Journal of Public Economics 232 (2024) 105082

Contents lists available at ScienceDirect

Journal of Public Economics

journal homepage: www.elsevier.com/locate/jpube

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

Lucas Gortazar[a,b], Claudia Hupkau[c,d], Antonio Roldán-Monés[a,e,*]

[a] Universitat Ramon Llull, ESADE, Spain
[b] World Bank, United States of America
[c] Department of Economics, CUNEF Universidad, Spain
[d] Centre for Economic Performance, London School of Economics, United Kingdom
[e] Department of Social Policy, London School of Economics, United Kingdom

ARTICLE INFO

Dataset link: https://github.com/chupkau/mentores_replication

JEL classification:
C93
I24
I28
H75

Keywords:
Online tutoring
Mentoring
RCT
Mathematics
Child outcomes

ABSTRACT

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

1. Introduction

Intensive, in-person tutoring in one-on-one and small group settings has been shown to have substantial positive effects on learning at moderate cost (Nickow et al., 2020). The Covid-19 pandemic and associated lockdowns, which disrupted education in over 150 countries (Azevedo et al., 2021) and disproportionately affected disadvantaged children (Betthäuser et al., 2023), has brought tutoring programs center stage as a cost effective policy to close educational gaps that have widened during the pandemic.[1]

[☆] We thank seminar participants at LSE-CEP, Universidad Carlos III, Paris School of Economics, the World Bank, the Inter American Development Bank, Universidad Pablo Olavide, and Universidad Autonoma de Madrid. Pablo García-Guzmán, Natalia Collado, Ana Herrero, and Angel Martinez provided excellent research assistance. We thank Antonio Cabrales, Jonathan de Quidt, Luis Garicano, Nagore Iriberri, Matthew A. Kraft, Mónica Martinez-Bravo, Pedro Rey-Biel, and Jenifer Ruiz-Valenzuela for their comments. We are grateful to Empieza por Educar for helping with the implementation of the program, including Miriam Arriola, Zaida Pillado, Lara Crespo, Oscar Ferri, Carla Carmona, Paula Royo, Miguel Costa, Carles Lopez, Luis Oliver, Justo Quintanar, and Beatriz Morilla. We are also grateful to Iria Mata, Dirk Rosquillas, Miguel Ujeda, and Gonzalo Romero for their support in various layers of implementation. We are grateful to the tutors and the schools that participated in the intervention. We acknowledge financial support from Esade, Fundación COTEC, Fundación CEOE, United Way Spain, Banco Santander, Spain, UNIR, Huawei, Fundación Atlantic Copper, Edvolution and Google, Spain. We also want to thank Fátima Báñez and Cristina Garmendia for their essential support in the early stages of the project. This project underwent ethics review by the Research Ethics Committee of the Universitat Ramon Llull and obtained IRB approval from the same university. The experiment reported in this paper is registered at The American Economic Association's registry for randomized controlled trials, AEA RCT Registry ID: AEARCTR - 0007138.

[*] Corresponding author.

E-mail addresses: lucas.gortazar@esade.edu (L. Gortazar), claudia.hupkau@cunef.edu (C. Hupkau), a.roldan-mones@lse.ac.uk (A. Roldán-Monés).

[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 both face-to-face or online tuition.

https://doi.org/10.1016/j.jpubeco.2024.105082
Received 2 March 2023; Received in revised form 24 January 2024; Accepted 24 January 2024
Available online 4 March 2024

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

In this paper, we study the effectiveness of an intensive, eight-week math tutoring program on academic and socio-emotional outcomes of secondary school children in Spain. It offered free, 100-percent online after-school tutoring to pupils aged 12 to 15 from very disadvantaged backgrounds. The program, called *Menπores*, has four key features.[2] First, the whole organization of the program and the tutoring sessions were implemented online. Second, the large majority of tutors delivering the program were paid-for, qualified math teachers. Third, the tutoring sessions were done in groups of two students per tutor. Fourth, the program focused on math and social-emotional support (motivation, well-being, and work routines). This focus was chosen because our target population was teenage children aged 12 to 15, and evidence suggests that tutoring in mathematics tends to be more effective for students in higher grades, while literacy interventions have been shown to be more effective in pre-school and primary school settings (Nickow et al., 2020). Further, the focus on socio-emotional support was introduced to mitigate the detrimental effects of the pandemic and associated school closures on children's mental health (Newlove-Delgado et al., 2021), and because of the growing evidence as to the importance of socio-emotional skills in educational attainment and future labor market outcomes (Heckman et al., 2006; Kosse et al., 2020; Kosse and Tincani, 2020; Eisner et al., 2020).

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

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

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

In terms of non-cognitive outcomes, we find that the program raised students' aspirations: students in the treatment group were 13.5 (p-value: 0.022) percentage points more likely to state that they would like to go onto the academic track after compulsory schooling (i.e. *Bachillerato*), equivalent to a 31 percent increase compared to the control group mean. This result is important because aspirations have been shown to positively affect future educational achievement (Khattab, 2015), and because attending the academic track at upper secondary school is linked to higher earnings later in life thus potentially increasing social mobility (Matthewes and Ventura, 2022). We do not find a positive impact on stated intentions to go to university, possibly because the decision to go to university lies too far away in the future for the students in the intervention (the average age of participants was 13).

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

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

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

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

students liked more and felt more confident in Spanish. It is possible, for instance, that as a student becomes better in math, they gain a comparative advantage in that subject and lose interest in other subjects. It is also possible that improved results in math could motivate students overall and make them more motivated for other subjects. However, our results show no such spillover effects.

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

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

While attrition for the in-class survey was low – the response rate was 88 percent – and equal for the treatment and control group, it was more pronounced and 13 percentage points higher for the control group in the parent-survey, on which we rely to measure end-of-year academic outcomes. This raises concerns about the internal validity of our estimates and could bias our results. For instance, experimenter demand effects might cause parents of treated children to report more positive results. We address this concern in different ways and show that results on parent-reported outcomes are robust to using inverse probability weights and provide bounds to our estimates using Lee (2009)'s and Behaghel et al. (2009)'s approach. We find that bounds only include positive impacts on the final math grade (ITT-estimate = 0.852, bounds = [0.304,1.317]), whether the student passed the subject (ITT-estimate = 0.205, bounds=[0.143,0.357]) and only negative impacts on whether the student had to repeat the school year (ITT-estimate = −0.089, bounds = [−0.270,−0.060]). This gives reassurance that when taking into account sample attrition, the main conclusions from our analysis continue to hold.

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

Our study contributes to the understanding of whether online tutoring can work as an effective tool for closing learning gaps for disadvantaged students. The closest to our research is the online tutoring program implemented in Italy in Spring 2020 by Carlana and La Ferrara (2021). They find large positive effects on student achievement (+0.26 SD) and positive effects on socio-emotional skills, aspirations, and psychological well-being. Kraft et al. (2022) also implement an online tutoring program for middle school students with college volunteers. They find positive but insignificant effects on math and reading.[3]

Our program departs from these studies in three fundamental ways. First, they were delivered by volunteer university students, while *Menπores* used mostly paid-for, qualified secondary school teachers. Second, our tutoring was implemented in groups of two students, instead of one-on-one. Third, and more importantly, these programs were implemented in exceptional circumstances. For the case of Carlana and La Ferrara (2021), the program took place during the harshest lockdown period in Italy from April to June 2020 (when all kids were at home and schools were closed).[4] In the case of Kraft et al. (2022), in early 2021 in the US, when schooling was still highly disrupted. Our program, instead, was implemented one year after the onset of the pandemic, several months after schools were fully re-opened in Spain. In that sense, we believe our results show the effectiveness of online tutoring in normal times, when tutoring can be considered a complement rather than a substitute for regular schooling.

Our contribution is relevant both in terms of policy and for further academic research. Governments are investing large amounts of money in tutoring programs (both in face-to-face and online formats). Our evidence suggests that this money is well spent. The intervention costs approximately €300 per student, and has a positive impact of 0.26 SD on our standardized math test, translating into a 0.087 SD increase per €100 spent.[5] This compares favorably with summer schools analyzed in Cooper et al. (2000), with a cost-effectiveness of 0.066 SD per €100 spent (based on an impact of 0.23 SD and a cost of €350 per student). It also compares favorably with increasing instruction time by one hour per day, which according to Higgins et al. (2012) costs €1,020 for an increase of 0.24 SD in test scores, resulting in a cost-effectiveness rate of 0.0235 SD per €100 spent.

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

[3] 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.

[4] 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).

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

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

The rest of the paper is organized as follows. Section 2 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 Sections 6 and 7, respectively. Section 8 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, just before the start of our intervention, only 0.5 percent of classes in Spain were operating remotely due to quarantines.

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

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

3. Study design

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

3.1. The program Menπores

Our online tutoring program, called *Menπores*, was an intensive intervention consisting of three 50-minute sessions per week over a period of eight weeks. The target population were students in Grades 7 and 8 (grades 1 and 2 of secondary school, students aged 12 to 15), attending schools in highly disadvantaged neighborhoods. We chose this target for two reasons. First, disadvantaged students were disproportionately affected by learning loss during the pandemic (Haelermans et al., 2021; Blainey and Hannay, 2021) and most likely to benefit from the intervention. The need to invest and experiment with remedial programs which could facilitate catch up for the learning loss of these students was and still is a priority in education policy in many countries (World Bank, 2021a,b). Second, evidence suggests that tutoring in mathematics tends to be more effective for students in higher grades (Nickow et al., 2020), and budget, logistical and time constraints meant that we could deliver tutoring only in one subject area and only in secondary schools.

Tutoring sessions were delivered online mostly by qualified math teachers in groups of two students per tutor. We decided to concentrate hiring efforts on qualified math teachers for several reasons: First, existing evidence on face-to-face tutoring shows that they are significantly more effective than non-professionals or volunteer tutors (Nickow et al., 2020). Second, while we had initially planned a second treatment arm with tutoring delivered by volunteer university students as in Carlana and La Ferrara (2021), we were neither able to recruit sufficient participating students nor sufficient volunteer tutors in the short time-frame we were operating in.[6] The timing of our intervention (towards the end of the academic year, when university students tend to be more busy because of final examinations) and the fact that life in Spain had largely gone back to normal by March 2021 (students were no longer locked inside their homes as they had been between March 2020 to May 2020) are possible explanations for the low response to our call.

The group composition was fixed throughout the program, with the same students attending meetings with the same tutor in each session. The students in each tutoring group of two were from the same class or grade from the same school. This was done in order to increase the power of our experimental design as well as to guarantee that students knew each other and would find it easier to connect and accommodate. We decided to go for a two-on-one student–tutor ratio for three reasons. First, the pedagogic team in charge of implementation suggested that being in a group with another child had the potential to generate mutual motivation and peer pressure not to abandon the program. Second, existing evidence for face-to-face programs in Nickow et al. (2020) shows that two-on-one tutoring is nearly as effective as one-on-one tutoring. Moreover, evidence from a two-on-one math tutoring program in Chicago (Guryan et al., 2023) shows that this design can be highly effective even for older (secondary school) children. Third, this design is relevant from a scalability perspective, as it significantly reduces cost per student.[7]

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

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.[8] Its core

[6] 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.

[7] 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.

[8] 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.

activity is based on a highly selective model of teacher training, identifying teacher candidates with top academic and socio-emotional skills that are relevant for the teaching profession, as well as an interest in the profession and in social change.

The academic content of the tutoring program was based on the national mathematics curriculum and covered the expected knowledge from 1st and 2nd graders in secondary schools in Spain. Additionally, the program aimed at providing psycho-social and socio-emotional support to students. This was done for several reasons. First, to potentially mitigate the detrimental effects of the pandemic and associated school closures on children's mental health (Newlove-Delgado et al., 2021). Second, there is growing evidence as to the importance of socio-emotional skills in educational attainment and future labor market outcomes (Heckman et al., 2006; Kosse et al., 2020; Kosse and Tincani, 2020; Eisner et al., 2020). There was therefore an explicit mandate for tutors to spend time in the sessions providing such support and reserve at least ten out of the 50 min to discuss any issues, fears or concerns the children might be facing at home or at school. The pedagogical approach of the sessions was inspired by the *No Excuses* methodology, which has been shown to be effective in raising academic and non-cognitive outcomes in the context of urban US charter schools for vulnerable children (Dobbie and Fryer, 2013). This methodology emphasizes high expectations, increased instructional time, individualized support, continuous feedback and intensive data collection on student progress to guide instruction. We provide details on how tutors were trained in these aspects in Section 3.4.

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

3.3. Recruitment of schools and participants

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

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

Emails were sent and calls were made to gauge interest to around 32 schools. We had calculated that in order to reach our target sample size of 400 enrollments, we would need to get about 20 schools on board.[9] After the initial emails and calls, recruitment efforts were intensified among those schools that showed an interest (i.e., those that replied to emails or answered calls and consulted with the governing bodies of their schools to see whether there was support for participation). Recruitment ended when the target number of schools had been reached, as time, budgetary and operational constraints meant that we could not deliver tutoring to more than about 200 students. Eventually, 18 schools signed participation agreements.[10]

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

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

3.4. Selection and training of tutors

Our implementation partner ExE designed and implemented the selection and training for tutors based on their longstanding experience with teacher selection. A key criterion for selection was to hold a post-graduate (Master's) degree in Teacher Training in a scientific specialization (math, physics, chemistry or biology), which is a formal requirement to teach mathematics in secondary education in Spain. While holding a Master's degree in teacher training was a desired characteristic, it was not binding. Other skills, such as motivation for the program, having taught in low-income schools, and prior teaching experience, were also considered. Advertisement of the positions was done through various channels, including online hiring portals, ExE's own network of current teachers and alumni, and other teachers whom they work with. A total of 199 applicants which met the minimum pre-requisites were sent a formal application form, and applied. Out of these, 110 candidates were sent a link for an online interview. Out of the 110 candidates interviewed we hired 37 professional tutors. In parallel, we recruited a small number of university students as volunteer tutors and ended up including eight such tutors in the program.[11]

Before the start of the program, tutors received between 15 to 20 h of online training through ExE's teacher training platform. Training included two remote training modules and two online webinars with

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

[10] 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.

[11] As mentioned previously, we had initially planned a third treatment arm with volunteer mentors only. Although we advertised the program at five large public and private universities, we only received 50 applications. We attribute the small number of applications to the fact that the program required a high level of time commitment and coincided with end of term examinations at university. This was also the reason why most of the applicants finally decided to drop out of the process before the start of tutoring sessions. After initial screening and interviews, we were able to include only eight volunteer tutors, who completed the entire application process. We decided to keep these tutors in our pool and included them in the randomization. This allowed us to fulfill the initial commitment to schools to provide tutoring to around 200 students. We include students tutored by volunteers in all the results presented. Results are very similar when students that were taught by volunteer tutors are excluded. We discuss these in Section 7.3.

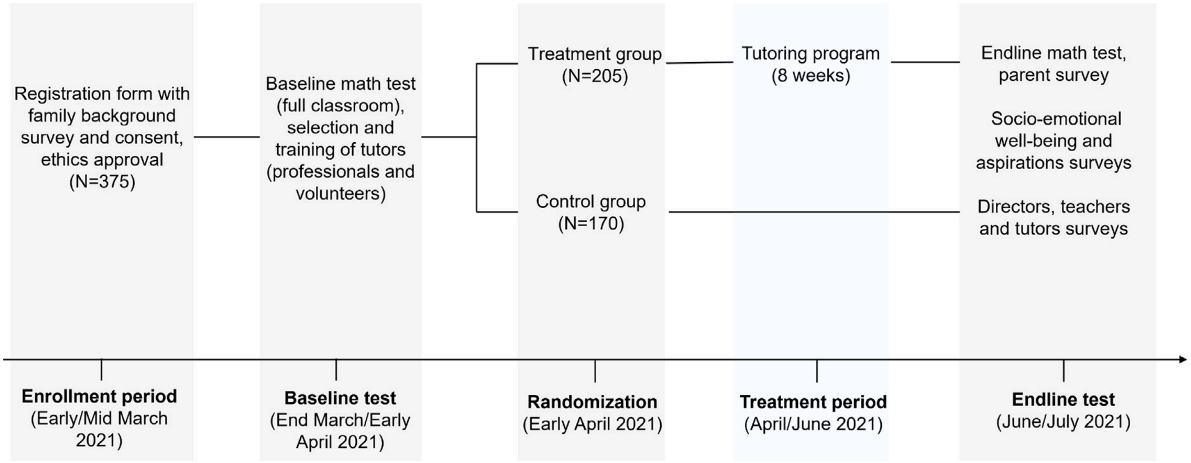
Registration form with family background survey and consent, ethics approval (N=375)
Baseline math test (full classroom), selection and training of tutors (professionals and volunteers)
Treatment group (N=205)
Control group (N=170)
Tutoring program (8 weeks)
Endline math test, parent survey
Socio-emotional well-being and aspirations surveys
Directors, teachers and tutors surveys
Enrollment period (Early/Mid March 2021)
Baseline test (End March/Early April 2021)
Randomization (Early April 2021)
Treatment period (April/June 2021)
Endline test (June/July 2021)

Fig. 1. Timeline of the Menπores program implementation.

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

Fig. 1 shows the timeline of the intervention. Planning and design took place between January and March 2021, and the registration period for parents and children started in early March 2021, lasting for about two weeks. A total of 375 complete registrations with valid consents were received during this time window. After registrations were closed, baseline tests and surveys were administered in all participating classrooms, that is, in all classes where at least one student had registered 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 six students. Volunteer tutors were assigned only one group. Randomization of tutors was stratified by geographic area, where those tutors based in Catalonia who indicated they spoke Catalan were assigned to students based in Catalonia, and those based in Madrid to those who were based in the region of Madrid.

3.7. Implementation

Students and tutors were able to organize their own schedule and agree on weekly meeting times.[12] 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.[13] 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.

Fig. 2 shows the distribution of total minutes and total number of tutoring sessions attended by students. Only seven students (3.4 percent of those assigned to the treatment group) actively dropped out of the program before it began. Among students assigned to treatment, the median number of minutes of tutoring received was 952, representing

[12] 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.

[13] 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.

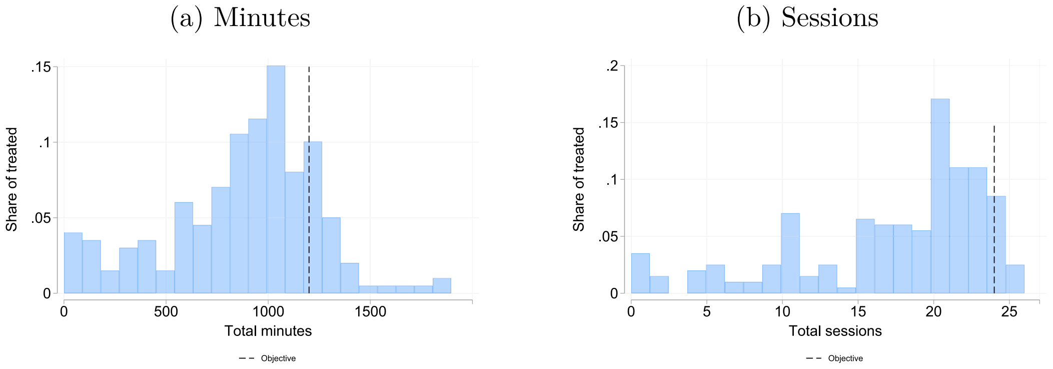

Fig. 2. Distribution of total minutes and number of sessions of tutoring attended
Note: This figure shows histograms of the total number of tutoring minutes attended (left panel) and the number of sessions attended (right panel). The data comes from electronic records of connection times to online meetings from 199 out of 205 students in the treatment group. This includes seven students who dropped out of the program before it started (zero minutes and zero sessions), and excludes six students for whom we could not match meeting data.

80 percent of the target number of minutes (1200). The median number of sessions attended was 20, corresponding to 83 percent of the envisaged number of sessions (24).[14]

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.[15]

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.[16] Because of the timing of the baseline test – right before the Easter holiday and after grading for the first term had finished – students were not missing regular math content to do the test. We explicitly instructed teachers not to mention the program *Menπores* while they ran the tests, so that students who had registered would not associate this assessment to their likelihood of being selected.

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

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

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

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

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

[15] 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.

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

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

We also conducted an online and phone survey of parents of study participants in early July, when the school year was over. Parents were asked two key questions about their child's academic outcomes: The final math grade, measured on a 1 to 10 numerical scale, obtained by their child at the end of the year, and whether the child would have to repeat the school year.[17] 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

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

Balancing between treatment and control group characteristics of the randomization sample is shown in Table 1. The table shows the control group mean and standard deviation (column 2), and the treatment-control difference estimated from a regression of the characteristic as the dependent variable on a treatment dummy and block fixed effects (column 3). Finally, it shows the normalized difference between treatment and control group, equal to the coefficient reported in column 3 divided by the standard deviation of the control group (column 4). Note that balance variables were not pre-registered. However, they represent all pre-determined characteristics that were collected during the registration process and the baseline child survey.

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

5. Empirical strategy

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

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

where Y denotes the outcome for student i, in block b at time $t \in \{pre, post\}$, i.e. either before or after the intervention. The α_b's are block fixed-effect (indicating the classroom of the student). $Treat_i$ is a dummy variable equal to one if the individual is in the treatment group, and $Post_t$ is a dummy variable equal to one for the period after the end of the program. The vector X represents a set of pre-determined student and parent characteristics that include student age, grade, gender, region fixed-effects, a dummy indicating school meal eligibility, a set of dummy variables indicating baseline math grade categories (fail, pass, good), a set of dummy variables indicating the frequency of online lessons during school closures in April and May 2020, a dummy indicating whether the student had a tablet or computer at home before the program, a dummy indicating whether the student was receiving other tutoring before the program, categorical variables indicating the number of people below age 18 at home, the language spoken at home, parental education, household income, an indicator for whether the responding parent is a single parent, and a dummy variable indicating whether the parent is of Spanish origin. Finally, ϵ_{ib} is an error term. The coefficient of interest, δ, corresponds to the intention-to-treat (ITT) estimate, which measures the effect of being assigned to participate in Menπores. Standard errors in this specification are clustered at the individual level.[18]

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.

The results presented here deviate in two ways from what had been pre-registered. First, we had not initially registered some of the academic outcomes (math grade, whether the course was passed and whether the school year was repeated). The reason for this was that even though we included a clause for consent to linking the student data to administrative records, there was a high chance that this data would not be made available. This is due to the fact that in Spain, the process for access to administrative data is not yet well established, and often depends on factors that are outside researchers' control.[19] Indeed, the data owners – the ministries of education (*Consejeria de Educacion*) in Madrid and Catalonia – did not give us access to this data after the end of the program, stating legal reasons. Because we believed these outcomes were extremely important and we had funding available at the end of the program, we decided to solicit the information on academic outcomes through a survey to parents. Additionally, our pre-registration did not include results on self-perceived affinity and ability in math and language. The reason was that we included these outcomes in the base- and endline surveys after pre-registration, but we considered it relevant to look at them as we thought they could potentially constitute mechanisms which might explain persistent effects of the program.

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

We look at many outcome variables, which raises the risk for false positives. To adjust for the fact that we are testing multiple hypotheses and may incorrectly reject null hypothesis of no effects, we calculate Romano–Wolf step-down adjusted p-values, which control for

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

[18] 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 9–12.

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

Table 1
Balancing between treatment and control group.

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

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

the family-wise error rate and allow for dependence among p-values. To do so, we group our outcomes into five families: (1) academic achievement; (2) self-perceived ability and affinity; (3) aspirations; (4) school attitudes and motivation and (5) socio-emotional outcomes.

6. Results

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

6.1. Selective attrition

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

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

To check whether there is selective attrition at endline, Table 2 shows balancing conditional on having an endline observation in the in-class math test and questionnaire. None one of the characteristics, except for the number of children aged 18 and below in the household, are significantly different between treatment and control group. The p-value for an F-test of the null hypothesis that baseline characteristics are jointly the same for treatment and control group is equal to 0.765. Additionally, we find no significant difference in missingness between treatment and control group for the outcomes measured through the in-class test and student questionnaire (see Table A1 in the online appendix).

The second set of academic achievement variables – end-of-year math grades, whether the math course was passed and whether the child had to repeat the grade – were collected via an online and phone survey to parents at the end of the school year. Attrition for this survey was higher, with 62 percent of parents responding at endline overall (233 out of 375). Missingness for the outcomes measured through the parent survey was significantly higher for control group students, whose parents were 13 percentage points less likely to respond at endline (see Table A1 in the online appendix). In Table 3, we show balancing conditional on having an endline observation in the parent-reported outcomes. There are no statistically significant differences in most baseline characteristics between treatment and control group. However, we reject the null hypothesis that baseline characteristics are jointly equal between treatment and control group due to some characteristics showing significant differences. For instance, we find that treatment students whose parents responded at endline were more

Table 2
Balancing table — endline respondents to in-class child questionnaire and test.

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

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

likely to be recipients of free school meals, and their parents were more likely to hold only compulsory schooling or below (versus high school diploma or above), suggesting slightly negative selection into responding at endline among the treatment group. While these analyses suggest differential attrition at endline between treatment and control group, it is reassuring that we observe no differences in terms of baseline academic achievement between the two groups. In robustness checks (Section 7.1), we will use inverse probability weights and provide Lee (2009) and Behaghel et al. (2009) bounds to treatment effects to account for non-random sample selection in these outcomes.

6.2. Academic outcomes

Table 4 summarizes the results for the impact of the intervention on academic outcomes. The dependent variable in column 1 is the score on the math test, standardized by grade level.[20] The difference-in-difference estimate indicates that treatment students improved their score by 0.26 SD more than control students, which is significant at the 10 percent level.[21]

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

In online appendix Table A2, we show results from parent-reported academic outcomes collected from a survey we implemented one year later (in autumn 2022). The response rate for the follow up questionnaire was only 45 percent (168 out of 375), out of which 120 had also

[20] 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.

[21] 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.

Table 3
Balancing table — children of endline respondents to parent survey.

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

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

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

6.3. *Self-perceived affinity and ability*

In Table 5 we present results on outcomes measuring self-perceived ability and affinity towards math. Columns 1 and 3 show, respectively, that the program did not increase the likelihood of pupils stating that they thought they were good at math or that they liked math. These results are surprising in light of the positive impact of the intervention on actual achievement (both externally graded math tests and teacher-assessed outcomes). Given the positive relationship between perceived ability and outcomes (Spinath et al., 2006), the failure to raise students' self-image in mathematics may have limited the potential longer-term effects of our program. For comparison, we also asked the same questions about Spanish language, to check whether there was some sort of crowding out (i.e., students shifting preferences towards math or away from math in favor of other subjects). It is also possible that improved results in math could motivate students overall and make them more motivated for other subjects. However, we do not find evidence that this happened (see columns 2 and 4).

6.4. *Aspirations, perseverance, effort and motivation*

We now look at the impact of the program on aspirations, perseverance, and motivation. Column 1 of Table 6 shows that treatment group students were 13.5 percentage points more likely than control group students to state that they would like to go on to complete a *bachillerato* (academic high school track), the pre-requisite for entering university in Spain, after completing compulsory education. This corresponds to a 31 percent higher probability than the control group, and the result is robust to adjusting for multiple hypothesis testing. We believe the impact of the program on raising aspirations to choose the academic track are important for several reasons: First, because aspirations have been shown to positively affect future educational achievement (Khattab, 2015). Second, attending the academic track at upper secondary school is linked to higher earnings later in life and may thus increase social mobility (Matthewes and Ventura, 2022). We do not find an increase in the likelihood of stating that students plan to go on to higher education (college/university) after-school (column 2). While choosing the academic track at upper secondary school tends to be highly correlated with planning to go to university, the fact that we do not find an impact here might be because this is a decision that lies very far in the future for the students in our intervention, who were on average just 13 years old.

Table 4
Impact on academic outcomes.

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

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

Table 5
Impact on self-perceived ability and affinity.

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

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at student level (columns 1–2) and heteroskedasticity-robust SEs (columns 3–4) in parenthesis. Romano–Wolf step-down-adjusted p-values for multiple hypothesis testing reported in brackets (based on 10,000 replications). The table shows the δ coefficients from Eq. (1) in columns 1–2 and from Eq. (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 4.

Table 6
Impact on aspirations and motivation.

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

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. Heteroskedasticity-robust SEs in parenthesis. Romano–Wolf step-down adjusted p-values for multiple hypothesis testing reported in brackets (based on 10,000 replications). The table shows the δ coefficients from Eq. (2). All regressions include block FEs and the same controls as reported in the notes to Table 4.

Table 7
Impact on socio-emotional outcomes.

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

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at student level in parenthesis. Romano–Wolf step-down adjusted p-values for multiple hypothesis testing reported in brackets (based on 10,000 replications). The table shows coefficients from Eq. (1). All regressions include block FEs and the same controls as reported in the notes to Table 4.

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

Given the programs specific focus on student motivation using the growth mindset approach developed by Dweck (1986), we tested whether the program positively affected student motivation, effort, and perseverance, which would be potential mechanisms driving also the increase in academic achievement. In column 3 of Table 6, we assess whether program assignment had any impact on grit, a measure of perseverance and conscientiousness. We do not find evidence that this was the case. It is likely that our program was too short to be able to change this outcome. In fact, recent research suggests that grit is a highly heritable personality trait with limited malleability (Rimfeld et al., 2016).

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

6.5. Well-being and socio-emotional outcomes

In the aftermath of the Covid-19 pandemic, there were considerable concerns about the longer-run effects of the lockdown and school closures on children's mental health (Newlove-Delgado et al., 2021). We expected the intervention to have a positive impact on socio-emotional well-being due to the positive group dynamics and the presence of an adult reference as a tutor and mentor (Kosse et al., 2020). Table 7 shows the ITT-estimates of the impact of the program on measures of well-being and locus of control. We find no impact of the program on overall subjective well-being measured by the well-being index (column

Table 8
Heterogeneous effects.

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

Notes: The table shows the coefficient on the treatment and the interaction between treatment assignment and different measures of heterogeneity in the composition of the group and baseline performance for students in the study sample, i.e. those who registered for Menπores. Estimates are from our ITT specification (Eq. (1) for standardized test scores and Eq. (2) for the remaining outcomes), including a dummy for treatment assignment interacted with indicators for each group with appropriate main effects added, including block fixed effects and our usual set of baseline covariates, as indicated in the notes to Table 4.

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

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

6.6. Tutor, teacher and parent feedback

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

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

Mathematics teachers and headteachers of the participating schools rated the impact of the program positively. More than 70 percent of teachers and 57 percent of headteachers surveyed agreed or strongly agreed that the program had been useful for their pupils. Some 69 percent of teachers believed that the program was a good support for their teaching. Finally, 71 percent thought that the program should continue, which is also shared by 100 percent of headteachers surveyed. More than 40 percent of surveyed mathematics teachers believed that the fact that pupils participated in the program helped them to work better, and another 42 percent believed that the coordination meetings with the mentors were useful. Some teachers said that they were overwhelmed by the additional workload during the program (due to coordination with tutors and administering base- and endline tests). In the open-ended responses, several teachers and headteachers suggested to start the program before April and make it longer.

We also analyze what tutors perceived to be the main obstacles to students attending the sessions. Fig. 3 shows a summary of the results from this analysis. Around 40 percent of tutors mentioned clashes with other extracurricular activities as a common cause for

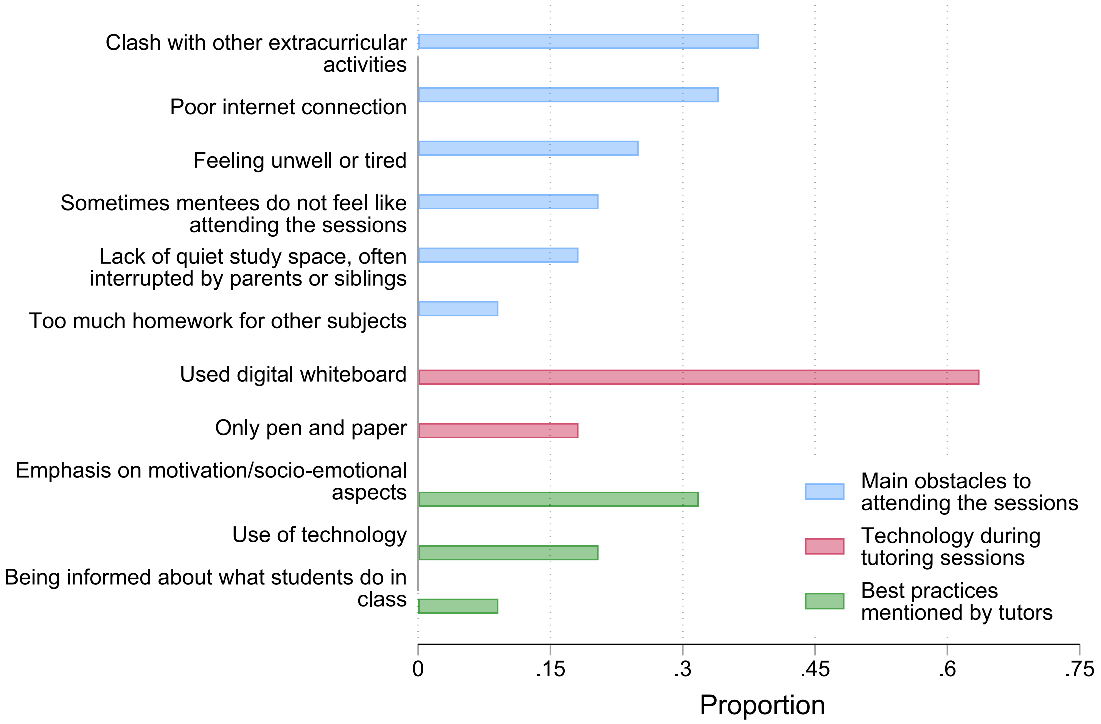

Fig. 3. Best practices mentioned by tutors at endline

Note: This figure shows information collected through surveys to tutors at the end of the program (N = 44 out of a total of 45 tutors). They show the share of tutors who mentioned different types of best practices, technology used and obstacles encountered.

non-attendance. Poor internet connections and feeling tired or unwell were also important. Almost 20 percent of tutors mentioned a lack of quiet study space and interruptions by parents or siblings a problem for effectively delivering tutoring sessions.

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

6.7. Heterogeneous effects and mechanisms

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

Fig. 4 shows the interaction effects of the different characteristics with the treatment variable (Table 8 shows the full set of coefficients on the treatment dummy and the interactions terms). While interaction terms tend not to be significant, which would be expected due to our small sample size, the signs and magnitudes of the coefficients provide several interesting insights. For the standardized test, the program seems to have been less effective for those in the bottom 50% of the baseline ability distribution. This is consistent with results in Guryan et al. (2023), who find that their program had positive treatment effects on math test scores for all but the bottom quartile in baseline achievement. The program seems slightly more effective in terms of math test scores if the student and the tutor were of the same gender, although this interaction is very imprecisely estimated. The gender match among the students in the group or whether students were of similar baseline ability does not seem to have mattered for impacts on math test scores. This latter finding is in contrast with those in Duflo et al. (2011), who find positive effects of ability tracking.

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

One of the findings from our study is that despite the fact that tutors had a clear mandate to address socio-emotional aspects, such as motivation and emotional well-being, we find little or no impact of our intervention on these outcomes. While the program might simply have been too short to have an effect on such outcomes, it raises the question of whether tutors were actually implementing socio-emotional support in their sessions. While we do not have monitoring data on this aspect, we have data on open-ended questions to tutors at endline regarding best practices that they would recommend for future editions of the program. About 32 percent of tutors mentioned the emphasis on motivation and socio-emotional aspects as important for the success of the intervention (see Fig. 3). The fact that almost a third of tutors mentioned this explicitly, and that this factor seems to have been more

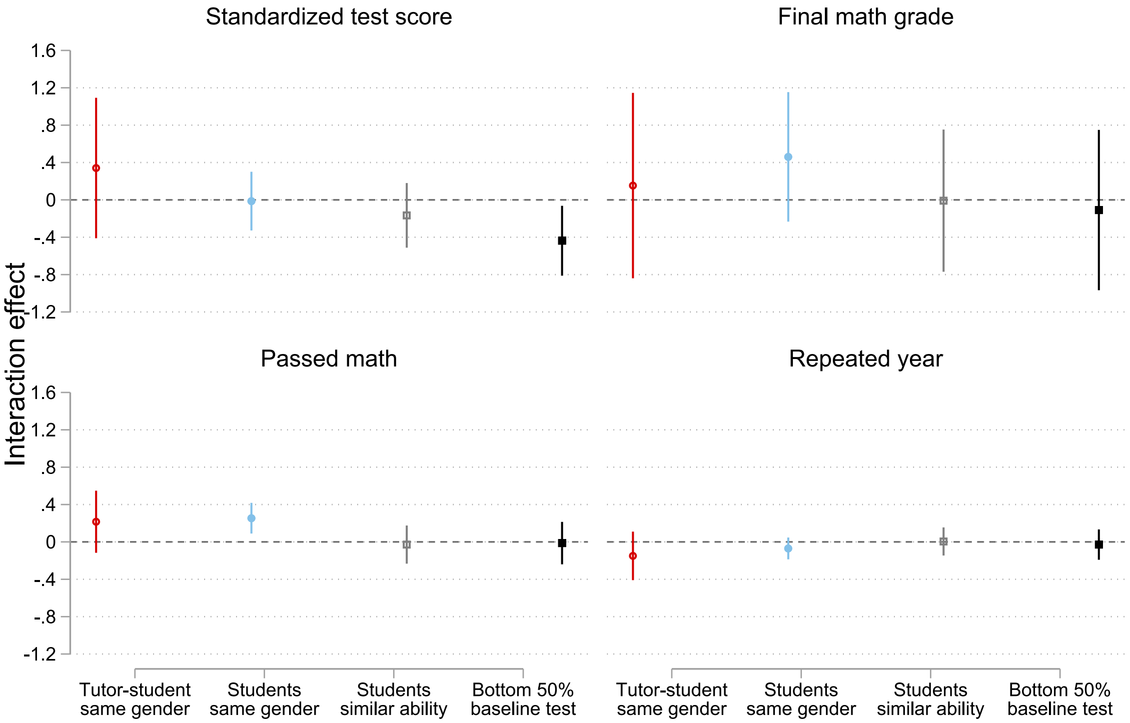

Fig. 4. Heterogeneous effects by group composition and baseline performance
Note: This figure shows the coefficients from the interaction terms shown in Table 8. These are estimated by regressing the outcome on a treatment dummy and the interaction between the treatment dummy and a dummy indicating (1) whether the student and tutor were of the same gender; (2) whether the students in the group were of the same gender; (3) whether the students in the group were of similar ability at baseline and (4) whether the students were both in the bottom 50% of the baseline test score distribution. All regressions include the full set of controls as specified in the notes to Table 4 and block fixed effects. The plot shows 90% confidence intervals.

important than, for instance, the use of technology (such as digital whiteboards) during sessions (mentioned by 20 percent of tutors) or being informed about what students were working on during their regular math classes (mentioned by 9 percent of tutors) suggests that indeed tutoring sessions were not merely focused on academic content. In future research it will be important to keep track and monitor more explicitly to what extent socio-emotional support is actually implemented, or randomly vary the degree of this support in order to understand its importance for program effectiveness.

7. Discussion of results and robustness checks

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

7.1. Selective attrition

As discussed in Section 6.1, the differential response rate to the endline parent survey among treatment and control group parents raises concerns about whether our estimates might be biased due to selective attrition. To quantify how much this might matter for our results, column 4 in Table 9 shows the estimated impact of program assignment on academic outcomes using inverse-probability weights. We can see that the effects size on our standardized math test (Panel A) is virtually identical when using these weights compared to our main result (reported in column 3). Panels B to D are outcomes that rely on parental responses to the endline survey. Effect size estimates are virtually identical, if anything, slightly higher, for all three outcomes when using inverse-probability weights. This is in line with the potential (downward) bias in effect sizes predicted from our analysis of selective attrition.

While the above robustness check is reassuring, the substantial attrition for parent-reported outcomes, which are also the most precisely estimated, raises concerns about experimenter demand effects, with parents of treated children possibly being more likely to report positive outcomes. Given that we do not have administrative data on these outcomes, we try to address this concern by calculating bounds on our estimates on parent-reported outcomes using Lee (2009)'s and Behaghel et al. (2009)'s approach.[22] The results of this exercise, shown in online appendix Table A3, indicate that bounds only include positive impacts on the final math grade (ITT-estimate = 0.852, bounds = [0.304,1.317]), whether the student passed the subject (ITT-estimate = 0.205, bounds = [0.143,0.358]) and only negative impacts on whether the student had to repeat the school year (ITT-estimate = −0.089, bounds = [−0.270,−0.060]). These results give reassurance that when taking into account sample attrition, the main conclusions from our analysis continue to hold.

In column 4 of Tables 10 to 12 we show the effect size estimates using inverse-probability weights for all non-academic results. Again, column 3 reports our main results for comparison. For each set of outcomes, the point estimates are very similar to each other.

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

7.2. Alternative specifications

The results shown thus far correspond to those using our preferred, full specification. In columns 1–3 of Tables 9 to 12, we additionally

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

Table 9
Robustness — academic outcomes.

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

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

Table 10
Robustness — self-perceived ability and affinity.

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

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at the individual level (Panels A–B) or heteroskedasticity-robust SE (Panels C–D) in parenthesis. Controls for specifications in columns 1–5 are as described in the notes to Table 9.

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

When looking at academic outcomes in Table 9, results are very stable across specifications, with effect size estimates mostly increasing as we add more controls to take into account heterogeneity at baseline across treatment and control group. For non-academic outcomes reported in the remaining Tables 10 to 12 the same holds: estimates are remarkably stable across specifications and so are their statistical significance levels. We therefore conclude that our main results are robust to the inclusion or exclusion of specific control variables.

Our preferred specification for outcomes where we have repeated measures at base- and endline is a difference-in-difference model. In Table 13 we check how the estimates differ when we estimate these results using Eq. (2) (post estimator with lagged dependent variables) instead of Eq. (1) (DID). At the bottom of the table we present the difference in the ITT-effect estimates between the two strategies and the p-value of a test for equality of the coefficients of the $Treat \times Post$ and the $Treat$-dummy across the two models. The effects are not substantially different using either method and most coefficients are very similar in magnitude across the two specifications. For all but one of the outcomes – the well-being index – we cannot reject the null hypothesis that the coefficients are equal across the two models. For one of our main outcomes, the standardized test score, the estimated effect is around 0.09 SD lower in the lagged dependent variable specification and becomes insignificant at conventional levels. However, looking at the predictive margins of treatment over different values of the baseline score shown in Fig. 5, estimated from a model that includes the interaction of treatment and baseline test score, we can see that the average effect masks substantial heterogeneity. For medium to high values (above 0.5) of the standardized baseline test score, the effect of treatment is large and significant at the ten percent level (and ranges up to half a standard deviation).

7.3. Volunteer tutors

We had initially planned a third treatment arm, where we wanted to compare the effectiveness of volunteer tutors with that of our professional tutors. As explained in Section 3.1, although we did not achieve

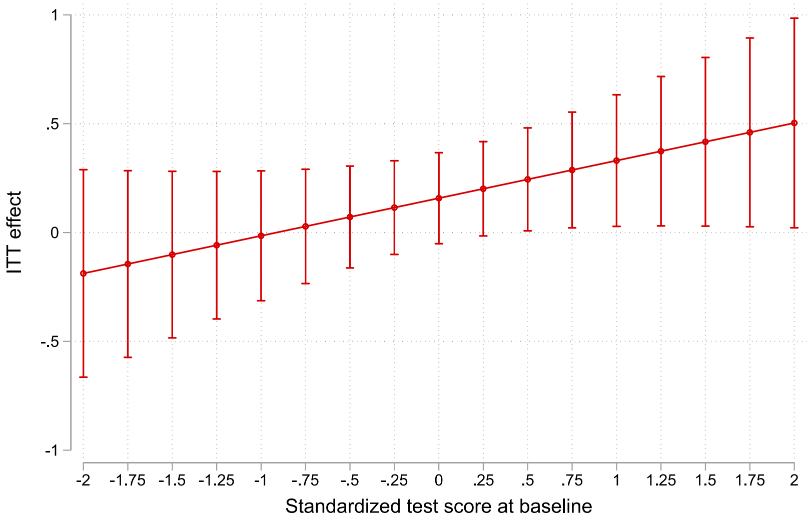

Fig. 5. Predictive margins of treatment assignment by baseline test score
Note: This figure shows the predicted standardized test scores post-treatment for treatment and control group students by standardized test score at baseline. These are estimated by regressing standardized test scores on a treatment dummy, the lag of the dependent variable, and the interaction between the two variables, plus the controls specified in the notes to Table 4 and block fixed effects. The plot shows 90% confidence intervals.

Table 11
Robustness — aspirations and motivation.

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

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. Heteroskedasticity-robust SE in parenthesis. Controls for specifications in columns 1–5 are as described in the notes to Table 9.

Table 12
Robustness — socio-emotional outcomes.

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

Notes: Significance levels are indicated by * < .1, ** < .05, *** < .01. SEs clustered at the individual level in parenthesis. Controls for specifications in columns 1–5 are as described in the notes to Table 9.

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

Table 13
Comparison of specifications.

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

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

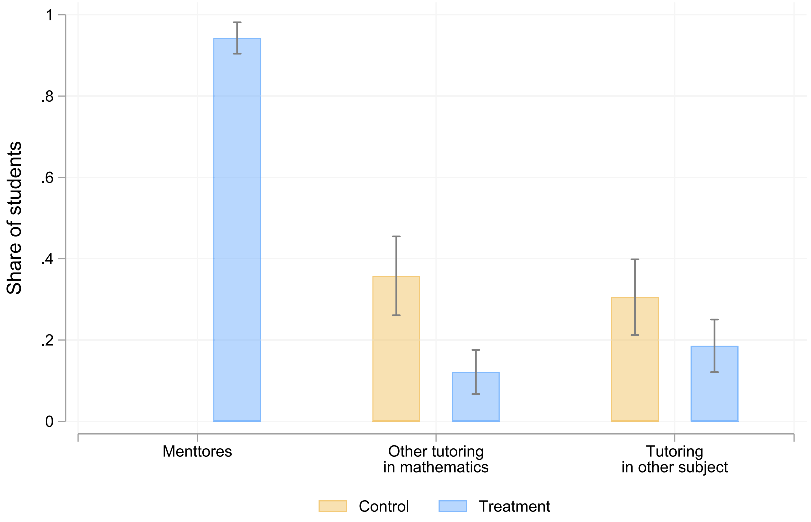

Fig. 6. Counterfactuals
Note: This figure shows the share or treatment and control group students who received (i) the *Menπores* program, (ii) another tutoring or academic support program in math, and (iii) another tutoring or academic support program in another subject. Sample of students whose parents responded to the endline survey.

7.4. Contamination of control group

In response to the pandemic school closures, many governments launched additional support programs to close learning gaps that emerged during lockdowns. Such competing programs were also launched in Spain around the same time as ours, which constituted a risk of contamination of the control group. To check whether this was likely a problem, during the endline survey we asked parents whether students had received any other tutoring or academic support program in math or other subjects during the period while *Menπores* was implemented. Indeed, as Fig. 6 shows, nearly 40 percent of control group students received some other tutoring or academic support in math, compared to only around 12 percent of the treatment group. The control group was also more likely to have received additional support in another subject, indicating that schools and/or parents might have compensated control group students with other offers. It is also possible that the process of initial identification of students in need for individualized support in math made parents more aware of the needs of their children and ended up prompting them to seek more support, especially if they were not selected for participation in Menπores. Overall, these findings suggest that our impact estimates could be interpreted as lower bounds.

7.5. External validity

Our program was specifically targeted at schools in disadvantaged areas, which means that our sample is not representative of the population of Spanish 7th and 8th grade students as a whole. To get a sense of how students at participating schools compare to our schools, online appendix Table A8 shows summary statistics of learning outcome indicators and socio-economic characteristics, separately for the entire population of schools in the region of Madrid and for those schools that participated in our experiment.[23] In column 1 we can see that the ESCS index, measuring student socio-economic status, of schools participating in the program is half a standard deviation below the regional average, approximately placed in the 25th percentile in the overall socio-economic distribution. Columns 2 to 5 show that students in participating schools were on average much lower performing, by between 73 (Spanish), 16 (Math), 90 (English) and 74 (Social subjects) percent of a standard deviation with respect to the regional average. Participating schools also have a lower overall share of children born to Spanish parents. Overall, students at our participating schools are on average lower performing and more disadvantaged than the average population of students in Madrid. This should be kept in mind when interpreting our results.

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

8. Conclusion

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

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

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

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

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

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

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

For future research, it will be relevant to explore in more detail the mechanisms driving our results: tutor characteristics and interactions with students, the type of training received or the number of students per tutor. It will also be important to explore whether the positive results of online tutoring shown here hold in different contexts: with primary school students, with variations in socio-emotional support or focusing on other subjects, such as reading. Also, it would be interesting to explore in more detail the potential benefits of small-group positive peer dynamics in online teaching. The remarkable academic effect of the program as well as the high attendance rates – the median number of completed sessions was 20 out of a target of 24 – indicate that our two-on-one design might have helped to mitigate some of the

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

shortcomings found in the literature in online education, such as a lack of perseverance and motivation (Escueta et al., 2020). Likewise, it would be interesting to explore the effect of introducing complementary technologies, such as adaptive software with high quality content, asynchronous interactions with tutors through chats or even more advanced AI bots, to support tutors in teaching and students in learning.

Declaration of competing interest

The authors declare that they have no known competing interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

The data and code for replication can be found at https://github.com/chupkau/menttores_replication.

Appendix A. Sample questions

A.1. Math test

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

- $3x + 5(x - 3) = 4x - 2(x - 5)$
  - □ $x = \frac{1}{2}$
  - □ $x = \frac{5}{6}$
  - □ $x = \frac{6}{25}$
  - □ $x = \frac{25}{6}$

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

- $x^2 + 2x - 15 = 0$
  - □ $x = -3, x = 5$
  - □ $x = 3, x = -5$
  - □ x does not belong to the set of real numbers
  - □ $x = 31, x = -33$

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

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

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

Locus of control. For each of the following questions, mark “Yes” or “No”:

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

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

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

Aspirations. What are your plans after you complete compulsory schooling?

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

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

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

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

□ Less than 15 min

□ 15–30 min

□ 30–60 min

□ 1–1.5 h

□ 1.5–2 h

□ 2–2.5 h

□ More than 2.5 h

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

Appendix B. Supplementary data

Supplementary material related to this article can be found online at https://doi.org/10.1016/j.jpubeco.2024.105082.

References

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

Angrist, Noam, Bergman, Peter, Matsheng, Moitshepi, 2022. Experimental evidence on learning using low-tech when school is out. Nat. Hum. Behav. 6 (7), 941–950.

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

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

Behaghel, Luc, Crépon, Bruno, Gurgand, Marc, Le Barbanchon, Thomas, 2009. Sample Attrition Bias in Randomized Experiments: a Tale of Two Surveys. Discussion Paper No. 4162, IZA Institute of Labor Economics.

Betthäuser, Bastian, Bach-Mortensen, Anders, Engzell, Per, 2023. A systematic review and meta-analysis of the evidence on learning during the COVID-19 pandemicdid students learn less during the COVID-19 pandemic? Reading and mathematics competencies before and after the first pandemic wave. Nat. Hum. Behav. 1–11.

Blainey, Katie, Hannay, Timo, 2021. The Impact of School Closures on Autumn 2020 Attainment. 2021 white papers, RS Assessment from Hodder Education.

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

Chalk, Karen, Bizo, Lewis A., 2004. Specific praise improves on-task behaviour and numeracy enjoyment: A study of year four pupils engaged in the numeracy hour. Educ. Psychol. Pract. 20 (4), 335–351.

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

Cooper, Harris, Charlton, Kelly, Valentine, Jeff C., Muhlenbruck, Laura, Borman, Geoffrey D., 2000. Making the most of summer school: A meta-analytic and narrative review. Monogr. Soc. Res. Child Dev. i–127.

Dee, Thomas S., 2005. A teacher like me: Does race, ethnicity, or gender matter? Am. Econ. Rev. 95 (2), 158–165.

Dobbie, Will, Fryer, Jr., Roland G., 2013. Getting beneath the veil of effective schools: Evidence from New York City. Am. Econ. J.: Appl. Econ. 5 (4), 28–60.

Duckworth, A.L., Quinn, P.D., 2009. Development and validation of the short grit scale (GRIT–S). J. Personal. Assess. 91 (2), 166—174.

Duflo, Esther, Dupas, Pascaline, Kremer, Michael, 2011. Peer effects, teacher incentives, and the impact of tracking: Evidence from a randomized evaluation in Kenya. Amer. Econ. Rev. 101 (5), 1739–1774.

Dweck, Carol S., 1986. Motivational processes affecting learning. Am. Psychol. 41 (10), 1040.

Dweck, Carol S., 1999. Self-Theories: Their Role in Motivation, Personality, and Development. Psychology Press.

Eisner, Manuel, Ribeaud, Denis, Sorrenti, Giuseppe, Zölitz, Ulf, 2020. The Causal Impact of Socio-Emotional Skills Training on Educational Success. Discussion Paper No. 13087, IZA Institute of Labor Economics.

Escueta, Maya, Nickow, Andre Joshua, Oreopoulos, Philip, Quan, Vincent, 2020. Upgrading education with technology: Insights from experimental research. J. Econ. Lit. 58 (4), 897–996.

Guryan, Jonathan, Ludwig, Jens, Bhatt, Monica P., Cook, Philip J., Davis, Jonathan M.V., Dodge, Kenneth, Farkas, George, Fryer, Jr., Roland G., Mayer, Susan, Pollack, Harold, Steinberg, Laurence, Stoddard, Greg, 2023. Not too late: Improving academic outcomes among adolescents. Amer. Econ. Rev. 113 (3), 738–765.

Haelermans, Carla, Jacobs, Madelon, van Vugt, Lynn, Aarts, Bas, Abbink, Henry, Smeets, Chayenne, van der Velden, Rolf, van Wetten, Sanne, 2021. A Full Year COVID-19 Crisis with Interrupted Learning and Two School Closures: The Effects on Learning Growth and Inequality in Primary Education. Maastricht University, Graduate School of Business and Economics. GSBE Research Memoranda No. 021.

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

Heckman, James J., Stixrud, Jora, Urzua, Sergio, 2006. The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. J. Labor Econ. 24 (3), 411–482.

Higgins, Steve, Kokotsaki, Dimitra, Coe, Robert, 2012. The Teaching and Learning Toolkit. Technical Report, Education Endowment Foundation and Sutton Trust.

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

Khattab, Nabil, 2015. Students' aspirations, expectations and school achievement: What really matters? Br. Educ. Res. J. 41 (5), 731–748.

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

Kosse, Fabian, Deckers, Thomas, Pinger, Pia, Schildberg-Hörisch, Hannah, Falk, Armin, 2020. The formation of prosociality: Causal evidence on the role of social environment. J. Polit. Econ. 128 (2), 434–467.

Kosse, Fabian, Tincani, Michela, 2020. Prosociality predicts labor market success around the world. Nature Commun. 11, 5298.

Kraft, Matthew A., List, John A., Livingston, Jeffrey A., Sadoff, Sally, 2022. Online tutoring by college volunteers: Experimental evidence from a pilot program. In: AEA Papers and Proceedings, vol. 112, pp. 614–618.

Lee, David S., 2009. Training, Wages, and Sample Selection: Estimating Sharp Bounds on Treatment Effects. Rev. Econom. Stud. 76 (3), 1071–1102.

Matthewes, Sönke Hendrik, Ventura, Guglielmo, 2022. On Track to Success? Returns to Vocational Education Against Different Alternatives. CVER Discussion Paper 038.

Newlove-Delgado, T., McManus, S., Sadler, K., Thandi, S., Vizard, T., Cartwright, C., Ford, T., Mental Health of Children and Young People group, 2021. Child mental health in England before and during the COVID-19 lockdown. Lancet Psychiatry 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, Oreopoulos, Philip, Quan, Vincent, 2020. 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.

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

Rimfeld, K., Kovas, Y., Dale, P.S., Plomin, R., 2016. True grit and genetics: Predicting academic achievement from personality. J. Personal. Soc. Psychol. 111, 780–789.

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

Santiago, Paulo, 2002. Teacher Demand and Supply: Improving Teaching Quality and Addressing Teacher Shortages. OECD Education Working Papers No. 1, OECD Publishing, Paris.

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

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

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

University College London, UCL Institute of Education, Centre for Longitudinal Studies, 2021. 1970 British Cohort Study Response Dataset, 1970–2016, SN: 5641. fourth ed., [data collection]. UK Data Service.

World Bank, 2021a. Accelerate Learning Recovery. Technical Report, World Bank Washington, D.C..

World Bank, 2021b. Remediating learning loss. Technical Report, World Bank Washington, D.C..