Centre for Economic Performance

Discussion Paper

ISSN 2042-2695

No. 1908
March 2023

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

Lucas Gortazar
Claudia Hupkau
Antonio Roldán

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

UKRI Economic and Social Research Council

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 by qualified math teachers in groups of two students during after-school hours. The intervention significantly increased standardized test scores (+0.26 SD) and end-of-year math 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.

Key words: online tutoring, mentoring, RCT, mathematics, child outcomes
JEL: C93; I24; I28; H75

This paper was produced as part of the Centre's Education and Skills Programme. The Centre for Economic Performance is financed by the Economic and Social Research Council.

We thank seminar participants at LSE-CEP, UC3M, UPO and UAM. 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 and Mónica Martinez-Bravo, Pedro Rey-Biel and Jenifer Ruiz-Valenzuela for their comments. We are grateful to Empieza por Educar for the implementation of the program, including Miriam Arriola, Zaida Pillado, Beatriz Morilla, Lara Crespo, Oscar Ferri, Carla Carmona, Paula Royo, Miguel Costa, Carles Lopez, Luis Oliver, and Justo Quintanar. We are also grateful to Iria Mata, Dirk Rosquillas, Miguel Ujeda, and Gonzalo Romero for their support in various layers of implementation. We are grateful to 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. AER RCT Registry ID: AEARCTR - 0007138. The project has obtained IRB approval from University Ramon Llull.
Lucas Gortazar, ESADE Center for Economic Policy and World Bank. Claudia Hupkau, CUNEF Universidad and Centre for Economic Performance, London School of Economics. Antonio Roldán, ESADE Center for Economic Policy and London School of Economics.

Published by
Centre for Economic Performance
London School of Economics and Political Science
Houghton Street
London WC2A 2AE

1 Introduction

Intensive, in-person tutoring in one-to-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.[1]

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 the pandemic have made online tutoring much more accessible to families from all backgrounds. However, very little evidence exists as regards to the effectiveness of online tutoring. 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 time cost and commuting times for both tutors and students (Kraft et al., 2022).

In this paper, we study the effectiveness of an intensive, eight-week tutoring program on academic and socio-emotional outcomes of secondary school children in Spain. It offered free, 100-percent remote (online) after-school tutoring to pupils aged 12 to 15 from very disadvantaged backgrounds. The program, called *Menπores*, has four key features. First, the whole organization of the program and the tutoring sessions were implemented online. Second, tutoring was carried out by 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).

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 in participating schools to identify students most in need for support in math and disseminate the program among them and

[1] 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 either face-to-face or online tuition.

their families. Among all students who signed up, we randomly assigned slightly more than half to the program. Randomization was blocked by classes 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 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. At the end of the program, we administered an endline survey of families to collect information on academic results at the end of the school year. We also collected very rich real-time data throughout the duration of the program capturing participation, connection time, and quality of the connection.

We find a positive and significant effect of program assignment on end-of-year math grades (+0.48 SD), equivalent to about six months of learning. Further, we find a significant increase of about 32 percent with respect to the control group mean in the likelihood of passing the subject. Using our standardized math test, we find an increase in the test score by 26.2 SD (equivalent to about three months of learning), which is significant at the 10 percent level. To put these numbers into context, Nickow et al. (2020)'s meta analysis of the effectiveness of in-person tutoring finds an overall pooled effect size of 0.37 SD. Further, we find a large and significant effect on grade retention: the program decreased the likelihood of repeating the school year by 9.4 percentage points, equivalent to a 78 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.

In terms of non-cognitive outcomes, we find that the program raised students' aspirations: Students in the treatment group were 13.6 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 33.2 percent increase compared to the control group mean. Students assigned to treatment were also 11.6 percentage points more likely to state that they exerted high effort always or most of the time at school, which corresponds to an increase by 21.5 percent when compared to the control group mean. We do not find an impact on student's self-perceived math competencies or the likelihood of stating they like mathematics. We neither find an effect on locus of control, grit, or on overall well-being.

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.

Our program departs from these studies in three fundamental ways. First, they were delivered by volunteer university students, while *Menπores* used paid-for, qualified secondary school teachers. Second, our tutoring was implemented in groups of two students, instead of one-to-one. Third, and more importantly, these program were implemented in exceptional circumstances. For the case of Carlana and La Ferrara (2021), during the harshest lockdown period in Italy from April to June 2020 (when all kids were at home and schools where closed).[2] In the case of Kraft et al. (2022), in early 2021 in the US, when schooling there was still highly disrupted. Ours, 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 in-class, regular teaching.

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.

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-

[2]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).

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 learning outcomes (in math), translating into a 0.087 SD increase per €100 spent. This compares favorably with summer schools (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 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 connection, 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 teachers.

In terms of cost, programs with volunteers are always going to be cheaper than programs with paid professionals. However, at large scale, volunteers are likely to face more practical and political economy limitations than programs with paid professional teachers. First, availability of large amounts of volunteers is likely to be a huge 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 teacher turnover. What we show here, however, is that our innovative two-to-one online design offers additional cost savings in relation to in-person programs and one-to-one online programs, while achieving very similar results.

It remains to be further explored whether the two-to-one online design may have other advantages, such as creating positive motivational dynamics and peer pressure not to abandon the program. Our program was completed by 96.6 percent of pupils of the treatment group, suggesting that indeed there may be positive peer dynamics at play. If this were true, it would help to address some of the key shortcomings found in the literature in online education, such as a lack of perseverance and motivation (Escueta et al., 2020). More research on this will be needed to better understand group dynamics in remote learning.

The rest of the paper is organized as follows. Section 2 describes the context of the intervention. In Section 3, we present the study design and in Section 4 we describe the data. The empirical strategy is presented in Section 5, and results and robustness checks are shown in Section 6. Section 7 concludes.

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, the latest date for which data is available (and closest to our intervention), only 0.5 percent of classes in Spain were closed (quarantined).

In primary school and the first two grades of lower secondary school (Grades 1 to 8, ages six 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. Some schools concentrated hours of instruction between 9 a.m. and 2 p.m., rather than the usual model of classes happening between 9 a.m. and 4 p.m., with a lunch break of 1.5 hours in between, in order to avoid lunch happening at school (a potential source of virus spread). While this did not lead to a reduction in instruction time, it meant that some of the students in our study potentially had one to two hours more free time in the afternoons compared to the pre-pandemic scenario.[3]

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 of students who enrolled in our program did not have the technology at home to attend online tutoring sessions.

3 Study Design

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

[3] In Catalonia, 43 percent of schools ended up operating in morning shifts while in Madrid 72.9 percent of schools did so. This meant an increase of 3 percentage points in Catalonia and an increase of 19.5 percentage points compared to pre-pandemic times in Madrid (Plataforma de Infancia, 2021).

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, 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 by qualified math teachers in groups of two students per tutor. We decided to hire 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 treatment arm delivered by university student tutors as volunteers 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.[4] 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 to guarantee that students knew each other and would find it easier to connect and accommodate. We decided to go for a two-to-one student-tutor ratio for three reasons. First, the pedagogic team in charge of implementation suggested that being in a group with another child

[4]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.

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-to-one tutoring is nearly as effective as one-to-one tutoring. Third, this design meant that we could deliver the tutoring to twice as many students as in a one-to-one setting, given our budget.[5]

A key element of the program was its online nature. The fact that face-to-face interactions were severely constrained by social distancing rules 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.

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. 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.[6]

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.

[5]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.

[6]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.

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 and support, 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 3.4.

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 participants from their pool of students and disseminate the program among them.

For recruitment of participant schools, we leveraged ExE's large school network 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 lagging behind in mathematics), and the fact that the program was to be evaluated scientifically through a randomized controlled trial. There were no strict eligibility rules. Instead, we relied on the knowledge of teachers and principals to identify students most in need for math tutoring.[7]

In the second step, parents of children identified by the school as in need were directed to an online registration form, which 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. In the registration process, parents were asked to give consent for their children's participation and the usage of data for research purposes. In the registration form we asked for detailed household and student characteristics and whether the student that was being registered had access to a tablet or other device to participate in the online sessions.

[7]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.

3.4 Selection and training of tutors

ExE designed and implemented the selection and training for tutors based on their long-standing 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. 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 46 tutors, which was the number required in order to provide tutoring to approximately 200 students.[8]

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.

3.5 Timeline

Figure 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, meaning in

[8]As mentioned previously, we had initially planned a third treatment arm with volunteer mentors. Although we advertised the program at four big 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 process. This allowed us to fulfill the initial commitment to schools to provide tutoring to around 200 students. Results when including volunteer tutors are basically unchanged compared to the main results presented in this paper. We discuss and present results including volunteer tutors in Section 6.5.

all classes where at least one student had registered 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.

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 usually 6 students. Randomization of tutors was stratified by geographic area, where those tutors who indicated they spoke Catalan were assigned to students based in Catalonia, and those who did not speak Catalan to those who were based in the region of Madrid.

Because of the small number of volunteer tutors we were able to recruit, we exclude these students from our main results that are provided in the following sections. All results are robust to the inclusion of students tutored by volunteers, and are discussed in Section 6.5.

3.7 Implementation

Students and tutors were able to organize their own schedule and agree on weekly meeting times.[9] 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.[10] 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 shows the distribution of total minutes and total number of tutoring sessions attended by students. Only seven students (3.4 percent of the students assigned to the treatment group) dropped out of the program before it began. Among those that did start the program, the median number of minutes of tutoring received was 960, representing 80 percent of the maximum envisaged number of minutes (1200). The median number of sessions attended was 20, corresponding to 83 percent of the maximum envisaged number of sessions (24).

[9]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.

[10]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.

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.

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.[11]

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.[12] 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.

[11] As noted in Section 3.7, 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.

[12] 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 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.

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, math and English. 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.

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 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 homework, and interest in language and math. To measure aspirations, we ask about the plans students have for after completion of 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 obtained by their children at the end of the year and whether the child will have to repeat the school year.[13] We also asked about whether their children had received any other remedial education support program (besides Menπtores in case of the treatment group).

4.3 Sample and balancing

For our analysis, we will work with different samples depending on the outcome measure. Of the overall 356 students that participated in the program who were tutored by professional tutors, we have 348 (98 percent) for whom we observe either the baseline or the endline math test score. The endline socio-emotional survey was completed at least partially by 353 students. Since students do not always answer all questions of the survey, the exact sample sizes vary depending on the outcome. We have end-of-year math grades and information on grade retention (measured only at endline) for 220 students (62 percent) whose parents responded to the parent endline survey.

Balancing between treatment and control group household characteristics is shown in Table 1. There are no significant differences between the treatment and control group. When looking at child level characteristics, for which we show balancing in Table 2, we find

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

that treated students are slightly less likely to be in public schools (versus publicly funded private schools, or *colegios concertados*), and slightly more likely to be born in Spain and to receive school meal support. We also find that children in the treatment group have slightly worse baseline performance in math. They are 6.24 percentage points more likely to have failed math in the first term of the academic year, and have a slightly worse baseline math grade. They also perform slightly worse on our baseline test, achieving on average 24.7 percent correct answers compared to 26 in the control group. While none of these differences are statistically significant, controlling for baseline performance in our empirical specification will be important to improve the precision of our estimates.

In Tables B1 and B2 we show the same balancing tables, but restricting the sample to children whose parents responded to the endline survey (representing 62 percent of participating students) where we solicit end-of-year math grades and whether the school year had to be repeated. Parents of the treatment group who responded to the endline phone survey are lower educated and are more likely to be of Spanish origin than those in the control group. However, none of these differences are statistically significantly different from zero. When it comes to the characteristics of children of parents who respond to the endline survey (see Table B2), we see that those in the treatment group are more likely to be recipients of school meal stipends and were more likely to have failed math in the first term. Overall, selection looks very similar when comparing the whole sample and the sub-sample of students whose parents responded to the endline survey. The only exception is our baseline test, where treatment students in the latter group performed slightly better than control students. Again, none of these differences are statistically significant.

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:

$$Y_{ibt} = \alpha_b + \beta Treat_{it} + \gamma Post_t + \delta Treat_i \times Post_t + \lambda X_i + \epsilon_{ib} \tag{1}$$

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 student and parent-level control variables that include student age, grade, gender, region, a dummy indicating school meal eligibility, baseline math grade, a set of dummy variables indicating the frequency of online lessons during school closures between April and June 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. 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.[14] We also report Romano-Wolf p-values for step-down adjusted standard errors to account for multiple hypothesis testing.

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

$$Y_{ib} = \alpha_b + \delta Treat_i + \lambda X_i + \epsilon_{ib} \tag{2}$$

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. We also report Romano-Wolf stepdown adjusted p-values to account for multiple hypothesis testing.

6 Results

In this section we present our main results and perform a set of robustness checks.

[14]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 7-10. With the exception of grade repetition, all results remain significant when clustering at block level is employed.

6.1 Academic outcomes

Table 3 summarizes the results for the impact of the intervention on academic outcomes. The dependent variable in Column 1 is the score on the standardized math test, standardized by grade level.[15] The difference-in-difference estimate indicates that treatment students improved their score by 0.262 SD more than control students, which is significant at the 10 percent level.[16]

Columns 2 to 4 show treatment effect estimates for teacher-assessed outcomes reported by parents. 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 0.48 SD (Column 3) compared to the control group. Treatment students are also 21.6 percentage points (Column 4) more likely to have passed the subject (math), corresponding to a 32 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 9.4 percentage points less likely to have to repeat the school year (Column 5), corresponding to a 78 percent drop in the repetition probability compared to the control group.

In Table 4 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 liked math. 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). Again, we do not find evidence that this happened (Columns 2 and 4).

In appendix Table B3, 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 up questionnaire was only 47 percent (168 out of 356), and results are mostly insignificant. However, the magnitude and direction of coefficient estimates 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 0.52 points higher

[15]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.

[16]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.

(+34 percent of an SD). Treatment students were also 11.1 percentage points less likely to have repeated the school year that started after the end of the program. We find a small (not significant) negative effect on parent's beliefs about their children's plans to attend the academic upper secondary route (doing the *bachillerato*), and a 6.2 percentage points increase in the likelihood of parents stating that they believe their children will go to college after school. 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.

6.2 Aspirations, perseverance, effort and motivation

We now look at the impact of the program on aspirations, perseverance and motivation. Column 1 of Table 5 shows that treatment group students were 13.6 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 33.2 percent higher probability than the control group.

We do not find an increase in the likelihood of stating that students plan to go on to 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 Column 3 of Table 5, 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 the duration of the program was too short to be able to change these personality traits.

Column 4 shows the impact of program assignment on self-perceived effort at school. Students in the treatment group were 11.6 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 21.5 percent higher probability than in the control group. Column 5 shows the effect of program assignment on our school motivation index. We do not find a significant impact on this outcome.

6.3 Well-being and socio-emotional outcomes

Table 6 shows the difference-in-difference 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). 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 0.2 SD increase (Column 2), which is however not significant at conventional levels.

Column 3 shows that the intervention had no significant impact on our measure of locus of control. If anything, students in the treatment group show a slightly lower level of internal locus of control, which is however not statistically significant. While this result is counter-intuitive, as 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, corresponding to a more external locus of control.

6.4 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 or lessons for a potential scale up of the program in the future.

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

Tutors were satisfied with the training received. When asked about the training and support received before and during the program, 89 percent of the tutors agreed or strongly agreed that the training plan was useful. Some 89 percent also agreed or strongly agreed that the training on the platform was adequate and sufficient. Some 68 percent agreed or strongly agreed that the webinars were useful and sufficient. Moreover, 86 percent of tutors consider the combination of online training and webinars to be adequate.

Tutors also answered several questions about their satisfaction with the project. Some 82 percent agreed or strongly agreed that participating in the project has helped them better understand the social reality facing students. Some 95.4 percent of tutors agreed or strongly agreed that the program has enriched them personally and professionally.

Mathematics teachers and headteachers of the participating schools rated the impact of the program positively. More than 70 percent of the teachers and 57 percent of the headteachers surveyed agreed or strongly agreed that the program has been useful for their pupils. Some 69 percent of the teachers believe that the program is a good support for their teaching. Finally, 71 percent think that the program should continue, which is also shared by 100 percent of the headteachers surveyed. More than 40 percent of surveyed mathematics teachers believe that the fact that pupils participated in the program helped them to work better, and another 42 percent believe 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), which might explain the lower proportion of positive responses. In the open-ended responses, several teachers and headteachers mentioned the need to start these programs before April and make it longer.

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.

6.5 Discussion of results and robustness checks

We now discuss several issues around the internal and external validity of our main results. The choice of specification or selective attrition might also cause bias in our estimates.

We therefore perform several robustness checks to see whether our results remain robust to different specifications, and taking into account selective attrition.

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 received any other tutoring or academic support program in math or other subjects during the period while *Menπores* was implemented. Indeed, as Figure 3 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. Overall, these findings suggest that our impact estimates could be interpreted as lower bounds.

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, Table B4 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.[17] 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

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

than the average population of students in Madrid. This should be kept in mind when interpreting our results.

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. While 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. All results presented so far exclude these students, but in Tables B5 to B8 we present our main results including all students that were part of the randomization. As can be seen, results are essentially identical to our main specifications, apart from slightly higher point estimates in our main effects when volunteer tutors are excluded. This is consistent with evidence showing that professional tutors are more effective than volunteers (Nickow et al., 2020), but we cannot draw any strong conclusions from this evidence given the extremely small sample sizes.

Alternative specifications The results shown thus far correspond to those using our preferred, full specification. In Columns 1-3 of Tables 7 to 10, we additionally show regressions using alternative specifications. Column 1 shows the most basic specifications, including only the treatment dummies and block fixed effects. In Column 2 we add demographic characteristics (age, gender, grade, autonomous community, whether eligible for school meal subsidy, baseline math grade, 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 speaks Spanish at home, household income, 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 7, 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 8 to 10 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.

Attrition Out of all children who initially signed up for the program, 93.3 percent took part in the baseline survey and math test, and 87.7 percent did so in the endline survey and math test.[18] To check whether there is selective attrition at endline, in Table 11 we run regressions where the outcome variable is a dummy equal to one if the child participated in the endline survey and math test, and zero otherwise. Column 1 shows that children in the treatment group were 4.5 percentage points more likely to respond, but the coefficient is not significant. In Column 2, we add control variables measured at baseline to see whether they correlate with attrition. Students in grade eight (versus those in grade seven) were more likely to respond at endline, but keeping grade constant, the association between age and responding at endline is negative. We also find a negative association between baseline math grade and the likelihood of responding at endline. Compared to the omitted category of having failed math in the first term of the academic year, students who passed math were less likely to respond at endline. Other child characteristics, such as gender or whether the child is receiving a school meal subsidy, show no significant association with responding at endline.

In terms of parental and household characteristics, there is no significant association between parental education and responding at endline, even though coefficients tend to be negative, meaning that compared to the omitted category (primary education or no education), children of more highly educated parents were less likely to respond at endline. We observe a positive correlation between household earnings and the probability of responding at endline, but none of these coefficients is significant at standard levels.

Among the group assigned to treatment, we observe very similar associations between the likelihood of responding and child and family characteristics as for the whole sample (Column 3). The same holds for control group students (Column 4). The coefficients on the interaction terms between the characteristic and treatment status shown in Column 5 however show that having performed "good" in math in the first term (relatively to the omitted category of having failed math) was negatively associated with responding at endline for treatment group children, while this association was positive among control group children, suggesting slightly negative selection into responding at endline in the

[18] Recall that tests were done during class hours and the main reason why children might have been absent is because they happened not to be in school that day, which explains the low overall rate of attrition.

treatment group and slightly positive selection into responding at endline in the control group.

Given these results, it is not clear in which direction selection bias for the child survey at endline might go, as on the one hand we see a negative association between parental education and response probability and on the other hand a positive association between household income and responding at endline. We conclude that it is likely that, if anything, we have a slight positive selection bias in terms of child responses, which would mean that our effect size estimates would be downwardly biased.

In Table 12 we perform the same exercise as above, but now looking at the the likelihood of parents responding to the endline phone survey, which we undertook in order to obtain teacher-assessed outcomes, such as the final grade in math, and whether the child had to repeat the school year. Parents of children in the treatment group were 13.6 percentage points more likely to respond than those in the control group (the overall response rate was only 62 percent, much lower than the response rate for the child endline test and survey at school). More highly educated parents were more likely to respond, but this was not true for parents of children in the treatment group, meaning we might have some negative selection in responses within this group. We see a clear negative association between household income and responding at endline for the parent survey. This indicates that overall, selection of parents into responding at endline might be slightly negative, especially for the treatment group. This would potentially bias our impact estimates based on parent reported outcomes, like the end-of-year math grades and passing the school year, downwards.

To quantify how much this might matter for our results, Column 4 in Table 7 shows the estimated impact of program assignment on academic outcomes using inverse-probability weights to account for selective attrition. 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 C are outcomes that rely on parental responses to the endline survey. If anything, effect size estimates here are higher for all three outcomes when using inverse-probability weights, which is in line with the potential (downward) bias in effect sizes predicted from our analysis of selective attrition.

In Tables 8 to 10 we perform the same exercise as above for non-academic outcomes. Column 4 of each table shows the effect size estimates using inverse-probability weights, and Column 3 reports our main results for comparison. Effect size estimates are not substantially different from one another. To conclude, our estimates are robust to accounting for attrition and if anything, are slightly downwardly biased in the case of academic outcomes reported by parents.

7 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 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, effort and aspirations of socially disadvantaged students. The 8-week program significantly increased standardized test scores (+ 0.26 SD), end of year math grades (+0.48 SD) and the probability of passing the subject (by about 22 percent with respect to the control group mean), while reducing the probability of repeating the school year (by about 78 percent with respect to the control group). In terms of non-academic outcomes, the intervention contributed to improve self-reported effort and aspirations. The students that were selected to receive tutoring were 14 percentage points more likely to state that they would go to the academic track after compulsory schooling, and 12 percentage points more likely than the control group to state that they exerted high effort at school always or most of the time.

Our results have immediate policy relevance 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 were 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-to-one online programs. Beyond this, global private tutoring is growing rapidly after the pandemic and is projected to grow at an annual rate of 8.95 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 completion rate of the program, despite the intensity (three sessions per week) and the fact that most participants came from highly disadvantaged backgrounds. When thinking about implementing such a program at bigger scale, these considerations are important, as data driven monitoring is a key advantage of online programs.

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). The number of tutors needed for a large scale program should be at least on a maximum ratio of one tutor per 10 students: afternoon time slots are limited to three to four hours per day, hence, the number of students a tutor can handle in total (in groups of two) could not be larger than eight or 10 (4 or 5 groups). Another limitation is that students may have been more willing to engage in online after-school tutoring program because they were coming out the of the pandemic, and this effect may fade away in the coming years.

For future research, it will be relevant to explore in more detail the mechanisms driving our results: tutor characteristics and interactions with students, type of training received, the number of students per tutor or type of devices used. Due to the relatively small sample size, we are not generally powered to detect small effect size differences between different groups. 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 teacher training 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 success in attendance and completion rates indicate that our two-to-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 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 math learning.

References

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.

Betthäuser, Bastian, Anders Bach-Mortensen, and Per Engzell, "A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemic-Did 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.

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.

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.

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.

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," 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.

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.

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.

Plataforma de Infancia, "Un curso escolar atravesado por la pandemia: Análisis de la inversión en educación para la adaptación al contexto COVID-19," 2021. Accessed on 21/12/2021 from https://coleseguro.es/resultados/#downloads.

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

Santiago, Paulo, "Teacher demand and supply: Improving teaching quality and addressing teacher shortages," 2002.

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.

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 1: Timeline of the Menπores program implementation

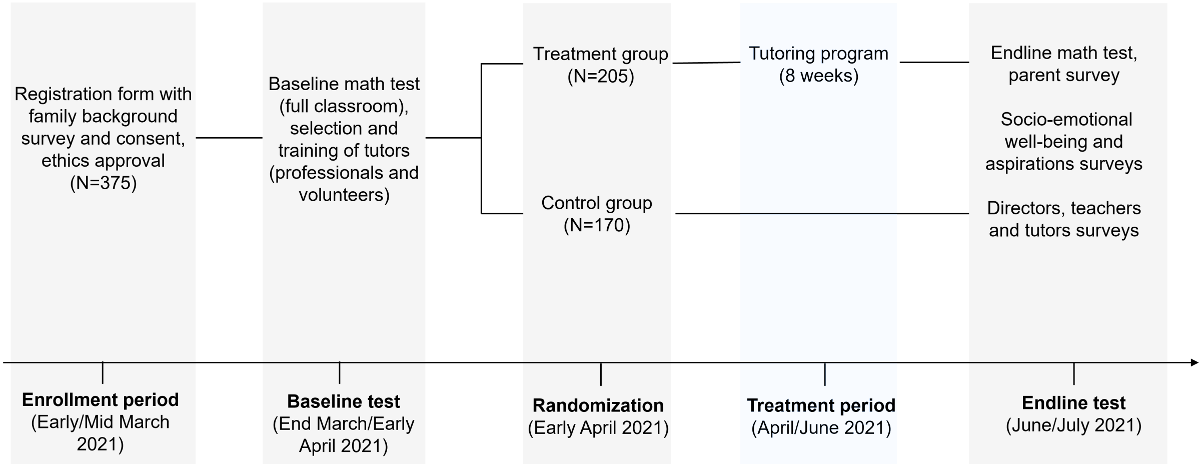
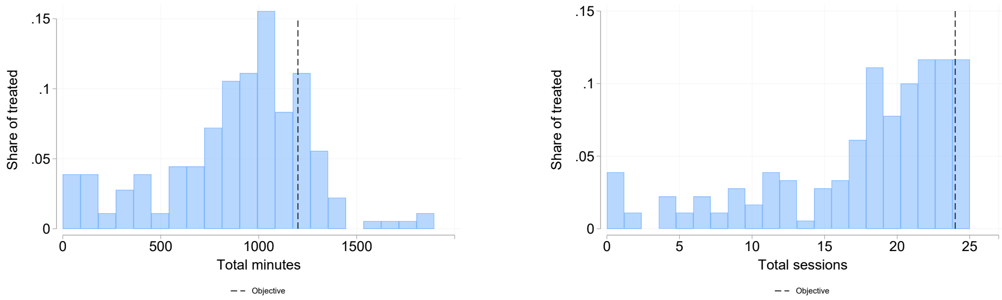


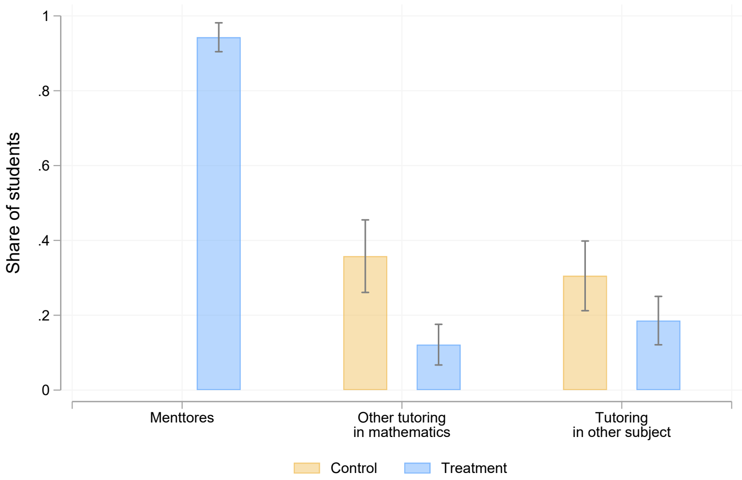
Figure 2: Distribution of total minutes and number of sessions of tutoring attended

(a) Minutes (b) Sessions



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 180 students out of 186 students in the treatment group tutored by professional tutors. This includes seven students who dropped out of the program before it started (zero minutes and sessions), and excludes six students for whom we could not match meeting data.

Figure 3: 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 1: Balancing table - parental characteristics

	(1) Treat	(2) Control	(3) Difference (2)-(1)	(4) p-value Col. (3)
Mother responded	77.96	76.47	1.49	0.74
Lives as couple	72.04	71.18	0.87	0.86
Spanish origin	55.38	51.76	3.61	0.50
Compulsory schooling or below	52.15	51.76	0.39	0.94
Income $<$ 1000 EUR	44.09	47.65	-3.56	0.50
Catalunya	26.34	31.18	-4.83	0.32
Madrid	73.66	68.82	4.83	0.32
HH size	4.08	4.11	-0.03	0.80
Nb. children age ≤ 18	2.04	1.88	0.16	0.16
Age of youngest child	9.76	9.96	-0.20	0.63
N	186	170		

The table shows summary statistics for the sample of students who registered to participate in Menπores. Column 4 reports the p-value of a t-test of the equality in means across the two groups.

Table 2: Balancing table - child characteristics

	(1) Treat	(2) Control	(3) Difference (1)-(2)	(4) p-value Col. (3)	(5) All Class
Age	13.02	13.04	-0.02	0.83	
Girl	47.85	48.82	-0.97	0.85	
Public school	22.04	25.29	-3.25	0.47	
Born in Spain	84.41	82.94	1.47	0.71	
Spanish spoken at home	86.02	81.18	4.85	0.22	
School meal stipend	9.14	7.65	1.49	0.61	
No tablet or computer	9.68	6.47	3.21	0.27	
Has access to internet	99.46	98.82	0.64	0.51	
Receiving academic support (at baseline)	19.89	18.24	1.66	0.69	
Grade 7	47.85	51.76	-3.92	0.46	
Grade 8	52.15	48.24	3.92	0.46	
Baseline child survey outcome					
Partial completion, 1+ maths question (%)	87.63	91.18	-3.54	0.28	
Completed survey fully (%)	47.31	48.24	-0.92	0.86	
Baseline performance					
Has failed math	52.29	46.04	6.24	0.29	
Has failed at least one subject	80.00	77.14	2.86	0.55	
Repeated grade at least once	24.73	26.47	-1.74	0.71	
Math grade (0-10) in first term	4.35	4.57	-0.22	0.35	
Test score (%)	24.55	25.89	-1.34	0.50	32.23
N	186	170			1,464

The table shows summary statistics for the sample of students who registered to participate in Menπores. Column 4 reports the p-value of a t-test of the equality in means across the two groups. Column 5 shows the average baseline test score among all students tested in participating schools, excluding those who registered to participate in Menπores.

Table 3: Impact on academic outcomes

	(1) Standardized test score	(2) Final math grade	(3) Passed math	(4) Repeated year
Post x Treat	0.262*			
	(0.148)			
RW p-value	[0.066]			
Treat		0.846***	0.216***	-0.094**
		(0.241)	(0.062)	(0.046)
RW p-value		[0.0248]	[0.0248]	[0.1186]
Constant	-0.377	9.361***	1.089**	0.374
	(0.597)	(2.451)	(0.482)	(0.350)
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. Romano-Wolf stepdown p-values for multiple hypothesis testing reported in brackets. The table shows the δ coefficients from Equation (1) in Column 1 - where we have a baseline measure of the dependent variable - and from Equation (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, 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 control for student age, grade, gender, region, a dummy indicating school meal eligibility, 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 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.

Table 4: Impact on self-perceived ability and affinity

	(1) Good at math	(2) Good at Spanish	(3) Likes math	(4) Likes Spanish
Post x Treat	-0.030	0.002		
	(0.048)	(0.067)		
RW p-value	[0.456]	[0.632]		
Treat			-0.052	-0.042
			(0.050)	(0.055)
RW p-value			[0.780]	[0.848]
Constant	-0.215	0.184	0.283	0.280
	(0.333)	(0.397)	(0.573)	(0.404)
Mean dep. var.	0.25	0.55	0.36	0.41
SD dep. var.	0.44	0.50	0.48	0.49
R^2	0.38	0.26	0.45	0.39
Obs.	628	629	351	353
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. Romano-Wolf stepdown p-values for multiple hypothesis testing reported in brackets. The table shows the δ coefficients from Equation (1) in Column 1 - where we have a baseline measure of the dependent variable - and from Equation (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 3.

Table 5: Impact on aspirations and motivation

	(1) Bachi llerato	(2) College	(3) Grit	(4) High effort	(5) Motivation school
Treat	0.136**	0.034	0.091	0.116**	0.011
	(0.059)	(0.050)	(0.063)	(0.054)	(0.019)
RW p-values	[0.202]	[0.848]	[0.590]	[0.236]	[0.848]
Constant	0.401	1.375***	2.485***	0.072	0.759***
	(0.530)	(0.325)	(0.511)	(0.404)	(0.192)
Mean dep. var.	0.41	0.81	3.04	0.54	0.75
SD dep. var.	0.49	0.39	0.50	0.50	0.16
R^2	0.46	0.43	0.40	0.47	0.43
Obs.	312	301	311	351	303

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. Heteroskedasticity-robust SEs in parenthesis. Romano-Wolf stepdown p-values for multiple hypothesis testing reported in brackets. The table shows the δ coefficients from Equation (2). All regressions include block FEs and the same controls as reported in the notes to Table 3.

Table 6: Impact on socio-emotional outcomes

	(1) Wellbeing index	(2) School satisfaction	(3) Locus of control
Post x Treat	0.037	0.265	-0.058
	(0.124)	(0.167)	(0.037)
RW p-values	[0.7168]	[0.1649]	[0.1649]
Constant	5.992***	4.624***	0.324*
	(0.758)	(1.078)	(0.188)
Mean dep. var.	6.27	5.47	0.61
SD dep. var.	1.62	1.36	0.31
R^2	0.64	0.30	0.29
Obs.	646	632	641
Unique ind.	348	348	348

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at student level in parenthesis. Romano-Wolf stepdown p-values for multiple hypothesis testing reported in brackets. The table shows the δ coefficients from Equation (1). All regressions include block FEs and the same controls as reported in the notes to Table 3.

Table 7: Robustness - academic outcomes

	(1) +Block FEs	(2) +Demog	(3) +SES	(4) +IPW	(5) Block cl.
	Panel A: Standardized test score				
Post x Treat	0.231	0.252*	0.262*	0.269*	0.262*
	(0.143)	(0.145)	(0.148)	(0.147)	(0.147)
Constant	-0.272**	-0.725**	-0.377	-0.264	-0.377
	(0.126)	(0.283)	(0.597)	(0.590)	(0.632)
R^2	0.22	0.26	0.31	0.31	0.31
Obs.	646	646	646	646	646
	Panel B: Final math grade				
Treat	0.797***	0.932***	0.846***	0.856***	0.846***
	(0.230)	(0.215)	(0.241)	(0.246)	(0.283)
Constant	5.469***	5.860***	9.361***	9.768***	9.361***
	(0.635)	(0.867)	(2.451)	(2.401)	(2.723)
R^2	0.36	0.51	0.63	0.66	0.63
Obs.	220	220	220	220	220
	Panel C: Passed math				
Treat	0.162***	0.194***	0.216***	0.220***	0.216***
	(0.062)	(0.055)	(0.062)	(0.062)	(0.072)
Constant	0.892***	0.792***	1.089**	1.105**	1.089*
	(0.068)	(0.169)	(0.482)	(0.485)	(0.555)
R^2	0.35	0.57	0.64	0.67	0.64
Obs.	220	220	220	220	220
	Panel D: Repeated year				
Treat	-0.094**	-0.090**	-0.094**	-0.095**	-0.094
	(0.037)	(0.040)	(0.046)	(0.044)	(0.059)
Constant	0.062	0.177	0.374	0.266	0.374
	(0.040)	(0.123)	(0.350)	(0.360)	(0.417)
R^2	0.43	0.51	0.58	0.60	0.58
Obs.	218	218	218	218	218

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: Child age, gender, grade, autonomous community, whether eligible for school meal subsidy, baseline math grade, whether had online classes during lockdown, whether had a device to connect to tutoring sessions available, whether received some for of academic tutoring at baseline. In Column 3 we further add controls relating to socio-economic status and parental characteristics; Whether speaks Spanish at home, dummies for household income, number of household members below age 18, dummies for parental education, a dummy for whether living in a single-parent household, and 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 8: 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	-0.017	-0.036	-0.030	-0.029	-0.030
	(0.047)	(0.047)	(0.048)	(0.048)	(0.050)
Constant	-0.003	0.116	-0.215	-0.173	-0.215
	(0.027)	(0.131)	(0.333)	(0.335)	(0.361)
R^2	0.26	0.34	0.38	0.38	0.38
Obs.	628	628	628	628	628
		Panel B: Good at Spanish			
Post x Treat	-0.002	0.004	0.002	0.019	0.002
	(0.064)	(0.066)	(0.067)	(0.068)	(0.075)
Constant	0.719***	0.637**	0.184	0.138	0.184
	(0.251)	(0.277)	(0.397)	(0.394)	(0.323)
R^2	0.17	0.21	0.26	0.27	0.26
Obs.	629	629	629	629	629
		Panel C: Likes math			
Treat	-0.048	-0.051	-0.052	-0.056	-0.052
	(0.047)	(0.048)	(0.050)	(0.043)	(0.053)
Constant	0.274	0.510*	0.283	0.408	0.283
	(0.248)	(0.287)	(0.573)	(0.489)	(0.589)
R^2	0.30	0.39	0.45	0.53	0.45
Obs.	351	351	351	351	351
		Panel D: Likes Spanish			
Treat	-0.038	-0.039	-0.042	-0.030	-0.042
	(0.052)	(0.051)	(0.055)	(0.049)	(0.066)
Constant	0.269	0.515*	0.280	0.198	0.280
	(0.247)	(0.263)	(0.404)	(0.343)	(0.304)
R^2	0.24	0.34	0.39	0.53	0.39
Obs.	353	353	353	353	353

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 Table 7.

Table 9: Robustness - aspirations and motivation

	(1) +Block FEs	(2) +Demog	(3) +SES	(4) +IPW	(5) Block cl.
	Panel A: Bachillerato				
Treat	0.136**	0.138**	0.136**	0.135**	0.136**
	(0.058)	(0.057)	(0.059)	(0.059)	(0.065)
Constant	0.242	0.391	0.401	0.451	0.401
	(0.354)	(0.375)	(0.530)	(0.525)	(0.385)
R^2	0.27	0.38	0.46	0.46	0.46
Obs.	312	312	312	312	312
	Panel B: College				
Treat	0.000	0.009	0.034	0.034	0.034
	(0.049)	(0.050)	(0.050)	(0.050)	(0.051)
Constant	1.000***	1.028***	1.375***	1.392***	1.375***
	(0.032)	(0.165)	(0.325)	(0.316)	(0.393)
R^2	0.26	0.35	0.43	0.43	0.43
Obs.	301	301	301	301	301
	Panel C: Grit				
Treat	0.066	0.066	0.091	0.086	0.091
	(0.059)	(0.064)	(0.063)	(0.063)	(0.067)
Constant	3.420***	3.211***	2.485***	2.489***	2.485***
	(0.135)	(0.171)	(0.511)	(0.503)	(0.571)
R^2	0.24	0.31	0.40	0.41	0.40
Obs.	311	311	311	311	311
	Panel D: High effort				
Treat	0.099*	0.098*	0.116**	0.094**	0.116**
	(0.054)	(0.051)	(0.054)	(0.044)	(0.056)
Constant	0.701***	0.892***	0.072	0.108	0.072
	(0.230)	(0.160)	(0.404)	(0.381)	(0.419)
R^2	0.20	0.39	0.47	0.69	0.47
Obs.	351	351	351	351	351
	Panel E: Motivation school				
Treat	0.000	0.001	0.011	0.011	0.011
	(0.017)	(0.019)	(0.019)	(0.019)	(0.022)
Constant	0.861***	0.893***	0.759***	0.770***	0.759***
	(0.069)	(0.102)	(0.192)	(0.189)	(0.167)
R^2	0.25	0.34	0.43	0.43	0.43
Obs.	303	303	303	303	303

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 Table 7.

Table 10: Robustness - socio-emotional outcomes

	(1) +Block FEs	(2) +Demog	(3) +SES	(4) +IPW	(5) Block cl.
	Panel A: Wellbeing index				
Post x Treat	0.059	0.012	0.037	0.042	0.037
	(0.127)	(0.122)	(0.124)	(0.129)	(0.128)
Constant	7.566***	7.481***	5.992***	5.867***	5.992***
	(0.358)	(0.569)	(0.758)	(0.748)	(0.652)
R^2	0.56	0.60	0.64	0.64	0.64
Obs.	646	646	646	646	646
	Panel B: School satisfaction				
Post x Treat	0.264*	0.290*	0.265	0.295*	0.265
	(0.160)	(0.162)	(0.167)	(0.170)	(0.177)
Constant	5.699***	5.686***	4.624***	4.849***	4.624***
	(0.584)	(0.770)	(1.078)	(1.096)	(1.097)
R^2	0.20	0.24	0.30	0.29	0.30
Obs.	632	632	632	632	632
	Panel C: Locus of control				
Post x Treat	-0.051	-0.057	-0.058	-0.062*	-0.058
	(0.035)	(0.036)	(0.037)	(0.037)	(0.038)
Constant	0.587***	0.611***	0.324*	0.305*	0.324
	(0.072)	(0.078)	(0.188)	(0.182)	(0.212)
R^2	0.18	0.24	0.29	0.30	0.29
Obs.	641	641	641	641	641

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 Table 7.

Table 11: Attrition between baseline and endline - child questionnaire and test

	(1) Full	(2) Full	(3) Treat	(4) Control	(5) Treat ×
Treat	0.045	0.035			
	(0.035)	(0.035)			
Child characteristics					
7th grade		0.104**	0.050	0.179**	-0.130
		(0.045)	(0.064)	(0.075)	(0.099)
Age		-0.101***	-0.107**	-0.100*	-0.007
		(0.032)	(0.043)	(0.051)	(0.067)
Catalonia		-0.052	-0.055	-0.086	0.032
		(0.049)	(0.067)	(0.083)	(0.106)
Girl		-0.051	-0.083*	-0.009	-0.073
		(0.035)	(0.050)	(0.058)	(0.076)
No FSM		0.057	0.121*	0.016	0.105
		(0.070)	(0.073)	(0.129)	(0.148)
Notable		-0.079*	-0.057	-0.101	0.043
		(0.040)	(0.049)	(0.071)	(0.086)
Sobresaliente		0.017	-0.076	0.133*	-0.209**
		(0.052)	(0.078)	(0.072)	(0.106)
Test score (%) at baseline		-0.041	-0.134	0.083	-0.217
		(0.101)	(0.157)	(0.133)	(0.206)
Had no online classes during lockdown (April-June 2020)		0.020	0.076	-0.134	0.210
		(0.061)	(0.070)	(0.115)	(0.134)
Has a laptop		0.130	0.157	0.164	-0.007
		(0.095)	(0.121)	(0.140)	(0.185)
Has a tablet or touchscreen laptop		0.107	0.144	0.100	0.044
		(0.091)	(0.118)	(0.142)	(0.184)
Received some form of academic support at baseline		-0.009	-0.034	0.016	-0.050
		(0.046)	(0.060)	(0.083)	(0.102)
Speaks Spanish at home		-0.038	-0.018	-0.078	0.060
		(0.060)	(0.075)	(0.097)	(0.123)
Immigrant background		-0.023	-0.031	-0.058	0.027
		(0.037)	(0.055)	(0.064)	(0.084)
Parental education					
Primary school		-0.015	-0.000	0.015	-0.015
		(0.052)	(0.076)	(0.088)	(0.116)
Compulsory schooling		-0.044	-0.034	-0.073	0.039
		(0.038)	(0.050)	(0.072)	(0.088)
Vocational Degree		-0.033	0.032	-0.098	0.130
		(0.042)	(0.055)	(0.071)	(0.090)
Household income					
De 500 a 999 euros		-0.049	-0.033	-0.085	0.051
		(0.061)	(0.088)	(0.083)	(0.121)
De 1.000 a 1.499 euros		-0.007	0.055	-0.063	0.118
		(0.057)	(0.085)	(0.085)	(0.121)
De 1500 a 1999 euros		0.028	0.033	0.036	-0.003
		(0.064)	(0.093)	(0.085)	(0.126)
mas de 2.000 euros		-0.021	-0.034	-0.038	0.004
		(0.067)	(0.095)	(0.113)	(0.147)
Parental characteristics					
1		-0.020	-0.022	-0.024	0.002
		(0.043)	(0.052)	(0.073)	(0.090)
Mean dep. var.	0.88	0.89	0.91	0.87	0.89
R^2	0.00	0.11	0.15	0.15	0.16
Obs.	356	332	173	159	332

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. Heteroskedasticity-robust standard errors in parentheses. The table reports coefficients from a linear probability model estimated by OLS. The dependent variable is a dummy equal to one if the child responded to the endline survey and math test, and zero otherwise. The full sample (Columns 1-2) includes all students who registered for the program and whose parents completed the registration form and consents and were included in the randomization sample. Column 3 restricts the sample to those assigned to the treatment, and Column 4 to those that were assigned to the control group. Column 5 shows the coefficients of the interaction term between treatment status and the variable shown in the left hand column.

Table 12: Attrition between baseline and endline - parent questionnaire

	(1) Full	(2) Full	(3) Treat	(4) Control	(5) Treat ×
Treat	0.136***	0.140**			
	(0.051)	(0.054)			
Child characteristics					
7th grade		0.050	0.046	0.008	0.038
		(0.068)	(0.103)	(0.099)	(0.143)
Age		-0.019	0.025	-0.030	0.055
		(0.043)	(0.060)	(0.060)	(0.085)
Catalonia		-0.096	-0.129	-0.088	-0.042
		(0.077)	(0.113)	(0.110)	(0.157)
Girl		-0.046	-0.020	-0.076	0.055
		(0.055)	(0.074)	(0.085)	(0.113)
No FSM		-0.027	0.112	-0.139	0.251
		(0.101)	(0.134)	(0.152)	(0.203)
Notable		0.070	0.143*	0.025	0.118
		(0.061)	(0.078)	(0.094)	(0.122)
Sobresaliente		0.108	0.044	0.077	-0.034
		(0.104)	(0.120)	(0.182)	(0.218)
Test score (%) at baseline		-0.317*	-0.091	-0.563**	0.471
		(0.165)	(0.225)	(0.241)	(0.330)
Had no online classes during lockdown (April-June 2020)		-0.059	-0.045	-0.058	0.013
		(0.084)	(0.121)	(0.147)	(0.190)
Has a laptop		-0.219**	-0.166	-0.376**	0.210
		(0.103)	(0.157)	(0.155)	(0.221)
Has a tablet or touchscreen laptop		-0.220**	-0.153	-0.387**	0.234
		(0.102)	(0.151)	(0.155)	(0.216)
Received some form of academic support at baseline		0.018	0.018	0.030	-0.011
		(0.068)	(0.100)	(0.104)	(0.144)
Speaks Spanish at home		-0.033	-0.189	0.083	-0.271
		(0.092)	(0.156)	(0.129)	(0.202)
Immigrant background		-0.005	-0.111	0.115	-0.226*
		(0.061)	(0.084)	(0.093)	(0.125)
Primary school		-0.127	0.038	-0.230*	0.268
		(0.090)	(0.117)	(0.132)	(0.176)
Parental education					
Compulsory schooling		-0.061	-0.056	-0.099	0.043
		(0.076)	(0.095)	(0.123)	(0.155)
Vocational Degree		-0.079	-0.035	-0.055	0.021
		(0.082)	(0.107)	(0.122)	(0.162)
Household income					
De 500 a 999 euros		-0.070	-0.162	0.058	-0.220
		(0.082)	(0.117)	(0.133)	(0.177)
De 1.000 a 1.499 euros		-0.093	0.046	-0.175	0.221
		(0.090)	(0.126)	(0.156)	(0.200)
De 1500 a 1999 euros		-0.256**	-0.474***	0.069	-0.543**
		(0.115)	(0.163)	(0.173)	(0.238)
mas de 2.000 euros		-0.147	-0.140	-0.100	-0.040
		(0.129)	(0.184)	(0.219)	(0.286)
Parental characteristics					
1		-0.179***	-0.220***	-0.163*	-0.057
		(0.063)	(0.082)	(0.097)	(0.127)
Mean dep. var.	0.62	0.62	0.69	0.55	0.62
R^2	0.02	0.11	0.21	0.20	0.22
Obs.	356	332	173	159	332

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. Heteroskedasticity-robust standard errors in parentheses. The table reports coefficients from a linear probability model estimated by OLS. The dependent variable is a dummy equal to one if the parent responded to the endline survey, and zero otherwise. The full sample (Columns 1-2) includes all children who registered for the program and whose parents completed the registration form and consents and were included in the randomization sample. Column 3 restricts the sample to those assigned to the treatment, and Column 4 to those that were assigned to the control group. Column 5 shows the coefficients of the interaction term between treatment status and the variable shown in the left hand column.

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?

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 Additional Tables

Table B1: Balancing table - parental characteristics (endline respondents)

	(1) Treat	(2) Control	(3) Difference (2)-(1)	(4) p-value Col. (3)
Mother responded	77.17	78.49	-1.33	0.82
Lives as couple	76.38	77.42	-1.04	0.86
Spanish origin	57.48	49.46	8.02	0.24
Compulsory schooling or below	52.76	48.39	4.37	0.52
Income $<$ 1000 EUR	43.31	51.61	-8.31	0.22
Catalunya	24.41	27.96	-3.55	0.56
Madrid	75.59	72.04	3.55	0.56
HH size	4.12	4.10	0.02	0.89
Nb. children age $\leq$ 18	2.00	1.82	0.18	0.15
Age of youngest child	9.86	10.22	-0.36	0.50
N	127	93		

Notes: The table shows summary statistics for the sample of students who registered to participate in Menπores and whose parents responded to the endline survey. Column 4 reports the p-value of a t-test of the equality in means across the two groups.

Table B2: Balancing table - child characteristics (endline respondents)

	(1) Treat	(2) Control	(3) Difference (1)-(2)	(4) p-value Col. (3)	(5) All Class
Age	13.02	12.98	0.05	0.68	
Girl	48.03	47.31	0.72	0.92	
Public school	22.05	22.58	-0.53	0.93	
Born in Spain	85.04	82.80	2.24	0.65	
Spanish spoken at home	85.04	83.87	1.17	0.81	
School meal stipend	9.45	4.30	5.15	0.15	
No tablet or computer	8.66	10.75	-2.09	0.60	
Has access to internet	99.21	97.85	1.36	0.39	
Receiving academic support (at baseline)	18.90	18.28	0.62	0.91	
Grade 7	46.46	52.69	-6.23	0.36	
Grade 8	53.54	47.31	6.23	0.36	
Baseline child survey outcome					
Partial completion, 1+ maths question (%)	88.98	92.47	-3.50	0.39	
Completed survey fully (%)	47.24	49.46	-2.22	0.75	
Baseline performance					
Has failed math	49.54	45.33	4.21	0.58	
Has failed at least one subject	78.64	75.32	3.32	0.60	
Repeated grade at least once	25.98	26.88	-0.90	0.88	
Math grade (0-10) in first term	4.45	4.65	-0.20	0.52	
Test score (%)	24.76	22.99	1.77	0.45	32.23
N	127	93			1,464

Notes: The table shows summary statistics for the sample of students who registered to participate in Menπores and whose parents responded to the endline survey. Column 4 reports the p-value of a t-test of the equality in means across the two groups.

Table B3: Impact on academic outcomes - One year later

	(1) Final math grade 1 year later	(2) Repeated year 1 year later	(3) Child plans bachillerato 1 year later	(4) Aspires to college 1 year later
Treat	0.517	-0.111	-0.041	0.062
	(0.369)	(0.102)	(0.124)	(0.111)
Constant	5.168*	-0.454	1.457*	0.237
	(3.054)	(0.503)	(0.779)	(0.664)
Mean dep. var.	5.26	0.17	0.50	0.78
SD dep. var.	1.52	0.38	0.50	0.42
R^2	0.63	0.62	0.70	0.67
Obs.	159	159	159	159

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at student level in parenthesis. The table shows the δ coefficients from Equation (1). All regressions include block FEs and the same controls as specified in the notes to Table B5.

Table B4: 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.

Table B5: Impact on academic outcomes - All tutors

	(1) Standardized test score	(2) Final math grade	(3) Passed math	(4) Repeated year
Post x Treat	0.252*			
	(0.146)			
RW p-value	[0.074]			
Treat		0.838***	0.201***	-0.091**
		(0.233)	(0.058)	(0.044)
RW p-value		[0.017]	[0.017]	[0.088]
Constant	-0.690	8.058***	0.975***	0.250
	(0.537)	(2.062)	(0.373)	(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 stepdown p-values for multiple hypothesis testing reported in brackets. The table shows the δ coefficients from Equation (1) in Column 1 and from Equation (2) for Columns 2-4. All regressions include block FEs and control for student age, grade, gender, region, a dummy indicating school meal eligibility, 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 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. Sample includes both students taught by professional as well as volunteer tutors.

Table B6: Impact on self-perceived ability and affinity - All tutors

	(1) Good at math	(2) Good at Spanish	(3) Likes math	(4) Likes Spanish
Post x Treat	-0.027	0.011		
	(0.049)	(0.066)		
RW p-value	[0.518]	[0.635]		
Treat			-0.040	-0.067
			(0.049)	(0.052)
RW p-value			[0.825]	[0.707]
Constant	-0.367	0.155	0.197	0.180
	(0.316)	(0.370)	(0.553)	(0.410)
Mean dep. var.	0.25	0.55	0.36	0.41
SD dep. var.	0.44	0.50	0.48	0.49
R^2	0.36	0.27	0.43	0.39
Obs.	658	659	368	370
Unique ind.	365	366		

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at student level (Columns 1-2) and robust SEs (Columns 3-4) in parenthesis. Romano-Wolf stepdown p-values for multiple hypothesis testing reported in brackets. The table shows the δ coefficients from Equation (1) in Columns 1-2 and from Equation (2) for Columns 3-4. All regressions include block FEs and the same controls as specified in the notes to Table B5.

Table B7: Impact on aspirations and motivation - All tutors

	(1) Bachi llerato	(2) College	(3) Grit	(4) High effort	(5) Motivation school
Treat	0.123**	0.021	0.075	0.104**	0.005
	(0.056)	(0.050)	(0.061)	(0.052)	(0.018)
RW p-value	[0.248]	[0.899]	[0.707]	[0.315]	[0.899]
Constant	0.563	1.275***	2.586***	0.136	0.739***
	(0.487)	(0.329)	(0.458)	(0.389)	(0.183)
Mean dep. var.	0.41	0.81	3.04	0.54	0.75
SD dep. var.	0.49	0.39	0.50	0.50	0.16
R^2	0.45	0.42	0.36	0.46	0.41
Obs.	328	315	327	368	317

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. Robust SEs in parenthesis. Romano-Wolf stepdown p-values for multiple hypothesis testing reported in brackets. The table shows the δ coefficients from Equation (2). All regressions include block FEs and the same controls as specified in the notes to Table B5.

Table B8: Impact on socio-emotional outcomes - All tutors

	(1) Wellbeing index	(2) School satisfaction	(3) Locus of control
Post x Treat	0.044	0.286*	-0.062*
	(0.120)	(0.164)	(0.036)
RW p-values	[0.544]	[0.136]	[0.054]
Constant	5.973***	5.080***	0.317*
	(0.711)	(1.064)	(0.181)
Mean dep. var.	6.27	5.47	0.61
SD dep. var.	1.62	1.36	0.31
R^2	0.63	0.29	0.29
Obs.	679	665	672
Unique ind.	367	367	367

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at student level in parenthesis. Romano-Wolf stepdown p-values for multiple hypothesis testing reported in brackets. The table shows the δ coefficients from Equation (1). All regressions include block FEs and the same controls as specified in the notes to Table B5.

CENTRE FOR ECONOMIC PERFORMANCE
Recent Discussion Papers

1907	Tito Boeri Matteo Gamalerio Massimo Morelli Margherita Negri	Pay-as-they-get-in: Attitudes towards migrants and pension systems
1906	Chrystalla Kapetaniou Christopher A. Pissarides	Productive robots and industrial employment: The role of national innovation systems
1905	Esther Ann Bøler Andreas Moxnes Karen Helene Ulltveit-Moe	Strapped for cash: the role of financial constraints for innovating firms
1904	Shumin Qiu Claudia Steinwender Pierre Azoulay	Who stands on the shoulders of Chinese (scientific) giants? Evidence from chemistry
1903	Jonathan Colmer Mary F Evans Jay Shimshack	Environmental citizen complaints
1902	Ester Faia Gianmarco I.P. Ottaviano Saverio Spinella	Robot adoption, worker-firm sorting and wage inequality: evidence from administrative panel data
1901	Federico Esposito Fadi Hassan	Import competition, trade credit and financial frictions in general equilibrium
1900	Fabrizio Leone	Multinationals, robots and the labor share
1899	Marcus Biermann Elsa Leromain	The indirect effect of the Russian-Ukraine war through international linkages: early evidence from the stock market
1898	Stephen J. Redding	The economics of cities: from theory to data
1897	Dan A. Black Jeffrey Grogger Tom Kirchmaier Koen Sanders	Criminal charges, risk assessment and violent recidivism in cases of domestic abuse

1896	Stephan Maurer Guido Schwerdt Simon Wiederhold	Do role models matter in large classes? New evidence on gender match effects in higher education
1895	Ihsaan Bassier	Collective bargaining and spillovers in local labor markets
1894	Paul C. Cheshire Christian A.L. Hilber Piero Montebruno Rosa Sanchis-Guarner	(In)convenient stores? What do policies pushing stores to town centres actually do?
1893	Kohei Takeda	The geography of structural transformation: Effects on inequality and mobility
1892	Martina Magli	The spillover effect of services offshoring on local labor markets
1891	Cheti Nicoletti Almudena Sevilla Valentina Tonei	Gender stereotypes in the family
1890	Rui Costa Swati Dhingra Stephen Machin	New dawn fades: Trade, labour and the Brexit exchange rate depreciation
1889	Amanda Dahlstrand	Defying distance? The provision of services in the digital age
1888	Jan David Bakker Nikhil Datta Richard Davies Josh De Lyon	Non-tariff barriers and consumer prices: evidence from Brexit
1887	Stefania Albanesi Claudia Olivetti Barbara Petrongolo	Families, labor markets and policy

The Centre for Economic Performance Publications Unit
Tel: +44 (0)20 7955 7673 Email info@cep.lse.ac.uk
Website: http://cep.lse.ac.uk Twitter: @CEP_LSE