrics*, 2021, *225* (2), 200–230.

Cerna, Lucie, "The nature of policy change and implementation: A review of different theoretical approaches," *Organisation for Economic Cooperation and Development (OECD) report*, 2013, pp. 492–502.

Christensen, Julian and Donald P. Moynihan, "Motivated reasoning and policy information: politicians are more resistant to debiasing interventions than the general public," *Behavioural Public Policy*, 2020, p. 1–22.

Cohen, Geoffrey, "Party Over Policy: The Dominating Impact of Group Influence on Political Beliefs," *Journal of personality and social psychology*, 12 2003, *85*, 808–22.

Crawford, Vicent P. and Joel Sobel, "Strategic Information Transmission," *Econometrica*, 1982, *50* (6), 1431–1451.

DellaVigna, Stefano and Devin Pope, "What Motivates Effort? Evidence and Expert Forecasts," *The Review of Economic Studies*, 06 2017, *85* (2), 1029–1069.

_ and Woojin Kim, "Policy diffusion and polarization across US states," Working Paper 30142, National Bureau of Economic Research 2022.

_ , _ , and Elizabeth Linos, "Bottlenecks for Evidence Adoption," Working Paper 30144, National Bureau of Economic Research June 2022.

Dercon, Stefan, "The Political Economy of Economic Policy Advice," Technical Report 2023-09, CSAE Working Paper WPS 2023.

Diamond, Emily and Jack Zhou, "Whose policy is it anyway? Public support for clean energy policy depends on the message and the messenger," *Environmental Politics*, 2022, *31* (6), 991–1015.

Druckman, James N and Mary C McGrath, "The evidence for motivated reasoning in climate change preference formation," *Nature Climate Change*, 2019, *9* (2), 111–119.

Druckman, James N., Samara Klar, Yanna Krupnikov, Matthew Levendusky, and John Barry Ryan, "How Affective Polarization Shapes Americans' Political Beliefs: A Study of Response to the COVID-19 Pandemic," *Journal of Experimental Political Science*, 2021, *8* (3), 223–234.

European Commission, "Staff Working Document - Supporting and connecting policymaking in the Member States with scientific research," 2022.

Favero, Nathen, Sebastian Jilke, Julia A. Wolfson, Chengxin Xu, and Matthew M. Young, "Messenger effects in COVID-19 communication: Does the level of government matter?," *Health Policy OPEN*, 2021, *2*, 100027.

Gaikwad, Nikhar and Gareth Nellis, "Do Politicians Discriminate against Internal Migrants? Evidence from Nationwide Field Experiments in India," *American Journal of Political Science*, 2021, *65* (4), 790–806.

Gao, Benjamin F. Jones Yian Yin Jian and Dashun Wang, "Coevolution of policy and science during the pandemic," *Science*, 2021, *371* (6525), 128–130.

Gentzkow, Matthew, Jesse Shapiro, and Daniel Stone, *Media Bias in the Marketplace* 01

_ , **Michael B Wong, and Allen T Zhang**, "Ideological bias and trust in information sources," *Unpublished manuscript*, 2018, *1* (1), 1–43.

Goldsmith-Pinkham, Paul, Peter Hull, and Michal Kolesar, "Contamination Bias in Linear Regressions," Working Paper 30108, National Bureau of Economic Research June 2022.

Grossman, Emiliano, "Media and policy making in the digital age," *Annual Review of Political Science*, 2022, *25*, 442–461.

Guilbeault, Douglas, Joshua Becker, and Damon Centola, "Social learning and partisan bias in the interpretation of climate trends," *Proceedings of the National Academy of Sciences*, 2018, *115* (39), 9714–9719.

Hall, Jonathan D. and Joshua M. Madsen, "Can behavioral interventions be too salient? Evidence from traffic safety messages," *Science*, 2022, *376* (6591), eabm3427.

Hinnosaar, Marit, Toomas Hinnosaar, Michael E. Kummer, and Olga Slivko, "Wikipedia Matters," *Journal of Economics & Management Strategy*, 2021, pp. 1–13.

_ , _ , _ , **and** _ , "Externalities in knowledge production: evidence from a randomized field experiment," *Experimental Economics*, April 2022, *25* (2), 706–733.

Hjort, Jonas, Diana Moreira, Gautam Rao, and Juan Francisco Santini, "How Research Affects Policy: Experimental Evidence from 2,150 Brazilian Municipalities," *American Economic Review*, May 2021, *111* (5), 1442–80.

Kahan, Hank Jenkins-Smith Dan M. and Donald Braman, "Cultural cognition of scientific consensus," *Journal of Risk Research*, 2011, *14* (2), 147–174.

Kremer, Michael, Sasha Gallant, Olga Rostapshova, and Milan Thomas, "Is Development Innovation a Good Investment? Which Innovations Scale? Evidence on social investing from USAID's Development Innovation Ventures," *Working paper*, 2019.

Kunda, Ziva, "Motivated Inference: Self-Serving Generation and Evaluation of Causal Theories," *Journal of Personality and Social Psychology*, 10 1987, *53*.

Lee, Nathan, "Do Policy Makers Listen to Experts? Evidence from a National Survey of Local and State Policy Makers," *American Political Science Review*, May 2022, *116* (2), 677–688.

Linos, Elizabeth, "Translating Behavioral Economics Evidence into Policy and Practice," Technical Report, APO: Analysis & Policy Observatory 2023.

Maclean, Johanna Catherine, John Buckell, and Joachim Marti, "Information Source and Cigarettes: Experimental Evidence on the Messenger Effect," NBER Working Papers 25632, National Bureau of Economic Research, Inc March 2019.

Majo-Vazquez, González-Bailóm, "Polarización en las audiencias de los medios en España," 2022.

Malmendier, Ulrike, Stefan Nagel, and Zhen Yan, "The Making of Hawks and Doves: Inflation Experiences on the FOMC," Working Paper 23228, National Bureau of Economic Research March 2017.

Masset, Edoardo, Marie Gaarder, Penelope Beynon, and Christelle Chapoy, "What is the impact of a policy brief? Results of an experiment in research dissemination," *Journal of Development Effectiveness*, 2013, *5* (1), 50–63.

Mehmood, Sultan, Shanheen Naseer, and Daniel L. Chen, "Training Policymakers in Econometrics," *NBER Working Paper*, 2024.

Merkley, Eric and Dominik A. Stecula, "Party Cues in the News: Democratic Elites, Republican Backlash, and the Dynamics of Climate Skepticism," *British Journal of Political Science*, 2021, *51* (4), 1439–1456.

Nakajima, Nozomi, "Evidence-based decisions and education policymakers," *Unpublished Paper*, 2021.

OECD, "Public Governance Reviews," *Building Capacity for Evidence-Informed Policy-Making. https://www.oecd.org/gov/building-capacity-for-evidence-informed-policy-making-86331250-en.htm*, 2020, p. 80.

_ , *OECD Tourism Trends and Policies 2022 https://www.oecd.org/cfe/tourism/oecd-tourism-trends-and-policies-20767773.htm* 2022.

Rodrik, Dani, "Understanding economic policy reform," in "Modern Political Economy and Latin America," Routledge, 2018, pp. 59–70.

Stroud, Natalie, "Polarization and Partisan Selective Exposure," *Journal of Communication*, 09 2010, *60*, 556 – 576.

Taber, Charles S. and Milton Lodge, "Motivated Skepticism in the Evaluation of Political Beliefs," *American Journal of Political Science*, July 2006, *50* (3), 755–769.

Toma, Mattie and Elizabeth Bell, "Understanding and Increasing Policymakers' Sensitivity to Program Impact," *Unpublished Paper*, 2022.

Vivalt, Eva and Aidan Coville, "How do policymakers update their beliefs?," *Journal of Development Economics*, 2023, *165*, 103121.

Wang, Shaoda and David Yang, "Policy Experimentation in China: The Political Economy of Policy Learning," *NBER Working Paper No. 29402*, 2021.

Wu, Kelly Zhao Meng-Jia and Francisca Fils-Aime, "Response rates of online surveys in published research: A meta-analysis," *Computers in Human Behavior Reports*, 2022, *7*, 100206.

Ideological Alignment and Evidence-Based Policy Adoption

Online Appendix

Jorge García-Hombrados

Marcel Jansen

Angel Martínez

Berkay Özcan

Pedro Rey-Biel

Antonio Roldán-Monés

A Endline survey

During April and May 2023, we conducted an online survey among an expanded sample of mayors, councilors, and officials of all municipalities to which we could assign an ideology.[48] In total, we sent invitations to 17,044 policymakers from 7,576 municipalities. Besides questions about their views on the relevance of scientific evidence for local public policies, the survey includes an online experiment to test how policymakers update their beliefs after receiving information about the true impact of a policy. As in our main experiment, we vary the ideological orientation of the institution that communicates the evidence. To avoid an explicit link with our main experiment, the survey invitations were sent by ESADE EcPol, a research-oriented think tank with a stated interest in evidence-based policies.[49] Below we describe the four modules of our endline survey along with the responses to a selection of the questions.

Common background questions

A total of 1,600 policymakers from 1,196 municipalities completed the survey, including 1,077 municipalities that were also included in the main experiment. Because one of the goals of the survey is characterizing policy-making in tourist municipalities and testing some of the assumptions of the main experiment, survey responses were presented in this section by treatment assignment in the main experiment, and also separately for those municipalities that were not tourist and therefore, not included in the main experiment. The 1,077 tourist municipalities that responded to the on-line survey were not a random sample of the tourist municipalities included in the first experiment. The descriptive statistics reported in Table A.I show that they are overall larger, with more comprehensive Wikipedia pages, although they have on average the same number of tourist accommodations per capita. While the survey respondents of the on-line survey were not representative, the number is large enough to provide valuable information about tourist municipalities in Spain.

Turning to the on-line survey, in the first question, the participants were asked to indicate their position within the municipality government by selecting one out of five possible options: mayor, councilor, administrative staff, tourism officers, and political advisor. Figure A.I shows the distribution of survey respondents among these categories by treatment status in the main intervention. Given the small sample size, the categories of tourism officers and political advisors are grouped together under the label "other". Among the municipalities of our main experiment, we observe no relevant differences in the distribution by treatment status. The majority of participants (50%-53%) are councilors in the municipality and members of the mayor's party or the ruling coalition.

[48] To assign the ideology of the municipality, we follow the same procedure used in the main experiment, which is described in Section 3.2.

[49] https://www.esade.edu/ecpol/en/

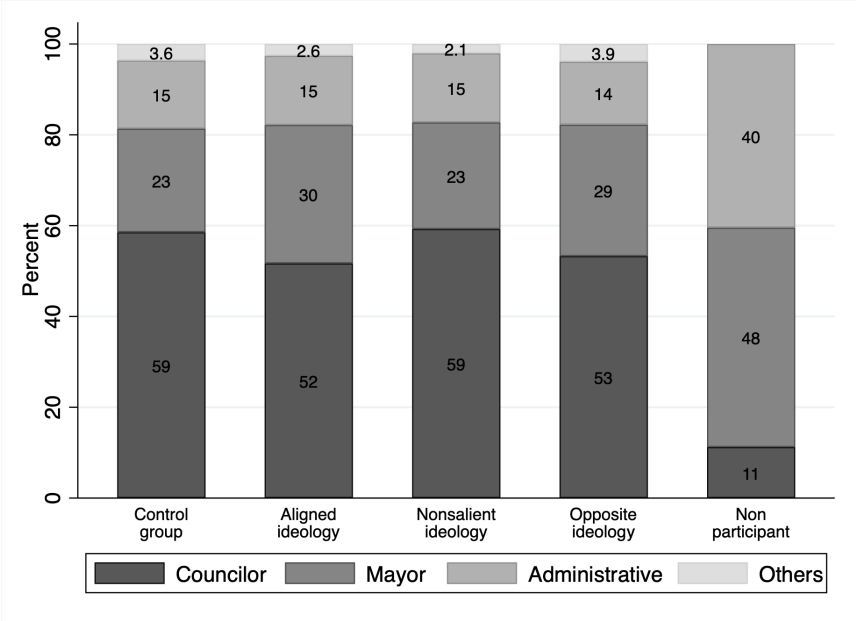
Table A.I: Difference in means: munici in both the on-line and main experiment vs participants only in the main experiment

	Mean Municipalities included in main experiment that responded online survey (1)	Mean Municipalities included in main experiment that did not respond online survey (2)	Difference (3)
Recommended Changes (0/1)	0.05	0.0302	-.017***
N words Sp	2,052.12	1,199.04	-853.09***
N words En	410.99	200.79	-210.20***
N images Sp	31.75	24.97	-6.78***
N images En	17.27	13.49	-3.77***
N languages	37.64	33.66	-3.98***
Tourist accom p/c	0.06	0.06	-0.00
Population	21,990.77	4,951.42	-17039.36***
N	1,077	4,601	-

Note: This table presents the mean and the difference in means between tourist municipalities that were included in the main experiment and also responded to the on-line survey, and tourist municipalities that were included in the main experiment but did not respond to the on-line survey.

The second largest group is mayors, who represent between 26% and 31% of the respondents, depending on treatment status. Finally, between 15% and 18% of the respondents are administrative staff, while less than 4% are tourism officers or political advisors. On the contrary, in the case of those municipalities that did not participate in our main experiment ("Nonparticipant"), others and administrative staff account for almost 90% of the respondents.

Figure A.I: Distribution of survey participants' occupations by treatment arm:



Note: All percentages are calculated excluding people who did not answer that question in the survey (n=1,323).

The second module includes questions that are relevant to our main intervention. Table A.II reports the responses to the first three questions regarding the respondents' views on the usefulness of scientific evidence in designing municipal public policies, the desirability of attracting more tourists to their municipality, and the effectiveness of Wikipedia as a means to achieve this objective. The answers are grouped by treatment status of the respondent's municipality in our main intervention, with columns (7) and (8) reporting the results for the municipalities that were excluded from our main experiment. Besides the tabulation of the responses, the table also includes a test for differences in means with respect to the municipalities in the control group.

The survey responses strengthen the credibility of our main intervention. The vast majority of respondents (over 80% across all treatment arms) consider scientific evidence to be quite or very useful for the design of municipal policies (Q2). Furthermore, around 70% of the respondents from the municipalities in our main experiment consider attracting more tourists a very desirable goal (Q3). Additionally, 60% of respondents believe that Wikipedia can be a fairly or very effective tool to achieve this goal (Q4). Importantly, we do not observe statistically significant differences across treatment arms relative to the control group.

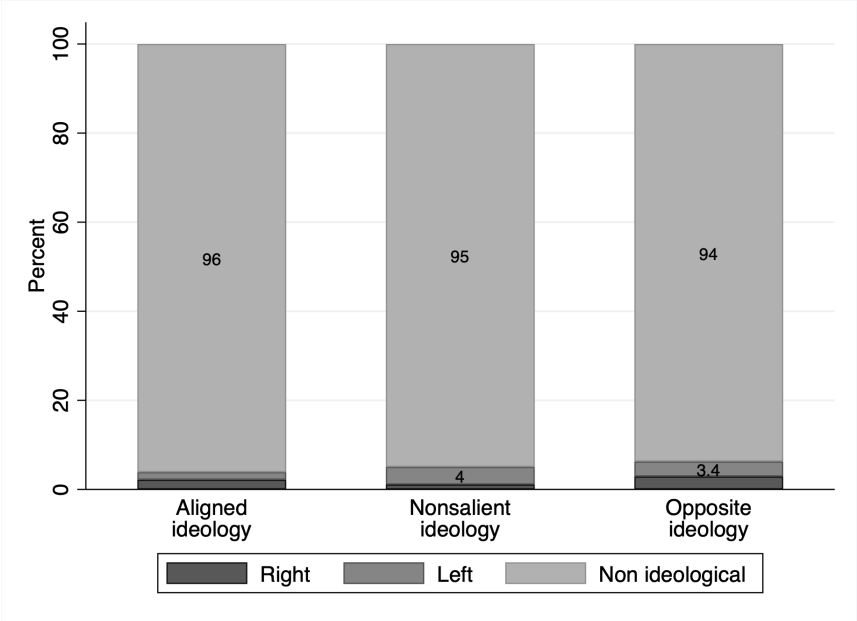
Table A.II: Responses to questions related to the main experiment:

	Aligned ideology		Nonsalient ideology		Opposite ideology		Non participant		Control group	
	(1) Perc	(2) Diff	(3) Perc	(4) Diff	(5) Perc	(6) Diff	(7) Perc	(8) Diff	(9) Perc	(10) Diff
Q2: Do you think that scientific evidence is useful for designing municipal public policies? (n=1,107)										
Very	36.22	-1.67	42.79	4.9	41.06	3.17	33.78	-4.1	37.89	-
Quite	47.68	3.58	44.23	.13	44.87	.77	41.89	-2.21	44.10	-
A bit	13.62	-1.91	10.58	-4.95	11.14	-4.38	16.22	.69	15.53	-
No	2.48	-.01	2.40	-.08	2.93	.45	8.11	5.62**	2.48	-
Q3: Do you think that increasing the number of tourists in your municipality would be a desirable goal? (n=1,108)										
Very	35.29	1.13	32.04	-2.12	36.73	2.57	37.33	3.17	34.16	-
Quite	38.39	7.33	37.86	6.81	37.03	5.97	34.67	3.61	31.06	-
A bit	21.36	.24	21.84	.73	18.37	-2.75	21.33	.22	21.12	-
No	4.95	-8.71***	8.25	-5.41*	7.87	-5.79**	6.67	-7	13.66	-
Q4: Do you think that having a good entry of your municipality on Wikipedia can be an effective way to attract tourism to your municipality? (n=1,109)										
Very	19.44	.81	20.19	1.56	21.11	2.48	16.00	-2.63	18.63	-
Quite	40.74	-2.74	40.38	-3.09	46.04	2.56	49.33	5.86	43.48	-
A bit	33.95	2.89	35.58	4.52	26.10	-4.96	26.67	-4.39	31.06	-
No	5.86	-.97	3.85	-2.99	6.74	-.09	8.00	1.17	6.83	-

Note: All percentages are calculated on the total number of participants who responded to that question in the survey. Columns (1), (3), (5), (7) and (9) only show the percentage of each response per treatment. Columns (2), (4), (6) and (8) show the percentage difference between that group and the Control group, which will be the reference group in the difference of means test that is carried out.

Another key assumption underlying our main intervention is the proposed policy's non-ideological nature, i.e., Wikipedia improvements to foster tourism. In the fifth question of the second module, we asked the participants to classify the proposed policy as either right-wing, left-wing, or neutral. The results, reported in Figure A.II, show that 94% to 96% of the participants consider the policy ideologically neutral, with no significant differences across treatment arms.

Figure A.II: The perceived ideology of the proposed policy

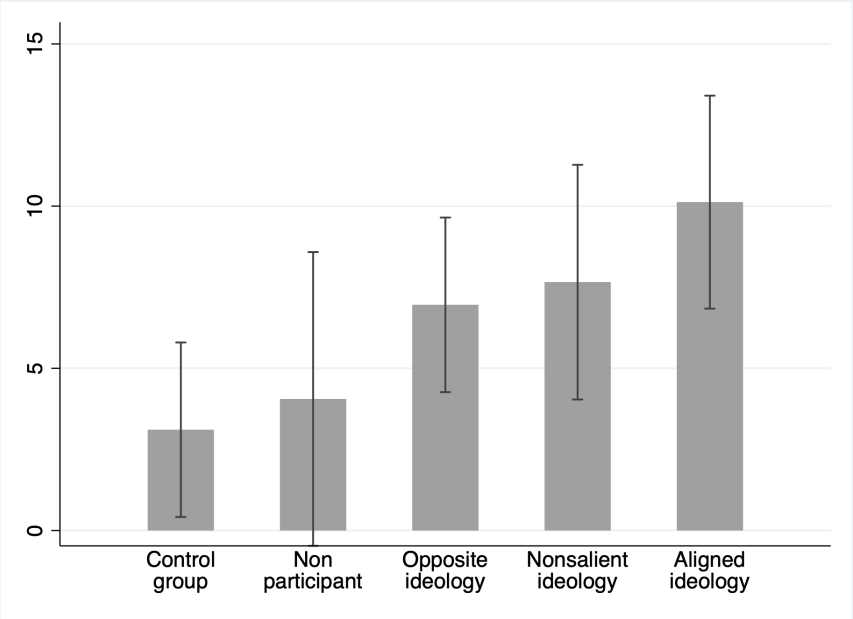


Note: All percentages are calculated excluding people who did not answer that question in the survey (n=1,105).

Next, participants were asked if they remembered receiving an email from TIDES related to our main experiment.[50] We included this question to check for contamination across treatments. Reassuringly, only 3% of the respondents in the control group and 4% of the respondents from excluded municipalities declare to have received an email from TIDES, as shown in Figure A.III. In contrast, the corresponding figure for the treated municipalities ranges from 6% for those municipalities that received information from a think tank or newspaper with an opposite ideology to 10% for the group that received information from an aligned source. Recall that the endline survey was sent almost six months after the date of the last mailing by TIDES. This helps to explain the relatively low recall rates.

[50] After an affirmative answer, the participants were subsequently asked to reveal their trust in the provided information, their updating of beliefs and actions to improve the municipalities Wikipedia page as well as the motives behind inaction for those who did not undertake any action. Unfortunately, however, the response rate for these questions was too low to draw any reliable conclusion.

Figure A.III: Percentage of participants who remember receiving a TIDES mailing:



Note: Figure displays percentage of individuals in each treatment arm that remember receiving a TIDES email and 95% confidence intervals. All percentages are calculated excluding people who did not answer that question in the survey (n=1,115).

Finally, in the last question of the survey, we asked the participants about their personal ideological affinity. A comparison between this self-reported ideology and the ideology assigned to their municipality in the main experiment reveals a high level of correspondence, but the correlation is not perfect. 20% of the respondents state a different ideology than the one assigned to their municipality. Moreover, the difference is more pronounced when the ruling party is right-wing. Nonetheless, the differences are much smaller when we restrict the comparison to mayors and councilors.

Survey experiment

The third part of the questionnaire included our survey experiment. For the sake of comparability, we use the same think tanks as in the main experiment to inform participants about the undesirable effects of messages on road panels to avoid speeding. Random treatment assignment was such that municipalities received the endorsement of this alternative policy from the same ideological spectrum as in the previous experiment. Municipalities that were not part of the main experiment and those assigned to the control group of the main experiment were equally randomized across all three treatments. Crucially, municipalities on both sides of the political spectrum received this new research evidence either from the same ideology or the opposite one. Figure A.IV displays the experimental design of the on-line survey. We then asked participants whether they would now

implement the policy, allowing us to study whether the political endorsement of the policy changed their beliefs about its effectiveness differently across treatments.

The results of the survey experiment have already been presented in Table 3.4 and discussed at length in Section 3.6. Below we present two further pieces of evidence. Table A.III reports the distribution of the answers to two questions of our survey experiment. The first question asked about prior beliefs, and the second post-treatment question was about belief updating and implementation. As expected, the vast majority of participants who responded to the first question indicated that they believed the policy would have the opposite effect to what the research presented to them immediately afterward. Specifically, 87% and 90% responded that the policy would reduce the target outcome, with minimal variations between treatment arms, while the scientific evidence provided shows that it increases the target outcome. Next, an inspection of the bottom panel shows that the percentage of respondents who state they believe the study and would refrain from implementing the counter-effective policy is significantly higher for those who received the information from non-salient or aligned think tanks than for those who are informed by a think tank on the opposite side of the political spectrum. Reversely, the percentage of those who disbelieve the study results and would insist on posting warnings against speeding is significantly higher among those who receive information from think tanks with an opposite ideology.

Finally, Table A.IV presents the results of a robustness check when we restrict the sample to respondents from the municipalities that were included in our main experiment. The results are virtually identical to those presented in Table 3.4.

Figure A.IV: Implementation and design of the on-line experiment

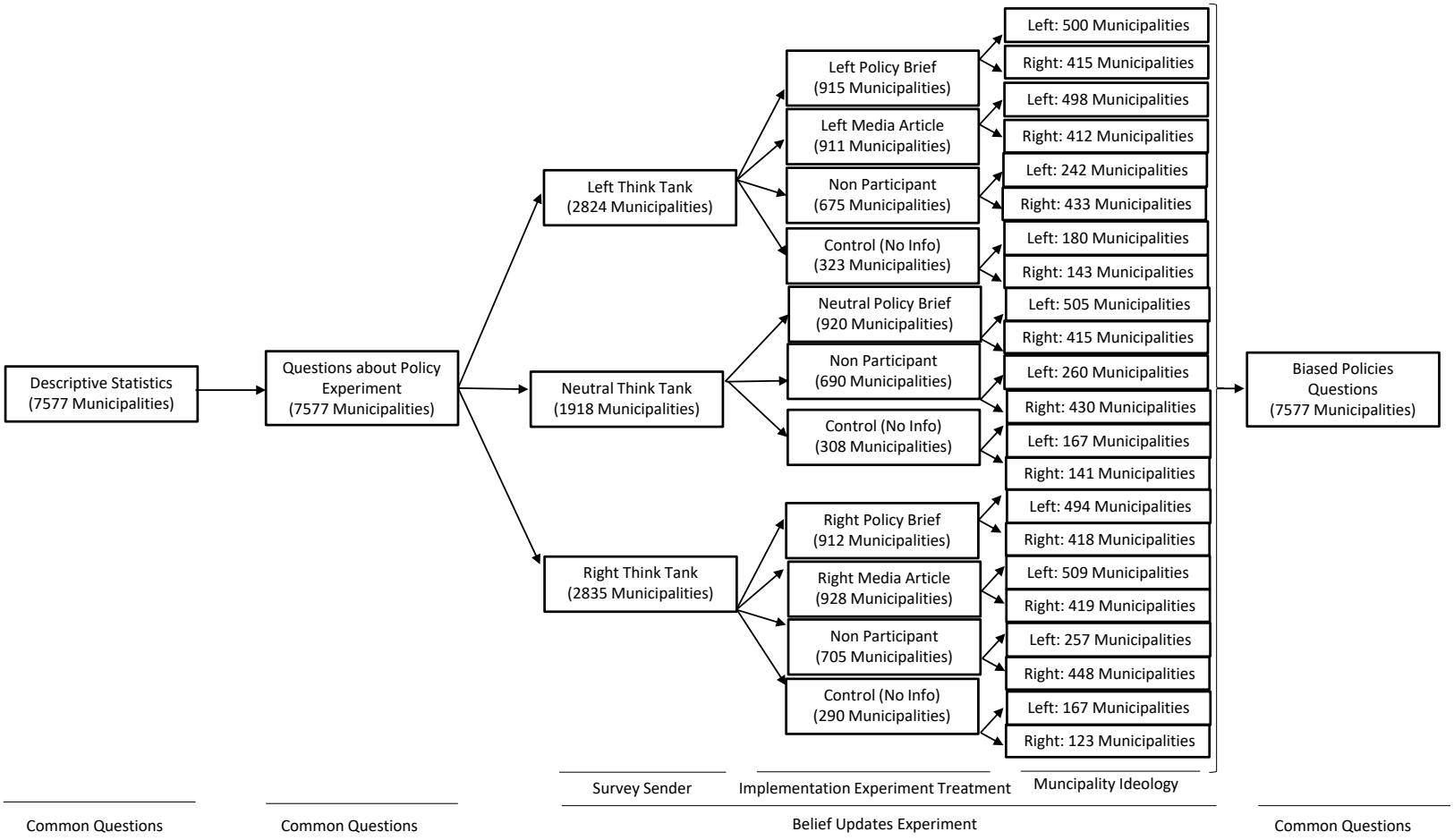


Table A.III: Responses to survey experiment questions by treatment arm:

	Aligned ideology		Nonsalient ideology		Opposite ideology	
	(1) Perc	(2) Diff	(3) Perc	(4) Diff	(5) Perc	(6) Diff
Q11: Do you think these messages may have any effect on the number of traffic accidents? (n=1,065)						
Substantial increase	0.51	-.24	0.00	-.74	0.74	-
Slight increase	2.27	.05	4.91	2.68*	2.23	-
No effect	8.33	.17	7.55	-.62	8.17	-
Slight reduction	58.33	-1.57	63.02	3.12	59.90	-
Substantial reduction	30.56	1.6	24.53	-4.43	28.96	-
Q12: If your goal was to reduce the number of road accidents, would you put this type of informational message on roadside panels? (n=955)						
Yes - Don´t believe study	16.76	-19.55***	15.90	-20.41***	36.31	-
Yes - Believe the study	39.66	10.06***	35.98	6.37	29.61	-
No - Don´t believe study	15.36	1.4	13.81	-.16	13.97	-
No - Believe the study	28.21	8.1**	34.31	14.2***	20.11	-

Note: All percentages are calculated on the total number of participants who responded to that question in the survey. Columns (1), (3) and (5) only show the percentage of each response per treatment. Columns (2), (4) and (6) show the percentage difference between that group and the Not Corresponding group, which will be the reference group in the difference of means test that is carried out.

Table A.IV: Treatment effects on update of beliefs only among participants in the main experiment

	Believe study results (1)	Follow study recommendations (2)	Believe study results & follow recommendations (3)
Panel A: All those who participated in the main experiment			
Aligned ideology	0.1955***	0.0966**	0.0782**
	(0.0366)	(0.0378)	(0.0340)
Nonsalient ideology	0.2362***	0.1216***	0.1452***
	(0.0408)	(0.0436)	(0.0391)
Reference group: Opposite ideology			
Mean dep var in reference group	0.6243	0.4128	0.2722
N	886	886	886
Panel B: Only individuals with pre-treatment beleifs not aligned with study results			
Aligned ideology	0.1938***	0.0954**	0.0738**
	(0.0374)	(0.0382)	(0.0343)
Nonsalient ideology	0.2345***	0.1054**	0.1286***
	(0.0417)	(0.0444)	(0.0393)
Reference group: Opposite ideology			
Mean dep var in reference group	0.4924	0.3364	0.2018
N	855	855	855

Note: The table reports the effects of the on-line experiment. Panel A reports the effects of the different treatment arms on the probability of updating beliefs about the effectiveness of the intervention for all individuals who answered the survey. Panel B reports the effects of the different treatment arms on the probability of updating beliefs about the effectiveness of the intervention for individuals that belief the intervention has no negative impacts. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

B Treatment arms: Policy briefs and newspaper articles

B.1 Text of the emails

This email was sent alternatively mentioning the publication of one of the five media through which the information was disseminated. Therefore, the variation in the paragraphs in which the respective medium or organization is mentioned is highlighted here.

Dear Mr/Ms. Councillor of "FINAL_NAME":

From the University Institute of Tourism and Economic Development (TIDES), we are contacting you to send you the results of a study that shows with data the beneficial effects of an effective, simple, and zero-cost intervention to increase tourism in "FINAL_NAME".

The research finds that simple changes to the Wikipedia page of municipalities like yours reported improvements of up to 33% in tourist income.

Fundación Alternativas*:[The study is summarised in the report published by the progressive ideas' laboratory Fundación Alternativas, directed by former socialist deputy Diego López Garrido]*

FAES*:[The study is summarised in the report published by the conservative think tank FAES, chaired by former popular president José María Aznar]*

LSE*:[The study is summarised in the following report published by researchers from the London School of Economics]*

El Mundo*:[The study is summarised in an article by the conservative media El Mundo, in its digital version, directed by Joaquín Manso]*

elDiario.es*:[The study is summarised in an article by the progressive digital media elDiario.es, directed by Ignacio Escolar]*

Insert the corresponding logo [Fundación Alternativas] [elDiario.es] [El Mundo] [FAES] [LSE]

Research shows that Wikipedia is a key instrument to promote tourism in Spanish municipalities

Insert the following text with the link to the article, depending on the media mentioned:

Fundación Alternativas:*[:Here, you can access the link to read the full report from the Alternativas Foundation]*

FAES:*[Here, you can access the link to read the full report from the FAES Foundation]*

LSE:*[Here, you can access the link to read the full report from researchers at the London School of Economics]*

El Mundo:*[Here, you can access the link to read the full El Mundo article]*

elDiario.es:*[Here, you can access the link to read the full elDiario.es article]*

Improvements to your municipality's Wikipedia page are straightforward- and free- and can generate more than 200,000 euros in tourist income in your municipality. Among the most effective

simple changes you could implement to your Wikipedia page are:

1. *Add photographs of your municipality.*
2. *Edit a simple English version of the exact text for foreign tourists.*
3. *Mention or expand the section dedicated to local festivals.*

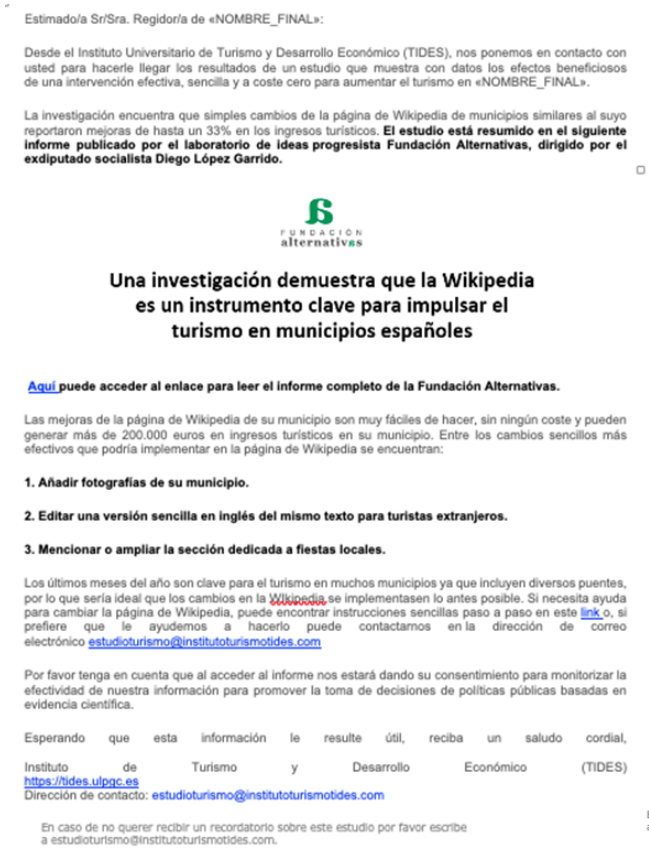
The last months of the year are critical for tourism in many municipalities since they include various bank holidays, so the changes in Wikipedia should be implemented as soon as possible. If you need help changing the Wikipedia page, you can find simple step-by-step instructions in this link, or, if you prefer us to help you do it, you can contact us at the email address estudioturismo@institutoturismotides.com

Please keep in mind that by accessing the report, you will be giving us your consent to monitor the effectiveness of our information to promote public policy based on scientific evidence.

Hoping that this information is helpful to you, receive a cordial greeting,

Instituto de Turismo y Desarrollo Económico (TIDES) https://tides.ulpgc.es Contact email: estudioturismo@institutoturismotides.com. If you do not want to receive a reminder about this study, please write to estudioturismo@institutoturismotides.com

Figure B.I: Emails PB:



Estimado/a Sr/Sra. Regidor/a de «NOMBRE_FINAL»:

Desde el Instituto Universitario de Turismo y Desarrollo Económico (TIDES), nos ponemos en contacto con usted para hacerle llegar los resultados de un estudio que muestra con datos los efectos beneficiosos de una intervención efectiva, sencilla y a coste cero para aumentar el turismo en «NOMBRE_FINAL».

La investigación encuentra que simples cambios de la página de Wikipedia de municipios similares al suyo reportaron mejoras de hasta un 33% en los ingresos turísticos. **El estudio está resumido en el siguiente informe publicado por el laboratorio de ideas progresista Fundación Alternativas, dirigido por el exdiputado socialista Diego López Garrido.**

FUNDACIÓN alternativas

Una investigación demuestra que la Wikipedia es un instrumento clave para impulsar el turismo en municipios españoles

Aquí puede acceder al enlace para leer el informe completo de la Fundación Alternativas.

Las mejoras de la página de Wikipedia de su municipio son muy fáciles de hacer, sin ningún coste y pueden generar más de 200.000 euros en ingresos turísticos en su municipio. Entre los cambios sencillos más efectivos que podría implementar en la página de Wikipedia se encuentran:

1. Añadir fotografías de su municipio.

2. Editar una versión sencilla en inglés del mismo texto para turistas extranjeros.

3. Mencionar o ampliar la sección dedicada a fiestas locales.

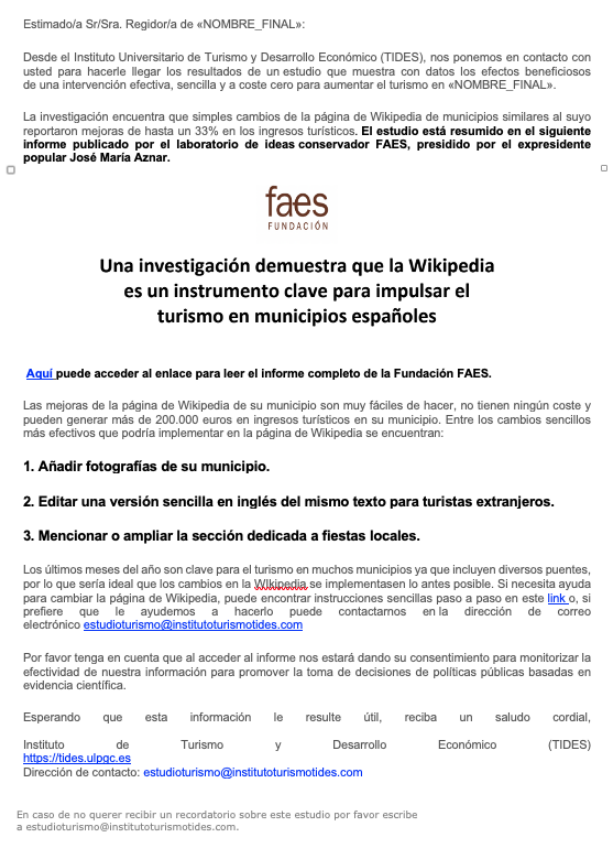
Los últimos meses del año son clave para el turismo en muchos municipios ya que incluyen diversos puentes, por lo que sería ideal que los cambios en la Wikipedia se implementasen lo antes posible. Si necesita ayuda para cambiar la página de Wikipedia, puede encontrar instrucciones sencillas paso a paso en este link o, si prefiere que le ayudemos a hacerlo puede contactarnos en la dirección de correo electrónico estudioturismo@institutoturismotides.com

Por favor tenga en cuenta que al acceder al informe nos estará dando su consentimiento para monitorizar la efectividad de nuestra información para promover la toma de decisiones de políticas públicas basadas en evidencia científica.

Esperando que esta información le resulte útil, reciba un saludo cordial,

Instituto de Turismo y Desarrollo Económico (TIDES)
https://tides.ulpgc.es
Dirección de contacto: estudioturismo@institutoturismotides.com

En caso de no querer recibir un recordatorio sobre este estudio por favor escribe a estudioturismo@institutoturismotides.com.



Estimado/a Sr/Sra. Regidor/a de «NOMBRE_FINAL»:

Desde el Instituto Universitario de Turismo y Desarrollo Económico (TIDES), nos ponemos en contacto con usted para hacerle llegar los resultados de un estudio que muestra con datos los efectos beneficiosos de una intervención efectiva, sencilla y a coste cero para aumentar el turismo en «NOMBRE_FINAL».

La investigación encuentra que simples cambios de la página de Wikipedia de municipios similares al suyo reportaron mejoras de hasta un 33% en los ingresos turísticos. **El estudio está resumido en el siguiente informe publicado por el laboratorio de ideas conservador FAES, presidido por el expresidente popular José María Aznar.**

faes FUNDACIÓN

Una investigación demuestra que la Wikipedia es un instrumento clave para impulsar el turismo en municipios españoles

Aquí puede acceder al enlace para leer el informe completo de la Fundación FAES.

Las mejoras de la página de Wikipedia de su municipio son muy fáciles de hacer, no tienen ningún coste y pueden generar más de 200.000 euros en ingresos turísticos en su municipio. Entre los cambios sencillos más efectivos que podría implementar en la página de Wikipedia se encuentran:

1. Añadir fotografías de su municipio.

2. Editar una versión sencilla en inglés del mismo texto para turistas extranjeros.

3. Mencionar o ampliar la sección dedicada a fiestas locales.

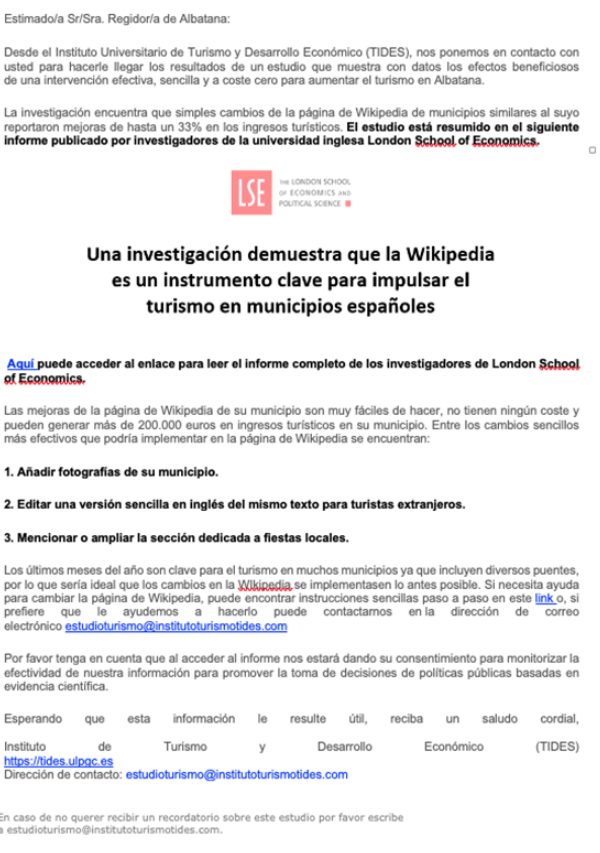
Los últimos meses del año son clave para el turismo en muchos municipios ya que incluyen diversos puentes, por lo que sería ideal que los cambios en la Wikipedia se implementasen lo antes posible. Si necesita ayuda para cambiar la página de Wikipedia, puede encontrar instrucciones sencillas paso a paso en este link o, si prefiere que le ayudemos a hacerlo puede contactarnos en la dirección de correo electrónico estudioturismo@institutoturismotides.com

Por favor tenga en cuenta que al acceder al informe nos estará dando su consentimiento para monitorizar la efectividad de nuestra información para promover la toma de decisiones de políticas públicas basadas en evidencia científica.

Esperando que esta información le resulte útil, reciba un saludo cordial,

Instituto de Turismo y Desarrollo Económico (TIDES)
https://tides.ulpgc.es
Dirección de contacto: estudioturismo@institutoturismotides.com

En caso de no querer recibir un recordatorio sobre este estudio por favor escribe a estudioturismo@institutoturismotides.com.



Estimado/a Sr/Sra. Regidor/a de Albatana:

Desde el Instituto Universitario de Turismo y Desarrollo Económico (TIDES), nos ponemos en contacto con usted para hacerle llegar los resultados de un estudio que muestra con datos los efectos beneficiosos de una intervención efectiva, sencilla y a coste cero para aumentar el turismo en Albatana.

La investigación encuentra que simples cambios de la página de Wikipedia de municipios similares al suyo reportaron mejoras de hasta un 33% en los ingresos turísticos. **El estudio está resumido en el siguiente informe publicado por investigadores de la universidad inglesa London School of Economics.**

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Una investigación demuestra que la Wikipedia es un instrumento clave para impulsar el turismo en municipios españoles

Aquí puede acceder al enlace para leer el informe completo de los investigadores de London School of Economics.

Las mejoras de la página de Wikipedia de su municipio son muy fáciles de hacer, no tienen ningún coste y pueden generar más de 200.000 euros en ingresos turísticos en su municipio. Entre los cambios sencillos más efectivos que podría implementar en la página de Wikipedia se encuentran:

1. Añadir fotografías de su municipio.

2. Editar una versión sencilla en inglés del mismo texto para turistas extranjeros.

3. Mencionar o ampliar la sección dedicada a fiestas locales.

Los últimos meses del año son clave para el turismo en muchos municipios ya que incluyen diversos puentes, por lo que sería ideal que los cambios en la Wikipedia se implementasen lo antes posible. Si necesita ayuda para cambiar la página de Wikipedia, puede encontrar instrucciones sencillas paso a paso en este link o, si prefiere que le ayudemos a hacerlo puede contactarnos en la dirección de correo electrónico estudioturismo@institutoturismotides.com

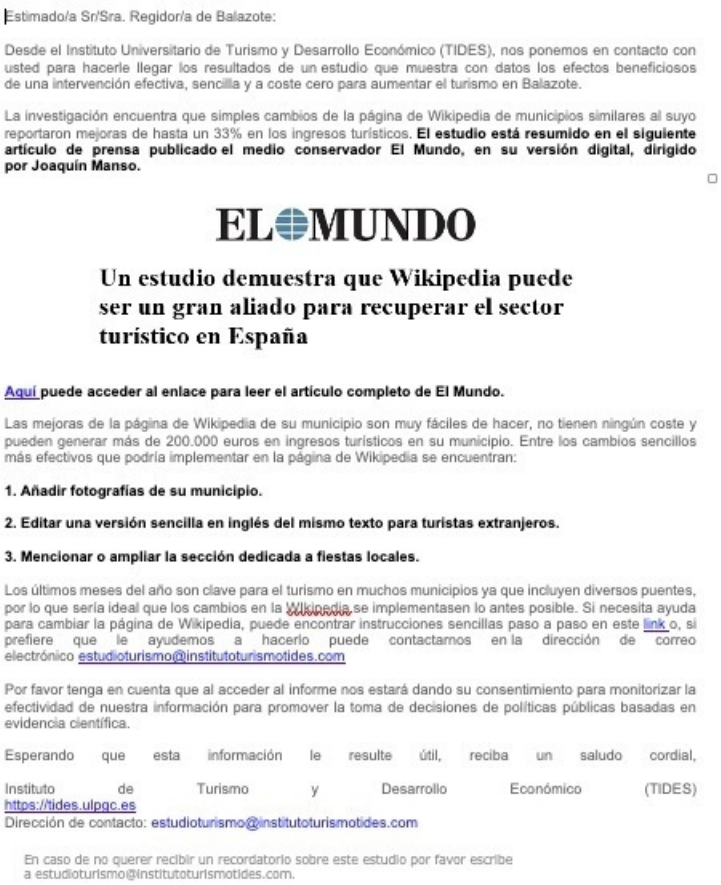
Por favor tenga en cuenta que al acceder al informe nos estará dando su consentimiento para monitorizar la efectividad de nuestra información para promover la toma de decisiones de políticas públicas basadas en evidencia científica.

Esperando que esta información le resulte útil, reciba un saludo cordial,

Instituto de Turismo y Desarrollo Económico (TIDES)
https://tides.ulpgc.es
Dirección de contacto: estudioturismo@institutoturismotides.com

En caso de no querer recibir un recordatorio sobre este estudio por favor escribe a estudioturismo@institutoturismotides.com.

Figure B.II: Emails Newspapers:



Estimado/a Sr/Sra. Regidor/a de Balazote:

Desde el Instituto Universitario de Turismo y Desarrollo Económico (TIDES), nos ponemos en contacto con usted para hacerle llegar los resultados de un estudio que muestra con datos los efectos beneficiosos de una intervención efectiva, sencilla y a coste cero para aumentar el turismo en Balazote.

La investigación encuentra que simples cambios de la página de Wikipedia de municipios similares al suyo reportaron mejoras de hasta un 33% en los ingresos turísticos. **El estudio está resumido en el siguiente artículo de prensa publicado el medio conservador El Mundo, en su versión digital, dirigido por Joaquín Manso.**

EL MUNDO

Un estudio demuestra que Wikipedia puede ser un gran aliado para recuperar el sector turístico en España

Aquí puede acceder al enlace para leer el artículo completo de El Mundo.

Las mejoras de la página de Wikipedia de su municipio son muy fáciles de hacer, no tienen ningún coste y pueden generar más de 200.000 euros en ingresos turísticos en su municipio. Entre los cambios sencillos más efectivos que podría implementar en la página de Wikipedia se encuentran:

1. Añadir fotografías de su municipio.

2. Editar una versión sencilla en inglés del mismo texto para turistas extranjeros.

3. Mencionar o ampliar la sección dedicada a fiestas locales.

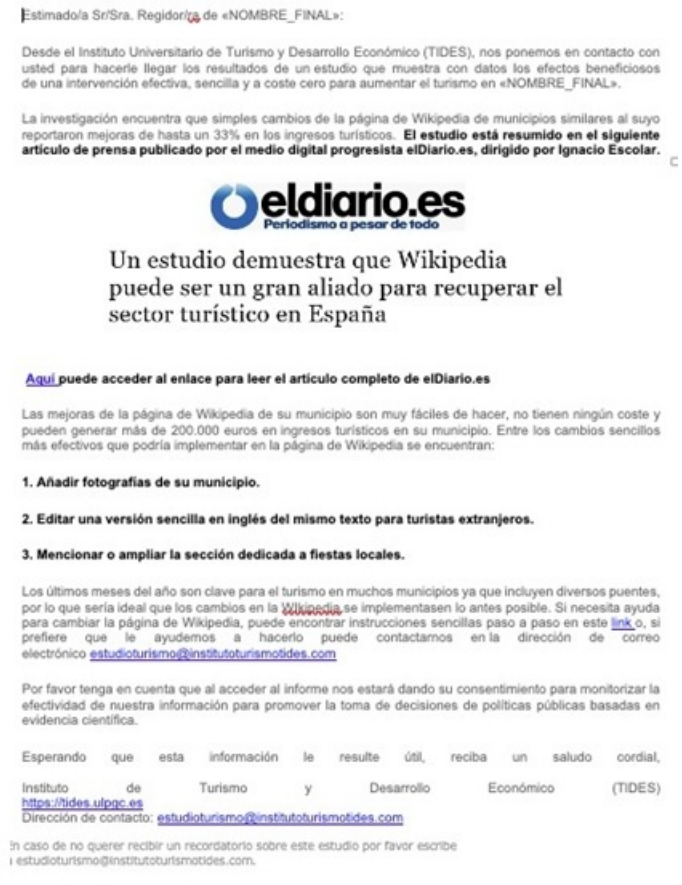
Los últimos meses del año son clave para el turismo en muchos municipios ya que incluyen diversos puentes, por lo que sería ideal que los cambios en la Wikipedia se implementasen lo antes posible. Si necesita ayuda para cambiar la página de Wikipedia, puede encontrar instrucciones sencillas paso a paso en este link o, si prefiere que le ayudemos a hacerlo puede contactarnos en la dirección de correo electrónico estudioturismo@institutoturismotides.com

Por favor tenga en cuenta que al acceder al informe nos estará dando su consentimiento para monitorizar la efectividad de nuestra información para promover la toma de decisiones de políticas públicas basadas en evidencia científica.

Esperando que esta información le resulte útil, reciba un saludo cordial,

Instituto de Turismo y Desarrollo Económico (TIDES)
https://tides.ulpgc.es
Dirección de contacto: estudioturismo@institutoturismotides.com

En caso de no querer recibir un recordatorio sobre este estudio por favor escribe a estudioturismo@institutoturismotides.com.



Estimado/a Sr/Sra. Regidor/a de «NOMBRE_FINAL»:

Desde el Instituto Universitario de Turismo y Desarrollo Económico (TIDES), nos ponemos en contacto con usted para hacerle llegar los resultados de un estudio que muestra con datos los efectos beneficiosos de una intervención efectiva, sencilla y a coste cero para aumentar el turismo en «NOMBRE_FINAL».

La investigación encuentra que simples cambios de la página de Wikipedia de municipios similares al suyo reportaron mejoras de hasta un 33% en los ingresos turísticos. **El estudio está resumido en el siguiente artículo de prensa publicado por el medio digital progresista elDiario.es, dirigido por Ignacio Escolar.**

elDiario.es
Periodismo a pesar de todo

Un estudio demuestra que Wikipedia puede ser un gran aliado para recuperar el sector turístico en España

Aquí puede acceder al enlace para leer el artículo completo de elDiario.es

Las mejoras de la página de Wikipedia de su municipio son muy fáciles de hacer, no tienen ningún coste y pueden generar más de 200.000 euros en ingresos turísticos en su municipio. Entre los cambios sencillos más efectivos que podría implementar en la página de Wikipedia se encuentran:

1. Añadir fotografías de su municipio.

2. Editar una versión sencilla en inglés del mismo texto para turistas extranjeros.

3. Mencionar o ampliar la sección dedicada a fiestas locales.

Los últimos meses del año son clave para el turismo en muchos municipios ya que incluyen diversos puentes, por lo que sería ideal que los cambios en la Wikipedia se implementasen lo antes posible. Si necesita ayuda para cambiar la página de Wikipedia, puede encontrar instrucciones sencillas paso a paso en este link o, si prefiere que le ayudemos a hacerlo puede contactarnos en la dirección de correo electrónico estudioturismo@institutoturismotides.com

Por favor tenga en cuenta que al acceder al informe nos estará dando su consentimiento para monitorizar la efectividad de nuestra información para promover la toma de decisiones de políticas públicas basadas en evidencia científica.

Esperando que esta información le resulte útil, reciba un saludo cordial,

Instituto de Turismo y Desarrollo Económico (TIDES)
https://tides.ulpgc.es
Dirección de contacto: estudioturismo@institutoturismotides.com

En caso de no querer recibir un recordatorio sobre este estudio por favor escribe a estudioturismo@institutoturismotides.com.

B.2 Text of the newspaper article

A study shows that Wikipedia can be a great ally in recovering the tourism sector in Spain

Improving a municipality's Wikipedia entry can mean an increase of up to 33% in hotel stays in that place. This was demonstrated by an experimental study developed with Spanish municipalities and published in the prestigious Journal of Economics and Management this year.

Tourism is one of our country's main economic sectors. In 2019, the sector represented 12.4% of GDP and 12.7% of employment. Furthermore, according to figures from the Ministry of Industry, Commerce and Tourism, Spain closed that same year with 83.7 million foreign tourist visits. However, COVID-19 changed the scenario radically: in 2020, international tourism decreased by 78% compared to 2019, and the tourism GDP in 2021 fell to 2017 levels, close to 6%.

In the Wikipedia Matters study, researchers Marit Hinnosaar, Toomas Hinnosaar, Michael Kummer and Olga Slivko carried out a curious experiment to analyse the impact of the information available on Wikipedia about tourist municipalities in Spain on tourism in those places. As part of the experiment, the authors improved the editing and content of Wikipedia pages referring to some Spanish municipalities chosen randomly from a sample of municipalities with tourism potential.

The experiment results showed that the improvement in Wikipedia content in different languages caused the municipalities in the sample whose pages were edited to increase the nights of accommodation of tourists native to the language in which the Wikipedia content was edited. During the tourist season, hotel nights increased by 9% in these municipalities compared to those with tourist potential but without edited pages. In the cities whose pages were more incomplete and briefer before the intervention, the increase reached 33%. That is, more detailed information captivates the attention of potential readers, and this has a direct effect on attracting tourists.

Official figures show that the average expenditure of a foreign tourist in Spain is 101 euros per day, and initial calculations suggest that improving a municipality's article on Wikipedia could generate, in that place, an approximate annual amount of 160 thousand euros of additional income. The results of this experiment demonstrate with evidence that improving the quality of the Wikipedia page is a simple and economical way to increase the visibility and number of tourists a city receives, which could substantially help the recovery of the tourism sector.

Figure B.III: Newspaper articles:

MEDIOS

MEDIOS

Un estudio demuestra que Wikipedia puede ser un gran aliado para recuperar el sector turístico en España

EL MUNDO

Actualizado Lunes, 7 marzo 2022 - 20:58

Comentar

Mejorar la entrada de Wikipedia de un municipio puede significar un aumento de hasta un 33% en las estadías hoteleras de ese lugar. Así lo demostró un estudio experimental desarrollado con municipios españoles y publicado este año en el prestigioso Journal of Economics and Management

Logotipo de Wikipedia.

El turismo es uno de los principales sectores económicos de nuestro país. En el año 2019, el sector representaba el 12,4% del PIB y el 12,7% del empleo. Además, según cifras del **Ministerio de Industria, Comercio y Turismo**, España cerró ese mismo año con 83,7 millones de visitas turísticas extranjeras. Sin embargo, la **Covid-19** cambió el escenario radicalmente: el año 2020 el turismo internacional se redujo un 78% respecto al 2019 y el PIB turístico del 2021 retrocedió a niveles del 2017, siendo cercano al 6%.

En el estudio titulado '**Wikipedia Matters**', los investigadores **Marit Hinnosaar, Toomas Hinnosaar, Michael Kummer** y **Olga Slivko** realizaron un curioso experimento para analizar el impacto que la información disponible en Wikipedia acerca de municipios turísticos de España tiene sobre el turismo en esos lugares. Como parte del experimento, los autores mejoraron la edición y contenido de las páginas de Wikipedia referidas a algunos municipios españoles elegidos al azar de entre una muestra de municipios con potencial turístico.

elDiario.es

Política Economía Igualdad Educación Clima Salud Internacional Desalambre Cultura Festival

La portada de mañana

ACCEDER

El choque del PSOE con sus socios por los desalojos amenaza la ley de vivienda

Por qué el BCE es prisionero de su objetivo del 2% de inflación

OPINIÓN | 'Sadoeconomía', por Antón Losada

Un estudio demuestra que Wikipedia puede ser un gran aliado para recuperar el sector turístico en España

Mejorar la entrada de Wikipedia de un municipio puede significar un aumento de hasta un 33% en las estadías hoteleras de ese lugar. Así lo demostró un estudio experimental desarrollado con municipios españoles y publicado este año en el prestigioso Journal of Economics and Management

Logotipo de Wikipedia. Pavlo Gonchar

elDiario.es 4 de abril de 2022 - 19:04h

El turismo es uno de los principales sectores económicos de nuestro país. En el año 2019, el sector representaba el 12,4% del PIB y el 12,7% del empleo. Además, según cifras del Ministerio de Industria, Comercio y Turismo, España cerró ese mismo año con 83,7 millones de visitas turísticas extranjeras. Sin embargo, la COVID-19 cambió el escenario radicalmente: el año 2020 el turismo internacional se redujo un 78% respecto al 2019 y el PIB turístico del 2021 retrocedió a niveles del 2017, siendo cercano al 6%.

En el estudio titulado Wikipedia Matters, los investigadores Marit Hinnosaar, Toomas Hinnosaar, Michael Kummer y Olga Slivko realizaron un curioso experimento para analizar el impacto que la información disponible en Wikipedia acerca de municipios turísticos de España tiene sobre el turismo en esos lugares. Como parte del experimento, los autores mejoraron la edición y contenido de las páginas de Wikipedia referidas a algunos municipios españoles elegidos al azar de entre una muestra de municipios con potencial turístico.

El Mundo:*[Here, you can access the link to read the full El Mundo article]*

https://www.elmundo.es/television/medios/2022/03/07/622663cffc6c832a3b8b45d5.html

elDiario.es:*[Here, you can access the link to read the full elDiario.es article]*

https://www.eldiario.es/economia/estudio-demuestra-wikipedia-gran-aliado-recuperar-sector-turistico-espana$_{18}$888694.*html*

B.3 Text of the Policy brief

Three organisations published this policy brief. All are written identically; the only difference is that each institution presented its version with its respective organisational logo plus a brief description of the organisation before showing the content of the policy brief.

[Logo Fundación Alternativas]

[About Fundación Alternativas]

[The Alternativas Foundation is an independent centre of thought and debate for political and social transformation chaired by Diego López Garrido. It was born in 1997 with the desire to be a channel for reflection, and its mission is to contribute to progressive theoretical and cultural thinking. www.fundacionalternativas.org]

[Logo FAES]

[About FAES]

[FAES is a private, non-profit foundation that has been working in the field of ideas since 1989. Chaired by José María Aznar, its objective is to nourish the thinking of the liberal reformist centre with political proposals that influence decision-making and impact public opinion. At the service of Spain and its citizens, its purpose is to create, promote and disseminate ideas based on political, intellectual, and economic freedom and strengthen the values of liberty, democracy, the rule of law, the free market and Western humanism]

[Logo LSE]

[About the London School of Economics]

[The London School of Economics and Political Science (LSE), founded in 1985, is a world-leading university specialising in teaching and research in the social sciences, with a global community of people rooted in London and ideas that transform the world]

Research shows that Wikipedia is crucial to promoting tourism in Spanish municipalities

- *A study published in the prestigious Journal of Economics and Management shows that an improvement in the content of the Wikipedia page of Spanish municipalities increased the number of hotel overnight stays by 9%.*
- *Wikipedia page improvements are straightforward, are done at 0 cost and can generate more than €200,000 of additional income to local coffers.*
- *Adding photographs of the municipality, improving information on local festivals, and translating into other languages, such as English, are some improvements that have proven effective.*

Improve Wikipedia, a pivotal policy to improve tourism in Spanish municipalities

Tourism is a crucial sector in Spain and the main economic activity in many municipalities, where it generates income worth approximately 51 billion euros each year. In the current context of return to normality, scarce resources and growing competition, promoting tourism is one of the main

political objectives in many municipalities. But what does the evidence tell us about the effectiveness of various policies in promoting tourism in your municipality? In this policy brief, we summarise the results of a recent study that shows a simple way to increase tourism in your municipality at zero cost.

The Wikipedia Matters study, published by Hinnosaar and co-authors in the prestigious Journal of Economics and Management, showed that an improvement in the content of the Wikipedia page of Spanish municipalities increased the number of hotel nights by 9% during the high season in the city. The increase in hotel overnight stays reached 33% for municipalities with Wikipedia entries that were significantly incomplete before being improved. Considering that each tourist spends an average of €136 per day, improving a municipality's page on Wikipedia can generate more than €200,000 of additional income at zero cost per year for local coffers.

Wikipedia is the world's largest collaborative encyclopaedia. It is consistently one of the first results displayed by the algorithms of the major online search engines when searching for information about anything. It is one of the most popular portals in the world. According to We Are Social's Digital 2022 report, Wikipedia was the 7th most viewed website in the world in 2021, with 66.9 billion visitors and content available in 300 languages.

What specific changes to the municipality's Wikipedia page help improve tourism?

Municipal officials of Spanish municipalities can edit the page about their locality on Wikipedia. But what changes to Wikipedia help increase tourism? The study points out some elementary changes:

1. Add photos of the municipality, either of places or celebrations, on the Wikipedia page of your city in Spanish, English, and other languages whose nationality receives tourists.

2. Include complete information about local festivals on the Wikipedia page of your municipality in Spanish, English, and other languages whose nationalities receive tourists.

3. Complete information in Spanish or other languages in which your municipality's Wikipedia page is concise on crucial aspects of the city, such as places to visit or transportation and communication routes.

Although the improvement in the entry is particularly effective in increasing tourism in municipalities with poorly developed Wikipedia pages in Spanish and other languages, changes as simple as adding 2 or 3 photographs or adding around 300 words expanding information on local festivals have proven effective even in those municipalities with extensive Wikipedia pages.

How to do it? A straightforward, fast, and free process

Finally, it is essential to highlight that changing your municipality's page on Wikipedia in Spanish and other languages is entirely free.

In the email, you will find attached a document in PDF format showing you step-by-step instructions on changing your municipality's page on Wikipedia. Improving your municipality's page on Wikipedia is an effective and cost-free way to increase tourism income in your municipality.

References

Hinoosar, M., Hinoosar, T., Kummer, M., Slivko, O. 2022. Wikipedia Matters, Journal of Economics and Management Strategy.

Figure B.IV: Policy Briefs:

FUNDACIÓN alternativas

Acerca de Fundación Alternativas

La Fundación Alternativas es un centro independiente de pensamiento y debate para la transformación política y social presidido por Diego López Garrido. Nació en 1997 con la voluntad de ser un cauce de reflexión y su misión es contribuir al pensamiento teórico y cultural progresista. www.fundacionalternativas.org

Una investigación demuestra que la Wikipedia es un instrumento clave para impulsar el turismo en municipios españoles

- **Un estudio publicado en la prestigiosa revista Journal of Economics and Management, muestra que una mejora en el contenido de la página de Wikipedia de los municipios españoles incrementó en un 9% el número de pernoctaciones hoteleras.**
- **Las mejoras de la página de Wikipedia son muy fáciles de hacer, se hacen a coste 0 y pueden generar más de 200.000€ de ingreso adicional a las arcas locales.**
- **Añadir fotografías del municipio, mejorar la información sobre fiestas locales y traducir a otros idiomas como el inglés son algunas de las mejoras específicas que han demostrado ser efectivas**

Mejorar la Wikipedia, política clave para mejorar el turismo en municipios españoles

El turismo es un sector clave en España y la principal actividad económica en muchos municipios, en los que genera ingresos por un valor aproximado de 51.000 millones de euros cada año. En el actual contexto de vuelta a la normalidad, recursos escasos y creciente competencia, uno de los objetivos políticos principales en muchos ayuntamientos es cómo fomentar el turismo. ¿Pero qué nos dice la evidencia sobre la efectividad de diversas políticas para fomentar el turismo en su municipio? En este policy brief le resumimos los resultados de un estudio reciente que muestra una forma sencilla de aumentar el turismo en su municipio a coste 0.

faes FUNDACIÓN

Acerca de FAES

FAES es una fundación privada sin ánimo de lucro que trabaja en el ámbito de las ideas desde 1989. Presidida por José María Aznar, su objetivo es nutrir el pensamiento del centro liberal reformista con propuestas políticas que influyen en la toma de decisiones y repercuten en la opinión pública. Al servicio de España y de sus ciudadanos, su propósito es crear, promover y difundir ideas basadas en la libertad política, intelectual y económica, así como fortalecer los valores de la libertad, la democracia, el Estado de derecho, el libre mercado y el humanismo occidental.

Una investigación demuestra que la Wikipedia es un instrumento clave para impulsar el turismo en municipios españoles

- **Un estudio publicado en la prestigiosa revista Journal of Economics and Management, muestra que una mejora en el contenido de la página de Wikipedia de los municipios españoles incrementó en un 9% el número de pernoctaciones hoteleras.**
- **Las mejoras de la página de Wikipedia son muy fáciles de hacer, se hacen a coste 0 y pueden generar más de 200.000€ de ingreso adicional a las arcas locales.**
- **Añadir fotografías del municipio, mejorar la información sobre fiestas locales y traducir a otros idiomas como el inglés son algunas de las mejoras específicas que han demostrado ser efectivas**

Mejorar la Wikipedia, política clave para mejorar el turismo en municipios españoles

El turismo es un sector clave en España y la principal actividad económica en muchos municipios, en los que genera ingresos por un valor aproximado de 51.000 millones de euros cada año. En el actual contexto de vuelta a la normalidad, recursos escasos y creciente competencia, uno de los objetivos políticos principales en muchos ayuntamientos es cómo fomentar el turismo. ¿Pero qué nos dice la evidencia sobre la efectividad de diversas políticas para fomentar el turismo en su municipio? En este policy brief le resumimos los resultados de un estudio reciente que muestra una forma sencilla de aumentar el turismo en su municipio a coste 0.

Tel: +44 (0)753 090 47 51
Email: b.ozcan@lse.ac.uk

Dr Berkay Özcan

Associate Professor
Doctoral Programme Director
Department of Social Policy
& School of Public Policy

LSE THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Acerca de la London School of Economics

La London School of Economics and Political Science (LSE), fundada en 1985, es una universidad líder a nivel mundial especializada en la docencia e investigación en ciencias sociales, con una comunidad global de personas arraigada en Londres y unas ideas que transforman el mundo.

Una investigación demuestra que la Wikipedia es un instrumento clave para impulsar el turismo en municipios españoles

- **Un estudio publicado en la prestigiosa revista Journal of Economics and Management, muestra que una mejora en el contenido de la página de Wikipedia de los municipios españoles incrementó en un 9% el número de pernoctaciones hoteleras.**
- **Las mejoras de la página de Wikipedia son muy fáciles de hacer, se hacen a coste 0 y pueden generar más de 200.000€ de ingreso adicional a las arcas locales.**
- **Añadir fotografías del municipio, mejorar la información sobre fiestas locales y traducir a otros idiomas como el inglés son algunas de las mejoras específicas que han demostrado ser efectivas**

Mejorar la Wikipedia, política clave para mejorar el turismo en municipios españoles

El turismo es un sector clave en España y la principal actividad económica en muchos municipios, en los que genera ingresos por un valor aproximado de 51.000 millones de euros cada año. En el actual contexto de vuelta a la normalidad, recursos escasos y creciente competencia, uno de los objetivos políticos principales en muchos ayuntamientos es cómo fomentar el turismo. ¿Pero qué nos dice la evidencia sobre la efectividad de diversas políticas para fomentar el turismo en su municipio? En este policy brief le resumimos los resultados de un estudio reciente que muestra una forma sencilla de aumentar el turismo en su municipio a coste 0.

B.4 Text of the instructions to change Wikipedia

Ten steps for editing content on Wikipedia

1. *Open the Wikipedia page of your town hall.*
2. *Click "Create an account" in the upper right corner.*

Figure B.V: Step 2 to edit Wikipedia:

3. *Complete the registration details.*

Figure B.VI: Step 3 to edit Wikipedia:

4. *Once completed, click on the "Create your account" box.*

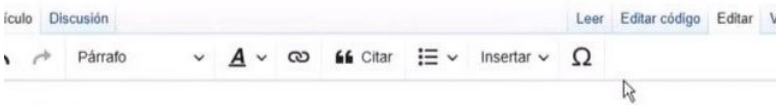
Figure B.VII: Step 4 to edit Wikipedia:

5. The "Edit" tab will appear once the account has been created.

Figure B.VIII: Step 5 to edit Wikipedia:

6. In the "Edit" section, you can modify and add new text to the page. You will see a toolbar like that of Word displayed. These tools will help you edit the content.

Figure B.IX: Step 6 to edit Wikipedia:



7. If after pressing the "Edit" tab (from the previous step), the "Code editing" version opens by default, select the "Visual editing" option to edit directly in the text.

Figure B.X: Step 7 to edit Wikipedia:

8. To add images, go to the "Insert" tab and select the "Multimedia" or "Gallery" option. Once there, select the picture you want to add and press "insert".

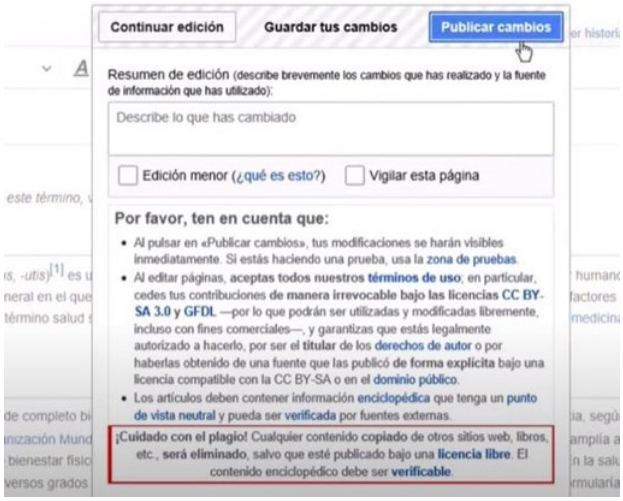
Figure B.XI: Step 8 to edit Wikipedia:

9. Once editing is complete, select the "Publish Changes" icon at the far right of the toolbar.

Figure B.XII: Step 9 to edit Wikipedia:

10. When you press "Publish Changes", a pop-up tab will appear with options such as "Review your changes", "Continue editing", "Save your changes", and finally "Publish changes". Once you press "Publish Changes", you will have finished publishing information to Wikipedia.

Figure B.XIII: Step 10 to edit Wikipedia:



C Heterogeneity of results

This appendix explores the heterogeneous effects of the different treatment arms by the ideology of the municipality's major, whether the mayor of the municipality belongs to one of the two main political parties, the population of the municipality, the number of words in the Wikipedia page of the municipality prior to our intervention, and whether the municipality belongs to a region with strong support for nationalist movements. To conduct these analyses, we expand the baseline specification 3.1 by adding interaction terms between the dummy variables indicating the treatment arms and the dummy variables that define the categories over which we want to estimate the heterogeneous effects of the treatments (e.g. ideology of the mayor).

First, we discuss whether the main effects of the treatments on policy adoption might differ for municipalities with a left- or a right-wing mayor. The results are reported in Table C.I. The probability of policy adoption is 38% higher among municipalities with a left-wing mayor. More importantly, the effect of receiving information from an ideologically aligned institution is nearly three times larger for left-wing municipalities, although the interaction term is not statistically significant at conventional confidence levels. Furthermore, the effect of receiving information from an ideologically nonsalient institution is relevant for left-wing policymakers but not for right-wing policymakers. On the other hand, the effect of receiving the summary of results from an institution with the opposite ideology is consistently 0 regardless of the ideology of the mayor.

It is important to interpret the results of this analysis cautiously, as the strength of the ideological alignment may not be symmetric for right- and left-wing mayors and institutions. To illustrate, consider the following example: Right-wing mayors assigned to receive the policy brief from an institution with the same ideology will receive a document endorsed by FAES. In contrast, left-wing mayors in the same treatment arm will receive a policy brief endorsed by Alternativas. If the prestige, authority, or trust associated with FAES differs for right-wing mayors compared to the prestige, authority, or trust of Alternativas for left-wing mayors, the observed results should not be solely attributed to differing ideological biases among right- and left-wing mayors, but also to this asymmetry. However, it's important to note that this consideration does not undermine the main findings reported in the paper, as each treatment arm includes both right- and left-wing municipalities, and ideology is used in the randomization as a stratification variable.

Second, we explore whether the effect of the different treatments differ based on the political affiliation of the municipality's mayor, specifically whether they belong to the PP or PSOE, the primary right-wing and left-wing political parties, or another political party. This is a reasonable consideration given the key roles of the directors of FAES and Alternativas, who are influential figures

and members of the Popular Party (PP) and the Socialist Party (PSOE), respectively.[51] While there is no formal affiliation at the moment, FAES and Alternativas wield considerable influence in shaping ideas and political and economic proposals for both PP and PSOE.

The findings of this analysis are presented in Table C.II. Overall, they indicate no discernible differential effects of the treatment arms on policy adoption based on whether the mayor belongs to either the PP/PSOE or a different political party. This result underscores that the primary effect identified in the paper—strong impacts on policy adoption when receiving information endorsed by an ideologically aligned institution—is not contingent on party discipline.

Third, the email sent to local policymakers not only includes the link to the policy report or newspaper article but also a link to a document providing step-by-step instructions for editing Wikipedia. While these instructions reduce implementation costs, one may wonder whether some non-negligible implementation costs might persist in small municipalities lacking equipment or personnel with the skills to edit the Wikipedia page. To explore this possibility, we investigate the heterogeneous effects of the treatments based on the population of the municipality, which arguably proxies for implementation capacity in the municipality. For this analysis, we divide the sample into tertiles of the population distribution and examine whether the effect differs across tertiles. The results of this analysis are reported in Table C.III. The findings indicate that the main effect on adoption is driven by municipalities in the top tertile of the population. These municipalities are likely those with a higher implementation capacity.

Fourth, we examine the heterogeneous effect of the treatment arms based on the length of the municipality's page on Spanish Wikipedia at the time before our intervention. Arguably, the word count serves as a good proxy for the completeness of the municipality's entry on Wikipedia. Although we recommended implementing changes that were easily executable even for very comprehensive web pages (e.g., adding photographs, expanding information on festivals, etc.), it is easier to make improvements on more incomplete pages. On the other hand, shorter pages might also correlate with a lower capacity in the municipality to implement changes or a lower interest in tourism. For this analysis, we divide the sample into municipalities above and below the median length of the Wikipedia page in the sample. We then expand the main specification, including interaction terms between the treatment arms and a dummy variable indicating whether the municipality has a Wikipedia page with an above-median number of words. The results of this analysis are reported in Table C.IV. While the coefficients are not statistically significant, their magnitude suggests that the effects of receiving information endorsed by an institution with the same ideology are two times

[51] José María Aznar, the president of FAES, is a former president of the PP and served as the Spanish prime minister between 1996 and 2004. Diego López Garrido, the president of Alternativas, is a prominent politician and former MP for PSOE.

larger in municipalities with lengthier Wikipedia pages. While this result may seem counterintuitive (as editing more comprehensive webpages is, in principle, more costly), it is in line with the idea that municipalities with higher implementation capacity tend to be more responsive to ideological alignment, provided the length of Wikipedia pages serves as a proxy for implementation capacity.

Fifth, we investigated whether the treatment effects might be weaker for municipalities in which the mayor belongs to a strong nationalist party. We define these parties as those promoting independence from Spain: Esquerra Republicana de Catalunya (ERC), Junts per Catalunya, Candidatura d'Unitat Popular (CUP), Partido Nacionalista Vasco (PNV), Euskal Herria Bildu, and Bloque Nacionalista Galego (BNG). Although we address local politicians in Galicia, Catalonia, and the Basque Country in Spanish and their regional language in the email, local politicians from these parties might perceive both right and left-wing national-level think tanks and media as ideologically opposite. The results of this heterogeneity analysis are reported in Table C.V. Contrary to expectations, the results suggest that, if anything, the effect on policy adoption of receiving a summary of research from an institution with an aligned ideology is larger for these municipalities.

Finally, we analyzed whether the treatment effects are different for the 60 municipalities included in the experiment conducted by Hinnosaar et al. (2021). Unlike most of the municipalities in our sample, these 60 municipalities were big towns or cities. In the experiment conducted by Hinnosaar et al. (2021), the changes in Wikipedia pages were conducted by the authors and their team without informing the treated municipalities. Furthermore, the results received very limited coverage from mass media in Spain, so they were unlikely to interact with our experiment.[52] The results of this heterogeneous analysis are reported in Table C.VI. They show that the impact of ideological alignment is larger among this subsample of 60 municipalities. This result is consistent with the results presented earlier in this section, illustrating that treatment effects are larger for more populated municipalities with more complete Wikipedia pages.

[52]Only the small online news portals La Informacion and Tourinews have published brief articles about the study results. https://www.lainformacion.com/management/turismo-editar-pueblos-wikipedia/2815402/https://www.tourinews.es/resumen-de-prensa/curiosidades/wikipedia-clave-reactivar-turismo-rural_4461915_102.html

Table C.I: Heterogeneous effects of the treatment arms by ideology

	Study period		Placebo period	
Dep. var: Recommended change in Wikipedia (0/1)	(1)	(2)	(3)	(4)
Panel A: Effect of treatment arms relative to control				
Aligned ideology - Policy brief	0.0239**	0.0235**	-0.0015	-0.0020
	(0.0092)	(0.0092)	(0.0058)	(0.0058)
Aligned ideology - Policy brief x Right	-0.0157	-0.0147	-0.0083	-0.0075
	(0.0163)	(0.0165)	(0.0084)	(0.0085)
Opposite ideology - Policy brief	0.0043	0.0037	-0.0035	-0.0037
	(0.0108)	(0.0109)	(0.0099)	(0.0100)
Opposite ideology - Policy brief x Right	-0.0054	-0.0039	0.0143	0.0148
	(0.0156)	(0.0157)	(0.0148)	(0.0150)
Nonsalient ideology - Policy brief	0.0236*	0.0236*	0.0022	0.0021
	(0.0119)	(0.0119)	(0.0081)	(0.0081)
Neutral ideology - Policy brief x Right	-0.0311*	-0.0309*	-0.0096	-0.0094
	(0.0161)	(0.0162)	(0.0106)	(0.0107)
Aligned ideology - Newspaper	0.0297**	0.0293**	-0.0016	-0.0019
	(0.0123)	(0.0124)	(0.0075)	(0.0075)
Aligned ideology - Newspaper x Right	-0.0285	-0.0275	0.0055	0.0060
	(0.0174)	(0.0175)	(0.0112)	(0.0112)
Opposite ideology - Newspaper	0.0100	0.0100	0.0002	0.0003
	(0.0092)	(0.0092)	(0.0068)	(0.0068)
Opposite ideology - Newspaper x Right	-0.0130	-0.0128	-0.0007	-0.0006
	(0.0151)	(0.0152)	(0.0113)	(0.0114)
Mean dep var in control	0.0255	0.0255	0.0202	0.0202
Strata FE	No	Yes	No	Yes
N	5,678	5,678	5,678	5,678
Panel B: Pooled effects relative to control				
Any treatment	0.0183**	0.0180**	-0.0008	-0.0010
	(0.0070)	(0.0070)	(0.0058)	(0.0058)
Any treatment x Right	-0.0187	-0.0179	0.0003	0.0007
	(0.0125)	(0.0126)	(0.0086)	(0.0087)
Aligned ideology	0.0268***	0.0264***	-0.0016	-0.0019
	(0.0084)	(0.0084)	(0.0060)	(0.0060)
Aligned ideology x Right	-0.0221	-0.0211	-0.0014	-0.0007
	(0.0147)	(0.0149)	(0.0084)	(0.0084)
Opposite ideology	0.0072	0.0068	-0.0016	-0.0017
	(0.0085)	(0.0085)	(0.0066)	(0.0066)
Opposite ideology x Right	-0.0092	-0.0083	0.0068	0.0071
	(0.0130)	(0.0130)	(0.0111)	(0.0114)
Policy brief	0.0173**	0.0169**	-0.0010	-0.0012
	(0.0080)	(0.0080)	(0.0072)	(0.0073)
Policy brief x Right	-0.0174	-0.0165	-0.0012	-0.0007
	(0.0132)	(0.0133)	(0.0095)	(0.0096)
Newspaper	0.0198***	0.0196**	-0.0007	-0.0008
	(0.0074)	(0.0074)	(0.0058)	(0.0058)
Newspaper X Right	-0.0207	-0.0201	0.0024	0.0027
	(0.0131)	(0.0132)	(0.0096)	(0.0097)

Note: Panel A reports the heterogeneous effects of the different treatment arms on the probability of conducting a recommended change on the Wikipedia page by ideology. To explore the heterogeneous effects of the treatment arms, we interact the treatment indicators with a dummy variable indicating whether the mayor of the municipality belongs to a right- or left-wing political party. Estimates in columns (1) and (2) examine the effect of the different arms for right- and left-wing municipalities between May and December 2022, the study period. Estimates in columns (3) and (4) examine the effect of the different arms for right- and left-wing municipalities between May and December 2019, a placebo period before the start of the intervention. Regressions in columns (1) and (3) do not include strata fixed-effects and regressions in columns (2) and (4) are estimated with strata fixed-effects. Panel B reports the pooled effects relative to the control group. The mean adoption rate among municipalities with a right-wing mayor is 2.8% and for municipalities with left-wing mayors is 3.8%. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

Table C.II: Heterogeneous effects of the treatment arms by whether the mayor belongs to the PP/PSOE or to a different political party

	Study period		Placebo period	
Dep. var: Recommended change in Wikipedia (0/1)	(1)	(2)	(3)	(4)
Panel A: Effect of treatment arms relative to control				
Aligned ideology - Policy brief	0.0187	0.0190	-0.0067	-0.0066
	(0.0138)	(0.0140)	(0.0065)	(0.0065)
Aligned ideology - Policy brief x PP/PSOE	-0.0028	-0.0032	0.0020	0.0020
	(0.0170)	(0.0172)	(0.0085)	(0.0086)
Opposite ideology - Policy brief	-0.0069	-0.0068	0.0061	0.0066
	(0.0101)	(0.0101)	(0.0103)	(0.0105)
Opposite ideology - Policy brief x PP/PSOE	0.0132	0.0132	-0.0045	-0.0051
	(0.0144)	(0.0145)	(0.0142)	(0.0143)
Nonsalient ideology - Policy brief	0.0093	0.0095	0.0062	0.0063
	(0.0162)	(0.0164)	(0.0077)	(0.0077)
Nonsalient ideology - Policy brief x PP/PSOE	0.0003	0.0002	-0.0125	-0.0125
	(0.0191)	(0.0193)	(0.0104)	(0.0104)
Aligned ideology - Newspaper	0.0157	0.0161	-0.0002	0.0000
	(0.0136)	(0.0138)	(0.0080)	(0.0080)
Aligned ideology - Newspaper x PP/PSOE	0.0014	0.0009	0.0017	0.0014
	(0.0179)	(0.0182)	(0.0108)	(0.0109)
Opposite ideology - Newspaper	-0.0100	-0.0097	0.0063	0.0066
	(0.0084)	(0.0084)	(0.0091)	(0.0092)
Opposite ideology - Newspaper x PP/PSOE	0.0210	0.0207	-0.0094	-0.0097
	(0.0132)	(0.0133)	(0.0113)	(0.0114)
Mean dep var in control	0.0255	0.0255	0.0202	0.0202
Strata FE	No	Yes	No	Yes
N	5,678	5,678	5,678	5,678
Panel B: Pooled effects relative to control				
Any treatment	0.0054	0.0056	0.0023	0.0026
	(0.0086)	(0.0087)	(0.0062)	(0.0063)
Any treatment x PP/PSOE	0.0066	0.0063	-0.0045	-0.0048
	(0.0120)	(0.0121)	(0.0083)	(0.0083)
Aligned ideology	0.0172	0.0176	-0.0035	-0.0033
	(0.0126)	(0.0128)	(0.0061)	(0.0061)
Aligned ideology x PP/PSOE	-0.0007	-0.0011	0.0019	0.0017
	(0.0156)	(0.0159)	(0.0082)	(0.0082)
Opposite ideology	-0.0084	-0.0082	0.0062	0.0066
	(0.0080)	(0.0079)	(0.0083)	(0.0084)
Opposite ideology x PP/PSOE	0.0171	0.0169	-0.0070	-0.0074
	(0.0117)	(0.0117)	(0.0108)	(0.0109)
Policy brief	0.0070	0.0072	0.0019	0.0021
	(0.0108)	(0.0109)	(0.0067)	(0.0068)
Policy brief x PP/PSOE	0.0036	0.0034	-0.0050	-0.0052
	(0.0137)	(0.0138)	(0.0092)	(0.0093)
Newspaper	0.0029	0.0032	0.0030	0.0033
	(0.0076)	(0.0078)	(0.0065)	(0.0066)
Newspaper X PP/PSOE	0.0112	0.0108	-0.0039	-0.0041
	(0.0119)	(0.0120)	(0.0089)	(0.0090)

Note: Panel A reports the heterogeneous effects of the different treatment arms on the probability of conducting a recommended change on the Wikipedia page by whether the mayor belongs to the PP/PSOE parties or to a different political party. To explore the heterogeneous effects of the treatment arms, we interact the treatment indicators with a dummy variable indicating whether the mayor of the municipality belongs to either the PP/PSOE, or to a different political party. Estimates in columns (1) and (2) examine the effect of the different arms for right- and left-wing municipalities between May and December 2022, the study period. Estimates in columns (3) and (4) examine the effect of the different arms for right- and left-wing municipalities between May and December 2019, a placebo period before the start of the intervention. Regressions in columns (1) and (3) do not include strata fixed-effects and regressions in columns (2) and (4) are estimated with strata fixed-effects. Panel B reports the pooled effects relative to the control group. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

Table C.III: Heterogeneous effects of the treatment arms by the population of the municipality

Dep. var: Recommended change in Wikipedia (0/1)	Study period (1)	(2)	Placebo period (3)	(4)
Panel A: Effect of treatment arms relative to control				
Aligned ideology - Policy brief	0.0320**	0.0322**	0.0096	0.0097
	(0.0146)	(0.0148)	(0.0088)	(0.0088)
Aligned ideology - Policy brief x Mid tertile population	-0.0162	-0.0166	-0.0288**	-0.0289**
	(0.0211)	(0.0212)	(0.0121)	(0.0121)
Aligned ideology - Policy brief x Bottom tertile population	-0.0291	-0.0293	-0.0160	-0.0160
	(0.0180)	(0.0181)	(0.0096)	(0.0097)
Opposite ideology - Policy brief	0.0061	0.0067	-0.0002	0.0002
	(0.0185)	(0.0186)	(0.0129)	(0.0130)
Opposite ideology - Policy brief x Mid tertile population	0.0000	-0.0007	0.0094	0.0091
	(0.0212)	(0.0213)	(0.0202)	(0.0204)
Opposite ideology - Policy brief x Bottom tertile population	-0.0125	-0.0133	0.0001	-0.0001
	(0.0207)	(0.0208)	(0.0151)	(0.0153)
Nonsalient ideology - Policy brief	0.0066	0.0071	0.0002	0.0001
	(0.0153)	(0.0154)	(0.0121)	(0.0121)
Nonsalient ideology - Policy brief x Mid tertile population	0.0026	0.0020	-0.0099	-0.0099
	(0.0220)	(0.0221)	(0.0146)	(0.0146)
Nonsalient ideology - Policy brief x Bottom tertile population	0.0060	0.0055	0.0030	0.0032
	(0.0205)	(0.0206)	(0.0141)	(0.0142)
Aligned ideology - Newspaper	0.0466**	0.0470**	0.0091	0.0093
	(0.0192)	(0.0194)	(0.0115)	(0.0116)
Aligned ideology - Newspaper x Mid tertile population	-0.0498**	-0.0504**	-0.0187	-0.0190
	(0.0229)	(0.0231)	(0.0146)	(0.0147)
Aligned ideology - Newspaper x Bottom tertile population	-0.0404*	-0.0408*	-0.0060	-0.0060
	(0.0206)	(0.0208)	(0.0136)	(0.0137)
Opposite ideology - Newspaper	0.0160	0.0166	0.0160	0.0163
	(0.0156)	(0.0157)	(0.0102)	(0.0103)
Opposite ideology - Newspaper x Mid tertile population	-0.0163	-0.0171	-0.0321**	-0.0324**
	(0.0192)	(0.0193)	(0.0140)	(0.0141)
Opposite ideology - Newspaper x Bottom tertile population	-0.0193	-0.0199	-0.0160	-0.0162
	(0.0189)	(0.0190)	(0.0121)	(0.0122)
Mean dep var in control	0.0255	0.0255	0.0202	0.0202
Strata FE	No	Yes	No	Yes
N	5,678	5,678	5,678	5,678
Panel B: Pooled effects relative to control				
Any treatment	0.0215*	0.0220*	0.0069	0.0071
	(0.0127)	(0.0128)	(0.0083)	(0.0083)
Any treatment x Mid tertile population	-0.0160	-0.0166	-0.0160	-0.0162
	(0.0163)	(0.0164)	(0.0119)	(0.0120)
Any treatment x Bottom tertile population	-0.0191	-0.0196	-0.0070	-0.0070
	(0.0149)	(0.0150)	(0.0094)	(0.0095)
Aligned ideology	0.0393**	0.0397**	0.0094	0.0095
	(0.0158)	(0.0160)	(0.0084)	(0.0084)
Aligned ideology x Mid tertile population	-0.0331*	-0.0335*	-0.0238**	-0.0240**
	(0.0194)	(0.0195)	(0.0112)	(0.0112)
Aligned ideology x Bottom tertile population	-0.0348**	-0.0351**	-0.0110	-0.0110
	(0.0170)	(0.0172)	(0.0096)	(0.0097)
Opposite ideology	0.0110	0.0116	0.0079	0.0082
	(0.0145)	(0.0146)	(0.0095)	(0.0097)
Opposite ideology x Mid tertile population	-0.0081	-0.0089	-0.0113	-0.0116
	(0.0169)	(0.0169)	(0.0153)	(0.0155)
Opposite ideology x Bottom tertile population	-0.0159	-0.0166	-0.0079	-0.0081
	(0.0168)	(0.0168)	(0.0110)	(0.0112)
Policy brief	0.0149	0.0153	0.0032	0.0033
	(0.0131)	(0.0132)	(0.0099)	(0.0100)
Policy brief x Mid tertile population	-0.0045	-0.0051	-0.0098	-0.0099
	(0.0172)	(0.0173)	(0.0137)	(0.0138)
Policy brief x Bottom tertile population	-0.0118	-0.0123	-0.0043	-0.0043
	(0.0165)	(0.0166)	(0.0111)	(0.0112)
Newspaper	0.0314**	0.0320**	0.0125	0.0128
	(0.0136)	(0.0137)	(0.0086)	(0.0087)
Newspaper X Mid tertile population	-0.0332*	-0.0338*	-0.0254**	-0.0257**
	(0.0172)	(0.0173)	(0.0120)	(0.0121)
Newspaper X Bottom tertile population	-0.0299**	-0.0305**	-0.0110	-0.0111
	(0.0145)	(0.0146)	(0.0102)	(0.0103)

Note: Panel A reports the heterogeneous effects of the different treatment arms on the probability of conducting a recommended change on the Wikipedia page by the population of the municipality. We interact the treatment indicators with a set of dummy variables indicating whether the population is in the mid or the bottom tertile of the population distribution. The interaction term with the top tertile of the population distribution is the reference category. The mean population in municipalities in the bottom tertile is 349 inhabitants, in the medium tertile 1,617 inhabitants and in the top tertile 22,593 inhabitants. Estimates in columns (1) and (2) examine the effect of the different arms for municipalities with larger and smaller populations between May and December 2022, the study period. Estimates in columns (3) and (4) examine the effect of the different treatment arms for municipalities with larger and smaller populations between May and December 2019, a placebo period before the start of the intervention. Regressions in columns (1) and (3) do not include strata fixed-effects and regressions in columns (2) and (4) are estimated with strata fixed-effects. Panel B reports the pooled effects relative to the control group. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

Table C.IV: Heterogeneous effects of the treatment arms by the length of the Wikipedia page

Dep. var: Recommended change in Wikipedia (0/1)	Study period (1)	(2)	Placebo period (3)	(4)
Panel A: Effect of treatment arms relative to control				
Aligned ideology - Policy brief	0.0144	0.0163	-0.0012	0.0000
	(0.0113)	(0.0113)	(0.0079)	(0.0080)
Aligned ideology - Policy brief x Below the median of words	0.0041	0.0013	-0.0089	-0.0108
	(0.0162)	(0.0161)	(0.0126)	(0.0131)
Opposite ideology - Policy brief	0.0095	0.0101	0.0048	0.0058
	(0.0121)	(0.0123)	(0.0101)	(0.0101)
Opposite ideology - Policy brief x Below the median of words	-0.0142	-0.0157	-0.0035	-0.0052
	(0.0145)	(0.0147)	(0.0133)	(0.0132)
Nonsalient ideology - Policy brief	0.0113	0.0116	0.0034	0.0040
	(0.0112)	(0.0114)	(0.0105)	(0.0105)
Nonsalient ideology - Policy brief x Below the median of words	-0.0042	-0.0041	-0.0115	-0.0125
	(0.0154)	(0.0154)	(0.0154)	(0.0154)
Aligned ideology - Newspaper	0.0287**	0.0295**	0.0090	0.0098
	(0.0141)	(0.0142)	(0.0092)	(0.0094)
Aligned ideology - Newspaper x Below the median of words	-0.0257	-0.0265	-0.0170	-0.0180
	(0.0179)	(0.0178)	(0.0135)	(0.0138)
Opposite ideology - Newspaper	0.0109	0.0099	0.0170*	0.0174*
	(0.0124)	(0.0126)	(0.0097)	(0.0098)
Opposite ideology - Newspaper x Below the median of words	-0.0120	-0.0106	-0.0320**	-0.0329**
	(0.0166)	(0.0168)	(0.0125)	(0.0127)
Mean dep var in control	0.0255	0.0255	0.0202	0.0202
Strata FE	No	Yes	No	Yes
N	5,678	5,678	5,678	5,678
Panel B: Pooled effects relative to control				
Any treatment	0.0151*	0.0157*	0.0064	0.0072
	(0.0088)	(0.0090)	(0.0069)	(0.0070)
Any treatment x Below the median of words	-0.0108	-0.0116	-0.0144	-0.0157
	(0.0124)	(0.0124)	(0.0110)	(0.0112)
Aligned ideology	0.0216*	0.0228*	0.0039	0.0049
	(0.0114)	(0.0115)	(0.0071)	(0.0073)
Aligned ideology x Below the median of words	-0.0108	-0.0126	-0.0130	-0.0144
	(0.0151)	(0.0151)	(0.0117)	(0.0121)
Opposite ideology	0.0101	0.0100	0.0108	0.0115
	(0.0099)	(0.0100)	(0.0079)	(0.0079)
Opposite ideology x Below the median of words	-0.0130	-0.0131	-0.0177	-0.0191*
	(0.0137)	(0.0138)	(0.0109)	(0.0110)
Policy brief	0.0118	0.0128	0.0023	0.0032
	(0.0090)	(0.0091)	(0.0080)	(0.0080)
Policy brief x Below the median of words	-0.0051	-0.0066	-0.0077	-0.0093
	(0.0122)	(0.0123)	(0.0119)	(0.0120)
Newspaper	0.0203*	0.0202*	0.0128	0.0134*
	(0.0105)	(0.0106)	(0.0078)	(0.0080)
Newspaper X Below the median of words	-0.0194	-0.0191	-0.0245**	-0.0254**
	(0.0148)	(0.0147)	(0.0116)	(0.0120)

Note: Panel A reports the heterogeneous effects of the different treatment arms on the probability of conducting a recommended change on the Wikipedia page by whether the Wikipedia page was above or below the median number of words before the experiment. To explore the heterogeneous effects of the treatment arms, we interact the treatment indicators with a dummy variable indicating whether the municipality's page in the Spanish Wikipedia is above or below the median number of Words in Wikipedia. Estimates in columns (1) and (2) examine the effect of the different arms for municipalities with longer and shorter pages in Wikipedia between May and December 2022, the study period. Estimates in columns (3) and (4) examine the effect of the different arms for municipalities with longer and shorter pages in Wikipedia between May and December 2019, a placebo period before the start of the intervention. Regressions in columns (1) and (3) do not include strata fixed-effects and regressions in columns (2) and (4) are estimated with strata fixed-effects. Panel B reports the pooled effects relative to the control group. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

Table C.V: Heterogeneous effects of treatment arms by belonging to nationalist party

Dep. var: Recommended change in Wikipedia (0/1)	Study period (1)	(2)	Placebo period (3)	(4)
Panel A: Effect of treatment arms relative to control				
Aligned ideology - Policy brief	0.0092	0.0089	-0.0074	-0.0075
	(0.0096)	(0.0097)	(0.0060)	(0.0059)
Aligned ideology - Policy brief x Nationalist party	0.0301*	0.0313*	0.0077	0.0080
	(0.0174)	(0.0176)	(0.0120)	(0.0119)
Opposite ideology - Policy brief	-0.0005	-0.0007	0.0038	0.0040
	(0.0098)	(0.0099)	(0.0096)	(0.0096)
Opposite ideology - Policy brief x Nationalist party	0.0092	0.0098	-0.0036	-0.0039
	(0.0126)	(0.0129)	(0.0113)	(0.0112)
Nonsalient ideology - Policy brief	0.0095	0.0091	-0.0031	-0.0030
	(0.0107)	(0.0107)	(0.0065)	(0.0066)
Nonsalient ideology - Policy brief x Nationalist party	-0.0008	0.0012	0.0033	0.0029
	(0.0145)	(0.0142)	(0.0107)	(0.0108)
Aligned ideology - Newspaper	0.0132	0.0123	0.0009	0.0007
	(0.0108)	(0.0108)	(0.0068)	(0.0068)
Aligned ideology - Newspaper x Nationalist party	0.0132	0.0168	-0.0005	0.0003
	(0.0154)	(0.0151)	(0.0131)	(0.0132)
Opposite ideology - Newspaper	0.0051	0.0055	-0.0017	-0.0013
	(0.0099)	(0.0098)	(0.0068)	(0.0069)
Opposite ideology - Newspaper x Nationalist party	-0.0050	-0.0061	0.0063	0.0049
	(0.0120)	(0.0119)	(0.0124)	(0.0129)
Panel B: Pooled effects relative to control				
Any treatment	0.0073	0.0070	-0.0015	-0.0014
	(0.0081)	(0.0081)	(0.0054)	(0.0054)
Any treatment x Nationalist party	0.0093	0.0105	0.0026	0.0025
	(0.0101)	(0.0101)	(0.0088)	(0.0090)
Aligned ideology	0.0112	0.0106	-0.0033	-0.0034
	(0.0086)	(0.0087)	(0.0052)	(0.0052)
Aligned ideology x Nationalist party	0.0217*	0.0240**	0.0036	0.0042
	(0.0120)	(0.0120)	(0.0102)	(0.0103)
Opposite ideology	0.0023	0.0024	0.0010	0.0013
	(0.0085)	(0.0086)	(0.0069)	(0.0070)
Opposite ideology x Nationalist party	0.0021	0.0019	0.0013	0.0005
	(0.0109)	(0.0110)	(0.0096)	(0.0100)
Policy brief	0.0060	0.0058	-0.0022	-0.0022
	(0.0086)	(0.0087)	(0.0064)	(0.0064)
Policy brief x Nationalist party	0.0128	0.0140	0.0025	0.0024
	(0.0121)	(0.0121)	(0.0095)	(0.0095)
Newspaper	0.0092	0.0089	-0.0004	-0.0003
	(0.0085)	(0.0085)	(0.0058)	(0.0059)
Newspaper X Nationalist party	0.0040	0.0052	0.0029	0.0026
	(0.0112)	(0.0110)	(0.0107)	(0.0109)
Mean dep var in control	0.0255	0.0255	0.0202	0.0202
Strata FE	No	Yes	No	Yes
N	5,678	5,678	5,678	5,678

Note: Panel A reports the heterogeneous effects of the different treatment arms on the probability of conducting a recommended change on the Wikipedia page by whether the mayor belongs to a political party that promotes independence from Spain. To explore the heterogeneous effects of the treatment arms, we interact the treatment indicators with a dummy variable indicating whether the mayor of the municipality belongs to one of the following political parties: Esquerra Republicana de Catalunya (ERC), Junts per Catalunya, Candidatura d'Unitat Popular (CUP), Partido Nacionalista Vasco (PNV), Euskal Herria Bildu, and Bloque Nacionalista Galego (BNG). Estimates in columns (1) and (2) examine the effect of the different arms for nationalist and non-nationalist municipalities between May and December 2022, the study period. Estimates in columns (3) and (4) examine the effect of the different arms for nationalist and non-nationalist municipalities between May and December 2019, a placebo period before the start of the intervention. Regressions in columns (1) and (3) do not include strata fixed-effects and regressions in columns (2) and (4) are estimated with strata fixed-effects. Panel B reports the pooled effects relative to the control group. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

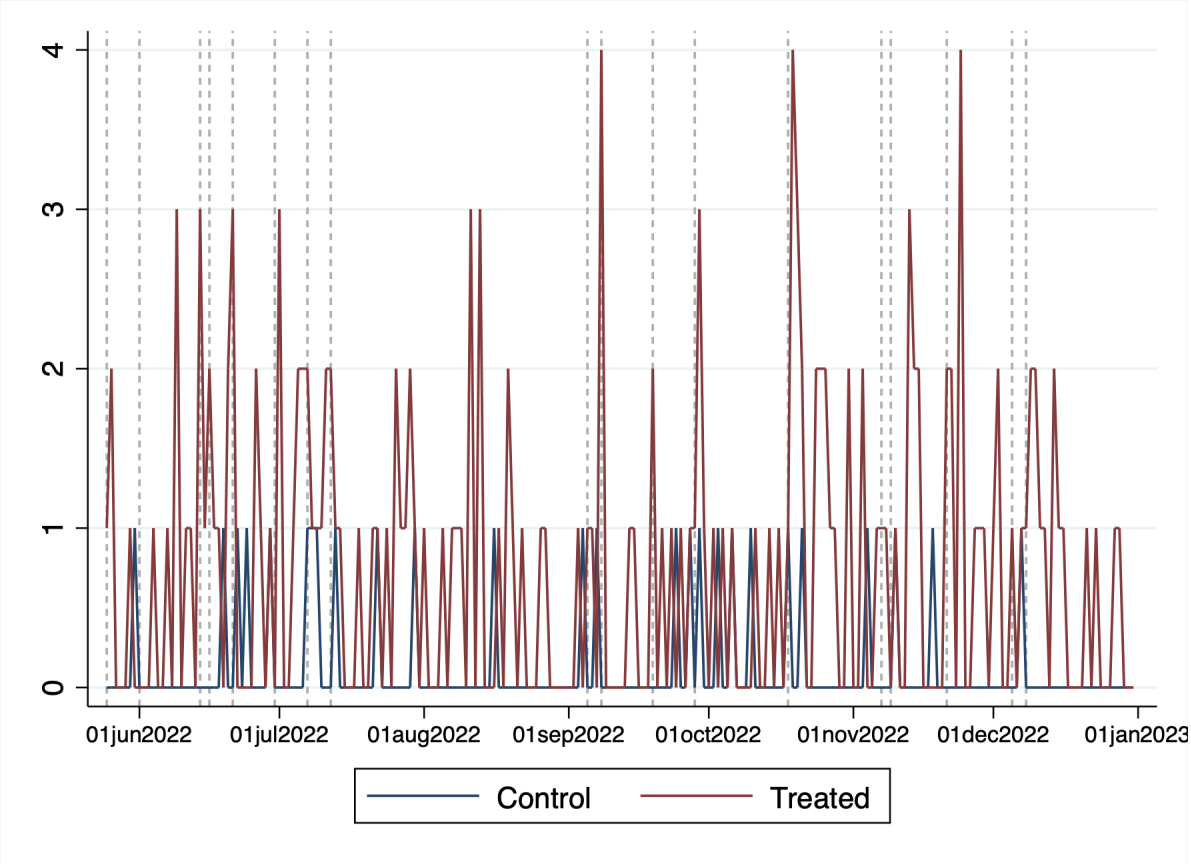
Table C.VI: Heterogeneous effects of the treatment arms by whether the municipality is included in the experiment conducted in Hinnosaar et al. (2021)

	Study period		Placebo period	
Dep. var: Recommended change in Wikipedia (0/1)	(1)	(2)	(3)	(4)
Panel A: Effect of treatment arms relative to control				
Aligned ideology - Policy brief	0.0135*	0.0136*	-0.0056	-0.0055
	(0.0077)	(0.0078)	(0.0043)	(0.0043)
Aligned ideology - Policy brief x Hinosaar	0.4865***	0.4727***	0.0056	-0.0056
	(0.1190)	(0.1219)	(0.0043)	(0.0109)
Opposite ideology - Policy brief	0.0018	0.0020	0.0029	0.0032
	(0.0078)	(0.0079)	(0.0074)	(0.0075)
Opposite ideology - Policy brief x Hinosaar	-0.0018	-0.0156	-0.0029	-0.0181
	(0.0078)	(0.0144)	(0.0074)	(0.0121)
Nonsalient ideology - Policy brief	0.0096	0.0100	-0.0022	-0.0021
	(0.0087)	(0.0088)	(0.0055)	(0.0055)
Nonsalient ideology - Policy brief x Hinosaar	-0.0096	-0.0174	0.0022	0.0009
	(0.0087)	(0.0116)	(0.0055)	(0.0059)
Aligned ideology - Newspaper	0.0136	0.0137	0.0008	0.0009
	(0.0086)	(0.0086)	(0.0056)	(0.0057)
Aligned ideology - Newspaper x Hinosaar	0.3198*	0.3043*	-0.0008	-0.0053
	(0.1678)	(0.1588)	(0.0056)	(0.0099)
Opposite ideology - Newspaper	0.0030	0.0033	-0.0055	-0.0054
	(0.0075)	(0.0076)	(0.0053)	(0.0053)
Opposite ideology - Newspaper x Hinosaar	0.0970	0.0861	0.5055**	0.5108**
	(0.0913)	(0.0840)	(0.2024)	(0.1987)
Panel B: Pool effects relative to control				
Any treatment	0.0083	0.0085	-0.0019	-0.0018
	(0.0061)	(0.0062)	(0.0042)	(0.0042)
Any treatment x Hinosaar	0.1406**	0.1258**	0.1083***	0.1044***
	(0.0599)	(0.0495)	(0.0380)	(0.0378)
Aligned ideology	0.0135*	0.0137*	-0.0024	-0.0023
	(0.0070)	(0.0070)	(0.0042)	(0.0042)
Aligned ideology x Hinosaar	0.3865***	0.3714***	0.0024	-0.0052
	(0.1080)	(0.0992)	(0.0042)	(0.0075)
Opposite ideology	0.0024	0.0026	-0.0013	-0.0011
	(0.0064)	(0.0065)	(0.0053)	(0.0053)
Opposite ideology x Hinosaar	0.0531	0.0408	0.2790***	0.2751***
	(0.0475)	(0.0391)	(0.0952)	(0.0942)
Policy brief	0.0083	0.0085	-0.0016	-0.0015
	(0.0067)	(0.0067)	(0.0048)	(0.0049)
Policy brief x Hinosaar	0.0988**	0.0857*	0.0016	-0.0045
	(0.0482)	(0.0457)	(0.0048)	(0.0065)
Newspaper	0.0083	0.0085	-0.0023	-0.0023
	(0.0065)	(0.0065)	(0.0046)	(0.0046)
Newspaper X Hinosaar	0.2022	0.1858	0.2655***	0.2671***
	(0.1293)	(0.1193)	(0.0988)	(0.0999)
Mean dep var in control	0.0255	0.0255	0.0202	0.0202
Strata FE	No	Yes	No	Yes
N	5,678	5,678	5,678	5,678

Note: Panel A reports the heterogeneous effects of the different treatment arms on the probability of conducting a recommended change on the Wikipedia page by whether the municipality is included in the experiment conducted in Hinnosaar et al. (2021). To explore the heterogeneous effects of the treatment arms, we interact the treatment indicators with a dummy variable indicating whether the municipality is included in the experiment conducted in the latter paper. Estimates in columns (1) and (2) examine the effects on recommended changes between May and December 2022, the study period. Estimates in columns (3) and (4) examine the effects of the different arms on recommended changes between May and December 2019, a placebo period before the start of the intervention. Regressions in columns (1) and (3) do not include strata fixed-effects and regressions in columns (2) and (4) are estimated with strata fixed-effects. Panel B reports the pooled effects relative to the control group. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

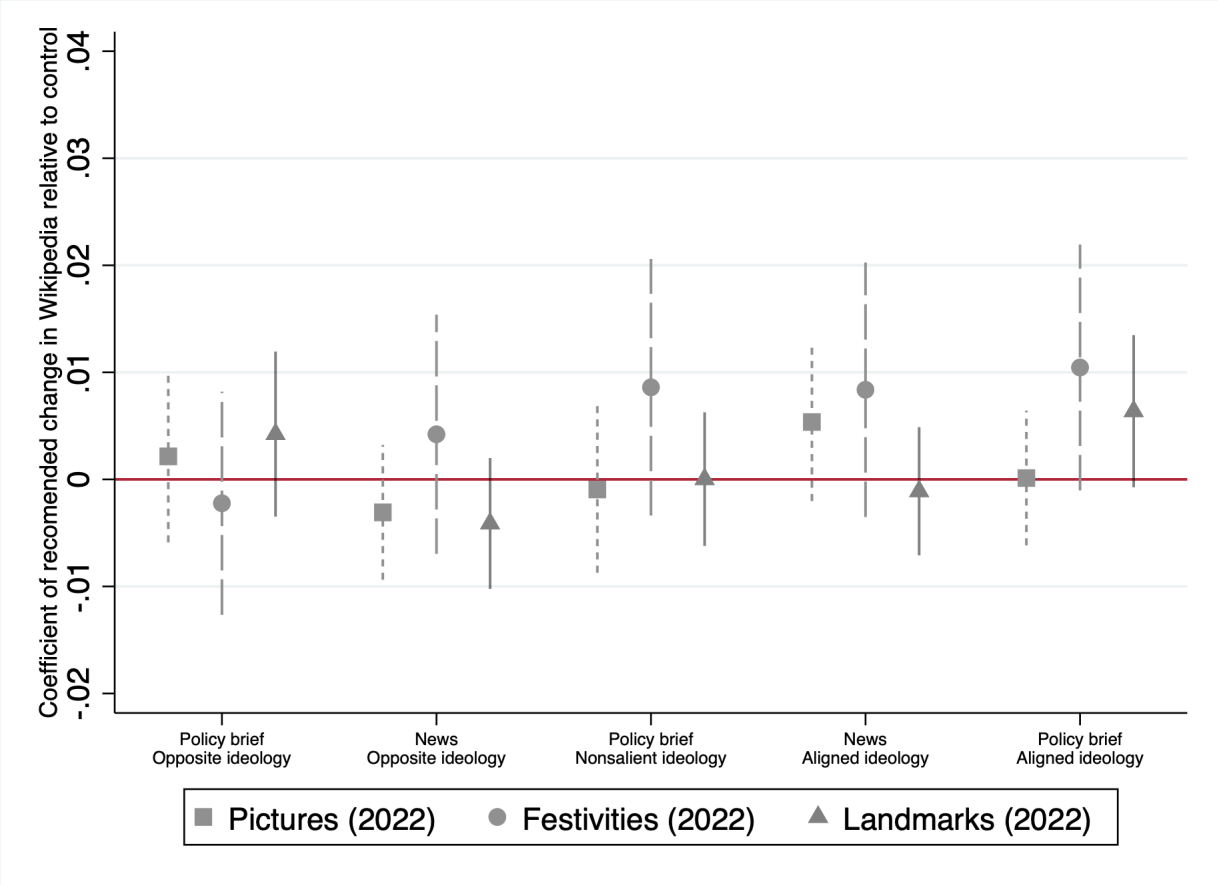
D Additional tables and graphs

Figure D.I: Timing of changes in Wikipedia over the study period



Note: The figure displays the number of recommended changes in the Spanish Wikipedia page of control and treatment municipalities over the study period. The vertical dashed lines show the timing of the reminder emails sent. The sum of the days on which mailings were made and the day after represent 16.7% of the days in our intervention period, but these account for 30% of the changes in the treatment group.

Figure D.II: Heterogeneity of results by types of recommended change in Wikipedia



Note: The figure displays the point estimates and 95% confidence intervals for the effect of the different treatment arms relative to the control group on the probability of doing different types of changes in the municipalities' page in the Spanish Wikipedia during the study period. The taxonomy of changes includes: changes related to local festivities, changes related to pictures, and changes related to landmarks. The estimates were conducted with strata fixed-effects.

Table D.I: P-values of the different treatments:

	Without strata fixed effect (1)	With strata fixed effect (2)
Panel A: All individuals		
Any treatment vs Control group	0.128	0.126
Aligned ideology vs Control group	0.029	0.030
Opposite ideology vs Control group	0.643	0.631
Nonsalient ideology vs Control group	0.276	0.269
Aligned ideology vs Opposite ideology	0.022	0.023
Aligned ideology vs Nonsalient ideology	0.344	0.355
Nonsalient ideology vs Opposite ideology	0.348	0.345
Policy brief vs Control group	0.169	0.166
Newspaper vs Control group	.124	.124
Newspaper vs Policy brief	0.819	0.823

Note: The table reports the p-values of t-test conducted after regressions estimated without strata fixed effects in column (1) and with strata fixed effects in column (2). The p-values corresponds to the effects reported in Table 3.2.

Table D.II: Spatial spillovers: Effect of distance to the nearest municipality in each treatment arm on the probability of changing Wikipedia for municipalities in the control group

	Study Period	Placebo Period
	Recommended changes in Wikipedia (1)	Recommended changes in Wikipedia (2)
Dist (in miles) to nearest municipality treated with...		
Aligned ideology - Policy brief	0.0007	-0.0001
	(0.0007)	(0.0005)
Opposite ideology - Policy Brief	0.0002	0.0000
	(0.0009)	(0.0007)
Nonsalient ideology - Policy Brief	0.0003	0.0007
	(0.0006)	(0.0005)
Aligned ideology - Newspaper	-0.0002	-0.0005
	(0.0007)	(0.0005)
Opposite ideology - Newspaper	-0.0003	0.0002
	(0.0004)	(0.0006)
Mean outcome	0.0255	0.0202
N	941	941

*Note: For the sample of control municipalities, the table reports the effect of distance to the closest municipality in each treatment arm on the probability of a recommended change in Wikipedia during the study period in column (1) and the placebo paeriod in column (2). Te mean distance in miles from control municipalities to the closest in municipality in the aligned policy brief treatment arm is 2.39, in the opposite policy brief treatment arm is 2.38, in the nonsalient policy brief treatment arm is 2.40, in the aligned newspaper treatment arm is 2.3802, and in the opposite newspaper treatment arm is 2.37. heteroskedasticity-consistent standard errors are reported in parentheses.***p<0.01;**p<0.05;*p<0.1.*

Table D.III: Effects of the treatment arms on the probability of making a recommended change in Wikipedia (Goldsmith-Pinkham et al., 2022)

	Study period		Placebo period	
	(1)	(2)	(3)	(4)
Panel A: Effect of treatment arms relative to control				
Aligned ideology - Policy brief	0.0168**	0.0169**	-0.0054	-0.0053
	(0.0081)	(0.0082)	(0.0043)	(0.0059)
Opposite ideology - Policy brief	0.0019	0.0020	0.0030	0.0032
	(0.0078)	(0.0073)	(0.0074)	(0.0066)
Nonsalient ideology - Policy brief	0.0094	0.0097	-0.0022	-0.0022
	(0.0086)	(0.0078)	(0.0054)	(0.0062)
Aligned ideology - Newspaper	0.0167*	0.0167**	0.0009	0.0010
	(0.0091)	(0.0082)	(0.0056)	(0.0065)
Opposite ideology - Newspaper	0.0041	0.0043	-0.0001	0.0001
	(0.0075)	(0.0074)	(0.0055)	(0.0064)
Mean dep var in control	0.0255	0.0255	0.0202	0.0202
Strata FE	No	Yes	No	Yes
N	5,678	5,678	5,678	5,678
Panel B: Pooled effects relative to control				
Any treatment	0.0098	0.0099	-0.0007	-0.0006
	(0.0063)	(0.0058)	(0.0042)	(0.0049)
Aligned ideology	0.0167**	0.0168**	-0.0022	-0.0022
	(0.0075)	(0.0068)	(0.0042)	(0.0054)
Opposite ideology	0.0030	0.0032	0.0015	0.0016
	(0.0065)	(0.0064)	(0.0054)	(0.0056)
Policy brief	0.0094	0.0095	-0.0015	-0.0014
	(0.0067)	(0.0062)	(0.0048)	(0.0052)
Newspaper	0.0104	0.0105	0.0004	0.0005
	(0.0067)	(0.0066)	(0.0046)	(0.0055)

Note: The table replicates the main results of the study reported in Table 3.2 but using the estimation method presented in Goldsmith-Pinkham et al. (2022). This method is used to account for contamination when estimating the effect of mutually exclusive treatments with control variables. Estimates in columns (1) and (2) examine the effect of the different arms on recommended changes between May and December 2022. These are the main results of the study. Estimates in columns (3) and (4) examine the effect of the different arms on recommended changes between May and December 2019, a placebo period before the start of the intervention. Regressions in columns (1) and (3) do not include strata fixed-effects and regressions in columns (2) and (4) are estimated with strata fixed-effects. *Any treatment* yields the pooled effect of receiving the information across all treatment groups relative to not receiving any information. *Aligned ideology* yields the pooled effect of receiving the summary of study results endorsed by an institution (newspaper or think tank) with the same ideology relative to not receiving any information. *Opposite ideology* yields the pooled effect of receiving the summary of study results endorsed by an institution (newspaper or think tank) with the opposite ideology relative to not receiving any information. *Policy brief* yields the pooled effect of receiving the summary of study results through a policy brief relative to not receiving any information regardless of the ideology of the think tank. *Newspaper* yields the pooled effect of receiving the summary of study results through a newspaper article regardless of the ideology of the newspaper relative to not receiving any information. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

Table D.IV: Effects of the treatment arms on the number of contact emails targetted in the municipality

	(1)	(2)
Aligned ideology - Policy brief	-0.0380	-0.0324
	(0.0661)	(0.0659)
Nonsalient ideology - Policy brief	0.0113	0.0045
	(0.0592)	(0.0589)
Aligned ideology - Newspaper	-0.0621	-0.0648
	(0.0700)	(0.0695)
Opposite ideology - Newspaper	0.0420	0.0409
	(0.0647)	(0.0655)
Reference group: Opposite ideology - Policy brief		
Mean dep variable	2.9226	2.9226
Strata FE	No	Yes
N	4,652	4,652

Note: Panel A reports the effects of the different treatment arms on the probability of conducting a recommended change on the Wikipedia page. Estimates in columns (1) and (2) examine the effect of the different arms on the number of emails targeted in each treatment group. The reference groups are the municipalities that received the information endorsed by an institution with an opposite ideology. Estimates in Column (1) do not include strata fixed-effects and regressions in column (2) are estimated with strata fixed-effects. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

Table D.V: Treatment effects on the probability of opening an email

	At the municipality level		At the email level	
Dep var: Opening email (0/1)	(1)	(2)	(3)	(4)
Aligned ideology - Policy brief	-0.0051	-0.0039	0.0083	0.0079
	(0.0239)	(0.0241)	(0.0212)	(0.0213)
Nonsalient ideology - Policy brief	0.0215	0.0216	-0.0002	-0.0043
	(0.0244)	(0.0245)	(0.0205)	(0.0203)
Aligned ideology - Newspaper	-0.0095	-0.0092	-0.0174	-0.0175
	(0.0263)	(0.0264)	(0.0219)	(0.0219)
Opposite ideology - Newspaper	-0.0257	-0.0248	-0.0204	-0.0197
	(0.0245)	(0.0247)	(0.0212)	(0.0214)
Reference group: Opposite ideology - Policy brief				
Mean dep variable	0.5716	0.5716	0.3796	0.3796
Strata FE	No	Yes	No	Yes
N	4,736	4,736	11,288	11,288

Note: The estimates presented in the table yield the differences across the different treatment groups and the arm that received the policy brief endorsed by a think tank with opposite ideology regarding the probability of opening the email containing the intervention. The latter group is the omitted category in the regressions since the control group did not receive the intervention email. Estimates reported in columns (1) and (3) are estimated without strata-fixed effects, and columns (2) and (4) are estimated with strata-fixed effects. The outcome variable, whether an email with the intervention is opened, is measured at the municipality level in columns (1) and (2). Because we had more than one email address in some municipalities, we estimated the effects in columns (3) and (4) with the outcome variable measured at the email address level. Standard errors in parentheses are clustered at the randomization strata level..***p<0.01;**p<0.05;*p<0.1

Table D.VI: Power calculations for dichotomous outcomes: Minimum detectable effect size (MDE)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Baseline prob.	Any treatment vs Control	Newspaper vs Policy brief	Newspaper vs Control	Policy brief vs Control	Same ideology vs Opposite ideology	Same ideology vs Control	Opposite ideology vs Control	Any group vs Control
0.4	0.049	0.041	0.055	0.052	0.045	0.055	0.055	0.064
0.1	0.033	0.026	0.036	0.034	0.029	0.036	0.036	0.042
0.03	0.02	0.016	0.023	0.021	0.018	0.023	0.023	0.026

Note: The table reports the minimum detectable effect size (MDE) with a probability of 80% for different treatment arms comparisons. The calculations are for dichotomous outcomes and we assume different baseline probabilities for the outcomes. We believe these were reasonable probabilities for the outcomes: opening the email received, clicking on the link to the policy brief/newspaper, and changing the Wikipedia.

Table D.VII: Power calculations

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Baseline prob.	Any treatment vs Control	Newspaper vs Policy brief	Newspaper vs Control	Policy brief vs Control	Same ideology vs Opposed ideology	Same ideology vs Control	Opposed ideology vs Control	Any group vs Control
Ln words	0.043	0.036	0.048	0.045	0.039	0.048	0.048	0.055
Ln image	0.078	0.065	0.087	0.082	0.071	0.087	0.087	0.101

Note: The table reports the minimum detectable effect size (MDE) with a probability of 80% for different treatment arms comparisons. The calculations are for continuous outcomes for two outcomes for which we have baseline information: the number of words and the number of images.

Table D.VIII: Effects of the different treatment arms on the number of clicks

	At the municipality level		At the email level	
	(1)	(2)	(3)	(4)
Panel A: Number of clicks				
Aligned ideology - Policy brief	0.0465	0.0478	0.0176	0.0179
	(0.1107)	(0.1112)	(0.0463)	(0.0448)
Nonsalient ideology - Policy brief	0.2793	0.2789	0.1151	0.1142
	(0.2295)	(0.2306)	(0.0920)	(0.0915)
Aligned ideology - Newspaper	-0.1191	-0.1198	-0.0484	-0.0490
	(0.1188)	(0.1196)	(0.0492)	(0.0461)
Opposite ideology - Newspaper	0.1496	0.1517	0.0614	0.0670
	(0.2078)	(0.2103)	(0.0854)	(0.0847)
Reference group: Opposite ideology - Policy brief				
Mean dep variable	0.5674	0.5674	0.2366	0.2366
Strata FE	No	Yes	No	Yes
N	4,736	4,736	11,288	11,288
Panel B: Number of clicks on policy brief/newspaper				
Aligned ideology - Policy brief	0.0143	0.0150	0.0044	0.0039
	(0.0574)	(0.0576)	(0.0238)	(0.0230)
Nonsalient ideology - Policy brief	0.1462	0.1459	0.0602	0.0591
	(0.1141)	(0.1146)	(0.0456)	(0.0454)
Aligned ideology - Newspaper	-0.0717	-0.0721	-0.0292	-0.0298
	(0.0609)	(0.0613)	(0.0251)	(0.0238)
Opposite ideology - Newspaper	0.0664	0.0676	0.0270	0.0294
	(0.1005)	(0.1017)	(0.0414)	(0.0409)
Reference group: Opposite ideology - Policy brief				
Mean dep variable	0.3155	0.3155	0.1312	0.1312
Strata FE	No	Yes	No	Yes
N	4,736	4,736	11,288	11,288
Panel C: Number of clicks on instructions to edit Wikipedia				
Aligned ideology - Policy brief	0.0321	0.0328	0.0132	0.0140
	(0.0559)	(0.0562)	(0.0234)	(0.0227)
Nonsalient ideology - Policy brief	0.1332	0.1330	0.0550	0.0552
	(0.1162)	(0.1168)	(0.0467)	(0.0464)
Aligned ideology - Newspaper	-0.0474	-0.0477	-0.0192	-0.0193
	(0.0595)	(0.0599)	(0.0247)	(0.0229)
Opposite ideology - Newspaper	0.0831	0.0841	0.0344	0.0376
	(0.1090)	(0.1104)	(0.0448)	(0.0445)
Reference group: Opposite ideology - Policy brief				
Mean dep variable	0.2519	0.2519	0.1054	0.1054
Strata FE	No	Yes	No	Yes
N	4,736	4,736	11,288	11,288

Note: The estimates presented in the table yield the effect of the different treatment arms on the number of clicks through the links in the email relative to the group of individuals that receive the summary of results endorsed by a think tank with an opposite ideology. The latter group is the omitted category in the regressions since the control group did not receive the intervention email, and we cannot measure clicks for them. In Panel A, the outcome variable is the number of clicks to the two links included in the email. In Panel B, the outcome variable is the number of clicks to the link that provided the results summary (either the policy brief or the newspaper). In Panel C, the outcome variable is the number of clicks to the link that provided the step-by-step instructions to change the Wikipedia page. The outcome variable is measured at the municipality level in columns (1) and (2). Because we had more than one email address in some municipalities, we estimate the effects in columns (3) and (4) with the outcome variable measured at the email address level. Estimates reported in columns (1) and (3) are estimated without strata-fixed effects, and columns (2) and (4) are estimated with strata-fixed effects. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

E Instructions for independent coders

List of inclusion criteria in the different categories of changes:

1. **Recommended Changes**: They can be of three types; it is enough for only one of them to be met for it to score as 1 in this variable.

- Changes in the festivals section: Any addition of an additional festival, the incorporation of dates in festivals that are already included, or the incorporation of relevant information about the activities carried out in said festivals will be considered a change in the parties section. In the same way, it will also be counted if what is done is to write a web link to a page where information of this type is collected (tourist information or the city council's page).
- Incorporation of new images: This type of change is especially difficult to detect since the change in bytes is small and may go unnoticed. The kind of image that is added is essential and can be known by the name of the file that is uploaded. Those that show the coat of arms or the logo of the city council, as well as those that show party logos or photos of politicians, will not be considered valid images, or graphs on the city council's outstanding debt or its demographic evolution. Yes, those that include photographs of municipal buildings, beaches, the environment, or that show a tourist attraction will be considered valid image changes.

Figure E.I: Inclusion of photos in Wikipedia:

- Edit a version in English: Any addition of text in English to the text, if it concerns festivals, the municipality's heritage, or tourism.

2. **Not recommended but credible changes**: They can be from at least three categories. Again, it is enough for one of them to be met to consider it as a change.

- Changes in the heritage section: In this case, any incorporation of new monuments or buildings that are part of the architectural or historical heritage of the municipality is considered a change, as well as the expansion of information on the buildings/monuments that are already included. Likewise, this also applies to artistic heritage. It will also be considered a change if, in any of the

buildings and works of art, information is added that it is regarded as an asset of tourist/cultural interest by a public institution.

- Changes in the nature/environment section: In this case, those that refer to the natural heritage of the municipality will be considered as changes, especially when reference is made to its value in terms of tourist attraction, either due to the inclusion of routes, or other possible outdoor activities. This section includes expanding or incorporating sections on the municipality's beaches. Changes related to recording temperatures, precipitation or similar data would not be considered. It will also be considered a change if, in any place of tourist interest, the information is added that it is viewed as an asset of tourist/cultural interest by some public institution.
- Changes in the gastronomy/tourism section: In this case, creating a gastronomy or tourism section is considered a change most municipalities do not have. Similarly, the incorporation of a list of hospitality establishments in the gastronomy section or the incorporation or expansion of information on places of tourist interest in the tourism section would also be considered a change. It will also be considered a change if, in any area of tourist interest, the information is added that it is viewed as an asset of tourist/cultural interest by some public institution.

3. **Not recommended or credible changes**: These are outside previous sections and will be the most common. The most common, which should be distinct from any of the earlier categories, are changes in geography, administration, politics, and history, which would only enter as changes in any of the previous categories if any of the previously mentioned requirements are met.

What types of users can we find on Wikipedia?

- Changes made by an IP
- Changes made by a user without a page built (User in red)
- Changes made by a Wikipedia user with a built page (Blue User)

Practical instructions for Wikipedia classification

1. First, you must enter the general page of Wikipedia and write the name of the corresponding municipality in the search engine. As the title is the one used by Wikipedia, the first tab that will appear will be the municipality's website, the one you must click.

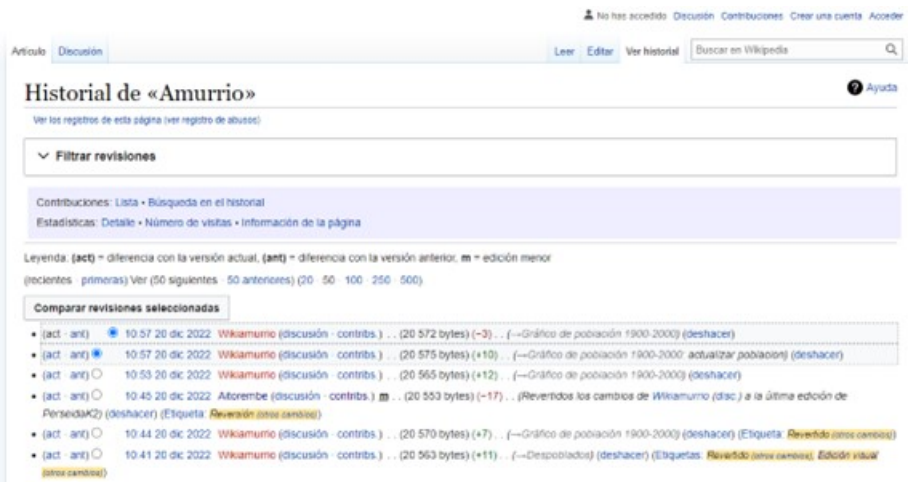
Figure E.II: Step 1 to classify Wikipedia changes:

2. Once you are on the Wikipedia page of the municipality in question, you must immediately access, without reading any of the page's content, the change history section, which is in the screen's upper right position.

Figure E.III: Step 2 to classify Wikipedia changes:

3. Next, you must go down in history until you reach the last change made before May 25, 2022. Next, you must open each of the changes from that date until December 31, 2022, and review each according to the previously established criteria. If any changes can be classified as recommended or non-recommended but credible, the URL associated with the main change that motivates its registration must be copied and pasted.

Figure E.IV: Step 3 to classify Wikipedia changes:



4. Once a type 1 or 2 change has been located, it is necessary to classify what kind of person makes the change. To do this, the following classification system must be followed:

- First, copy the URL associated with the change in question and paste it in the corresponding column.
- Indicate if the author of the change is a user with a page developed within Wikipedia. These users have their name in blue and may (or may not) have a very generated Wikipedia page, as is the following case:

Figure E.V: Step 4 to classify Wikipedia changes:

- If the author of the change is an IP address, it is necessary to indicate it with a 1 in the variable showed for this (or with a zero otherwise) and then answer the following questions, whose information can be accessed by clicking the IP address responsible for the change in question, as can be seen in the next image:

1. Has that IP made changes in other municipalities?

2. If they have done so, were these municipalities all in the same province as the study municipality?

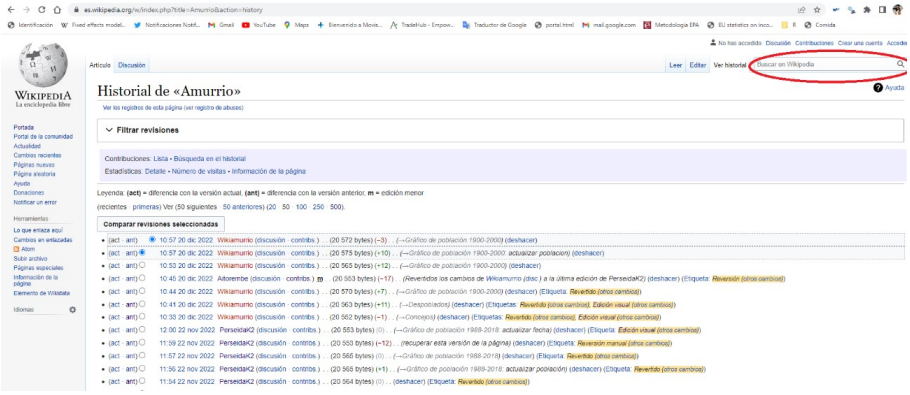
3. If this IP has only made changes in the municipality, did it change before the period analysed?

Figure E.VI: Step 5 to classify Wikipedia changes:

5. Once the changes in the 2022 period have been reviewed, the changes in the same period of 2019 must be looked over, and, if applicable, please note the changes in the respective row of the municipality corresponding to 2019 in the Excel template.

6. Once in which at least one change of type 1 and another of type 2 has been identified in each period considered, from May 25 to December 31, 2019, and 2022, respectively, regardless of whether the same user makes them, it is not necessary to continue reviewing said Wikipedia page. In this case, the name of the following municipality must be written in the Wikipedia search engine, located in the list at the top right of the screen.

Figure E.VII: Step 6 to classify Wikipedia changes:



F The Effect of Treatment Arms on Other Wikipedia Outcomes

This appendix examines the effect of the different treatment arms on secondary Wikipedia outcomes. These include the length in words and the number of images on the page of the municipality in the Spanish Wikipedia, the length in words and the number of images on the page of the municipality in the English Wikipedia, and the number of languages in which there exists a municipalities' Wikipedia page. We collect information for these outcomes at the start (May 2022) and at the end of the study period (January 2023). We define the dependent variable as the difference in the variable between these two temporal points. In this section, we also report the effect of the different treatment arms on the probability of conducting a recommended change on the English page of the municipality in Wikipedia.

The results reported in this appendix should be interpreted with caution because most of these Wikipedia outcomes are poor indicators of the adoption of the recommended policy. First, most of the Wikipedia changes registered during the study period are minor edits to the text or images unrelated to the changes we recommend, which may increase or decrease the length of the text and add measurement error to the outcome variable, biasing estimates towards zero.[53] Second, a large proportion of changes along the recommended guidelines, either in images or in text, were reverted after some time by other Wikipedia users (more than 30%).[54] While recommended changes during the study period were identified even if they were reverted, the change in the number of words or images between the end and the beginning of the study period would not capture changes that were reverted. These reasons require the results reported in this appendix to be interpreted with caution, particularly when compared with the results reported in Section 3.5, which uses the probability of conducting a change in Wikipedia along the recommended guidelines at any point during the study period as the indicator of policy adoption. Finally, while we mention in the summary of evidence the importance of improving the Wikipedia page in other languages, the low baseline levels of change in the English Wikipedia page suggest that implementation costs might be larger than for the Spanish page, constraining policy adoption. This is not a surprising result given that, according to the Spanish census, only 15% of Spanish people speak "good" English, and many small municipalities have very limited resources.[55] Thus, the variables measuring recommended changes in the webpage

[53] They were very small amendments to the history sections, slight edits in the text such as the removal of articles in sentences (not necessarily grammar errors), adding references, and other minor changes unrelated to the changes recommended by the study.

[54] There are regular Wikipedia users who, in addition to making changes to enrich the Wikipedia pages of municipalities, also revert changes they consider inappropriate. Unfortunately, this was the case for a considerable number of Wikipedia changes that aligned with our recommendations.

[55] See for example https://ine.es/jaxi/Tabla.htm?tpx=55481&L=0

of the municipality in the English Wikipedia and the number of languages in which the municipality has a page in Wikipedia might be unrealistic outcomes for most of the municipalities in the sample.

We first investigate the pooled effect of receiving information about study results on these secondary Wikipedia outcomes. The results are reported in Tables F.I and F.II. They show insignificant effects of information provision on changes in the probability of changing the English page of Wikipedia along the recommended guidelines, in the number of words in the municipalities' Wikipedia page in Spanish, in the number of images in the municipalities' Wikipedia page in Spanish, in the number of languages in which there exists a municipalities' Wikipedia page, and in the number of images in the municipalities' Wikipedia page in English. Only the positive effect of information provision on the change in the number of words in the municipality Wikipedia page in English is statistically significant at the 10% confidence level.

We then examine the effect of ideological alignment between the informing institution and policymakers on the secondary Wikipedia outcomes. The results reported in Table F.I show that the effect of ideological alignment on recommended changes identified in Section 3.5 does not translate into a larger number of words or images on the Spanish page of the municipality. As discussed above, these results should, however, be interpreted with caution as the vast majority of changes in Wikipedia during the study period were unrelated to our intervention (introducing noise in the estimation), which bias the estimates towards 0. Moreover, some of the recommended changes were reverted before the study period's end, which made variations in the length of the Wikipedia page and in the number of images during the study period poor proxies of policy adoption. Similarly, Panel B of Table F.I and Table F.II show overall no effects of ideological alignment on the probability of changing the English Wikipedia page along the recommended guidelines and on other Wikipedia-related outcomes in the municipality's English Wikipedia page. The low rate of changes in the control group and the lack of treatment effects are consistent with the hypothesis of high implementation costs for most municipalities to change Wikipedia in other languages.

The effect of receiving a "Nonsalient ideology - Policy brief" on the Wikipedia outcomes examined in this section is reported in Tables F.I and F.II. Once again, the results reveal insignificant effects on most of the Wikipedia outcomes examined in this section of receiving information from a researcher from an ideologically nonsalient institution. Only the coefficient measuring the effect on the number of words on the municipality's English page is marginally significant at 10 percent confidence levels.

The effect of receiving a summary of evidence from an "ideologically opposite institution" on other Wikipedia-related outcomes is reported in Tables F.I and F.II. We find no statistically significant effects of receiving information from an ideologically opposite institution on any of the Wikipedia outcomes analysed in this section.

Finally, the results reported in Tables F.I and F.II show that the effects of policy briefs and newspapers on the Wikipedia-related outcomes examined in this section are similarly small.

Table F.I: Effects of treatment arms on other Wikipedia outcomes

	Δ N words		Δ N images		Δ Languages	
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A: Spanish page						
Effect of treatment treatment arms relative to control						
Aligned ideology - Policy brief	0.14	0.19	-0.06	-0.06	0.00	0.00
	(2.61)	(2.63)	(0.04)	(0.04)	(0.00)	(0.00)
Opposite ideology - Policy brief	-1.58	-1.55	-0.04	-0.04	0.00	0.00
	(2.51)	(2.55)	(0.04)	(0.04)	(0.00)	(0.00)
Nonsalient ideology - Policy brief	0.22	0.29	-0.01	-0.01	0.00	0.00
	(2.02)	(2.03)	(0.04)	(0.04)	(0.00)	(0.00)
Aligned ideology - Newspaper	-0.36	-0.31	-0.02	-0.02	0.00	0.00
	(1.88)	(1.89)	(0.04)	(0.04)	(0.00)	(0.00)
Opposite ideology - Newspaper	-0.45	-0.42	-0.01	-0.01	0.00	0.00
	(1.70)	(1.71)	(0.03)	(0.03)	(0.00)	(0.00)
Mean dep var in control	8.02	8.02	0.19	0.19	0.00	0.00
Strata FE	No	Yes	No	Yes	No	Yes
Control Dep. var	No	Yes	No	Yes	No	Yes
N	5,669	5,669	5,669	5,669	5,669	5,669
Pooled effects relative to control						
Any treatment	-0.41	-0.36	-0.03	-0.03	0.00	0.00
	(1.87)	(1.88)	(0.02)	(0.02)	(0.00)	(0.00)
Aligned ideology	-0.11	-0.06	-0.04	-0.04	0.00	0.00
	(2.13)	(2.15)	(0.03)	(0.03)	(0.00)	(0.00)
Opposite ideology	-1.01	-0.99	-0.03	-0.02	0.00	0.00
	(2.01)	(2.03)	(0.03)	(0.03)	(0.00)	(0.00)
Policy brief	-0.41	-0.36	-0.04	-0.03	0.00	0.00
	(2.07)	(2.09)	(0.03)	(0.03)	(0.00)	(0.00)
Newspaper	-0.40	-0.36	-0.02	-0.02	0.00	0.00
	(1.66)	(1.67)	(0.03)	(0.03)	(0.00)	(0.00)
Panel B: English page						
Effect of treatment treatment arms relative to control						
Aligned ideology - Policy brief	1.09**	1.08**	0.01	0.01		
	(0.53)	(0.53)	(0.02)	(0.02)		
Opposite ideology - Policy brief	0.51	0.49	-0.01	-0.01		
	(0.37)	(0.37)	(0.02)	(0.02)		
Nonsalient ideology - Policy brief	0.73*	0.73*	0.00	0.00		
	(0.43)	(0.43)	(0.02)	(0.02)		
Aligned ideology - Newspaper	0.90**	0.89**	0.02	0.02		
	(0.37)	(0.37)	(0.02)	(0.02)		
Opposite ideology - Newspaper	0.32	0.32	-0.03	-0.02		
	(0.49)	(0.49)	(0.02)	(0.02)		
Mean dep var in control	-2.00	-2.00	0.13	0.13		
Strata FE	No	Yes	No	Yes		
Control Dep. var	No	Yes	No	Yes		
N	5,663	5,663	5,663	5,663		
Pooled effects relative to control						
Any treatment	0.71*	0.70*	-0.00	-0.00		
	(0.36)	(0.36)	(0.02)	(0.02)		
Aligned ideology	0.99**	0.99**	0.01	0.01		
	(0.40)	(0.41)	(0.02)	(0.02)		
Opposite ideology	0.41	0.41	-0.02	-0.02		
	(0.39)	(0.39)	(0.02)	(0.02)		
Policy brief	0.77**	0.77**	0.00	0.00		
	(0.36)	(0.36)	(0.02)	(0.02)		
Newspaper	0.61	0.61	-0.01	-0.00		
	(0.38)	(0.39)	(0.02)	(0.02)		

Note: The table reports the effect of the different treatment arms on other Wikipedia outcomes. Panel A reports the effects on the number of words in the municipality's Spanish Wikipedia page, the number of images in the municipality's Spanish Wikipedia page, and the number of languages in which the municipality has a page in Wikipedia. Panel B reports the effects on the number of words in the municipality's English Wikipedia page, and the number of images in the municipality's English Wikipedia page. The dependent variables are defined in changes between the variable measured at the end of the study period and the variable measured at the beginning of the study period. The table also presents the pooled effects of different treatment arms relative to the control group. *Any treatment* yields the pooled effect of receiving the information across all treatment groups relative to not receiving any information. *Aligned ideology* yields the pooled effect of receiving the summary of study results endorsed by an institution (newspaper or think tank) with the same ideology relative to not receiving any information. *Opposite ideology* yields the pooled effect of receiving the summary of study results endorsed by an institution (newspaper or think tank) with the opposite ideology relative to not receiving any information. *Policy brief* yields the pooled effect of receiving the summary of study results through a policy brief relative to not receiving any information regardless of the ideology of the think tank. *Newspaper* yields the pooled effect of receiving the summary of study results through a newspaper article regardless of the ideology of the newspaper relative to not receiving any information. Regressions in columns (1), (3), and (5) do not include strata fixed-effects, and regressions in columns (2), (4), and (6) are estimated with strata fixed-effects. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

Table F.II: Treatment effects on the probability of making a recommended change in Wikipedia:

English Wikipedia page	Study period		Placebo period Placebo	
	(1)	(2)	(3)	(4)
Aligned ideology - Policy brief	-0.0021	-0.0021	-0.0021	-0.0020
	(0.0033)	(0.0033)	(0.0026)	(0.0026)
Opposite ideology - Policy brief	-0.0000	-0.0000	-0.0011	-0.0010
	(0.0037)	(0.0037)	(0.0032)	(0.0032)
Nonsalient ideology - Policy brief	-0.0032	-0.0032	-0.0032*	-0.0031*
	(0.0031)	(0.0031)	(0.0017)	(0.0017)
Aligned ideology - Newspaper	-0.0011	-0.0011	-0.0032	-0.0031
	(0.0032)	(0.0032)	(0.0023)	(0.0023)
Opposite ideology - Newspaper	-0.0043	-0.0043	0.0032	0.0033
	(0.0028)	(0.0028)	(0.0035)	(0.0035)
Mean dep var in control	0.0053	0.0053	0.0042	0.0042
Strata FE	No	Yes	No	Yes
N	5,678	5,678	5,678	5,678
Pooled effects relative to control				
Any treatment	-0.0021	-0.0021	-0.0013	-0.0012
	(0.0028)	(0.0029)	(0.0021)	(0.0021)
Aligned ideology	-0.0016	-0.0016	-0.0027	-0.0026
	(0.0029)	(0.0029)	(0.0023)	(0.0023)
Opposite ideology	-0.0021	-0.0022	0.0010	0.0011
	(0.0031)	(0.0031)	(0.0030)	(0.0031)
Policy brief	-0.0018	-0.0018	-0.0021	-0.0020
	(0.0030)	(0.0030)	(0.0021)	(0.0022)
Newspaper	-0.0027	-0.0027	-0.0000	0.0001
	(0.0028)	(0.0028)	(0.0024)	(0.0024)

Note: Panel A reports the effects of the different treatment arms on the probability of conducting a recommended change on the English Wikipedia page. Estimates in columns (1) and (2) examine the effect of the different arms on recommended changes between May and December 2022. These are the main results of the study. Estimates in columns (3) and (4) examine the effect of the different arms on recommended changes between May and December 2019, a placebo period before the start of the intervention. Regressions in columns (1) and (3) do not include strata fixed-effects and regressions in columns (2) and (4) are estimated with strata fixed-effects. Panel B reports the pooled effects relative to the control group. *Any treatment* yields the pooled effect of receiving the information across all treatment groups relative to not receiving any information. *Aligned ideology* yields the pooled effect of receiving the summary of study results endorsed by an institution (newspaper or think tank) with the same ideology relative to not receiving any information. *Opposite ideology* yields the pooled effect of receiving the summary of study results endorsed by an institution (newspaper or think tank) with the opposite ideology relative to not receiving any information. *Policy brief* yields the pooled effect of receiving the summary of study results through a policy brief relative to not receiving any information regardless of the ideology of the think tank. *Newspaper* yields the pooled effect of receiving the summary of study results through a newspaper article regardless of the ideology of the newspaper relative to not receiving any information. Standard errors in parentheses are clustered at the randomization strata level.***p<0.01;**p<0.05;*p<0.1.

G The monetary cost of ideological misalignment

Using data on overnight stays and average tourist expenditure from Spain, Hinnosaar et al. (2021) estimates that improving Wikipedia leads to an average annual increase in revenue of 160,000 euros by municipality. Assuming this figure as a valid estimate of the effect of policy adoption on revenues for municipalities in our sample and neglecting general equilibrium effects, we present in this subsection a basic calculation for the monetary cost of ideological misalignment between the policymaker and the informing institution in the context of the policy recommended. We calculate the cost of ideological misalignment as the difference in the expected value of providing information from an ideologically aligned institution compared to providing information from an ideologically opposite institution.

Our estimations, as reported in Section 3.5, show that policy adoption among municipalities receiving the summary of evidence from an ideologically aligned institution is 4.81%. In comparison, policy adoption among municipalities receiving the summary of evidence from an ideologically opposite institution is 2.8%. The difference in policy adoption between these groups is 1.38 percentage points, corresponding to an increase in policy adoption of nearly 48%. The economic cost of ideological misalignment between the informing institution and the policymaker in the policy recommendation is then calculated as the difference in the probability of policy adoption multiplied by the revenues generated by the intervention, as calculated by Hinnosaar et al. (2021), amounting to 2,192 euros per municipality per year.

H Effect of the treatments on tourism

Using an RCT design, Hinnosaar et al. (2021) found that improvements in Wikipedia increased overnight stays in the municipality by 9%, although they did not find any significant effect on the number of tourists. Then, using information from average tourist expenditure per day, the authors estimate that improving the municipality's Wikipedia page increases tourist revenues by 160,000 euros per year on average. While we do not have information on the number of overnight stays to proxy for tourist expenditure, La Caixa Bank provides us with monthly-level information on card payments from tourists at the zip code level.[56] La Caixa is the third-largest retail bank in Spain, and it manages a database of tourist card payments on La Caixa terminals. Tourist payments are defined as those conducted by individuals in places located 50 km away from their residences.

We use this information to investigate how changes in the Wikipedia pages of municipalities in our sample impact tourist expenditure conducted on La Caixa terminals. The *cleanest* identification strategy would be using an assignment to a group that is informed by a think tank/newspaper with the same ideology as an instrumental variable for Wikipedia changes. The main challenge in our case is that while the effect of being assigned to treatment groups that receive the information from an institution with the same ideology on changing Wikipedia is statistically significant at 5%, the t-statistics (t=2.1) is well below the conventional threshold required for instrumental variables (t>3.33). Thus, the results of this approach are likely affected by the problem of *weak* instruments.

Thus, rather than using assignment in the experiment as instrumental variables as a source of exogenous variation, we exploit the exact timing of the Wikipedia change in either treatment or control groups to estimate the effect of Wikipedia changes using two different designs. First, we use the doubly robust difference-in-differences estimator for staggered treatments developed by Callaway and Sant'Anna (2021). Second, we use the synthetic difference-in-differences estimator developed by (Arkhangelsky et al., 2021). The results of these analyses are reported in Table H.III, showing no effects of Wikipedia changes on tourist expenditure on La Caixa terminals for all and national tourists.

[56] The database on overnights used by Hinnosaar et al. (2021) only includes information on overnights for a sample of 133 large municipalities. These are very different from the vast majority of the municipalities that we use in our analytical sample.

Table H.III: The effect of changes in Wikipedia on tourist expenditure (in thousand euros)

	(1) All tourist expenditure	(2) All tourist expenditure	(3) National tourist expenditure	(4) National tourist expenditure
Wikipedia change	37.4	-5.8	0.9	-3.4
	(35.8)	(36.3)	(27.3)	(42.4)
Estimation methods	Callaway & Sant'Anna	Synthetic control	Callaway & Sant'Anna	Synthetic control
Mean dep var	507.4	307.1	196.1	139.8
N	960	53,832	960	53,832

Note: The dependent variable in columns 1 and 2 is the monthly expenditure in thousands of euros of all tourists in the municipality. The dependent variable in columns 3 and 4 is the monthly expenditure in thousands of euros of national tourists in the municipality. Estimations in Columns 1 and 3 use the doubly robust difference-in-differences estimator with staggered treatment adoption developed in Callaway and Sant'Anna (2021). Estimations in Columns 2 and 4 use the synthetic difference-in-differences method developed in Arkhangelsky et al. (2021). ***p<0.01;**p<0.05;*p<0.1

Our results on tourism do not dismiss the robust results reported in Hinnosaar et al. (2021). Thus, we do not believe that the intervention recommended is ineffective. First, we do not observe all tourist expenditures, but only those conducted in retails with a La Caixa terminal. Furthermore, it excludes payments conducted at origin such as on-line hotel bookings or travel tickets. Thus, the extent to which the latter measure represents a credible proxy for total tourist expenditure might depend on the specific municipality. The existence of measurement error biases the results downwards, which may lead to the observed lack of effects. Second, the editorial team's improvements in the Wikipedia pages of treated municipalities in the experiment conducted by Hinnosaar et al. (2021) were substantial. In contrast, the changes induced by our policy briefs were, in most cases, minor, and this could have decreased the effect of these changes on tourism. Finally, we only have information for a maximum of 7 months, and most of the changes were conducted after the high tourist season. On the other hand, Hinnosaar et al. (2021) explores the effect of Wikipedia changes on tourist expenditure during the high tourist season one year after the changes were conducted.

For these reasons, the back-of-the-envelope calculations reported in Section 3.5 on the cost of political ideological bias are calculated using the effects estimated in Hinnosaar et al. (2021).

I Deviations from the Pre-Analysis Plan

A pre-analysis plan for the main experiment was registered in the American Economic Association registry for randomized controlled trials (AEARCTR-0008967). This appendix reports and discusses

the deviations from the pre-analysis plan.

I.1 Post-treatment survey

In April-May 2023, after the end of the study period, we conducted an online survey that targeted all mayors and appointed local policymakers of municipalities to which we could assign an ideology. In total, we invited 17,044 policymakers from 7,576 municipalities. 1,600 policymakers from 1,196 municipalities responded to the survey. The endline survey was aimed at understanding attitudes toward research evidence from policymakers. Furthermore, we include in the end-line survey an online experiment to test how policymakers update their beliefs in response to evidence endorsed by organizations with aligned or opposite ideologies. This online experiment was introduced to investigate whether belief updates could drive the effects of ideological endorsement on policy adoption. A description of the survey sampling strategy, questionnaire, and results is provided in Appendix A. The survey includes informed consent.

The pre-analysis plan briefly describes the intention to conduct a post-treatment survey:

"In October/November 2022, after the end of the touristic season in Spain, we will run a second online survey targeting all mayors in Spain. This survey will be sent by the Federación Española de Municipios y Provincias (FEMP), a public organization that holds the contact details of all Spanish municipal governments and communicates with them regularly to coordinate joint actions of Spanish municipalities. The survey includes an informed consent approved by the IRB. The names of the researchers involved in the study and the organizations used later to send the policy brief or news will not be included in the survey questionnaire or the email.

In this post-treatment survey, we will collect more detailed information on the socioeconomic and demographic characteristics of the mayor, the mayor's attitudes towards evidence-based policy making, perceptions about public policies to increase tourism, and the tourist activity during the tourism season of 2022. The post-treatment survey will be used for descriptive purposes. If the survey response rate is sufficiently high, we will use the information gathered to explore the effects of the different experimental treatments on the mayors' beliefs about how to increase tourism in their municipality."

We deviate from the initial plans in two ways. First, the survey was conducted in April-May 2023, after the end of the study period. The goal was to prevent the survey - which asks policymakers about attitudes towards evidence-based policy making- from influencing policy adoption. Second, the survey was not sent by FEMP but rather by ESADE using Qualtrics software. We do not use FEMP as initially planned because the pre-treatment survey sent by ESADE has very low response rates. Third, we did not target all mayors, but all appointed local policymakers that share ideology

with the mayor in municipalities for which we could identify the governing party's ideology. To expand the sample of the online experiment, we extend the survey to local policymakers rather than restrict it to mayors. Finally, the survey includes an online experiment to test how policymakers update their beliefs in response to evidence endorsed by organizations with aligned or opposite ideologies. A detailed description of the online experiment is provided in Appendix A and Section 3.6.

I.2 Initial survey

We conducted the survey as described in the pre-analysis plan. However, the response rate was very low: less than 100 municipalities responded to the survey, including 80 municipalities in the analytical sample. The goal of the survey was to get a better understanding of Spanish mayors, particularly those included in the sample. Given the low response rates, the initial survey cannot be used for these purposes, and therefore, the results of this survey are not reported in the paper. The database is, however, available upon request.

I.3 Stratification

The pre-analysis plan reports that the randomization was stratified by (a) political party (Partido Popular, PSOE, other right-wing parties, and other left-wing parties), (b) population in the municipality (tertile in the distribution of population of the Spanish municipalities in the sample), (c) importance of tourism in the municipality (tertile in the distribution of population of the Spanish municipalities in the sample), and (d) the extension of Wikipedia page (tertile in the distribution of the variable number of words in the Wikipedia's page of the municipality for the Spanish municipalities in the sample).

However, the stratification was conducted based only on the first three because the correlation between the extension of the Wikipedia page and the population in the municipality was very high (0.4185), and using both variables created almost *empty cells* in the randomization process. The pre-analysis plan wrongly described the randomization process as based on four rather than three characteristics.

I.4 Study period

Initially, the study period or the period in which we were going to study changes in Wikipedia was from May 25th, 2022, to September 30th, 2022. That is the summer season in Spain. Some policymakers advised us that in many Spanish municipalities (particularly in the Canary Islands), the high season for tourism starts in the autumn. Furthermore, many municipalities in Spain (par-

ticularly rural areas) experience high demand for tourism on key dates after the summer: October 12th, November 1st, or December 6th and 8th. Thus, we decided to extend the study period from May 25th, 2022, to December 31st, 2022.

I.5 Other changes

The final sample includes a total of 5678 municipalities rather than 5677:

The pre-analysis states that we will investigate the heterogeneous effects of the treatment by whether the mayor belongs to either PP-PSOE (the main right- and left-wing political parties in Spain) or to a different political party. This heterogeneity analysis is reported in the Appendix. Additionally, we have also explored the heterogeneity of the effects by whether the mayor belongs to a right- or a left-wing party, by whether the mayor belongs to a pro-independence regionalist party, by the population of the municipality (as a proxy for capacity), by the initial length of the municipality page in Wikipedia, and by whether the municipality is included in the experiment conducted by Hinnosaar et al. (2021). While not initially included in the pre-analysis plan, we believe these dimensions of heterogeneity are crucial to better understanding which municipality characteristics help to explain the effect of ideological alignment on policy adoption.

During the experiment, a total of 18 reminder emails were sent. Following the pre-analysis plan, the first 9 reminders were sent with the support of the marketing enterprise M-DIRECTOR, which allows tracking the clicks through the links in the emails. Unexpectedly, a non-negligible share of emails arrived in the spam folder. To maximize the reach of our emails, the last 6 reminders were sent through an Outlook account (keeping the same sender). The main drawback is that we cannot track clicks through the links in the last emails. These outcomes are therefore analyzed using only the reminder rounds sent with M-DIRECTOR.

Conclusion

The two decades that preceded the Great Recession were marked by an era of great optimism in economic and political progress. Rapid developments in technology offered a new world of opportunities for economic development and productivity growth. The internet was expected to lead to unstoppable democratic expansion. The dominating intellectual paradigm in economics and policy studies showed an almost blind trust in the possibilities of human rationality. In policy studies, new sophisticated econometric methods and the increasing availability of data promised an almost mechanical guarantee for social progress.

Yet, after the Global Financial Crisis, the foundations of the old liberal economic order started to shake and new vulnerabilities appeared. Economic instability and democratic recess became the norm in many advanced and developing economies.

Technology seemed to be related to many of the unfolding challenges. New internet platforms and social networks were associated to rising disinformation and political polarisation. At the same time, globalisation and technological disruptions were creating increasingly visible divisions between winners and losers. The rise of information technologies had allowed for a greater integration of supply chains in global markets and substantial gains in efficiency. This contributed to propel China's and other developing countries economic growth. But Chinese import competition also had had a severe negative impact in many regions in advanced economies, especially in labor-intensive industrial areas (Autor et al., 2013).

By facilitating the offshoring of many core tasks previously performed by middle-skill workers, this process led to profound changes in the value of certain skills (Autor et al., 2003). Researchers documented a process of job-polarization: high skilled workers benefited from globalisation and automation. Computers helped enhanced their productivity, liberating time to dedicate to more value-added tasks, while many routine and manual jobs were displaced, exacerbating inequalities. Acemoglu and Restrepo (2022) estimate that between 50% and 70% of changes in the US wage structure over the last four decades are accounted for by the relative wage declines of worker groups specialised in routine tasks in industries experiencing rapid automation.

The rise of populist movements can be partly attributed to the economic insecurities and social discontent stemming from these labour-market disruptions. Although the electoral implications of these changes are not the focus of my research, adverse economic shocks, such as trade or technological shocks, have been shown to lead to ideological realignments and changes in political and economic preferences. Autor et al. (2020), for instance, show that electoral districts in the US more exposed to trade exhibited more ideological polarisation. Rodrik and Tella (2020) find that support for government intervention rises sharply in response to shocks and is heavily biased towards trade protection. Becker et al. (2017), analysing Brexit, find that within cities, areas with more deprivation in terms of education, income and employment were more likely to vote "Leave".

Over my time in politics, as a spokesperson on economic affairs at the Spanish Parliament for a centrist party, I experienced the difficulties of defending an evidence-based policy agenda in a time of rising political discontent. While our policies were recognised to be solid and socially sensible, voters did not seem to care much about policy programs. After some time in politics, I realised that my view of the policy process was very naif: when dealing with politics, individuals seemed to be driven by different motives than the dispassionate and predicable *homo economicus* I had studied in economic schools. Furthermore, I became convinced that political systems were unprepared to respond to the massive democratic and economic shocks that accelerating technological changes were bringing about.

New innovations in machine learning and artificial intelligence technologies were expected to cause a profound impact on the economy.(Frey and Osborne, 2017) predicted that automation was going to displace 47 percent of jobs in the US over two decades. The sudden rise of GenAI platforms was going to inevitably accelerate the trend. Researchers predicted that because of their new capabilities to learn and replicate human and "tacit knowledge" ((Brynjolfsson et al., 2023), GenAI systems would extend their impact to high-skilled professions that had previously been shielded from automation.

As an economist, I felt the need to have a deeper understanding of the effects of technological shocks and their impact on inequality. But also, to explore new and innovative policies incorporating a more sophisticated understanding of the human brain when dealing with policy. The seemingly unstoppable rise of populist and extremist parties highlighted the limitations of traditional technocratic approaches to policy. Well-meaning economists, I thought, needed to reassess new behavioural ways to approach policy-making and get a better understanding of the role of persuasion in politics if they wanted to succeed in changing reality.

In sum, I decided to start the PhD journey to address a fundamental intellectual *décalage* between my experience in politics and the depoliticized view of technology and the human brain I had learned in my formative years as an economist.

In my first chapter, I contribute to understanding key dimensions of the interaction between GenAI systems and inequality, testing a novel set of skills in a unique setting. My findings complement previous research, indicating that the effects of GenAI are likely to vary substantially depending on the context and the skills that are demanded. In tasks requiring higher-order skills, such as persuasion, negotiation or social-perceptiveness, I show that GenAI has stronger complementarities with top performers than with low-performers.

Given that higher order skills are necessary for managerial and high-earning jobs of all kinds, these results - if replicated - could have deep potential implications for our understanding of the social effects of these technologies. Research in this area should continue exploring the productivity impact of these technologies in different contexts and in interaction with different tasks.

My second chapter explores a novel pedagogic design to analyse how online learning can be effective to reduce inequalities. Previous research had demonstrated that in-person individual tutoring or tutoring in small groups can have very positive learning effects. However, very little evidence existed regarding tutoring in 100% online environments. The established view was that online learning had several limitations that made it ineffective. Our paper shows a new very cost-effective way of improving learning outcomes with children from very vulnerable environments. We hypothesise that the positive student dynamics of the 2-students-to-1-mentor design might help combat common attrition problems in online learning. Future research should focus on digging deeper to understand the different mechanisms behind effective online learning as well as the effectiveness of these designs at scale given existing constraints, such as the availability of good quality teachers. Pilots at a larger scale could allow for a more precise evaluation of the impact of teacher characteristics that could make these programs more effective. In all cases, these results already provide robust evidence of a cost-effective way to combat educational inequality, as online settings reduce significantly the cost of providing tutoring, while getting to geographically left-behind areas that tend to have poorer access to good quality teachers.

Finally, my third paper explores a key question for scientific communication and evidence-based policy diffusion. The research question – i.e. how the ideology of the messenger and the format in which the evidence is presented affect the implementation of a politically neutral policy - is of especial relevance given the rising concerns related to the diffusion of scientific information on climate change or the Covid-19 pandemic. Previous research had found that individuals as well as politicians tend to behave as motivated thinkers when dealing with political issues: they tend to give more weight to information that conforms with their values, beliefs or trusted networks. Ideology, such as religion, activate "identity-protective" ways of learning. This limits individuals willingness or capacity to incorporate new evidence, thus helping to explain the constraints for evidence-based

policy expansion.

With our research we contribute to understanding how political alignment between different knowledge brokers, such as think tanks or media outlets, and politicians matters for policy adoption. A key feature that makes our contribution unique is that we run a nation-wide RCT with real policymakers and we can track the policy until its final implementation. This is a significant addition to the literature on improving the diffusion of scientific findings to improve policy implementation of evidence-based policies.

Future research should focus on identifying strategies to overcome political, cultural, and behavioural barriers to evidence-based policy diffusion and adoption. Specifically, research should test how leveraging trusted networks to enhance the credibility and acceptance of scientific findings could help improve belief updating and policy implementation. Moreover, future research should contribute to deepen our understanding of the impact of differential characteristics associated with the messenger of the policy. Integrating more behavioural insights and a deeper understanding of (politically) motivated reasoning dynamics when dealing with policy could help improve the effectiveness of policy communication to the benefit of all.

These three papers provide a novel perspective on the relationship between technology and social policy. They share a common methodology. And they also point to a common idea: the effects of technology on inequality are not necessarily negative: they depend on the institutional and political contexts in which the technology is deployed and the policies designed to compensate for their potentially negative effects. In an age of permanent disruption, an ongoing challenge for offering effective answers will require a better understanding of how the human brain behaves when dealing with policy. I hope this work has helped shedding some light on those questions.

Bibliography

Acemoglu, Daron and Pascual Restrepo, "Tasks, Automation, and the Rise in U.S. Wage Inequality," *Econometrica*, 2022, *90* (5), 1973–2016.

Autor, David, David Dorn, Gordon Hanson, and Kaveh Majlesi, "Importing Political Polarization? The Electoral Consequences of Rising Trade Exposure," *American Economic Review*, October 2020, *110* (10), 3139–83.

Autor, David H., David Dorn, and Gordon H. Hanson, "The China Syndrome: Local Labor Market Effects of Import Competition in the United States," *American Economic Review*, October 2013, *103* (6), 2121–68.

_ , 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.

Becker, Sascha O, Thiemo Fetzer, and Dennis Novy, "Who voted for Brexit? A comprehensive district-level analysis," *Economic Policy*, 10 2017, *32* (92), 601–650.

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

Frey, Carl Benedikt and Michael A. Osborne, "The future of employment: How susceptible are jobs to computerisation?," *Technological Forecasting and Social Change*, 2017, *114*, 254–280.

Rodrik, Dani and Rafael Di Tella, "Labour Market Shocks and the Demand for Trade Protection: Evidence from Online Surveys," *Economic Journal*, 2020, *130*, 1008–1030.