Apart but Connected: Online Tutoring and Student Outcomes *

Michela Carlana †, Eliana La Ferrara ‡

This version: September 2023. Latest version HERE

Abstract

The demand and supply of online tutoring has surged in response to the COVID-19 outbreak to address learning gaps. This paper evaluates the effectiveness of the Tutoring Online Program (TOP), which was first implemented in 2020 in response to the pandemic and later replicated in 2022, matching university tutors with disadvantaged middle school students. Using original survey data collected from students, parents, teachers and tutors, as well as administrative data, we find that the program substantially increased students' test scores (by 0.23 SD in 2020 and 0.16 SD in 2022) and decreased failure rate (by 24% in 2020 and 21% in 2022). This positive impact was particularly pronounced among students who received high-dosage and individualized tutoring, as opposed to small group instruction. TOP significantly improved students' socio-emotional skills, aspirations, and psychological well-being when implemented during school closure. Lastly, the program had a positive effect on the tutors themselves, enhancing their measured levels of empathy. This study sheds light on the valuable role that online tutoring programs can play in addressing educational disparities, while underlining the potential for positive impacts on both students and their mentors.

*We thank seminar participants at several universities and webinars for helpful comments. Micol Morellini, Vrinda Kapoor, Angelica Bozzi, Marco Cappelluti, Isabela Duarte, Agnese Gatti, Gaia Gaudenzi, Matteo Fossi, Federica Mezza, Chiara Soriolo, Amy Tan, and Monia Tomasella provided excellent research assistance. We are grateful to the schools that took part in the intervention for their collaboration, to the team of pedagogical experts guided by Giulia Pastori and Andrea Mangiatoridi and including Anna Maria Carletti, Paola Catalani, Silvia Negri, Doris Valente, Stefania Zacco, and Monica Zanon. We also thank our team of tutor supervisors (Angela Caloia, Leila Pirbay, Ilaria Ricchi, and Giulia Zaratti). La Ferrara acknowledges financial support from the Invernizzi Foundation. Carlana acknowledges support from the Malcolm Wiener Center for Social Policy at Harvard Kennedy School. La Ferrara and Carlana acknowledge support from JPAL PPE. AER RCT Registry ID: AEARCTR-0005607. The project has obtained IRB approval from Bocconi University and Harvard University.

†Harvard Kennedy School, NBER, CEPR, LEAP, and IZA (e-mail: michela_carlana@hks.harvard.edu).

‡Harvard Kennedy School, NBER, CEPR, and LEAP (e-mail: elaferrara@hks.harvard.edu).

1 Introduction

In response to the COVID-19 outbreak, schools have closed in over 190 countries (UNESCO, 2020; Stantcheva, 2022). School closure has created massive learning losses for children (Engzell et al., 2020; Lichand et al., 2022; Singh et al., 2022; Carlana et al., 2023b), estimated in around 0.14 standard deviations in achievement test scores and equivalent to 0.5 years of schooling (Patrinos et al., 2022). The pandemic has also had adverse psychological and social effects for children and adolescents, leading to higher depression and lower development of socio-emotional skills (Orgilés et al., 2020; Golberstein et al., 2020). The combination of these effects risks having long term consequences on the human capital of the cohorts affected by school closures.

To mitigate learning losses and accelerate learning, countries around the world have been promoting online and in-person tutoring (Robinson and Loeb, 2021; Carlana et al., 2023a; Parker, 2023). This paper reports the results of a novel policy experiment launched in Italy in 2020 and then replicated in 2022. In 2020, Italian schools were closed from the beginning of March until the summer – more than 1/3 of the entire school year. In response to this, our research team designed and implemented an innovative online tutoring program: TOP ("Tutoring Online Program"). The program targeted middle school students (grade 6 to 8) from disadvantaged backgrounds in terms of socioeconomic status, linguistic barriers, or learning difficulties, who were identified by school principals among those lagging behind during distance learning. The program was offered to middle schools from all over Italy on a voluntary basis, and it was completely free.

TOP has two defining features. First, tutoring is entirely *online*.While tutoring has shown promising results when done in person by teachers and paraprofessionals (Nickow et al., 2020), such mode of delivery was impossible during lockdown and it often makes it difficult to achieve the most disadvantaged students who live in areas where there is limited availability of qualified tutors. All interactions in our program occurs through personal computers, tablets or smartphones. Second, the tutors in TOP are not trained professionals but *volunteer university students*, trained and supported by pedagogical experts. While teachers and professionals are certainly qualified, the skills required for tutoring differ from those for classroom interaction (Cook et al., 2015). The choice of volunteer tutors has advantages in terms of budget (mobilizing resources to hire professionals may not allow for a rapid response and large scale implementation), and possibly also in terms of the quality of inter-personal interaction, as TOP leverages the intrinsic motivation of university students to be volunteers, as well as for the development of soft skills for tutors themselves.

For each round of implementation of the program, the demand from middle school students and their families could not be fully met due to budgetary and administrative constraints: for TOP2020, 530 was the maximum number of tutors to whom we could offer training and pedagogical support, while for TOP2022 we recruited enough tutors to support 607 students. In order to equip tutors with a basic set of pedagogical skills and to help them with potential problems in the relationship with children, we worked with a team of education experts to design an online self-training module for tutors and to hold regular group meeting and on-demand one-to-one sessions with expert educators. We collected baseline data from students, parents and tutors before the start of the tutoring, and follow-up data from students, parents, tutors and teachers at the end of the school year. Thanks to the over-subscription and random allocation of tutors to students, we can estimate the causal impact of the program not only on academic outcomes, bu also aspirations, socio-emotional skills, and psychological well-being.

We find sizeable and significant improvements for students who were assigned an online tutor compared to those who were not. Performance in a math standardized test that we administered at endline improved by 0.23 standard deviations (SD) for TOP2020 and 0.16 SD for TOP2022. Furthermore, the failure rate decreased by 24% in 2020 and 21% in 2022. The effects are particularly strong for math, which is the subject on which the majority of tutoring sessions focused, but positive also for Italian and English. These are remarkable effects given that the median length of the online tutoring was around 6 weeks.

Our experimental design allows us to assess more precisely the channels of the impact following the conceptual framework on the impact of tutoring interventions (Nickow et al., 2020). First, we focus on the "dosage" of the intervention. While the majority of students in TOP2020 received 3 hours of online tutoring per week, a random subset of those students who needed help in more than one subject (143 out of 427 treated students) was assigned a tutor who gave his or her availability for 6 hours per week. We find that performance gains double with the hours of tutoring and also persists one year after the intervention only for high-dosage tutoring. Second, we investigate the impact of tailored instruction by randomizing students to individual vs. small group tutoring in TOP2022. We do find that the effect is qualitatively substantially larger in terms of performance for individual tutoring. Finally, tutor/tutee relationship may go beyond academic content, but interact with it. We investigate the impact on three set of additional outcomes: aspirations, socio-emotional skills, and psychological well-being. We find extremely positive effects on these three dimensions only when TOP is implemented during school closure, with limited interactions allowed outside of the family environment.

In terms of demographics and socio-economic background, we do not detect significant differences in impact between boys and girls, nor immigrants and natives – except for the effect on psychological well-being which is entirely driven by immigrants in TOP2020. Interestingly, tutor characteristics such as gender, GPA, degree program and pro-social attitudes do not systematically affect the effectiveness of the tutoring.

Finally, we can estimate how the experience of being a TOP tutor during the pandemic affected the tutors themselves. We can do so because we randomly selected the university students to whom we offered the job from the pool of those who applied to be volunteers.[1] Four months after the end of the program, we find that volunteers who were included in the TOP program have significantly higher empathy than those who were not. The effect corresponds to a 0.27 SD increase. We instead do not find significant effects on tutors' beliefs regarding the relative role of luck versus hard work in determining success in life.

Our paper contributes to several strands of literature. A robust body of empirical work shows that in-person tutoring is highly effective for improving academic outcomes. Recent meta analyses find that the impacts are sizeable (a pooled effect size of 0.37 SD in Nickow et al. (2020)), and robust across a wide array of contextual factors (Fryer Jr, 2017). The importance of small group or individual tutoring has been underlined for students who struggle (Ander et al., 2016) and in order to teach at the right level (Banerjee et al., 2015). Also, the tutor-student relationships are often close to a mentorship connection that may affect the development of cognitive as well as social skills, such as prosociality (Kosse et al., 2020). On the other hand, tutoring is much costlier than classroom instruction and it may not be easy to arrange individual, in-person tutoring in the presence of geographical constraints. Also, tutoring may sometimes be attached with the stigma of being identified as a student in-need and pulled out from regular classes (Coie and Krehbiel, 1984; Richmond, 2015). We contribute to this literature by providing evidence on large-scale *online* tutoring, based on volunteer tutors supported and trained by pedagogical experts. Our model allows to substantially reduce the cost of tutoring –one of the biggest barriers to large-scale implementation– but also to efficiently reach students located in disadvantaged areas through virtual learning. Finally, online tutoring is less observable from peers than in-person tutoring, which may reduce the sense of stigma possibly attached with this intervention.

Our results are of course directly relevant to the debate on effective strategies to mitigate the effect of Covid-19 on education. The existing evidence suggests that students who lag

[1] As mentioned above, we could not accept all tutor applicants because we were constrained in the number of hours of support for tutors that we could pay for. The randomization was conditional on the characteristics that we use to allocate tutors to students, notably subject of tutoring, number of hours per week, and previous tutoring experience and training.

behind the most during the pandemic are from low-income families with limited access to technology, and that they receive less support from parents and lower quality of remote learning from schools (Bacher-Hicks et al., 2020; Chetty et al., 2020). Different forms of remote learning instruction have been adopted around the world and, although the evidence on interventions increasing access to computers and internet is mixed (Escueta et al., 2017; Malamud and Pop-Eleches, 2011), the impact of digital technology may differ during school closure compared to normal school years.

To the best of our knowledge, very few policy experiments have attempted to use remote tools to improve learning outcomes during the pandemic and that replicated after the school closure. Angrist et al. (2022) evaluate two low-tech interventions in Botswana that use SMS text messages and direct phone calls to support parents in the education of their children. The combined intervention resulted in a 0.12 SD improvement in student outcomes and led parents to update their beliefs about their children's learning level. Hardt et al. (2022) evaluate a remote peer mentoring intervention at a German university during the pandemic, where peers met online to discuss self-organization. They find positive effects on motivation and exam registration, though not on earned credits. Hassan et al. (2022) evaluate the impact of randomizing an over-the-phone learning support intervention on primary school students in 200 Bangladeshi villages. They found a positive effect on overall learning of 0.75 SD. In the US context, Kraft et al. (2022) conduct a pilot randomized controlled trial in which they assign a (volunteer) college student tutor to a middle school student. They find positive but statistically insignificant results. Our work contributes to this literature by evaluating the effects of an innovative and low cost online tutoring program targeting teenage students who had been adversely affected by school closures, and by showing impacts on learning outcomes as well as soft skills and psychological well-being.

Finally, recent work on organizations highlights the power of intrinsic motivation and social recognition for improving public service delivery (e.g., Ashraf et al., 2014; Gauri et al., 2019). In particular, Levitt et al. (2016) underline that such behavioral aspects can be leveraged to improve educational performance. While we cannot directly speak to this question, as we did not vary the recruitment method or the incentives provided to tutors, the fact that our tutors self-selected into volunteering for the TOP program and their intrinsic motivation may have contributed to the effectiveness of our intervention. Furthermore, we do provide evidence that volunteering as a tutor increased empathy compared to those university students that applied but were not assigned to a student.

2 Background

2.1 COVID-19 and School Closure in Italy

While many countries have tried to mitigate learning losses by switching to remote instruction and using asynchronous or synchronous platforms, the implementation of these tools has varied substantially. Even within the same country, schools in wealthier areas have shown higher prevalence of synchronous learning and online participation of students (Malkus, 2020; Chetty et al., 2020). High-income students also have access to better homeschooling inputs, including technology, help from parents (Agostinelli et al., 2022; Bacher-Hicks et al., 2020), which exacerbates educational inequalities.

Italy has been the first country after China hardly hit by the COVID-19 pandemic, with around 80,000 deaths as of January 2021, one third of which were concentrated in the region of Lombardy. All school buildings closed on March 5th, 2020. Since then, school reopening has been repeatedly postponed until September 2020 and, even during the school year 2020-21, many schools had to offer temporary remote learning instruction, depending on regional or local outbreaks. As in most counties, the infrastructure for remote school learning substantially improved during the pandemic. In the Spring 2020 the Italian Ministry of Education allocated 70 million euro to buy tablets that students could temporarily borrow and 10 million to improve internet connection and online platforms of schools. This intervention facilitated access to devices and internet for disadvantaged students, but did not reach all students in need.

As of March 2020, teachers' digital competencies were still somewhat limited, with less than 50 percent of the teachers using any digital tool in their daily lectures (Agcom, 2019). When schools closed due to COVID-19, the response from teachers was extremely heterogeneous, but by the end of the Spring 2020 most students in Italy received some synchronous interactions with teachers.

3 Intervention and Experimental Design

3.1 The Tutoring Online Program (TOP)

3.1.1 TOP 2020: Project Description

Two weeks after the school closure in Italy, we started the process to design and implement a new program, the "Tutoring Online Program" (henceforth, TOP), as an attempt to provide immediate response to the emergency situation. TOP has two defining features.

First, tutoring is entirely *online* through personal computers, tablets or smartphones. Second, the tutors in TOP are *volunteer university students*, trained and supported by pedagogical experts. TOP 2020 included support for remedial education of either 3 hours or 6 hours per week, randomized at student level.

Schools self-select to participate in the program: they answer a brief baseline survey expressing their interest and send us a list of students including up to three pupils for each class who "may need TOP the most in terms of their learning level and family environment". For each child, the form indicates the preferred subjects of the tutoring (one or more among math, Italian, and English), and contact details of parents and teachers. We clarified that we could not guarantee the tutoring to all applicants and that, if the number of requests exceeded the number of tutors that we could mobilize and support, we would randomly assign tutors to students, in order to give every applicant the same chances.

Out of 1,594 students identified as in need of the program from 76 schools, 1,059 students and their parents completed the baseline survey by the end of enrollment period.To assess the representativeness of our self-selected sample of schools, in Appendix Table A.I we compare the provinces with and without schools that took part in the TOP program. While the breakdown by education level of the population is quite comparable (the differences are significant but extremely small), schools in North of Italy – the areas strongly hit by the pandemic early on – were more likely to apply, most likely as they foresaw that schools would have not reopened until the end of the year.

We recruited volunteer university students as tutors through the of three large Italian universities in Milan.[2] Applicant tutors had to complete a baseline survey, indicating among other things the subjects in which they would feel comfortable tutoring, and their time availability. The number of applications from volunteers reached 2,000 by the end of the enrollment period, far exceeding our expectations. As our volunteers were not trained professionals, we hired a team of pedagogical experts to train and support the tutors.[3] Within a few weeks, they set up an online learning platform with a self-training program that included slides and videos. The topics included: how to approach students; tools and online platforms for effective online tutoring; learning disorders; and tips to help students in math, Italian and English. The platform also included a supervised forum where tutors could ask questions and share their experiences. Finally, the pedagogical team organized regular group meetings with around 20 tutors, as well as one-on-one meetings on demand

[2]The three universities were Bicocca, Bocconi and Statale, which approximately enroll 33,000, 14,000 and 61,000 students respectively.

[3]The team was led by prof. Giulia Pastori and prof. Andrea Mangiatordi , both at Bicocca University .

to offer support in specific circumstances.[4] Despite being the most expensive component of TOP (see section 8 for cost estimates), the tutor training was an important component of TOP and it ensured that our volunteers, even without professional training, could offer a high quality support to their tutees and could receive professional advice and support in case of need.

The implementation of TOP in 2020 was entirely done by the research team. The timeline of the project is include in Appendix Figure A.1. The tutoring lasted from mid-April to the end of May 2020.[5] In June 2020, we collected the endline surveys. After each meeting, tutors were required to record some information about the session using a management tool, including whether there were issues with the connection and whether he/she had exerted effort during the tutoring session.

On average, treated students had 14 tutoring meetings over the course of the program, for a total of 17 hours over 34 days. The distribution of the number of meetings and tutoring days is presented in Appendix Figure A.2. Less than 5 percent of students chose not to start the tutoring (hence have zero meetings). During the tutoring, the subject covered by the great majority of students was math, which was covered by 78 percent of the students. The entire distribution of subjects covered in the meetings, as reported in tutors' registries, is displayed in Panel A of Figure 1.

3.1.2 TOP 2022: Project Description

Most of the aspects regarding the program descriptions are identical compared to TOP 2020, with few exceptions. First, both the tutor training and the tutoring sessions happened within a unique platform adapted for our specific project (WeSchool). Second, all students were offered three hours per week of tutoring in up to two subjects between math, Italian, and English and we randomized whether students were exposed to individual or group tutoring. Third, the program was implemented by an NGO supported by the research team with the objective of bringing the program to scale in the medium term.[6]

The timeline of the implementation in 2022 is available in Panel B of the Appendix

[4]Based on the information reported by our tutors at endline, around 80 percent of them used the training platform, 50 percent watched the videos and followed the online training, 8 percent used the forum, and 36 percent and 12 percent joined at least one small group or individual meeting, respectively.

[5]Some tutors voluntarily decided to support the students during the summer and in the following academic year.

[6]The Bank Foundation that financed and helped organizing the program from 2022 is Cariplo Foundation and the implementing partner of TOP is CIAI (Centro Italiano Aiuti all'Infanzia). Given the geographic focus of the Foundation financing the project in 2022, schools in TOP 2022 were located in Lombardy and part of Piedmont (Verbania e Novara).

Figure A.1. TOP 2022 involved other students compared to TOP 2020. The recruitment of schools and students was managed by the implementing partner following a protocol similar to TOP 2020, but focusing on the provinces in the North in the geographic focus of the partner Foundation. University students were recruited as tutors with a campaign similar to TOP 2020, but expanding the number of universities involved.[7] The training of tutors was led by the same team of pedagogical experts as in TOP 2020, in close collaboration with supervisors from the implementing partner team.

On average, treated students had 15 tutoring meetings over the course of the program, for a total of 19 hours. During the tutoring, the subject covered by the great majority of students was math, which was covered by 63 percent of the students, is displayed in Panel B of Figure 1.

3.2 Randomization

The randomization proceed in two steps. First, for each wave of the program, students in our sample were divided into two groups: a treatment group that received tutoring and a control group that did not. In order to guarantee that students could start as soon as possible, we processed applications on a rolling basis by creating 'blocks'. We stratified the randomization at the block level, where blocks were created depending on the timing of baseline completion.[8] Second, in each wave we randomize an additional treatment to investigate more deeply the mechanisms and the identify cost-effective interventions. For TOP2020, among the 530 treated students, teachers identified 427 as needing help in more than one subject. We randomly assigned one third of these 427 students to an 'intense' version of the program with 6 hours of tutoring per week instead of 3. This will allow us to estimate the impact of higher treatment dosage. For TOP2022, among the 607 treated students, we randomized 265 to group tutoring with other students in the same class and/or school-grade. This additional randomization will allow us to study the effectiveness of targeted individual instruction compared to the small group setting. Given the logistical concerns in setting up the meetings between the volunteers and the group of students, we decided to create groups of only two students.

[Insert Table I]

[7]The universities involved include also Cattolica University, Politecnico di Milano, Insubria University, IULM, University of Pavia, University Carlo Cattaneo, University of Brescia, on top of Bicocca, Bocconi and Statale, approximately with a total of XX students.

[8]In TOP2022 we had an extremely high number of applications from families, compared to the number of tutors available. Hence, we randomly selected a sample of children to be included in the experimental randomization based on the supply of tutors per subject. Then, we perform the standard randomization within 'blocks' as in TOP 2020.

Table I shows that the treatment and control groups do not differ according to baseline characteristics from administrative data and data collected from students and parents, including gender, immigration status, learning disorders, interest in the subjects taught, and parental education and occupation. In the Appendix, we provide an additional set of balance tables, comparing also high vs. low tutoring dosage for TOP2020 and group vs. individual tutoring for TOP2022 (Appendix Tables A.III and A.IV). The table shows that the samples are overall well-balanced, with only limited number of statistically significant differences consistent with random sampling. We also control for these baseline characteristics in all regressions.

Attrition rates were different for the treatment and control group, which is not surprising given that students who received a tutor remained engaged with the program, while control students had to be contacted after not receiving a tutor. As shown in Appendix Table A.VII, on average 91 percent of students in treatment and 60 percent of students in the control group completed the endline test score. [9] Compared to students in grade 8, younger students were more likely to complete the test score, as during the period of the endline, grade 8 students were involved in the final middle school exam, and control students may have been relatively less motivated to devote time to the survey. While there are no other substantial differences in the sample who completed the endline survey in 2022, in 2020 we find that children from families with higher level of education or with higher familiarity with computers are slightly more likely to complete the endline test, with the effect being driven mainly by students in the control group (column 4, Appendix Table A.VII). Appendix Tables A.V and A.VI reports the balance table restricting the sample only to students who completed the final survey in 2020 and 2022, respectively. Overall, most characteristics are balanced between treatment and control group. If anything in 2020, compared to the full sample shown in Appendix Table A.III, the control group is positively selected in terms of parental education (as highlighted above). Given the direction of imbalance in response rates, one may expect an underestimate of the treatment effect. Nonetheless, we will present different robustness checks, including inverse probability-weighted estimates of treatment effects and the inclusion of different sets of controls when the outcomes come from survey data (Appendix Table A.XXI). The administrative outcomes do not suffer from the issue related to attrition.

Finally, Appendix Table A.IX and A.X compares the sample with the characteristics of the tutor and tutees in TOP2020 and TOP2022.

[9] As shown in Appendix Table A.VII, 46 and 73 percent of the control group and 88 and 95 percent of the treatment group, respectively in 2020 and 2022.

Tutor allocation

We assigned tutors to students following a step-by-step procedure. First, we restricted the sample of tutors to those currently enrolled in university and fluent in Italian. Second, we divided tutors into different groups depending on their expertise in the various subjects (math, Italian, English or combinations of these). Within each group, we randomly assigned tutors to tutees.[10]

While for TOP2022 we assigned all available tutors, for TOP2020 some volunteers were not matched. Appendix Table A.XI reports the differences in characteristics of volunteers between those assigned and not to a student and Appendix Table A.X compares the characteristics of tutors in the two waves.[11]

The great majority of tutors are female (70 percent in 2020 and 73 percent in 2022), born in Italy (98 percent in 2020 and 10 percent in 2022), they were moved by a desire to help others when applying to TOP (and they have previous experience as volunteers. In terms of degree programs, about 34 percent of the tutors in 2020 and 26 in 2022 attend a STEM major or medical school, around 28 percent an economics/business major in both years, 14 percent in 2020 and 42 percent in 2022 a humanities major, and only 7 percent and 2 percent a major in education, respectively in 2020 and 2022.

Appendix Table A.XI and A.XII also show summary statistics separately for tutors that offered their availability for 3 vs. 6 hours per week in 2020 and group vs. individual tutoring in 2022. Tutors who made themselves available for 6 hours are less likely to come from economics or business and more likely to come from humanities; they are also more likely to be born outside Italy and to have training to work with immigrants. Although students are randomly assigned to a high vs low-intensity treatment, we should keep in mind that they get a 'package' of different tutor characteristics when assigned to 6 vs. 3 hours of tutoring.

3.3 Data: Surveys and Administrative Data

We build a unique dataset merging the baseline surveys of parents, students, and tutors with endline data coming from (i) the results of a standardized test administered by us

[10]Given the high number of volunteers in 2020, we decided to further restrict the sample to tutors with previous tutoring experience and/or specific training (e.g., to support students with learning disorders or immigrants). In 2020, we also considered the time availability (3 vs. 6 hours per week) while in 2022 we considered the availability for group tutoring. We also considered the special training for immigrants and for students with learning disorders, but this was a very limited number of tutors (less than 5% in both years).

[11]Notice that 530 students were assigned to the treatment in TOP2020, but 7 dropped out before starting the tutoring, therefore we only assigned 523 tutors.

and taken by the students, and (ii) surveys of parents, students, tutors and teachers of the classes in which our treated and control students were enrolled (iii) administrative data from the Italian Ministry of Education including teacher assigned grades. We report the summary statistics of our main outcomes measured at endline in Appendix Table A.II.[12] The survey questions and more details on the outcomes are reported in the data appendix online (Appendix B).

On top of the baseline information, we asked all the volunteers that had applied to be tutors to complete a very short endline survey in September 2020, six months after the start of the program. Almost all the tutors who were recruited into TOP completed the endline survey, while only around one third of those who were not assigned a student did so.[13] Appendix Table A.XI shows the difference in observable characteristics among the tutors who participated in TOP ('treated' tutors) and the others ('control' tutors). Once we account for the criteria used to assign students to tutors (e.g., tutors from STEM are over-represented in treatment because math was the subject most in demand by the students), very few significant differences appear. This will allow us to investigate how participation in TOP affected some outcomes measured at the tutor level.

The first outcome is empathy. We collected two standard questions on a 4-point likert scale, asking respondents if they (i) "find it easy to put themselves in somebody else's shoes"; and (ii) "are able to make decisions without being influenced by people's feelings". We sum all points and divide by 8 to obtain a variable ranging from 0 to 1.

The second set of outcomes concerns views on the role of hard work and effort to achieve success in life. The index we build aggregates answers to three separate questions on (i) income differences and effort; (ii) the importance of hard work versus luck and connections; and (iii) the prospects of getting a well-paid job after studying hard, independent of family background. We aggregate the variables in a similar way as described above.

3.4 Empirical Strategy

To assess the impact of TOP on the various outcomes we collected, we estimate the following OLS regression:

$$Y_{ir} = \alpha_r + \beta Treated_i + \gamma X_i + \varepsilon_{ir} \quad (1)$$

[12]The variables labeled as 'outcomes reported by parent' or 'outcomes reported by teachers' do not refer to parents or teachers themselves, but to the answers that parents/teachers gave about a given child.

[13]TOP tutors received a longer questionnaire in June that included the questions on their experience during TOP, and then again in September the same short questionnaire that control students received.

where Y_{ir} is the relevant outcome for student i who was assigned to treatment or control in randomization round r; α_r denotes randomization round fixed effects; $Treated_i$ is an indicator for whether the student was assigned a tutor in the TOP program; X_i is a vector of student level controls measured at baseline, including: gender, immigrant status, grade in which the student is enrolled, mothers' and father's education, mother's and father's employment type, learning disability, interest in the different subjects, perseverance, belief on the importance of luck, and familiarity with computers; ε_{ir} is an error term. We estimate robust standard errors. We also correct for multiple hypothesis testing using the Westfall-Young stepdown adjusted p-values, which also control the family-wise error rate (FWER) and allow for dependence amongst p-values.

4 Students' Academic Performance

In TableII we study the impact of TOP2020 on academic performance in math. In column 1 of Panel A, we find that the teacher-assigned grade increased by 0.12 points for treated students, which corresponds to an increase by 0.11 SD compared to the control group. The failure rate, i.e. the probability of getting a score below 6, decreased by 4.3 percentage points for treated students, corresponding to a 24% reduction compared to the control group mean (column 3, Panel A). Finally, column 5 of Panel A shows that TOP2020 improved the performance in the standardized test score of treated students by 0.23 SD compared to the control group. This is an impressive result if we take into account two factors. First, the median duration of tutoring was five weeks. Second, tutors did not specifically prepare the students for this type of test (multiple choice tests are not typically assigned as homework in Italian schools), but rather focused on helping students find a method for studying and doing regular homework. The effects are significant at conventional levels: in square bracket, we report the p-values adjusted for multiple hypothesis testing using both FWER adjusted p-values (based on 10,000 replications) and sharpened q-value.

Finally, while the administrative grades are available for the entire sample, the endline test score is available only for part of the sample. Hence, as a first robustness, columns 2 and 4 of Table II reports the impact on grades and failure rate for the sample restricted to the same observations as the test score. The results are qualitatively and quantitatively unchanged.

[Insert Table II]

Panel A of Table III reports symmetrically the same results analyzing the impact of

TOP2022. Overall, the average impact of the program on the math performance was similar: it improved teacher-assigned grades by 0.09 points corresponding to an increase by 0.09 SD compared to the control group, it decreased failure rate by 0.4 percentage points or 21% compared to the control group, and it improved performance in the standardized test score by 0.16 SD compared to students who did not receive the support from the program. The effect are statistically significant at conventional level, with the exception of the teacher-assigned grades.

[Insert Table III]

There are two key differences between TOP2020 and TOP2022. First, in-person interactions were limited and school were remote in 2020, while students interacted in-person with their teachers and peers in the school building in 2022. Second, conditional on being assigned to the treatment, students were randomized to three vs. six hours of individual tutoring in 2020 and group vs. individual tutoring in 2022. Panel A of Tables II and III shows the overall effect of the program. To investigate the differential impact of online tutoring during and after the pandemic, we disentangle the effect on three hours of individual tutoring in Panel B of both tables with the coefficient on *"Treatment"*.[14] To ease the interpretation, Figure 2 reports the impact of three hours of individual tutoring in standard deviations on the three academic outcomes, during and after the COVID-19 pandemic.

The three hours of individual tutoring have an impact statistically indistinguishable during and after the school closure due to the pandemic on the academic outcomes. In terms of performance in the standardized test score with an impact around 0.15-0.21 SD and it is statistically significant at 5 percent level for both editions. The effects on grades and failure rate are slightly smaller in magnitude compared to the test score with effects between 0.07 and 0.15 SD. By splitting the sample and focusing only on the three hours of individual tutoring we loose part of the significance, but the results overall qualitatively suggest a positive impact on all these dimensions. This impact is comparable in magnitude to the average impact of large-scale in-person tutoring interventions, as reported in the meta-analysis by Nickow et al. (2020).[15]

[Insert Figure 2]

[14] We dig deeper into the different effect of dosage and customization of learning in the mechanisms section 5.

[15] We consider large-scale tutoring interventions involving more than 400 observations and implemented mainly by non-professional tutors. Compared to our intervention, most of the previous tutoring experiments that were causally evaluated and included in the meta-analysis by Nickow et al. (2020) focus on elementary school children.

The results so far represent the impacts on the academic outcomes in math given that around 70% of students focused on this subject during the tutoring sessions and the share is comparable for both editions (see Figure 1). Between 30 and 55% of students received support from their tutoring in Italian and English. In Panel A and B of Appendix Tables A.XIII and A.XIV, we report the effect on grades, failure rate, and performance in standardized test scores in all subjects. Impacts are positive across the board, but slightly smaller in magnitude and less precisely estimated. This is not surprising, as most students focused on math during the tutoring sessions.

5 Evidence on the Mechanisms

In this section, we build on our conceptual framework to explore the mechanisms through which online tutoring may improve learning outcomes. We explore specific design features of our program to investigate the causal effect, as well as heterogeneity on students' outcomes based on baseline characteristics.

5.1 Instruction time and tutoring dosage

One of the most prominently considered mechanisms in the literature is the increase in instruction time (Nickow et al., 2020; Fryer Jr, 2017). As described in section 3.2, the experimental design of TOP2020 allows us to investigate how the impact of the program varies with the intensity of the tutoring. Among all the students who needed help in more than one subject, we randomly chose which students would be matched with a tutor who had offered to help for 3 versus 6 hours per week.[16]

We start by assessing how participation in the 3 versus 6 hours program affected the time devoted to homework and the quality of their participation in regular (online) classes offered by their schools. Figure 4 shows that being assigned to 6 hours of individual tutoring per week significantly increases the time devoted to homework as reported by students in our endline surveys. The effect of 3 hours of tutoring is modest, both in TOP2020 and in TOP2022. Appendix Table A.XVI quantifies this effect and consider a broader set of outcomes which includes not only homework, but also attendance to online classes. For each outcome we report both the overall treatment effect and standard errors (columns 1 and), the impact of the 3 hours of tutoring and the standard error (column 4

[16]Overall, 27 percent of the *treated* students received a tutor for 6 hours per week (the mean of this variable in the full sample of treated and control students is 0.135). 427 students out of the 530 treated were in need of help in more than one subject, and 1/3 of them received the 6-hours tutoring.

and 5), and the additional impact of the intense tutoring with its standard error (column 6 and 7), including the same specifications as in the main Table II.

[Insert Figure 4]

Consistent with the data in Figure 4, we find that treatment increased the time devoted to homework, as reported by both students and parents: the average effect is about $17-19$ minutes per day for the intense tutoring and $6-7$ minutes per day for the 3 hours of tutoring. Also, the regularity of homework completion as reported by teachers is significantly higher for treated students, but does not vary with the intensity of the tutoring. This evidence is consistent as the objective of the program, both for the tree and six hours of tutoring, is helping students in doing their homework.

During the lockdown, classes were offered online and students were supposed to connect every day and attend them. Compliance with this requirement was not always full, though. In Panel B of Appendix Table A.XVI, we find that the probability of regularly attending online classes, as reported by the children and the parents, respectively, is uncorrelated with treatment. This is not true, however, when we consider teachers' reports (column 8): students in the TOP program are 10.1 percentage points more likely to attend classes regularly – a 18 percent increase over the control group mean. The discrepancy is not surprising if one observes the difference in average values of the dependent variable reported by the three categories of respondents: children and parents report regular attendance in 83 and 88 percent of the cases, respectively, while –for the same student– teachers report it only in 57 percent of the cases. It is possible that reporting bias by children and parents may introduce too much noise for us to detect a treatment effect.

Overall, these results indicate that instruction time was substantially higher for students exposed to the intense tutoring, while students exposed to the three hours of tutoring per week had some positive but more limited improvement in the interaction mainly coming from teachers during school hours.

In Panel B of Table II, we regress our main academic outcomes on the treatment dummy and on an indicator for whether the student got a 6-hour tutor ("Intense Treatment"). Column 1, 3 and 5 shows that the impact of the standard, 3-hours/week version of the program is a 0.11 point increase in grades, a 2.7 percentage points reduction in failure rate, and a 0.16 standard deviations increase in the performance in test scores, respectively. Having a tutor for 6 hours a week further improves the outcomes for students, leading to an overall improvement by 0.14 points in the teacher-assigned grades, a reduction by 8.5 percentage points in the failure rate, and an improvement by 0.44 SD in the test score.

This is a remarkable effect, but the magnitude of the effect of the intense treatment

compared to the less intense treatment is not surprising in light of the results on the instruction time, but also the meta-analysis in Nickow et al. (2020): they show that the average effect of tutoring on learning almost doubles going from 1-2 days per week to 4-5 days (from 0.24 to 0.41 standard deviations).

5.2 Targeting Instruction: Group vs. Individual Tutoring.

One of the other leading mechanisms through which tutoring may impact students' academic outcomes is the customization of learning. A well-established literature has underlined the importance to "teaching at the right level" to increase effectiveness of instruction (Banerjee et al., 2015): teachers have more limited ability to support students if they span a wide range of skill levels and they need support in different subjects.

TOP2022 investigates the impact of being randomly assigned to individual vs. group tutoring. The groups were formed to ease the customization of learning for the tutor: students were matched in pairs with a schoolmates in the same grade who need help in the same subjects according to their own teachers. The intensity of the tutoring was 3 hours for both individual and group tutoring in 2022. Despite some higher scheduling challenges for tutors, as shown in Panel B of Figure 4, the difference in the time devoted to homework is minimal for group vs. individual tutoring.

Panel B of Table III reports the impact of group tutoring. Summing up the coefficients of *"Treatment"* and *"Group Treatment"*, we observe that the impact on students' outcomes is overall positive on their math academic outcomes, but small in magnitude and not statistically significant at conventional levels. The impact on the test score is 0.21 SD for individual tutoring (p-value= 0.03), while for group tutoring 0.08 SD (p-value= 0.262). While the impact on grades is statistically indistinguishable from zero for both treatments, the reduction in failure rate is 7 percentage points for individual tutoring (p-value=0.022), while for group tutoring it is 5.7 percentage points (p-value=0.105).

Overall, the results suggests that online tutoring is more effective when students receive personalized individual support with a greater degree of focus compared to group interactions. Despite a constant "quantity" of instruction time, "quality" of the feedback received may vary in an individual and group setting leading to less positive outcomes without individual customization of learning.

5.3 Mentorship: impact on aspirations

The human connection generated by the tutoring can be an additional channel inducing students to improve their academic outcomes. Mentoring relationships have been shown

to have positive long-term implications for children's educational outcomes and life trajectories (Resnjanskij et al., 2023; Kosse et al., 2020; Falk et al., 2023). However, it is unclear whether similar positive impacts may be generated when students are connected with a tutor in an online setting.

[Insert Table IV and V]

In column 1 of Table IV and V, we estimate the impact of the program on students' aspirations and perceived ability to achieve educational goals in TOP2020 and TOP2022, respectively. The overall effect points toward a positive impact on aspirations when the online tutoring happens during the pandemic with limited interactions of students outside their home. The aspiration index improves of 0.162 SD (p-value = 0.05) for individual tutoring and 0.264 SD (p-value = 0.026) compared to the control group. The first two bars of Figure 3 compare the effect of three hours of online tutoring during the pandemic and when schools are open. The effect is smaller in magnitude and not statistically significant at conventional levels when the program is implemented during normal school activities, suggesting that their aspirations were more likely to get shaped by in-person interactions when possible.

[Insert Figure 3]

Panel A of Appendix Table A.XV and A.XVIII estimate the impact on each question used to create the aspiration index, with information reported by the students, their parents, and their teachers. The direction of the effects are in line with the results presented above, with positive impacts of TOP2020 and small and insignificant effects of TOP2021. Starting from long-term goals, TOP2020 students and their parents appear more likely to report that in the future they plan to enroll in university and teachers are more likely to say that they should do so. A similar consideration applies to self-efficacy: we asked students and their parents, aside from their intentions, how much they thought the student would be *able* to attend university. Perceptions here are low on average (only 21 percent of students and 29 percent of parents in the control group responded in the affirmative). Finally, a more immediate choice for our students (in terms of time horizon) concerns the high school track in which they plan to enroll after middle school: vocational, technical or academic. Treated students are 6.5 percentage points less likely to say that they plan to attend the least prestigious track, that is, vocational. The effect corresponds to more than 20 percent decrease compared to students in the control group. When adjusting for multiple hypothesis testing, the effects are not statistically significant at conventional levels. We collect the same outcomes for TOP2022 and find smaller impacts in magnitude and indistinguishable from zero on all dimensions.

5.4 Impact on socio-emotional skills and beliefs

We next test whether the program affected students' socio-emotional skills, in particular their reactions in the face of obstacles and their perceived ability to control what happens in their lives. This could be another important channel through which tutoring online may affect long-term educational achievements of students.

Column 2 of Table IV and V reports the main results using a summary index of socio-emotional skills as described in section B.1. The evidence suggests a positive impact of online tutoring on socio-emotional skills when implemented during school closure and no effect in 2022, as also clearly emerges from Figure 3 comparing the three hours of individual tutoring.

Next, we dig deeper into the different components of the socio-emotional index to understand the driver of such result. To measure perseverance we collect two set of measure: (i) the index of 'grit' proposed by Duckworth and Quinn (2009) (ii) we gave students a logic task and after they completed it, we asked if they wanted to get another one of the same level of difficulty, a more difficult one, or if they wanted to stop. Panel B of Appendix Table A.XV and A.XVIII shows that the effect is statistically insignificant and quantitatively negligible

While perseverance and grit do not appear significantly affected by the interaction with a tutor, students' 'locus of control' does when in-person interactions are limited (p-value adjusted for multiple hypothesis testing $= 0.07$). Students in TOP2020 believe to a greater extent that they (rather than fate or luck) can control the outcome of events in their lives. The magnitude of the treatment effect corresponds to a 0.19 SD increase over the control group mean.

A possible interpretation of this finding is that when students who work with a tutor realize the positive results on the academic front, they understand that success in school is not a matter of luck. To investigate the difference between the impact on the locus of control for TOP2020 and TOP2022, we analyze how beliefs on own ability are affected in the two editions of TOP. The results are reported in Panel C of Appendix Table A.XIII and A.XIV. Despite similar effects on the performance in the standardized test score, students only during the pandemic update their beliefs on their own ability, with strong results robust also after accounting for multiple hypothesis testing. Students in TOP2020 may have then extrapolated this belief to life in general as emerged from the "locus of control". Students in remedial education programs are often exposed to strong negative stereotypes from teachers and peers on their ability as emerges from anecdotical evidence from TOP supervisors, as well as from text analysis of students' baseline answers on their

relationship with school. The isolation from the pandemic may have shielded them from the negative pressure in schools allowing them to update their beliefs on their own ability.

In Panel B of Appendix Table A.XIII and A.XIV, we estimate the effect of the program on the beliefs held also by parents and teachers about the number of correct responses given by students in the test. While teachers positively update similarly their beliefs in both editions reflecting actual improvements of students' performance, parents' beliefs mirrors the same pattern we observe for students' own beliefs with an improvement during the edition of TOP2020 and no effect for TOP2022.[17]

5.5 Mental health: impact on well-being

An important goal of our program, in addition to the academic component, was to help students navigate the psychological difficulties that the lockdown and isolation due to Covid-19 may have created. The tutor represented, among other things, someone to talk to outside one's own immediate family –a different voice and a connection with the outside world.

The last column of Table IV and V reports the main results using a index of well-being of students, including measure of depression and happiness (details on the creation of the index in Appendix B.1). Consistently with the results reported above on the "soft" components, there is a positive and significant effect on the well-being of students from the interaction with a tutoring during the lock-down but no effect during normal interactions as shown also in the last two bars of Figure 3 comparing three hours of online tutoring in the two editions.

In Panel C of Appendix Table A.XIII and A.XIV, we estimate the impact of the program on two measures of psychological well being: Frühe et al. (2012) Children's Depression Screener and a self-reported index of happiness.[18] While the mean of both measures in the control group is comparable in TOP2020 and TOP2022 reflecting a long-lasting mental discomfort associated with Covid-19, only students who participated in the program during the lockdown report fewer symptoms of depression. The positive effect of TOP2020 is qualitatively similar but insignificant when reported by the parents. Correspondingly, happiness increases: this time the effect from parents' responses is more precisely estimated (but not very different in magnitude from that based on students'

[17] We collected also an alternative measure of beliefs for both students and teachers. We asked them how they would rank themselves/the student compared to their classmates, on a scale from 1 to 10. The 10-point scale is akin to the grading scale used in Italian schools, where 6 indicates a pass. We see that the treatment effects on these dimensions are similar compared to the beliefs in the performance in the test score further corroborating our evidence.

[18] Both variables have been normalized so that they range from 0 to 1, as explained in Appendix B.1.

answers).

These results suggest that TOP played an important role in mitigating potential mental health problems associated with the pandemic and with the strict regime of lockdown, but this may not be the driving channel in improving learning outcomes of students as the positive impact on academic performance emerges even without such mental health relief.

6 Additional Results and Robustness

6.1 Heterogeneous Treatment Effects and Causal Forest

Appendix Table C reports the heterogeneous treatment effects based on baseline characteristics of students and their tutors.

To complement the above analysis with a more systematic approach, we estimate heterogeneous treatment effects using a causal forest algorithm. We follow Athey and Imbens (2016) and Wager and Athey (2018) and apply their method to understand who benefits most from the tutoring. Online Appendix C.4 describes our methodology in detail.

In a nutshell, we estimate the Conditional Average Treatment Effect (CATE), including in the causal forest demographics (e.g., gender, immigrant dummy, parental education and occupation, etc.) and other controls (e.g., school grades and interest for different subjects, familiarity with computers, etc.). We use the predictions on the expected treatment effect for each individual, given the covariates, to investigate treatment heterogeneity. We divide the sample in two groups: top and bottom half of the predictions.

Appendix Table A.XVII reports the mean of each baseline characteristics for the students above and below the median of predicted impact. Overall, the results are consistent with our analysis in the previous sub-sections. Students who have learning disorders, lower initial grades and parents with less skilled occupations (e.g., blue collar mothers) are over-represented among the students with the highest predicted impact on performance. Immigrants and students with blue collar mothers are over-represented among the students with the highest enhancement in their well-being.

Overall, the most disadvantaged children seem to have benefited the most from the tutoring. However, heterogeneity depending on parents' or students' characteristics is not stark. It is worth emphasizing that the sample of students included in TOP had been already selected by school principals and teachers among the ones deemed most in need of the tutoring intervention. This may have potentially led to lower heterogeneity in treatment effects.

6.2 Robustness

Appendix Table A.XX provides a robustness analysis of the main results presented so far. First, in columns 1 and 2, we present the OLS estimates and standard errors without including the baseline controls: we find that the results are very similar to the main results reported in the previous tables.

Second, in columns 3 and 5, we re-estimate the effect of treatment on our main outcomes choosing the set of control variables in a systematic way with double post LASSO procedure, following Belloni et al. (2012). We include all baseline characteristics that are sufficiently correlated with treatment (after imposing the LASSO penalty given that the regression includes many variables) and the variables that are sufficiently correlated with control (after imposing the LASSO penalty) (Ludwig et al., 2017).[19] We list the controls selected using LASSO for each outcome in Appendix Table A.XXI. Including the variables picked by LASSO in the second step makes no substantial difference in most results, with the exception of the Aspiration index, where the estimated effect is still positive, but smaller in magnitude and not significant at conventional levels.

Finally, in the last two columns we present inverse probability-weighted estimates of treatment effects. The estimated effects are almost unchanged and, if anything, slightly higher due to the minor imbalances presented in Table A.V, with the control group being more positively selected compared to the treatment group in terms of parental background. These different robustness checks provide a consistent and overall positive picture of the impact of TOP on student outcomes.

6.3 Long-term Outcomes

Appendix Table A.XIX reports preliminary evidence on the long term impact of the intervention, focusing only on some cohort of students that participate in TOP2020 as we do not have realized outcomes for all. First, we observe test score data in grade 8 from INVALSI for the students who participate in TOP2020 during grade 7. Column 1 shows that the average effect of participating in the online tutoring is 0.072 SD more than one year after the end of the intervention compared to the control group, even if not statistically significant at conventional level due to the limited sample for which we

[19] The double post LASSO procedure is based on three steps. First, we fit a LASSO regression predicting the dependent variable and we select all variables with non-zero estimated coefficient after the introduction of a penalty term that shrinks the estimated regression coefficients towards zero to reduce over-fitting. Second, we fit a LASSO regression predicting the treatment variable and following the same procedure of step one. Finally, we fit a linear regression of the outcome variable on the treatment variable including the covariates selected in the first or second step.

observe this outcome. The effect is fully driven by students that participate in the intense six hours of tutoring, with a positive effect of 0.244 SD (p-value = 0.07).

Column 2 shows the impact of attending a vocational track for students who participate in TOP2020 during grade 7 or grade 8. Although the effect is not statistically significant at conventional level, the evidence qualitatively suggest a reduction by 4.6 percentage points in the probability of attending a vocational track, a 9.4% reduction compared to the control group. The effect of the intense tutoring is larger in magnitude, with a reduction by 6.2 percentage points.

7 Results on Tutors

The primary purpose of the project was to improve outcomes for students who were the direct beneficiaries of the intervention. However, the volunteering experience of being a TOP tutor may have affected tutors' capacity to empathize, as well as their perception of the relative importance of hard work versus luck to achieve success in life. We collected a short questionnaire (as described in Appendix B.4.2) from volunteers who applied to the TOP tutoring program, independently on whether they were assigned a student or not. As mentioned in Section 3.2, the assignment of tutors to students was random, conditional on a set of baseline characteristics used for the allocation of tutors (e.g., subject and time availability). Around half of the respondents who completed the endline questionnaire had been randomly assigned to a student. Appendix Table A.XI shows that the characteristics of tutors who experienced TOP and who did not experience TOP are overall balanced once we take into account the allocation criteria.

[Insert Table VI]

Table VI shows the impact of tutoring on the two key outcomes (the Empathy index and the Hard Work index) described in Appendix B.4.2, when controlling for the factors used in the assignment of tutors to students (time and subject availability, previous training and tutoring experience, and regular enrollment in university).

We find that participating in TOP increased tutors' empathy by 3.4 percentage points, a 0.27 SD increase compared to volunteers who were not assigned a student to supervise. We do not find any economically or statistically significant effect on tutors' perceptions of the role of hard work to achieve success in life.[20]

[20]For completeness, we present the ordered logit results for the individual questions used to build the empathy and hard work indexes in Appendix Table A.XXII. Panel A shows the results using the endline collected in September, six months after the beginning of the intervention, while Panel B shows the results

8 Conclusion

School closure due to the COVID-19 outbreak has created massive learning losses and adverse psychological effects for children, especially the most vulnerable and those from a low socioeconomic background (Agostinelli et al., 2022; Carlana et al., 2023b), exacerbating pre-existing inequalities. In this paper, we show that online tutoring can be an effective tool to help students improve their academic performance during and after the pandemic. We exploit over-subscription by schools and students to an innovative online tutoring program in Italy, "TOP", to evaluate its impact using a randomized control trial. We find that the one-on-one support provided virtually by volunteer university students increased performance in a standardized test by 0.21 in the Spring 2020 and 0.15 SD in the Spring 2022. The effect is more than doubled by intense tutoring providing six hours or support instead of three hours per week, while group tutoring has a small and statistically insignificant effect. When in extreme distress and social isolation, online tutoring has been shown effective also in improving aspirations, psychological well-being, and development of socio-emotional skills.

In-person tutoring, especially when implemented by professionals and teachers and/or for several days per week, has proved to be highly effective in several contexts (Nickow et al., 2020; Fryer Jr, 2017). However, these programs are widely viewed as "too costly to be undertaken on a large scale" (Ander et al., 2016). Our TOP intervention allows to achieve sizeable results on learning and other life outcomes, keeping the costs extremely contained. The program leverages volunteer university students as tutors, mainly moved by intrinsic motivation and supported by a team of pedagogical experts. Volunteer tutors represent a viable and effective solution to reach a large number of students in need of support. The overall cost of the program per pupil was around 50 euro, covering the organizational and pedagogical support.[21] Even when schools are opened, virtual tutoring implemented by volunteer university students provide an effective tool to help vulnerable children and prevent inequalities to emerge, in a cost-effective way.

References

Agcom (2019). Educare digitale lo stato di sviluppo della scuola digitale un sistema complesso ed integrato di risorse digitali abilitanti. *Studio del Servizio Economico-*

when imputing the value for the first endline collected in June for those tutors who did not complete the second endline. The results are qualitatively and quantitatively very similar with both samples.

[21] This excludes the research costs, namely the incentives to complete the endline survey for families, and the salaries of the enumerators who supervised the endline test score data collection.

Statistico Agcom.

Agostinelli, F., Doepke, M., Sorrenti, G., and Zilibotti, F. (2022). When the great equalizer shuts down: Schools, peers, and parents in pandemic times. *Journal of Public Economics*, 206:104574.

Alan, S., Boneva, T., and Ertac, S. (2019). Ever failed, try again, succeed better: Results from a randomized educational intervention on grit. *The Quarterly Journal of Economics*, 134(3):1121–1162.

Ander, R., Guryan, J., and Ludwig, J. (2016). Improving academic outcomes for disadvantaged students: Scaling up individualized tutorials. *The Hamilton Project – Brookings*.

Angrist, N., Bergman, P., and Matsheng, M. (2022). Experimental evidence on learning using low-tech when school is out. *Nature Human Behaviour*.

Ashraf, N., Bandiera, O., and Jack, B. K. (2014). No margin, no mission? a field experiment on incentives for public service delivery. *Journal of Public Economics*, 120:1–17.

Athey, S. and Imbens, G. (2016). Recursive partitioning for heterogeneous causal effects. *Proceedings of the National Academy of Sciences*, 113(27):7353–7360.

Bacher-Hicks, A., Goodman, J., and Mulhern, C. (2020). Inequality in household adaptation to schooling shocks: Covid-induced online learning engagement in real time. *Journal of Public Economics*, 193:104345.

Bandura, A., Freeman, W., and Lightsey, R. (1999). *Self-efficacy: The exercise of control*. Springer.

Banerjee, A., Banerji, R., Berry, J., Duflo, E., Kannan, H., Mukherji, S., and Walton, M. (2015). Teaching at the right level: Evidence from randomized evaluations in India. *NBER Working Paper*, 22746.

Belloni, A., Chen, D., Chernozhukov, V., and Hansen, C. (2012). Sparse models and methods for optimal instruments with an application to eminent domain. *Econometrica*, 80(6):2369–2429.

Carlana, M., Gaudenzi, G., and La Ferrara, E. (2023a). Italy: Tutoring online program (top): A successful gobal experience.

Carlana, M., La Ferrara, E., and Lopez, C. (2023b). Widening inequalities in education: the learning loss from covid-19 in italy. *AEA Papers and Proceedings*.

Carlana, M., La Ferrara, E., and Pinotti, P. (2022). Goals and gaps: Educational careers of immigrant children. *Econometrica*, 90(1):1–29.

Chetty, R., Friedman, J. N., Hendren, N., Stepner, M., et al. (2020). How did covid-19 and stabilization policies affect spending and employment? A new real-time economic tracker based on private sector data. *NBER Working Paper*.

Coie, J. D. and Krehbiel, G. (1984). Effects of academic tutoring on the social status of low-achieving, socially rejected children. *Child Development*, pages 1465–1478.

Cook, P. J., Dodge, K., Farkas, G., Fryer, R. G., Guryan, J., Ludwig, J., Mayer, S., Pollack, H., and Steinberg, L. (2015). Not too late: Improving academic outcomes for disadvantaged youth. *Institute for Policy Research Northwestern University Working Paper WP-15*, 1.

Davis, J. and Heller, S. B. (2017). Using causal forests to predict treatment heterogeneity: An application to summer jobs. *American Economic Review*, 107(5):546–50.

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

Engzell, P., Frey, A., and Verhagen, M. D. (2020). Learning inequality during the covid-19 pandemic. *SocArXiv*.

Escueta, M., Quan, V., Nickow, A. J., and Oreopoulos, P. (2017). Education technology: An evidence-based review. *NBER Working Paper*.

Falk, A., Kosse, F., and Pinger, P. (2023). Mentoring and schooling decisions: Causal evidence. *Journal of Political Economy*.

Frühe, B., Allgaier, A.-K., Pietsch, K., Baethmann, M., Peters, J., Kellnar, S., Heep, A., Burdach, S., von Schweinitz, D., and Schulte-Körne, G. (2012). Children's depression screener (ChilD-S): development and validation of a depression screening instrument for children in pediatric care. *Child Psychiatry & Human Development*, 43(1):137–151.

Fryer Jr, R. G. (2017). The production of human capital in developed countries: Evidence from 196 randomized field experiments. In *Handbook of Economic Field Experiments*, volume 2, pages 95–322. Elsevier.

Gauri, V., Jamison, J. C., Mazar, N., and Ozier, O. (2019). Motivating bureaucrats through social recognition: External validity—a tale of two states. *Organizational Behavior and Human Decision Processes*.

Genicot, G. and Ray, D. (2017). Aspirations and inequality. *Econometrica*, 85(2):489–519.

Golberstein, E., Wen, H., and Miller, B. F. (2020). Coronavirus disease 2019 (covid-19) and mental health for children and adolescents. *JAMA pediatrics*.

Hardt, D., Nagler, M., and Rincke, J. (2022). Can peer mentoring improve online teaching effectiveness? an rct during the covid-19 pandemic. *Labour Economics*, page 102220.

Hassan, H., Islam, A., Siddique, A., and Wang, L. C. (2022). Telementoring and home-schooling during school closures: A randomized experiment in rural bangladesh. *Munich Papers in Political Economy 13, TUM School of Governance at the Technical University of Munich*.

Kosse, F., Deckers, T., Pinger, P., Schildberg-Hörisch, H., and Falk, A. (2020). The formation of prosociality: causal evidence on the role of social environment. *Journal of Political Economy*, 128(2):434–467.

Kraft, M. A., List, J. A., Livingston, J. A., and Sadoff, S. (2022). Online tutoring by college volunteers: Experimental evidence from a pilot program. *AEA Papers and Proceedings*, 112:614–618.

La Ferrara, E. (2019). Aspirations, social norms, and development. *Journal of the European Economic Association*, 17(6):1687–1722.

Levitt, S. D., List, J. A., Neckermann, S., and Sadoff, S. (2016). The behavioralist goes to school: Leveraging behavioral economics to improve educational performance. *American Economic Journal: Economic Policy*, 8(4):183–219.

Lichand, G., Doria, C. A., Leal-Neto, O., and Fernandes, J. P. C. (2022). The impacts of remote learning in secondary education during the pandemic in brazil. *Nature Human Behaviour*, pages 1–8.

Ludwig, J., Mullainathan, S., and Spiess, J. (2017). Machine-learning tests for effects on multiple outcomes. *arXiv preprint arXiv:1707.01473*.

Malamud, O. and Pop-Eleches, C. (2011). Home computer use and the development of human capital. *The Quarterly Journal of Economics*, 126(2):987–1027.

Malkus, N. (2020). School districts' remote-learning plans may widen student achievement gap. *Education Next*, 20(3).

Nickow, A., Oreopoulos, P., and Quan, V. (2020). The impressive effects of tutoring on prek-12 learning: A systematic review and meta-analysis of the experimental evidence. *NBER Working Paper*, (w27476).

Orgilés, M., Morales, A., Delvecchio, E., Mazzeschi, C., and Espada, J. P. (2020). Immediate psychological effects of the covid-19 quarantine in youth from italy and spain. *Frontiers in psychology*, 11:579038.

Parker, D. (2023). Scaling tutoring: Five key factors to benefit more students.

Patrinos, H. A., Vegas, E., and Carter-Rau, R. (2022). An analysis of covid-19 student learning loss. *World Bank Policy Research Working Paper 10033*.

Resnjanskij, S., Ruhose, J., Wiederhold, S., and Woessmann, L. (2023). Can mentoring alleviate family disadvantage in adolscence? a field experiment to improve labor-market prospects. *Journal of Political Economy*.

Richmond, A. D. (2015). Academic task avoidance and achievement as predictors of peer status during the early primary school years. *Educational Psychology*.

Robinson, C. D. and Loeb, S. (2021). High-impact tutoring: State of the research and priorities for future learning. *National Student Support Accelerator*, 21(284):1–53.

Rosenthal, R. (1973). The pygmalion effect lives. *Psychology Today*.

Rotter, J. B. (1966). Generalized expectancies for internal versus external control of reinforcement. *Psychological monographs: General and applied*, 80(1):1.

Singh, A., Romero, M., and Muralidharan, K. (2022). Covid-19 learning loss and recovery: Panel data evidence from india. *NBER Working Paper*.

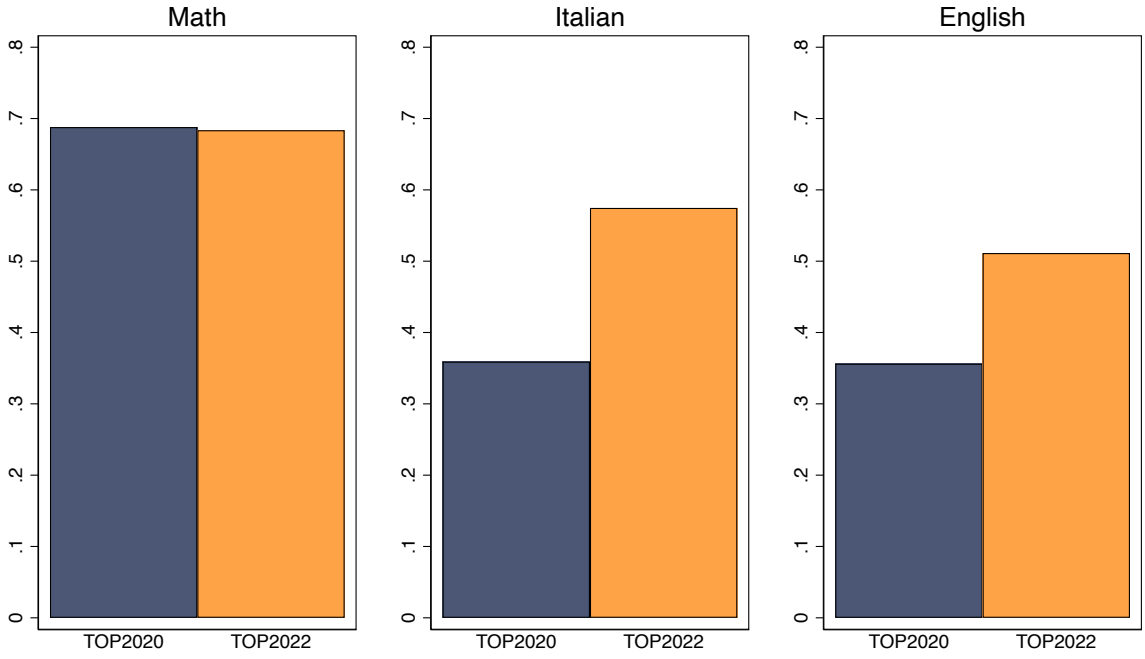
Stantcheva, S. (2022). Inequalities in the times of a pandemic. *Economic Policy*, 37(109):5–4.

UNESCO (2020). Covid-19 impact on education.

Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

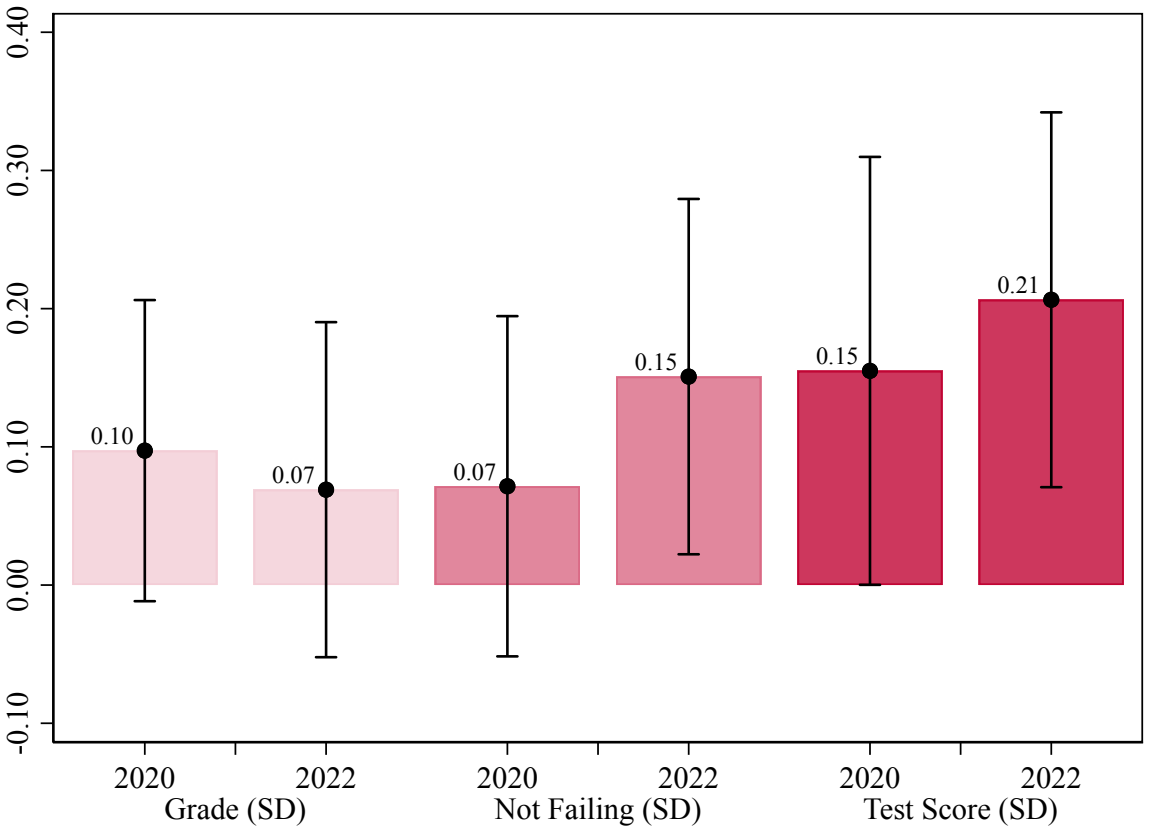
Tables and Figures

Figure 1. Main Subjects During Tutoring



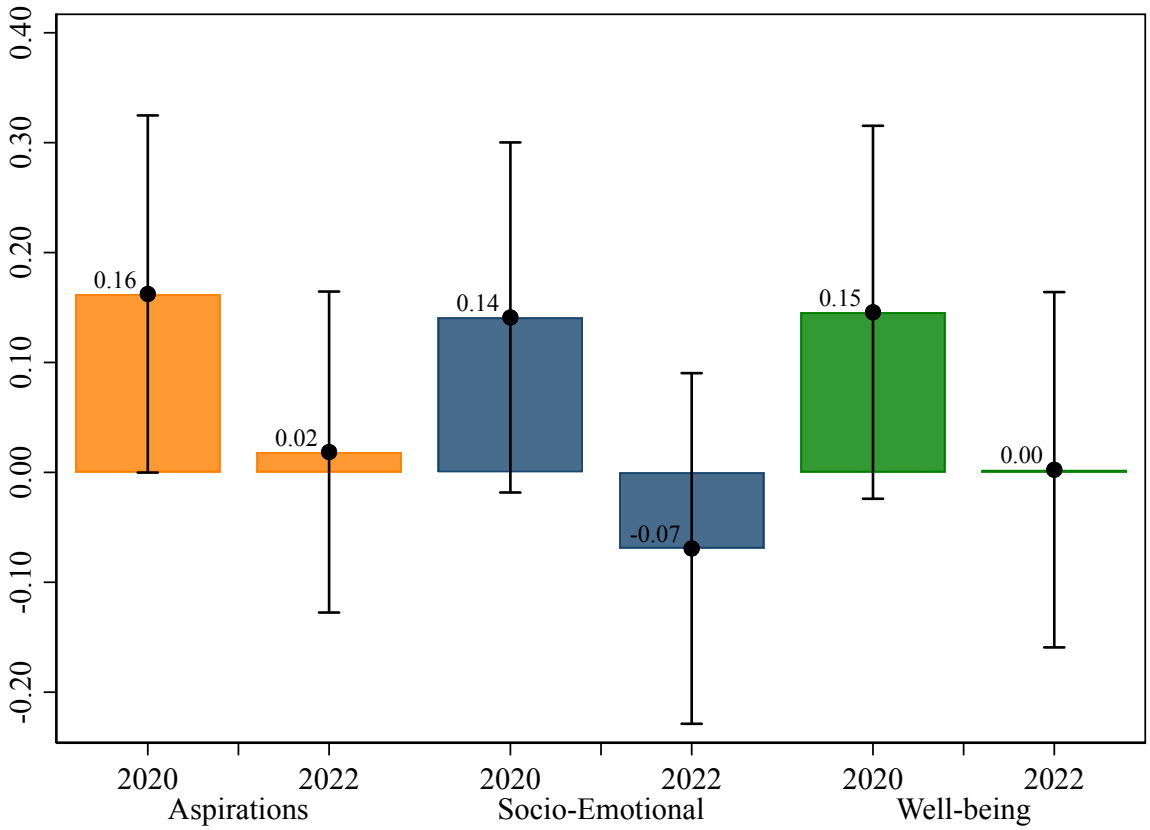
Notes: This Figure shows the main subjects done during the tutoring meetings, as reported by tutors at endline, for TOP 2020 and for TOP 2022. In TOP2022, we asked the tutors to focus on up to two subjects.

Figure 2. The impact of three hours of individual tutoring on math academic outcomes



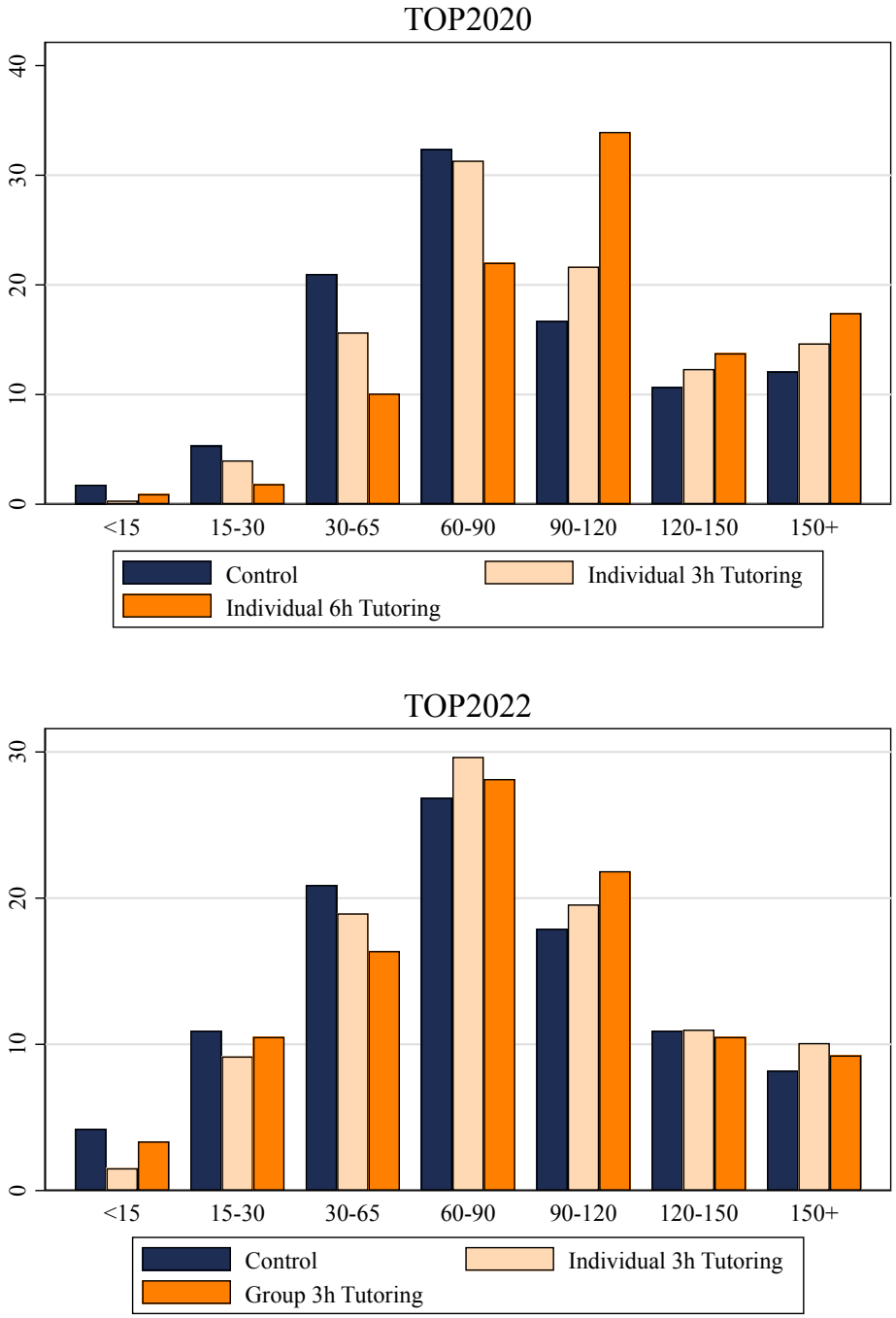
Notes: This Figure reports OLS estimates of the impact of 3 hours per week of individual tutoring on the main academic outcomes in the paper, reported in the headings. The bar shows 95% confidence intervals. The mean of the control group for each index is 0 and the standard deviation is 1. Controls included are the same as in Table II and III. The coefficient plotted are the equivalent on the effect of individual 3h tutoring as in Panel B.

Figure 3. The impact of three hours of individual tutoring on non-academic outcomes



Notes: This Figure reports four OLS estimates of the impact of 3 hours per week of individual tutoring on the main non-academic outcomes in the paper, reported in the headings. The bar shows 95% confidence intervals. The mean of the control group for each index is 0 and the standard deviation is 1. Controls included are the same as in Table IV and V.

Figure 4. Homework time



Notes: This Figure shows homework time per day in minutes, as reported by students at endline, for TOP 2020 in the top panel and for TOP 2022 in the bottom panel.

Table I. Balance Table (full sample)

	(1) Control	(2) Treatment	(3) P-value	(4) Std diff.
Students				
Survey Data				
Female	0.414	0.436	0.277	0.045
Immigrant	0.211	0.240	0.106	0.069
Has Learning Disorder (Baseline)	0.291	0.285	0.887	-0.013
Student is in 6th grade (Baseline)	0.293	0.322	0.127	0.063
Student is in 7th grade (Baseline)	0.353	0.356	0.960	0.006
Student is in 8th grade (Baseline)	0.351	0.318	0.112	-0.070
How much does the student enjoy math? (Baseline)	2.664	2.671	0.783	0.007
How much does the student enjoy Italian? (Baseline)	3.009	3.025	0.643	0.018
How much does the student enjoy English? (Baseline)	3.083	3.097	0.929	0.012
Child-reported homework time (Baseline)	81.937	84.834	0.194	0.066
Admin data				
Math Grade at baseline	6.309	6.277	0.447	-0.029
Italian Grade at baseline	6.462	6.519	0.221	0.059
English Grade at baseline	6.544	6.608	0.397	0.057
Parents				
Child lives in one-parent household (Baseline)	0.252	0.232	0.267	-0.047
Mother's highest degree: High School (Baseline)	0.455	0.420	0.163	-0.071
Mother's highest degree: College (Baseline)	0.118	0.117	0.733	-0.003
Father's highest degree: High School (Baseline)	0.366	0.387	0.315	0.043
Father's highest degree: College (Baseline)	0.077	0.074	0.813	-0.011
Mother has a white collar job (Baseline)	0.409	0.398	0.489	-0.022
Mother has a blue collar job (Baseline)	0.567	0.564	0.909	-0.006
Father has a white collar job (Baseline)	0.324	0.335	0.539	0.023
Father has a blue collar job (Baseline)	0.645	0.631	0.479	-0.029
Internet Access	0.969	0.962	0.238	-0.038
No access to device for tutoring (Baseline)	0.041	0.046	0.596	0.025
Family time spent helping the student (Baseline)	40.627	41.571	0.464	0.024
Observations	1,059	1,137		

Notes: This Table shows the characteristics of treated and control students in the initial sample at baseline for both TOP2020 and TOP2022. P-values for difference in means, controlling for the round fixed effects as in our main specification, are reported in the third column. The last column also reports the standardized difference between group averages. This table includes the entire baseline sample of students who were randomized to treatment and control groups. Additional balance tables separately for TOP2020 and TOP2022 are available in the Appendix Tables A.III and A.IV.

Table II. Estimation of the impact of TOP 2020 on math academic outcomes

	(1)	(2)	(3)	(4)	(5)
	Grades		Fail		Test Score
Sample	All	Endline	All	Endline	Endline
Panel A: Overall results					
Treatment	0.120**	0.156**	-0.043**	-0.054*	0.234***
	(0.056)	(0.072)	(0.022)	(0.029)	(0.075)
FWER	[0.057]		[0.057]		[0.006]
Sharpened q-value	[0.043]		[0.052]		[0.011]
Mean Dep:	6.37	6.35	0.18	0.18	-0.00
R^2	0.362	0.378	0.120	0.139	0.223
Panel B: Results of 3 vs. 6 hours of tutoring					
Treatment	0.108*	0.132*	-0.027	-0.034	0.155**
	(0.062)	(0.077)	(0.024)	(0.031)	(0.079)
Intense Treatment	0.030	0.077	-0.058*	-0.073**	0.284***
	(0.085)	(0.091)	(0.034)	(0.036)	(0.102)
Treat+Intense Treatment==0	0.083	0.026	0.009	0.005	0.000
Mean Dep:	6.37	6.35	0.18	0.18	-0.00
R^2	0.378	0.393	0.123	0.144	0.233
Obs	1059	712	1059	712	712

Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values (based on 10,000 replications) and sharpened q-value in square brackets. The dependent variable for each regression is listed in the column heading and it refers to math. Only 712 students completed the endline test score implemented by the research team (sample used for columns 2, 4, 5). "Mean Dep" at the bottom of the table is the mean of the dependent variables for students in the control group. Controls included are parental education and occupation, gender, immigration status, learning disorders, school grade, teacher assigned grades and test scores in math at baseline. In Panel B, we control for whether the students was identified for intense tutoring.

Table III. Estimation of the impact of TOP 2022 on math academic outcomes

	(1)	(2)	(3)	(4)	(5)
	Grades		Fail		Test Score
Sample	All	Endline	All	Endline	Endline
Panel A: Overall results					
Treatment	0.085	0.084	-0.064**	-0.069**	0.155***
	(0.055)	(0.061)	(0.027)	(0.029)	(0.059)
FWER	[0.150]		[0.030]		[0.030]
Sharpened q-value	[0.067]		[0.031]		[0.031]
Mean Dep:	5.97	6.01	0.31	0.30	-0.00
R^2	0.238	0.231	0.134	0.132	0.191
Panel B: Individual vs. Group tutoring					
Treatment	0.071	0.076	-0.070**	-0.078**	0.206***
	(0.063)	(0.070)	(0.030)	(0.033)	(0.069)
Group Treatment	0.032	0.019	0.013	0.021	-0.121
	(0.075)	(0.079)	(0.037)	(0.039)	(0.084)
Treat+Group Treatment==0	0.145	0.212	0.105	0.129	0.262
Mean Dep:	5.97	6.01	0.31	0.30	-0.00
Obs	1121	943	1121	943	953
R^2	0.238	0.231	0.134	0.132	0.193

Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values (based on 10,000 replications) and sharpened q-value in square brackets. The dependent variable for each regression is listed in the column heading and it refers to math. Only 712 students completed the endline test score implemented by the research team (sample used for columns 2, 4, 5). "Mean Dep" at the bottom of the table is the mean of the dependent variables for students in the control group. Controls included are parental education and occupation, gender, immigration status, learning disorders, school grade, teacher assigned grades and test scores in math at baseline.

Table IV. Estimation of the impact of TOP 2020 on non-academic outcomes

	(1)	(2)	(3)
	Aspirations	Socio-Emotional	Well-being
Panel A: Overall results			
Treatment	0.192**	0.153**	0.164**
	(0.078)	(0.075)	(0.083)
FWER	[0.048]	[0.084]	[0.084]
Sharpened q-value	[0.044]	[0.044]	[0.044]
R^2	0.334	0.158	0.073
Panel B: Results of 3 vs. 6 hours of tutoring			
Treatment	0.162*	0.141*	0.146*
	(0.083)	(0.081)	(0.086
Intense Treatment	0.102	0.044	0.065
	(0.119)	(0.107)	(0.118)
Treat+Intense Treatment==0	0.026	0.087	0.087
R^2	0.343	0.158	0.074
Mean Dep:	0.00	-0.00	-0.00
Obs	523	636	614

Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values (based on 10,000 replications) and sharpened q-value in square brackets. The dependent variable for each regression is listed in the column heading and it refers to math. Only 712 students completed the endline test score implemented by the research team (sample used for columns 2, 4, 5). "Mean Dep" at the bottom of the table is the mean of the dependent variables for students in the control group. Controls included are parental education and occupation, gender, immigration status, learning disorders, school grade, teacher assigned grades and test scores in math at baseline. In Panel B, we control for whether the students was identified for intense tutoring.

Table V. Estimation of the impact of TOP 2022 on non-academic outcomes

	(1)	(2)	(3)
	Aspirations	Socio-Emotional	Well-being
Panel A: Overall results			
Treatment	0.054	-0.034	0.037
	(0.065)	(0.070)	(0.073)
FWER	[0.771]	[0.842]	[0.842]
Sharpened q-value	[1.000]	[1.000]	[1.000]
R^2	0.218	0.059	0.060
Panel B: Group vs. Individual tutoring			
Treatment	0.018	-0.069	0.002
	(0.074)	(0.081)	(0.082
Group Tutoring	0.087	0.087	0.082
	(0.091)	(0.099)	(0.104)
Treat+Group Tutoring==0	0.217	0.851	0.391
R^2	0.218	0.060	0.061
Mean Dep:	0.00	0.00	-0.00
Obs	889	898	881

Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values (based on 10,000 replications) and sharpened q-value in square brackets. The dependent variable for each regression is listed in the column heading and it refers to math. Only 712 students completed the endline test score implemented by the research team (sample used for columns 2, 4, 5). "Mean Dep" at the bottom of the table is the mean of the dependent variables for students in the control group. Controls included are parental education and occupation, gender, immigration status, learning disorders, school grade, teacher assigned grades and test scores in math at baseline.

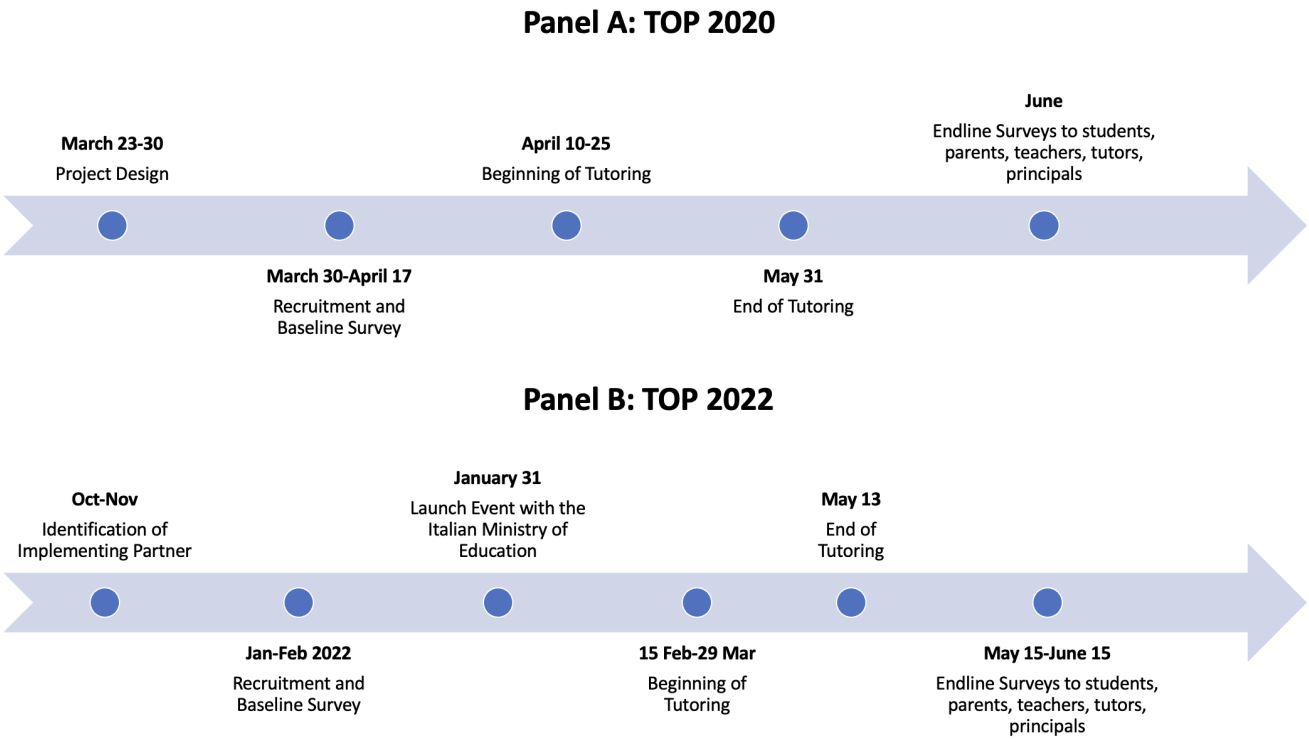
Table VI. Treatment Effect on Tutors' Outcomes

	(1)	(2)	(3)	(4)
	Emphaty Index		Hardwork Index	
TOP Tutors	0.019**	0.014*	-0.002	-0.002
	(0.008)	(0.008)	(0.008)	(0.009)
wyoung FWER p-value	[0.037]	[0.825]	[0.161]	[0.795]
Sharpened q-value	[0.087]	[0.153]	[0.711]	[0.711]
Randomization controls:	Yes	Yes	Yes	Yes
Tutor controls:	No	Yes	No	Yes
Mean Dep:	0.67	0.67	0.65	0.65
Obs	740	740	735	735

Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values in square brackets, based on 10,000 replications. The dependent variable for each regression is listed in the column heading . The randomization controls include whether the volunteer has tutoring experience and specific training (to support students with learning disorders or immigrants), their expertise in the subjects (math, Italian, English), their time availability (3 hours per week or 6 hours per week), whether they are on time in their university enrollment and if they confirmed their availability. The additional tutor controls include gender, university faculty, whether they are enrolled in a undergraduate or master, GPA, previous volunteering activities, whether they applied to TOP to help others (motivation), parental education, and familiarity with the computer. "Mean Dep" at the bottom of the table is the mean of the dependent variables for students in the control group.

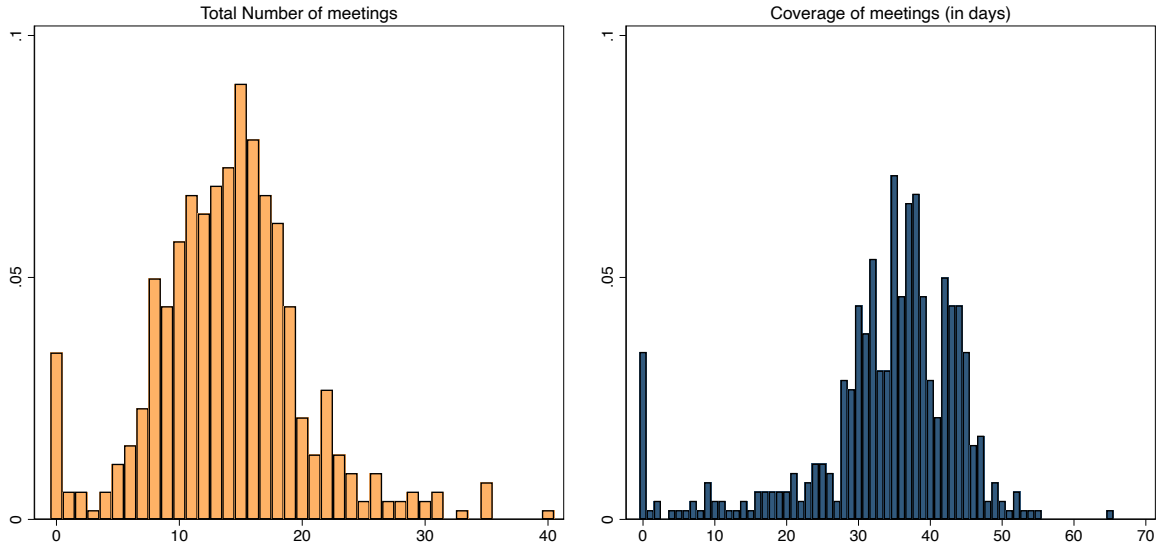
A Online Appendix: Additional Tables and Figures

Figure A.1. Timeline of TOP



Notes: This Figure shows the timeline for the implementation of TOP 2020 in Panel A and TOP 2022 in Panel B.

Figure A.2. Number of tutoring meetings and coverage in days



Notes: This Figure shows the total number of tutoring meetings (left panel) and the number of days from the beginning to the end of the tutoring (right panel). The data used in these graphs are reported by tutors in the registry. The Figure shows data from 522 treated students: 8 students did the tutoring but we do not have precise information from the tutors on the number of meetings. For TOP2022, the mean of hours of tutoring done is 36 as reported by the NGO CIAI and around 24 meetings.

Table A.I. Balance Table at province level

Variable	(1) No TOP	(2) TOP	(3) Diff.	(4) Std. Diff.
Macro-area: North	0.313	0.650	0.337***	0.501
	(0.467)	(0.483)	(0.095)	
Macro-area: Center	0.209	0.200	-0.009	-0.016
	(0.410)	(0.405)	(0.082)	
Macro-area: South and Islands	0.478	0.150	-0.328***	-0.529
	(0.503)	(0.362)	(0.091)	
Level of education: Elementary	0.305	0.287	-0.017***	-0.453
	(0.026)	(0.027)	(0.005)	
Level of education: Middle school	0.300	0.297	-0.003	-0.079
	(0.024)	(0.023)	(0.005)	
Level of education: Diploma	0.297	0.309	0.012**	0.354
	(0.028)	(0.020)	(0.005)	
Level of education: University	0.101	0.109	0.008**	0.296
	(0.015)	(0.023)	(0.004)	
Covid-19 cases March'20 (1000 inhabitants)	1.477	2.234	0.757*	0.266
	(1.799)	(2.199)	(0.391)	
Covid-19 cases April'20 (1000 inhabitants)	2.889	4.399	1.511**	0.333
	(3.024)	(3.376)	(0.631)	
Covid-19 cases May'20 (1000 inhabitants)	3.209	4.961	1.752**	0.343
	(3.418)	(3.800)	(0.712)	
Immigrants 2020	0.070	0.096	0.026***	0.592
	(0.033)	(0.030)	(0.006)	
Unemployment rate (2019)	11.678	8.316	-3.362***	-0.438
	(5.908)	(4.912)	(1.111)	
Observations	67	40	107	

Notes: This Table shows the characteristics of provinces that had at least one treated school ("TOP province") compared to provinces with no treated schools ("No TOP Province") for TOP2020 with a nation-wide recruitment. The first two columns show the mean for the two groups, the third column shows the difference in means, while the last column provides the standardized difference between group averages. In parenthesis, the first two columns show the standard deviations of the mean, while the third column shows the standard errors of the difference between treatment and control groups.

Table A.III. Balance Table TOP2020 (full sample)

	(1) Control	(2) Treatment	(3) 3h Tutoring	(4) 6h Tutoring	(5) P-value (1)-(2)	(6) P-value (1)-(3)
Students						
Survey Data						
Female	0.381	0.412	0.403	0.437	0.298	0.481
Immigrant	0.169	0.178	0.161	0.225	0.690	0.962
Has Learning Disorder (Baseline)	0.337	0.312	0.297	0.355	0.410	0.307
Student is in 6th grade (Baseline)	0.318	0.311	0.310	0.315	0.843	0.947
Student is in 7th grade (Baseline)	0.338	0.343	0.354	0.315	0.882	0.721
Student is in 8th grade (Baseline)	0.338	0.338	0.328	0.364	0.985	0.701
How much does the student enjoy math? (Baseline)	2.772	2.724	2.701	2.787	0.430	0.246
How much does the student enjoy Italian? (Baseline)	3.023	3.104	3.099	3.120	0.111	0.273
How much does the student enjoy English? (Baseline)	3.107	3.143	3.206	2.972	0.589	0.222
Child-reported homework time (Baseline)	84.935	87.723	87.777	87.577	0.287	0.348
Admin data						
Math Grade at baseline	6.270	6.150	6.187	6.049	0.062*	0.112
Italian Grade at baseline	6.331	6.397	6.423	6.324	0.260	0.368
English Grade at baseline	6.454	6.435	6.509	6.234	0.774	0.777
Parents						
Child lives in one-parent household (Baseline)	0.258	0.219	0.221	0.211	0.144	0.326
Mother's highest degree: High School (Baseline)	0.469	0.439	0.412	0.514	0.344	0.071*
Mother's highest degree: College (Baseline)	0.113	0.107	0.120	0.072	0.791	0.915
Father's highest degree: High School (Baseline)	0.358	0.380	0.384	0.370	0.452	0.472
Father's highest degree: College (Baseline)	0.084	0.063	0.065	0.058	0.201	0.245
Mother has a white collar job (Baseline)	0.407	0.408	0.417	0.383	0.992	0.876
Mother has a blue collar job (Baseline)	0.587	0.575	0.567	0.596	0.751	0.698
Father has a white collar job (Baseline)	0.331	0.343	0.348	0.330	0.672	0.763
Father has a blue collar job (Baseline)	0.660	0.655	0.649	0.670	0.828	0.875
Internet Access	0.941	0.932	0.935	0.923	0.511	0.699
No access to device for tutoring (Baseline)	0.027	0.029	0.036	0.007	0.847	0.331
Family time spent helping the student (Baseline)	49.062	48.553	49.003	47.319	0.846	0.947
Observations	529	530	387	143		

Notes: This Table shows the characteristics of treated and control students in the initial sample at baseline for TOP2020. P-values for difference in means, controlling for the round fixed effects as in our main specification, are reported in column (5). The p-values for the coefficient $treat$ and $treat_intense$ are reported in columns (6) and (7), respectively. This table includes the entire baseline sample of students who were randomized to treatment and control groups. Additional balance tables for TOP2020 are available in the Appendix Table A.V.

Table A.II. Summary statistics of main outcome variables

	TOP 2020			TOP 2022		
	Obs	Mean	S.D.	Obs	Mean	S.D.
Section I: Academic Outcomes and Beliefs						
Grade Math Endline	1059	6.39	1.09	1121	6.03	0.98
Fail Math Endline	1059	0.16	0.37	1121	0.28	0.45
Std Performance Math	712	0.10	1.03	953	0.09	0.96
Beliefs reported by Students						
Student's belief on share of correct answers	705	0.67	0.15	947	0.66	0.16
Student's self-rating	524	6.30	1.39	959	5.50	1.68
Beliefs reported by Parents						
Parent's belief on share of correct answers	756	0.71	0.14	1001	0.70	0.14
Beliefs reported by Teachers						
Teacher's belief on share of correct answers	792	0.49	0.20	1068	0.54	0.14
Teacher's rating of student	792	5.65	1.87	1069	5.42	1.37
Section II: Aspirations						
Std Aspirations Index	523	0.07	1.00	889	0.03	1.02
Outcomes reported by Students						
Student intends to go to university	674	0.39	0.49	961	0.32	0.47
Student believes they will likely graduate university	682	0.23	0.42	965	0.26	0.44
Student intends to choose professional track	681	0.28	0.45	966	0.22	0.41
Student intends to choose top-track school	681	0.15	0.36	966	0.15	0.36
Outcomes reported by Parents						
Parent believes child will attend university	765	0.35	0.48	1000	0.28	0.45
Parent believes the child can attend university	772	0.33	0.47	1002	0.36	0.48
Outcomes reported by Teachers						
Teacher believes child will attend university	839	0.14	0.34	1076	0.05	0.23
Section III: Socio-Emotional Skills						
Std Socio-emotional Index	636	0.07	0.95	898	-0.01	1.01
Outcomes reported by Students						
Student is willing to go to second diff. question	685	0.59	0.49	967	0.55	0.50
Student does not want to go to second diff. question	685	0.12	0.32	967	0.13	0.34
Grit Index - Student	673	0.69	0.13	958	0.66	0.13
Importance of luck vs. effort Index	685	0.72	0.11	961	0.70	0.11
Outcomes reported by Parents						
Grit Index - Parent	736	0.67	0.13	955	0.65	0.13
Section IV: Well-being						
Std Well-being Index	614	0.10	0.95	881	0.01	1.05
Outcomes reported by Students						
Student-reported depression scale	669	0.54	0.12	948	0.58	0.12
Student-reported Happiness Index	665	0.63	0.22	955	0.69	0.21
Outcomes reported by Parents						
Parent-reported depression scale	731	0.58	0.10	968	0.58	0.10
Parent-reported Happiness Index	741	0.62	0.21	969	0.68	0.18

Notes: This Table shows the summary statistics of all outcome variables, as reported in the endline questionnaire from students, parents, and teachers. All outcomes refer to children even when reported by parents or teachers. The table also includes the mean of the indices in the entire sample, standardized to have mean 0 and standard deviation 1 for the control group.

Table A.IV. Balance Table TOP2022 (full sample)

	(1) Control	(2) Treatment	(3) Individual Tutoring	(4) Group Tutoring	(5) P-value (1)-(2)	(6) P-value (1)-(3)	(7) P-value (3)-(4)
Students							
Survey Data							
Female	0.447	0.456	0.401	0.528	0.615	0.355	0.012**
Immigrant	0.253	0.293	0.310	0.272	0.077*	0.018**	0.098*
Has Learning Disorder (Baseline)	0.245	0.262	0.275	0.245	0.500	0.379	0.554
Student is in 6th grade (Baseline)	0.268	0.331	0.310	0.358	0.019**	0.220	0.161
Student is in 7th grade (Baseline)	0.368	0.367	0.363	0.374	0.941	0.785	0.694
Student is in 8th grade (Baseline)	0.364	0.301	0.327	0.268	0.026**	0.363	0.066*
How much does the student enjoy math? (Baseline)	2.556	2.624	2.769	2.437	0.262	0.066*	0.090*
How much does the student enjoy Italian? (Baseline)	2.994	2.955	2.935	2.981	0.404	0.289	0.539
How much does the student enjoy English? (Baseline)	3.059	3.057	3.059	3.054	0.718	0.835	0.337
Child-reported homework time (Baseline)	78.945	82.283	81.822	82.884	0.435	0.605	0.782
Admin data							
Math Grade at baseline	6.348	6.387	6.441	6.318	0.493	0.474	0.802
Italian Grade at baseline	6.592	6.625	6.628	6.621	0.541	0.556	0.912
English Grade at baseline	6.633	6.758	6.765	6.750	0.160	0.092*	0.378
Parents							
Child lives in one-parent household (Baseline)	0.246	0.244	0.228	0.264	0.898	0.553	0.377
Mother's highest degree: High School (Baseline)	0.441	0.403	0.423	0.378	0.304	0.818	0.227
Mother's highest degree: College (Baseline)	0.124	0.124	0.128	0.120	0.826	0.713	0.720
Father's highest degree: High School (Baseline)	0.375	0.392	0.435	0.337	0.504	0.119	0.056*
Father's highest degree: College (Baseline)	0.069	0.082	0.083	0.081	0.364	0.505	0.859
Mother has a white collar job (Baseline)	0.410	0.389	0.427	0.343	0.325	0.824	0.261
Mother has a blue collar job (Baseline)	0.548	0.554	0.509	0.608	0.638	0.726	0.167
Father has a white collar job (Baseline)	0.318	0.328	0.336	0.318	0.655	0.969	0.417
Father has a blue collar job (Baseline)	0.631	0.612	0.588	0.645	0.440	0.539	0.935
Internet Access	0.996	0.988	0.988	0.989	0.092*	0.146	0.652
No access to device for tutoring (Baseline)	0.055	0.061	0.056	0.068	0.610	0.926	0.349
Family time spent helping the student (Baseline)	32.272	35.572	36.121	34.868	0.146	0.227	0.982
Observations	530	607	342	265			

Notes: This Table shows the characteristics of treated and control students in the initial sample at baseline for TOP2022. P-values for difference in means, controlling for the round fixed effects as in our main specification, are reported in column (5). The p-values for the coefficient $treat$ and $treat_group$ are reported in columns (6) and (7), respectively. This table includes the entire baseline sample of students who were randomized to treatment and control groups. Additional balance tables for TOP2022 are available in the Appendix Table A.VI.

Table A.V. Balance Table TOP2020 (endline test score sample)

	(1) Control	(2) Treatment	(3) 3h Tutoring	(4) 6h Tutoring	(5) P-value (1)-(2)	(6) P-value (1)-(3)	(7) P-value (3)-(4)
Students							
Survey Data							
Female	0.373	0.418	0.418	0.421	0.222	0.230	0.878
Immigrant	0.197	0.176	0.162	0.214	0.508	0.438	0.637
Has Learning Disorder (Baseline)	0.317	0.312	0.303	0.336	0.896	0.851	0.809
Student is in 6th grade (Baseline)	0.359	0.315	0.315	0.315	0.279	0.374	0.749
Student is in 7th grade (Baseline)	0.335	0.340	0.341	0.339	0.912	0.923	0.986
Student is in 8th grade (Baseline)	0.298	0.336	0.335	0.339	0.313	0.425	0.708
How much does the student enjoy math? (Baseline)	2.834	2.728	2.686	2.840	0.094*	0.032**	0.093*
How much does the student enjoy Italian? (Baseline)	2.979	3.097	3.083	3.135	0.086*	0.172	0.521
How much does the student enjoy English? (Baseline)	3.153	3.147	3.211	2.976	0.862	0.609	0.055*
Child-reported homework time (Baseline)	83.950	89.414	89.941	88.000	0.116	0.113	0.651
Admin data							
Math Grade at baseline	6.281	6.155	6.195	6.048	0.096*	0.167	0.549
Italian Grade at baseline	6.421	6.401	6.420	6.349	0.701	0.600	0.656
English Grade at baseline	6.529	6.430	6.509	6.216	0.204	0.463	0.109
Parents							
Child lives in one-parent household (Baseline)	0.250	0.214	0.220	0.198	0.320	0.603	0.246
Mother's highest degree: High School (Baseline)	0.500	0.443	0.410	0.533	0.136	0.023**	0.014**
Mother's highest degree: College (Baseline)	0.113	0.102	0.116	0.066	0.570	0.835	0.223
Father's highest degree: High School (Baseline)	0.368	0.380	0.378	0.385	0.868	0.944	0.824
Father's highest degree: College (Baseline)	0.105	0.065	0.065	0.066	0.078*	0.065*	0.640
Mother has a white collar job (Baseline)	0.461	0.416	0.416	0.417	0.274	0.229	0.696
Mother has a blue collar job (Baseline)	0.533	0.568	0.566	0.571	0.376	0.321	0.742
Father has a white collar job (Baseline)	0.315	0.333	0.333	0.333	0.692	0.876	0.600
Father has a blue collar job (Baseline)	0.671	0.664	0.663	0.667	0.927	0.929	0.667
Internet Access	0.955	0.929	0.929	0.929	0.177	0.239	0.839
No access to device for tutoring (Baseline)	0.029	0.024	0.030	0.008	0.855	0.669	0.028**
Family time spent helping the student (Baseline)	47.591	49.650	50.775	46.675	0.624	0.438	0.346
Observations	245	467	340	127			

Notes: This Table shows the characteristics of treated and control students in the sample who completed the endline test score for TOP2020. P-values for difference in means, controlling for the round fixed effects as in our main specification, are reported in column (5). The p-values for the coefficient $treat$ and $treat_intense$ are reported in columns (6) and (7), respectively. Additional balance tables for TOP2020 are available in the Appendix Table A.III.

Table A.VI. Balance Table TOP2022 (endline test score sample)

	(1) Control	(2) Treatment	(3) Individual Tutoring	(4) Group Tutoring	(5) P-value (1)-(2)	(6) P-value (1)-(3)	(7) P-value (3)-(4)
Students							
Survey Data							
Female	0.454	0.458	0.403	0.533	0.801	0.285	0.013**
Immigrant	0.284	0.297	0.311	0.279	0.533	0.203	0.144
Has Learning Disorder (Baseline)	0.236	0.258	0.271	0.242	0.385	0.291	0.540
Student is in 6th grade (Baseline)	0.309	0.331	0.311	0.358	0.451	0.900	0.306
Student is in 7th grade (Baseline)	0.402	0.366	0.366	0.367	0.284	0.290	0.777
Student is in 8th grade (Baseline)	0.289	0.303	0.323	0.275	0.719	0.323	0.179
How much does the student enjoy math? (Baseline)	2.654	2.647	2.779	2.468	0.942	0.453	0.118
How much does the student enjoy Italian? (Baseline)	3.036	2.964	2.935	3.004	0.241	0.128	0.344
How much does the student enjoy English? (Baseline)	3.122	3.052	3.075	3.021	0.217	0.752	0.158
Child-reported homework time (Baseline)	81.388	83.178	82.102	84.651	0.794	0.848	0.440
Admin data							
Math Grade at baseline	6.409	6.423	6.460	6.372	0.774	0.757	0.903
Italian Grade at baseline	6.636	6.648	6.635	6.665	0.829	0.916	0.870
English Grade at baseline	6.682	6.769	6.781	6.753	0.397	0.159	0.196
Parents							
Child lives in one-parent household (Baseline)	0.232	0.230	0.222	0.242	0.951	0.695	0.508
Mother's highest degree: High School (Baseline)	0.422	0.416	0.433	0.393	0.963	0.555	0.279
Mother's highest degree: College (Baseline)	0.125	0.125	0.125	0.124	0.796	0.545	0.428
Father's highest degree: High School (Baseline)	0.367	0.406	0.442	0.356	0.219	0.097*	0.221
Father's highest degree: College (Baseline)	0.077	0.080	0.078	0.082	0.735	0.961	0.623
Mother has a white collar job (Baseline)	0.396	0.386	0.424	0.337	0.616	0.951	0.336
Mother has a blue collar job (Baseline)	0.554	0.556	0.514	0.608	0.852	0.667	0.249
Father has a white collar job (Baseline)	0.303	0.326	0.329	0.323	0.427	0.890	0.265
Father has a blue collar job (Baseline)	0.645	0.618	0.596	0.651	0.398	0.545	0.801
Internet Access	0.995	0.989	0.991	0.988	0.252	0.366	0.998
No access to device for tutoring (Baseline)	0.054	0.058	0.055	0.062	0.747	0.964	0.539
Family time spent helping the student (Baseline)	32.625	35.898	36.506	35.081	0.201	0.277	0.988
Observations	388	565	325	240			

Notes: This Table shows the characteristics of treated and control students in the sample who completed the endline test score for TOP2022. P-values for difference in means, controlling for the round fixed effects as in our main specification, are reported in column (5). The p-values for the coefficient $treat$ and $treat_group$ are reported in columns (6) and (7), respectively. Additional balance tables for TOP2022 are available in the Appendix Table A.IV.

Table A.VII. Attrition between baseline and endline

Dependent variable: Dummy for endline completion						
Sample	All	All	TOP2020		TOP2022	
	Full	Full	Treatment	Control	Treatment	Control
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment status	0.306***	0.299***				
	(0.017)	(0.017)				
Female		0.011	0.029	-0.016	0.012	-0.014
		(0.017)	(0.029)	(0.046)	(0.022)	(0.038)
Immigrant		0.046**	-0.016	0.098	0.029	0.112**
		(0.021)	(0.042)	(0.061)	(0.025)	(0.045)
Learning Disorder		-0.025	-0.020	-0.049	-0.006	-0.038
		(0.019)	(0.030)	(0.048)	(0.026)	(0.046)
Grade 6		-0.374**	-0.024	-0.064	-0.020	0.259***
		(0.170)	(0.134)	(0.238)	(0.029)	(0.050)
Grade 7		-0.395**	-0.045	-0.144	0.034	0.227***
		(0.169)	(0.140)	(0.233)	(0.047)	(0.059)
Math Grade at baseline		0.007	0.005	-0.011	0.023**	0.011
		(0.008)	(0.017)	(0.023)	(0.010)	(0.019)
Child lives in one-parent household		-0.030	-0.018	0.000	-0.061**	-0.012
		(0.021)	(0.037)	(0.054)	(0.029)	(0.049)
Motherâs education: High School		0.005	-0.001	0.017	0.042	-0.052
		(0.020)	(0.034)	(0.054)	(0.027)	(0.044)
Motherâs education: College		-0.042	-0.105	-0.130	0.058*	-0.070
		(0.033)	(0.065)	(0.085)	(0.034)	(0.074)
Fatherâs education: High School		0.018	0.016	0.040	0.022	-0.002
		(0.019)	(0.034)	(0.053)	(0.024)	(0.044)
Fatherâs education: College		0.055	0.084	0.197**	-0.048	0.048
		(0.037)	(0.069)	(0.097)	(0.048)	(0.076)
Mother's job: white collar		-0.057	0.054	0.014	-0.025	-0.192
		(0.050)	(0.150)	(0.304)	(0.049)	(0.123)
Mother's job: blue collar		-0.106**	-0.010	-0.147	-0.006	-0.182
		(0.050)	(0.150)	(0.303)	(0.050)	(0.124)
Father's job: white collar		-0.004	-0.128**	-0.462**	0.078	0.003
		(0.056)	(0.063)	(0.214)	(0.064)	(0.101)
Father's job: blue collar		0.046	-0.079*	-0.366*	0.102	0.022
		(0.055)	(0.047)	(0.210)	(0.064)	(0.098)
Dep var mean	.758	.758	.881	.463	.931	.463
Obs.	2196	2196	530	529	607	530
R^2	0.170	0.196	0.059	0.091	0.081	0.123

Notes: This Table reports the coefficients from a OLS regression. The outcome is a dummy which assumes value 1 if the student completed the endline survey. The sample in this table includes all students who completed the baseline survey and therefore are included in the TOP randomization samples. Column 3 and 4 restricts the sample to students in TOP2020 assigned to the treatment and control group, respectively. Column 5 and 6 restricts the sample to students in TOP2022 assigned to the treatment and control group, respectively. Robust standard errors in parentheses. ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively.

Table A.VIII. Regressions with interaction term

	(1) Grade in math (admin)	(2) Grade in Italian (admin)	(3) Grade in English (admin)
Grade in math (survey)	0.244***		
	(0.045)		
Treated	0.350	-0.034	0.431
	(0.373)	(0.288)	(0.441)
Treated=1 × Grade in math (survey)	-0.049		
	(0.062)		
Grade in Italian (survey)		0.175***	
		(0.038)	
Treated=1 × Grade in Italian (survey)		0.019	
		(0.047)	
Grade in English (survey)			0.261***
			(0.054)
Treated=1 × Grade in English (survey)			-0.099
			(0.076)
Observations	180	230	101
Mean grade in admin (Control)	6.37	6.49	6.55
Mean grade in survey (Control)	5.29	5.59	5.35
R-squared	.568	.572	.617

Notes: This Table reports the coefficients from an OLS regression. ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively.

Table A.IX. Balance Table: TOP2020 vs. TOP2022 (full sample)

	(1) TOP2020	(2) TOP2022	(3) P-value	(4) Std diff.
Students				
Survey Data				
Female	0.396	0.452	0.008***	0.113
Immigrant	0.173	0.274	0.000***	0.242
Has Learning Disorder (Baseline)	0.324	0.254	0.000***	-0.155
Student is in 6th grade (Baseline)	0.314	0.302	0.517	-0.026
Student is in 7th grade (Baseline)	0.341	0.368	0.190	0.056
Student is in 8th grade (Baseline)	0.338	0.331	0.715	-0.015
How much does the student enjoy math? (Baseline)	2.748	2.592	0.000***	-0.151
How much does the student enjoy Italian? (Baseline)	3.064	2.973	0.015**	-0.105
How much does the student enjoy English? (Baseline)	3.125	3.058	0.161	-0.060
Child-reported homework time (Baseline)	86.333	80.721	0.003***	-0.127
Admin data				
Math Grade at baseline	6.210	6.369	0.001***	0.146
Italian Grade at baseline	6.364	6.610	0.000***	0.256
English Grade at baseline	6.445	6.700	0.000***	0.225
Parents				
Child lives in one-parent household (Baseline)	0.238	0.245	0.708	0.016
Mother's highest degree: High School (Baseline)	0.454	0.421	0.121	-0.067
Mother's highest degree: College (Baseline)	0.110	0.124	0.311	0.043
Father's highest degree: High School (Baseline)	0.369	0.384	0.475	0.031
Father's highest degree: College (Baseline)	0.074	0.076	0.804	0.008
Mother has a white collar job (Baseline)	0.407	0.399	0.749	-0.016
Mother has a blue collar job (Baseline)	0.581	0.551	0.248	-0.060
Father has a white collar job (Baseline)	0.337	0.323	0.537	-0.030
Father has a blue collar job (Baseline)	0.657	0.621	0.108	-0.075
Internet Access	0.936	0.992	0.000***	0.306
No access to device for tutoring (Baseline)	0.028	0.058	0.000***	0.147
Family time spent helping the student (Baseline)	48.810	34.028	0.000***	-0.381
Observations	1,059	1,137		

Notes: This Table shows the characteristics of students in the initial sample at baseline for TOP2020 and TOP2022. P-values for difference in means are reported in the third column. The last column also reports the standardized difference between group averages.

Table A.X. Balance Table for tutors: TOP2020 vs. TOP2022

	(1) TOP2020	(2) TOP2022	(3) P-value	(4) Std diff.
Tutor is female	0.700	0.741	0.122	0.092
Tutor was born outside of Italy (Baseline)	0.017	0.097	0.000***	0.338
Raw tutor GPA (Baseline)	26.727	26.411	0.042**	-0.123
Tutor has prior volunteering experience (Baseline)	0.822	0.781	0.082*	-0.102
Tutor's motivation: wants to help / be useful'	0.831	0.565	0.000***	-0.575
Tutor has prior tutoring experience (Baseline)	0.958	0.781	0.000***	-0.515
Tutor has DSA training (Baseline)	0.038	0.020	0.074*	-0.109
Tutor has ELL training (Baseline)	0.038	0.013	0.010**	-0.161
Tutor studies economics (Baseline)	0.287	0.303	0.549	0.035
Tutor studies education (Baseline)	0.065	0.018	0.000***	-0.240
Tutor studies humanities (Baseline)	0.138	0.389	0.000***	0.564
Tutor studies STEM (Baseline)	0.337	0.270	0.016**	-0.146
Observations	530	607		

Notes: This Table shows the characteristics of students in the sample selected to be tutor at baseline for TOP2020 and TOP2022. For TOP2020, we selected a limited number of tutors considering the subject availability, their training certificates (learning disorders and immigrants). For TOP2022, we assigned all available volunteers. P-values for difference in means are reported in the third column. The last column also reports the standardized difference between group averages.

Table A.XI. Balance Table of Tutors for TOP2020

	(1) Control	(2) Treatment	(3) 3h Tutoring	(4) 6h Tutoring	(5) P-value (1)-(2)	(6) P-value (1)-(3)	(7) P-value (3)-(4)
Female	0.719	0.700	0.688	0.730	0.339	0.252	0.343
Tutor's Major - STEM	0.158	0.337	0.356	0.284	0.004***	0.000***	0.110
Tutor's Major - Humanities	0.158	0.138	0.113	0.206	0.174	0.015**	0.014**
Tutor's Major - Education	0.058	0.065	0.055	0.092	0.207	0.812	0.169
Tutor's Major - Economics	0.307	0.287	0.327	0.177	0.718	0.460	0.000***
Degree Type: Bachelor's	0.510	0.472	0.495	0.411	0.362	0.600	0.087*
GPA	26.386	26.011	26.402	24.954	0.809	0.946	0.014**
Volunteering Before	0.777	0.822	0.817	0.837	0.909	0.075*	0.586
Hours studying per day	5.332	5.076	5.149	4.879	0.970	0.116	0.193
Familiarity with computer	3.247	3.323	3.296	3.397	0.735	0.190	0.095*
Do you have younger siblings?	0.502	0.549	0.555	0.532	0.819	0.063*	0.639
Tutor's Mother: Completed high school	0.429	0.480	0.474	0.496	0.195	0.115	0.646
Tutor's Mother: Completed College	0.420	0.384	0.398	0.348	0.734	0.424	0.287
Tutor's Father: Completed high school	0.405	0.434	0.432	0.440	0.554	0.336	0.874
Tutor's Father: Completed College	0.399	0.384	0.382	0.390	0.203	0.535	0.870
Motivation TOP: help others	0.783	0.830	0.843	0.794	0.012**	0.005***	0.211
Observations	1,532	523	382	141			

Notes: This Table shows the characteristics of treated and control tutors in the initial sample at baseline for TOP2020. P-values for difference in means, controlling for the randomization conditions (subject and time availability, previous training and tutoring experience), are reported in column (5). The p-values for the coefficient *treat* and *treat_intense* are reported in columns (6) and (7), respectively. For the p-values in columns (6) and (7) we do not include the randomization controls to show the difference in the characteristics of tutors the tutees with the 3 hours and 6 hours treatment were exposed to.

Table A.XII. Balance Table of Tutors for TOP2022

	(1) Treatment	(2) Individual Tutoring	(3) Group Tutoring	(4) P-value (3)-(4)
Tutor is female	0.741	0.731	0.755	0.507
Tutor was born outside of Italy (Baseline)	0.097	0.101	0.091	0.700
Raw tutor GPA (Baseline)	26.411	26.321	26.531	0.377
Tutor has prior volunteering experience (Baseline)	0.781	0.775	0.789	0.683
Tutor's motivation: wants to help / be useful'	0.565	0.535	0.604	0.090*
Tutor has prior tutoring experience (Baseline)	0.781	0.775	0.789	0.683
Tutor has DSA training (Baseline)	0.020	0.009	0.034	0.040**
Tutor has ELL training (Baseline)	0.013	0.012	0.015	0.721
Tutor studies economics (Baseline)	0.303	0.269	0.347	0.039**
Tutor studies education (Baseline)	0.018	0.023	0.011	0.249
Tutor studies humanities (Baseline)	0.389	0.427	0.340	0.028**
Tutor studies STEM (Baseline)	0.270	0.260	0.283	0.533
Observations	1,532	607	342	265

Notes: This Table shows the characteristics of treated and control tutors in the initial sample at baseline for TOP2022. P-values for difference in means are reported in column (4).

Table A.XIII. Estimation of the impact of TOP 2020 on other academic outcomes and beliefs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Regression Model 1			Regression Model 2					
	Treatment	SE	FWER	Treatment	SE	Intense	SE	Mean Dep	Obs
Outcomes									
Panel A: Italian									
Grade	0.077	(0.050)	[0.360]	0.056	(0.054)	0.061	(0.078)	6.49	1059
Fail	-0.017	(0.016)	[0.560]	-0.013	(0.018)	-0.014	(0.025)	0.08	1059
Std Test Score Endline	0.173**	(0.079)	[0.210]	0.141*	(0.084)	0.119	(0.117)	0.00	712
Std Test Score Follow-up	0.073	(0.094)	[0.560]	0.024	(0.100)	0.188	(0.146)	0.00	342
Panel B: English									
Grade	0.006	(0.059)	[0.930]	0.006	(0.063)	-0.020	(0.100)	6.55	1059
Fail	-0.015	(0.019)	[0.790]	-0.009	(0.020)	-0.018	(0.032)	0.12	1059
Std Test Score Endline	0.189**	(0.086)	[0.160]	0.087	(0.091)	0.343***	(0.129)	0.00	516
Std Test Score Follow-up	0.040	(0.091)	[0.880]	0.000	(0.098)	0.147	(0.152)	-0.00	343
Panel C: Beliefs									
Students									
Beliefs on math test score	0.032**	(0.014)	[0.060]	0.026*	(0.015)	0.022	(0.020)	0.66	704
Beliefs on test score	0.032***	(0.012)	[0.010]	0.029**	(0.012)	0.012	(0.018)	0.65	705
Beliefs on grade	0.293**	(0.117)	[0.060]	0.332***	(0.125)	-0.126	(0.177)	6.17	524
Parents									
Beliefs on math test score	0.293**	(0.117)	[0.060]	0.332***	(0.125)	-0.126	(0.177)	6.17	524
Beliefs on test score	0.029**	(0.013)	[0.060]	0.028**	(0.014)	0.002	(0.019)	0.67	746
Teachers									
Beliefs on math test scor	0.024**	(0.010)	[0.060]	0.026**	(0.011)	-0.007	(0.015)	0.69	756
Beliefs on test score	0.227**	(0.098)	[0.000]	0.183*	(0.105)	0.152	(0.142)	4.61	704
Beliefs on grade	0.045***	(0.013)	[0.000]	0.047***	(0.014)	-0.010	(0.022)	0.48	792

Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values (based on 10,000 replications) in square brackets.

Table A.XIV. Estimation of the impact of TOP 2022 on other academic outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Regression Model 1			Regression Model 2					
	Treatment	SE	FWER	Treatment	SE	Group	SE	Mean Dep	Obs
Outcomes									
Panel A: Italian									
Grade	0.022	(0.045)	[0.850]	0.011	(0.054)	0.026	(0.065)	6.36	1121
Fail	0.001	(0.019)	[0.950]	0.004	(0.023)	-0.007	(0.028)	0.10	1121
Std Test Score Endline	0.047	(0.065)	[0.850]	0.151**	(0.077)	-0.247***	(0.094)	-0.00	952
Panel B: English									
Grade	0.087	(0.060)	[0.220]	0.069	(0.069)	0.042	(0.087)	6.44	1121
Fail	-0.057***	(0.021)	[0.030]	-0.056**	(0.023)	-0.004	(0.027)	0.16	1121
Std Test Score Endline	0.082	(0.062)	[0.220]	0.121	(0.074)	-0.092	(0.087)	-0.00	951
Panel C: Beliefs									
Students									
Belief on math test score	0.008	(0.013)	[0.920]	0.016	(0.016)	-0.017	(0.019)	0.64	946
Belief on own test score	0.007	(0.010)	[0.920]	0.018	(0.012)	-0.026*	(0.015)	0.66	947
Belief on grade	-0.014	(0.107)	[0.990]	-0.076	(0.127)	0.149	(0.159)	5.52	959
Parents									
Belief on math test score	0.013	(0.013)	[0.770]	0.013	(0.015)	0.000	(0.018)	0.63	998
Belief on test score	-0.000	(0.009)	[0.990]	-0.005	(0.010)	0.010	(0.013)	0.69	1001
Teachers									
Belief on math test score	0.041***	(0.014)	[0.010]	0.046***	(0.016)	-0.013	(0.020)	0.48	821
Belief on test score	0.015*	(0.009)	[0.450]	0.014	(0.010)	0.004	(0.013)	0.53	1068
Belief on grade	0.175**	(0.086)	[0.260]	0.262***	(0.100)	-0.201*	(0.121)	5.33	1069

Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values (based on 10,000 replications) in square brackets.

Table A.XV. Estimation of the impact of TOP 2020 on other academic outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Regression Model 1			Regression Model 2					
	Treatment	SE	FWER	Treatment	SE	Intense	SE	Mean Dep	Obs
Outcomes									
Panel A: Aspirations									
Students									
Aspirations University	0.064*	(0.037)	[0.319]	0.039	(0.039)	0.100*	(0.054)	0.36	674
Self-efficacy university	0.044	(0.032)	[0.523]	0.058	(0.036)	-0.051	(0.046)	0.21	682
High-school: vocational	-0.065*	(0.034)	[0.270]	-0.056	(0.037)	-0.035	(0.049)	0.31	681
High-school: top tier	0.008	(0.026)	[0.752]	-0.007	(0.028)	0.060	(0.040)	0.16	681
Parents									
Aspirations University	0.045	(0.034)	[0.523]	0.037	(0.036)	0.038	(0.050)	0.34	765
Self-efficacy university	0.082**	(0.032)	[0.071]	0.083**	(0.035)	-0.002	(0.048)	0.29	772
Teachers									
Aspirations University	0.029	(0.022)	[0.523]	0.019	(0.023)	0.022	(0.037)	0.14	839
Panel B: Socio-emotional									
Students									
Logic task: difficult	0.047	(0.038)	[0.472]	0.008	(0.041)	0.147***	(0.054)	0.56	685
Logic task: give-up	-0.033	(0.026)	[0.472]	-0.036	(0.028)	0.012	(0.037)	0.14	685
Grit	0.015	(0.010)	[0.416]	0.017	(0.011)	-0.005	(0.015)	0.68	673
Locus of control	0.024***	(0.009)	[0.029]	0.023**	(0.009)	0.005	(0.012)	0.71	685
Parents									
Grit	-0.006	(0.010)	[0.547]	-0.002	(0.010)	-0.013	(0.015)	0.67	736
Panel C: Wellbeing									
Students									
Depression	-0.018*	(0.009)	[0.139]	-0.017*	(0.010)	-0.005	(0.014)	0.55	669
Happiness	0.025	(0.018)	[0.259]	0.025	(0.018)	-0.001	(0.026)	0.61	665
Parents									
Depression	-0.012	(0.008)	[0.259]	-0.010	(0.009)	-0.007	(0.012)	0.59	731
Happiness	0.035**	(0.016)	[0.107]	0.032*	(0.017)	0.010	(0.022)	0.60	741

Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values (based on 10,000 replications) in square brackets.

Table A.XVI. Estimation of the impact of TOP 2022 on instruction time

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Regression Model 1			Regression Model 2					
	Treatment	SE	FWER	Treatment	SE	Intense	SE	Mean Dep	Obs
Outcomes									
Panel A: Homework time									
Minutes per day reported by students	10.190***	(3.492)	[0.003]	7.126*	(3.729)	11.940**	(4.989)	88.26	690
Minutes per day reported by parents	8.663***	(2.784)	[0.002]	6.027**	(3.006)	10.764***	(3.993)	81.71	778
Always homework reported by teachers	0.103***	(0.032)	[0.001]	0.101***	(0.034)	-0.011	(0.051)	0.28	851
Panel B: Online classes									
Always follows reported by students	0.024	(0.029)	[0.404]	0.033	(0.031)	-0.032	(0.043)	0.83	687
Always follows reported by parents	-0.002	(0.024)	[0.934]	0.007	(0.025)	-0.031	(0.037)	0.88	777
Always follows reported by teachers	0.107***	(0.032)	[0.001]	0.100***	(0.035)	0.024	(0.049)	0.57	859

Notes: The continuous dependent variable expressed in minutes per day in Panel A is constructed by assigning midpoint values to the intervals displayed in Figure 4. We asked students whether in the last month of school they had been following online classes regularly, and we posed the same question to parents regarding their children. The dependent variable in Panel B is a dummy taking value 1 if the answer is "Yes, every time there was an online class". Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values (based on 10,000 replications) in square brackets.

Table A.XVII. Causal Forest Heterogeneous Treatment Effect: Median of Predicted Impact by Student and Parents Characteristics

	Performance Index		Aspirations Index		Socio-Emotional Index		Well-being Index	
Students								
	p(50)	p(100)	p(50)	p(100)	p(50)	p(100)	p(50)	p(100)
Male	0.582	0.599	0.541	0.614	0.607	0.558	0.632	0.553
Immigrant	0.229	0.227	0.187	0.192	0.179	0.181	0.079	0.280
Learning disorders	0.224	0.396	0.300	0.297	0.315	0.325	0.307	0.328
Grade in maths	6.438	6.113	6.389	6.380	6.458	6.233	6.366	6.347
Grade in Italian	6.047	5.787	5.967	6.038	6.214	5.773	6.164	5.847
Grade in English	6.471	6.043	6.346	6.286	6.532	6.101	6.366	6.228
Grade 6	0.303	0.366	0.341	0.333	0.364	0.252	0.319	0.287
Grade 7	0.338	0.334	0.308	0.346	0.295	0.409	0.369	0.354
Grade 8	0.359	0.300	0.351	0.321	0.341	0.339	0.313	0.359
How much do you like maths?	2.800	2.670	2.692	2.851	2.911	2.584	2.855	2.659
How much do you like Italian?	3.117	2.945	3.046	3.097	3.044	3.014	2.940	3.127
How much do you like English?	3.250	2.987	3.123	3.159	3.264	3.004	3.107	3.138
Perseverance	0.834	0.796	0.826	0.818	0.850	0.804	0.849	0.815
Importance Luck (vs. Effort)	0.055	0.059	0.050	0.047	0.042	0.048	0.037	0.053
Familiarity with computer	3.280	2.967	3.304	3.064	3.425	2.971	3.359	3.027

Notes: This Table presents the average value of students and parents' characteristics for students with below median conditional treatment effect and above median conditional treatment effect. Student level conditional treatment effects are estimated using the Causal Forest methodology.

Table A.XVIII. Estimation of the impact of TOP 2022 on other academic non-outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Regression Model 1			Regression Model 2					
	Treatment	SE	FWER	Treatment	SE	Group	SE	Mean Dep	Obs
Outcomes									
Panel A: Aspirations									
Students									
Aspirations University	0.019	(0.030)	[0.949]	0.000	(0.035)	0.044	(0.042)	0.31	961
Self-efficacy university	-0.004	(0.029)	[0.985]	-0.004	(0.035)	-0.001	(0.041)	0.27	965
High-school: vocational	-0.034	(0.027)	[0.713]	-0.029	(0.031)	-0.012	(0.037)	0.24	966
High-school: top tier	0.038	(0.023)	[0.528]	0.031	(0.027)	0.015	(0.034)	0.13	966
Parents									
Aspirations University	0.032	(0.028)	[0.729]	0.028	(0.033)	0.008	(0.039)	0.27	1000
Self-efficacy university	-0.002	(0.030)	[0.985]	0.001	(0.035)	-0.007	(0.043)	0.36	1002
Teachers									
Aspirations University	-0.007	(0.014)	[0.950]	-0.018	(0.016)	0.025	(0.019)	0.06	1076
Panel B: Socio-emotional									
Students									
Logic task: difficult	-0.011	(0.034)	[0.924]	-0.003	(0.039)	-0.020	(0.047)	0.56	967
Logic task: give-up	0.007	(0.023)	[0.924]	0.001	(0.026)	0.013	(0.031)	0.13	967
Grit	-0.007	(0.009)	[0.889]	-0.015	(0.010)	0.018	(0.013)	0.66	958
Locus of control	0.006	(0.008)	[0.889]	0.008	(0.009)	-0.003	(0.011)	0.70	961
Parents									
Grit	-0.007	(0.008)	[0.889]	-0.011	(0.009)	0.008	(0.012)	0.65	955
Panel C: Wellbeing									
Students									
Depression	-0.006	(0.008)	[0.879]	-0.000	(0.009)	-0.013	(0.011)	0.58	948
Happiness	0.007	(0.014)	[0.926]	0.001	(0.016)	0.015	(0.020)	0.69	955
Parents									
Depression	-0.003	(0.007)	[0.926]	-0.000	(0.008)	-0.005	(0.010)	0.59	968
Happiness	0.001	(0.012)	[0.926]	0.000	(0.014)	0.002	(0.017)	0.68	969

Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values (based on 10,000 replications) in square brackets.

Table A.XIX. Estimation of the impact of TOP 2022 on other academic outcomes

	(1)	(2)	(3)
	Test Score	Vocational track choice	Vocational track recommendation
Sample: TOP2020 in	Grade7	Grade7 and 8	Grade6, 7 and 8
Outcome observed in	2021	2021-2022	2020-2021-2022
Panel A: Overall results			
Treatment	0.072	-0.046	-0.009
	(0.087)	(0.034)	(0.026)
FWER	[0.662]	[0.454]	[0.728]
Sharpened q-value	[1.000]	[1.000]	[1.000]
R^2	0.405	0.170	0.248
Panel B: Results of 3 vs. 6 hours of tutoring			
Treatment	0.013	-0.037	0.001
	(0.097)	(0.037)	(0.029
Intense Treatment	0.231	-0.025	-0.026
	(0.143)	(0.056)	(0.042)
Treat+Intense Treatment==0	0.070	0.253	0.529
R^2	0.410	0.177	0.262
Mean Dep:	-0.00	0.49	0.67
Obs	341	775	1008

Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values (based on 10,000 replications) and sharpened q-value in square brackets. The dependent variable for each regression is listed in the column heading and it refers to math. Only only students in grade 7 during TOP2020 participated in the test score implemented by INVALSI in grade 8 in 2021 (column 1) and the track choice is available only for students that participated in TOP2020 in grade 7 and 8 (column 2). "Mean Dep" at the bottom of the table is the mean of the dependent variables for students in the control group. Controls included are parental education and occupation, gender, immigration status, learning disorders, school grade, teacher assigned grades and test scores in math at baseline. In Panel B, we control for whether the students was identified for intense tutoring.

Table A.XX. Robustness checks with different sets of controls

	(1)	(2)	(3)	(4)	(5)	(6)
	No controls		LASSO controls		Inverse Probability Weighting	
	Coeff	Std. Error	Coeff.	Std. Error	Coeff.	Std. Error
1. Academic and Beliefs (average of all subjects)						
Performance Index	0.246***	0.081	0.269***	0.079	0.239**	0.101
Outcomes reported by Students						
Beliefs right perc.	0.026**	0.012	0.031***	0.011	0.032*	0.016
Overconfidence Students	-0.053	0.038	-0.053	0.038	0.021	0.054
Grade - Self Rate	0.225**	0.109	0.228**	0.100	0.302**	0.128
Outcomes reported by Parents						
Beliefs - Parent	0.024**	0.010	0.021**	0.010	0.020	0.012
Overconfidence Parents	-0.030	0.039	-0.030	0.039	-0.046	0.052
Outcomes reported by Teachers						
Beliefs right perc. (Teacher reported)	0.035**	0.014	0.033**	0.013	0.035**	0.018
Overconfidence Teachers	-0.023	0.045	-0.023	0.045	0.006	0.066
Teacher's grade - All subjects	0.302**	0.131	0.287**	0.117	0.323**	0.154
Outcomes from administrative data						
Grade in maths (2019/20)	0.048	0.067	0.063	0.051	NA	NA
Grade in Ita (2019/20)	0.034	0.056	0.034	0.045	NA	NA
Grade in Eng (2019/20)	-0.036	0.064	-0.037	0.051	NA	NA
Std Blind score maths	0.047	0.110	0.025	0.094	NA	NA
Std Blind score Ita	0.064	0.103	0.089	0.095	NA	NA
Std Blind score Eng	0.013	0.105	0.014	0.089	NA	NA
2. Aspirations						
Aspirations Index	0.111	0.090	0.019	0.057	0.227**	0.097
Outcomes reported by Students						
Wants to go to university	0.039	0.038	-0.002	0.028	0.098**	0.046
Thinks he/she can go to university at Endline	0.041	0.033	0.023	0.030	0.066	0.041
Child will choose professional school	-0.047	0.036	-0.002	0.028	-0.069	0.046
Child will choose top-track school	-0.012	0.028	0.005	0.020	0.000	0.034
Outcomes reported by Parents						
Child will go to university	0.014	0.035	0.008	0.024	0.022	0.043
Thinks he/she can go to university at Endline (Parent)	0.058*	0.034	0.054*	0.030	0.086**	0.041
Outcomes reported by Teachers						
Students intends to go to Uni	0.014	0.024	0.022	0.020	-0.001	0.029
3. Socio-Emotional Skills						
Socio-Emotional Skills	0.125	0.079	0.087	0.070	0.125	0.095
Outcomes reported by Students						
Grit	0.038	0.039	0.039	0.038	0.004	0.052
Grit: gave up	-0.034	0.026	-0.035	0.025	-0.012	0.034
Grit	0.013	0.010	0.011	0.009	0.017	0.013
Locus of control - Student	0.024***	0.009	0.022**	0.009	0.019*	0.011
Outcomes reported by Parents						
Grit (Parent)	-0.005	0.010	-0.007	0.010	-0.011	0.013
4. Well-being						
Well-being Index	0.156*	0.080	0.134*	0.076	0.087	0.102
Outcomes reported by Students						
Depression scale (Student reported)	-0.019**	0.009	-0.020**	0.009	-0.017	0.012
Happiness index at Endline (Student)	0.023	0.018	0.023	0.018	-0.009	0.023
Outcomes reported by Parents						
Depression scale (Parent reported)	-0.011	0.008	-0.009	0.007	0.005	0.010
Happiness index at Endline (Parent)	0.035**	0.016	0.035**	0.016	0.023	0.019

Notes: Randomization round fixed effects included in all regressions. The controls included for each regression and selected with LASSO are listed in Appendix Table A.XXI. The results with inverse probability weighting include all standard controls as in our main Tables.

Table A.XXI. LASSO selected variables

(a). Academic Outcomes and Beliefs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Performance	Students			Parents		Teachers		
		Beliefs	Overconfidence	Grade	Beliefs	Overconfidence	Beliefs	Overconfidence	Grade
Self-efficacy: university		✓		✓	✓		✓		✓
How much do you like maths?				✓					
Has learning disability					✓				
How much do you like English?					✓				
Homework time before lockdown (Square-Root)									✓

(b). Aspirations Index

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	Aspirations Index	Students				Parents		Teachers
		Aspirations	Self-efficacy	High school vocational	High school top track	Aspirations	Self-efficacy	Aspirations
Has learning disability	✓							
Homework time before lockdown	✓							
Self-efficacy: university	✓	✓	✓			✓	✓	✓
Goal: university (parent) - Enroll in a technical institute (istituto tecnico)	✓	✓				✓		
Goal: Professional institute	✓	✓		✓		✓	✓	
High school preference 1: natural sciences high-school	✓				✓			✓
High school preference 1: professional high-school	✓			✓				
Continue study length - Enroll in a technical institute (istituto tecnico)	✓	✓				✓		
Goal: vocational high-school	✓	✓		✓		✓		
Continue study length - Enroll in high school (liceo)		✓						
High school preference 2: vocational high-school				✓				
High school: scientific high-school					✓			
High school preference 2: natural sciences high-school					✓			
Goal: university (parent) - Work after finishing school						✓		
Goal: university (parent) - Enroll in high school (liceo)						✓		

(c). Socio-emotional Skills

	(1)	(2)	(3)	(4)	(5)	(6)
	Socio-Emotional Index	Students				Parents Grit
		Perservance: difficulty	Perseverance: give up	Grit	Locus of control	
Has learning disability	✓					
Self-efficacy: university	✓			✓	✓	✓
Barriers: own ability	✓			✓		✓
Who helps child with homework - during closure (Person 1) - I do	✓					
Goal: Professional institute						
Logic question: correct+Perseverance		✓				
Perseverance			✓			

(d). Well-being

	(1)	(2)	(3)	(4)	(5)
	Well-being Index	Students		Parents	
		Depression	Happiness	Depression	Happiness
Self-efficacy: university	✓				
Barriers: own ability	✓	✓		✓	
Two statements q2 - Fate - Happen will happen	✓				
Who helps child with homework - during closure (Person 1) - I do					
Two statements q2 - Fate - Take action		✓			
Locus of control 1				✓	

Notes: This Table shows the controls selected using LASSO for each outcome variable.

Table A.XXII. Treatment Effect on Tutors

	(1)	(2)	(3)	(4)	(5)	(6)
	Income as Incentive vs. Income equality	Hard work vs. Luck	Work to Natives over Immigrants	If Effort Well-paid job	Easy to put in others' shoes	Make decisions irrespective others' feelings
Panel A: Sample –second endline for all tutors						
Treatment	0.151	0.118	0.327	-0.279	0.499**	-0.200
	(0.195)	(0.200)	(0.232)	(0.223)	(0.238)	(0.228)
Mean Dep:	4.46	3.47	1.81	2.98	3.15	2.79
Obs	739	742	738	738	740	740
Panel B: Sample – second endline with imputed values from first endline						
Treatment	-0.026	-0.020	0.377*	-0.485**	0.373*	-0.361*
	(0.181)	(0.183)	(0.213)	(0.209)	(0.219)	(0.206)
Mean Dep:	4.46	3.47	1.81	2.98	3.15	2.79
Obs	933	936	933	936	937	937

Notes: This Table reports the coefficients from an ordered logit regressions. Panel A restricts the sample only to tutors that replied to the second short endline. Panel B imputes the values from the first endline for those who did not reply to the second endline. All columns include the controls that were used to assign the tutors to students: whether the volunteer has tutoring experience and specific training (to support students with learning disorders or immigrants), their expertise in the subjects (math, Italian, English), their time availability (3 hours per week or 6 hours per week), whether they are on time in their university enrollment and if they confirmed their availability. We also include the additional tutor controls (gender, university faculty, whether they are enrolled in a undergraduate or master, GPA, previous volunteering activities, whether they applied to TOP to help others (motivation), parental education, and familiarity with the computer). The mean of the dependent variable reported in the Table is for the control group of tutors. Robust standard errors in parentheses. ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively.

B Data Appendix

B.1 Description of Outcome Variables

B.1.1 Student achievement

One of our main outcomes of interest is student learning. In normal years, standardized test scores are collected in May/June from all Italian students in grade 8 by the Institute for the Evaluation of the Italian Schooling System (INVALSI). However, due to the pandemic, these tests were not administered in 2020. In collaboration with two expert middle school teachers, we designed a (shorter) standardized test very close in format to the national standardized one. Our test included seven multiple-choice questions in math, seven in Italian, and five in English.

The test was administered to treatment and control students by enumerators. The research team sent to each student the link to complete the test score, but they needed a password to access it. The enumerator called each parent to set a time for the test. During the test, the student was on a video call with the enumerator, he/she opened the link with the questionnaire in his/her own device and entered the password given in real time by the enumerator: at that point, the test could start. Enumerators were clearly instructed *not* to help children during the test. Once the student completed and submitted the test online, the enumerators were available to discuss any doubts and answer potential questions.

By design, during the course of our program, TOP tutors did not follow a specific curriculum but they helped students with the homework assigned by school teachers. For this reason, the test we administered covered the basic achievement expected from students of each grade. On average, treated and control students answered correctly 56 percent of the questions (as shown in Appendix Table A.V, line 1): 67 percent in math, 48 percent in Italian, and 50 percent in English. The assessment covered a wide range of competencies and very few students reached a ceiling in terms of correct answers.

B.1.2 Student, parent, and teacher surveys

We asked students, parents, and teachers to complete a questionnaire that we sent by email and/or SMS. The questions covered four main sets of outcomes: academic achievement and beliefs, educational aspirations, socio-emotional skills, and psychological well-being. Teachers were asked to complete the same question for each child in their class that was either treated or control in TOP. The main outcomes in our empirical analysis will be indexes built extracting the first principal component from the variables in each category,

standardized to have mean zero and standard deviation one in the control group.

Academic outcomes and beliefs. We asked children and parents their beliefs on the number of correct questions for each subject of the test described in Section B.1.1.[22] Notably, 64 percent of the students and 71 percent of the parents are 'overconfident', in the sense that they expect a higher number of correct answers than the ones actually obtained in the test. Teachers' beliefs tend to be closer to the actual performance: teachers expect their students to correctly answer 49 percent of the questions on average, and only 36 percent of them are overconfident about children's performance.[23]

To obtain a different measure of achievement from the standardized test score, we asked each teacher to assign a grade from 1 to 10 to every child in our study (treated or control) that was in one of their classes. The average grade was 5.65, that is just below the pass grade of 6 in the Italian school context. This is consistent with the target of our intervention being children who were struggling to keep up with school work. We also asked children how they would rate their own school performance on a 1 to 10 scale, and the average was 6.29, somewhat more optimistic than the teachers but definitely not high.

Aspirations. Low goals and ambitions may lead students into an "aspiration trap" (Genicot and Ray, 2017; La Ferrara, 2019). Children from disadvantaged background may underinvest in their education, dropping out from school or choosing easier and less profitable high school tracks (Carlana et al., 2022). We hypothesized that TOP may have a direct effect on students' aspirations, by providing an alternative role model (the tutor) that may induce them to revise their goals. In our survey, we collected information from students on their long term educational goals (e.g., attend university), and on their short term plans (e.g., the type of high school they wanted to enroll in).[24] Among the students in our sample, only 15 percent are interested in a top-tier academic high school, while around 1/3 are planning to attend a vocational high school.[25] As for long-term

[22]Children were asked about this at the end of the test. Parents were asked this question in their endline survey, which typically took place after the kid had taken the test (neither the test nor the child's answers were shared with the parent). Indeed, we expect an impact of TOP on academic outcomes, but also on the beliefs and expectations of parents and teachers (Rosenthal, 1973). Overall, the data show that students and their parents are overconfident on their performance, with an average expected share of correct answers equal to 67 percent (for students) and 71 percent (for parents), against an actual share of correct answers of only 56 percent in the test.

[23]We tried to interview the math, Italian and English teachers for each child. For cases where one of the teachers did not reply, we calculate the beliefs (and grade, to be described below) as the average for the subjects for which data is available.

[24]In Italy, after grade 8 students need to choose their high school track. The schooling system is organized in top-tier academic tracks (scientific and classical *lyceum*), other academic tracks (linguistic, pedagogical, and other types of *lyceum*), technical tracks (with technological or economic focus, e.g., accounting), and vocational tracks.

[25]On average, in Italy 32 percent of students are enrolled in a top tier track and 14 percent in a

goals, 39 percent of the students at endline tell us that they are considering university education, and the figure is similar for parents (35 percent). The share is instead much lower when we ask teachers until what level the student should continue to study: only 14 percent say 'university'. Finally, we also collected a measure of self-efficacy (Bandura et al., 1999), asking students (and parents) whether, aside from what they would like to do in the future, they think they (their children) would be capable of successfully attending university if they wanted to.

Socio-emotional skills. Social distancing and school closure can result in a lack of opportunities to develop not only cognitive, but also socio-emotional skills in the classroom (Alan et al., 2019). In our endline survey we collected several outcomes to capture socio-emotional skills. First, in order to measure perseverance, we asked students to answer a logic question. At the end of the question, we asked them whether they wanted to answer a new question with the same level of difficulty, with a higher level of difficulty or whether they wanted to give up. We use their choice as an outcome measure of perseverance in a real effort task. Second, we measure 'grit' following the Short Grit Scale developed by Duckworth and Quinn (2009). Starting from 8 questions on a 5-point scale, we add up all the points and divide by 40. The maximum score on this scale is 1 (extremely gritty), and the lowest is 0 (not at all gritty). We asked the same questions to children and parents, finding a high correlation among their answers (0.64). Third, we collected a measure of 'locus of control' to capture the extent to which students believe they can control the outcome of events in their lives or whether fate and luck determine the course of action (Rotter, 1966). To calculate the final score, we start from 4 questions on a 5-point scale, add up all the points and divide by 20. Also for this outcome, the maximum score is 1 (high locus of control), and the lowest is 0 (low locus of control).

Well-being. Last but not least, we want to understand if the interaction with the tutor may have helped students to feel less isolated, possibly overcoming depression, and happier. For this purpose, we collected two measures of psychological well-being from students and their parents. The first is the Children's Depression Screener (ChilD-S) developed by Frühe et al. (2012), which is calculated aggregating a battery of 9 questions. The answers are given on a 4-point likert scale; we add up all the points and divide by 36. Also on this outcome, the maximum score is 1 (high level of depression), and the minimum is 0 (no depression). The second measure is a proxy for happiness: we asked whether students were feeling happy or unhappy during the lockdown, on a scale from 1 to 10 (10 being the maximum happiness). The correlation between the depression measure

vocational track. As expected from the targeting, at baseline, the sample of students who applied to TOP tends to include more low-achieving and low-aspiring students.

reported by parents and the one reported by students is 0.56, while the one for happiness it is 0.53.

B.2 Achievement Test

- **Example of math question for grade 8:** a is a odd number grater than 3. Which of the following expression represents the first odd number following a?
  - a+1
  - 2a+1
  - 2a-1
  - a+2
- **Example of Italian question for grade 8:** which of the following words corresponds to the grammar analysis: name, male, singular, derivative
  - Libreria
  - Libresco
  - Libraio
  - Libricini
- **Example of English question for grade 8:** Correct the following sentence: "You go to the swimming pool in Sunday".
  - You go the swimming pool in Sunday
  - You goes to the swimming pool in Sunday
  - You go to the swimming pool on Sunday
  - You go on swimming pool on Sunday

B.3 Student Questionnaire

- **Beliefs on academic outcomes:**
  - **Self-grade** Overall, considering your school performance in all assignments (homework, oral test, written test) in the month of May, how would you rate yourself compared to your classmates for each of the following subjects

(Math/Italian/English)? Consider a scale from 1 to 10, where 10 are the high-performing students (top 2-3 students) in the class and 1 are the low-performing students in the class (bottom 2-3 students).

  - **Beliefs on academic outcomes:** How many questions do you expect to have answer correctly in MATH/ITALIAN/ENGLISH?

- **Aspirations:**

  - **Education Goals.** Thinking about your future, how long do you think you will continue to study? Multiple choice options: (1) I think I will start working as soon as I complete this school (2) I think I will continue studying and enroll in high school, and start working after obtaining a diploma (3) I think I will continue studying and enroll in a technical institute, and start working after obtaining a diploma (4) I think I will continue studying and enroll in a professional/vocational institute (such as cosmetology, auto mechanic, etc.) and then start working (5) I think I will continue studying and reach university.
  - **High-school goal.** Which high-school would you like to do? Up to two choices are possible. Multiple choice options with all sub-tracks of high school including the two top tier tracks (humanistic and scientific) and vocational high-school
  - **Self-efficacy.** Apart from what you would like to do in the future, do you think you will be able to go to university when you are older if you wish to do so? Multiple choice options: (1) Very much (2) Much (3) Somewhat (4) Slightly (5) Not at all

- **Socio-emotional skills:**

  - **Perseverance.** First, we ask students to answer a first logic question. Second, if they want to persevere, we ask them a second logic question.
    * Would you like to try and answer another logic question? Multiple choice options: (1) Yes, I'd like to try with a question as difficult as this one (2) Yes, but I'd like to try an easier question (3) No
  - **Grit (following Duckworth and Quinn (2009)).** 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. (5-points likert scale)

1. I like schoolwork best which makes me think hard, even if I make a lot of mistakes.
2. Setbacks discourage me.
3. If I think I will lose in a game, I do not want to continue playing.
4. If I set a goal and see that it's harder than I thought I easily lose interest.
5. When I receive a bad result on a test I spend less time on this subject and focus on other subjects that I'm actually good at.
6. I work hard in tasks.
7. I prefer easy homework where I can easily answer all questions correctly.
8. If I'm having difficulty in a task, it is a waste of time to keep trying. I move on to things which I am better at doing.

– **Locus of control.** For each of the following statements, give a score from 1 to 5 indicating whether you agree or disagree with the statement.

1. Many of the unhappy things in people's lives are partly due to bad luck
2. Trusting in fate has turned out better for me than making a decision to take a definite course of action.
3. In the case of the well-prepared student, there is rarely, if ever, such a thing as an unfair test.
4. When I make plans, I am almost certain that I can make them work

- **Well-being:**

– **Depression (following Frühe et al. (2012)).** For each item please mark whether you agree or disagree with the statement. (4 points likert scale)

1. I am happy
2. I worry a lot
3. I feel sad
4. I get upset quickly
5. I am not in the mood for anything
6. I often think I did something wrong
7. It's often hard for me to concentrate
8. I feel lonely
9. I enjoy a lot of things

- **Happiness.** Think about the period of lockdown during Covid-19. During this period, how happy or unhappy have you been overall? 1-10 scale going from very unhappy to very happy

- **Additional outcomes:**

  - **Homework.** Think about the month of May this year. On average, how much time did you devote to doing homework every day? Multiple choice options: (1) Less than 15 minutes (2) 15 30 minutes (3) 30 - 60 minutes (4) 1 hour - 1 hour and a half (5) 1 hour and a half - 2 hours (6) 2 hours - 2 hours and a half (7) More than 2 hours and a half
  - **Following online classes.** In the month of May, have you been following classes online? Multiple choice options: (1) Yes, everytime there was an online class (2) Yes, but not always (3) Sometimes (4) No.
  - **Like subjects** How much do you like the following subjects (Math/Italian/English)? Check one box for each subject. Multiple choice options: Very much/ Much/ Somewhat/ Slightly/ Not at all
  - **Difficult online classes.** How difficult do you find it to follow classes online and use your school's online platform during the month of May? Multiple choice options: Extremely difficult /Very difficult / Moderately difficult /Slightly difficult / Not at all difficult

- **Tutoring experience and satisfaction:** we included few questions only for treated students.

B.4 Parent Questionnaire

- **Beliefs on academic outcomes.** As part of the final questionnaire for the project, we will ask your child 7 (7/5) questions in math (Italian/English). These are multiple choice questions prepared by middle school teachers that collaborate with us. How many correct answers do you expect your child to get? We will not share your answers with your child.

- **Aspirations:**

  - **Education Goals.** Thinking about your child's future, how long do you think he/she will continue to study? Multiple choice options: (1) I think he/she should start working as soon as he/she completes compulsory schooling (2)I

think he/she should continue studying and enroll in high school, and start working after obtaining a diploma (3) I think he/she should continue studying and enroll in a technical institute, and start working after obtaining a diploma (4) I think he/she should continue studying and enroll in a vocational high-school (such as cosmetology, auto mechanic, etc.) and then start working (5) I think he/she should continue studying and reach university.

  - **Self-efficacy.** Do you think your child has the capability to attend and successfully graduate from university if he/she wanted to? Multiple choice options: (1) Very much (2) Much (3) Somewhat (4) Slightly (5) Not at all

- **Socio-emotional skills:**

  - **Grit (following Duckworth and Quinn (2009)).** Here are a number of statements that may or may not apply to your child. There are no right or wrong answers, so please just answer truthfully. Think mainly about your perception from the last month. (5 points likert scale)

    1. He/she likes schoolwork best which makes him/her think hard, even if he/she makes a lot of mistakes.
    2. Setbacks discourage him/her.
    3. If he/she thinks he/she will lose in a game, he/she does not want to continue playing.
    4. If he/she sets a goal and sees that it's harder than he/she thought he/she easily loses interest.
    5. When he/she receives a bad result on a test he/she spends less time on this subject and focuses on other subjects that he/she is actually good at.
    6. He/she works hard in tasks.
    7. He/she prefers easy homework where he/she can easily answer all questions correctly.
    8. If he/she is having difficulty in a task, he/she thinks it is a waste of time to keep trying. He/she moves on to things which he/she is better at doing.

- **Well-being:**

  - **Depression (following Frühe et al. (2012)).** For each item please mark whether you believe the statement is true for your child. (4 points likert scale)

    1. is happy

2. worries a lot
3. feels sad
4. gets upset quickly
5. is not in the mood for anything
6. often thinks he/she did something wrong
7. is often hard for him/her to concentrate
8. feels lonely
9. enjoys a lot of things

– **Happiness.** Think about the period of lockdown during Covid-19. During this period, how happy or unhappy would you say your child has been overall?

- **Additional outcomes:**
  - **Homework.** Think about the month of May. On average, how much time did your child devote to studying and doing homework every day? Multiple choice options: (1) Less than 15 minutes (2) 15 30 minutes (3) 30 - 60 minutes (4) 1 hour - 1 hour and a half (5) 1 hour and a half - 2 hours (6) 2 hours - 2 hours and a half (7) More than 2 hours and a half
  - **Following online classes.** In the month of May, did your child follow classes online? Multiple choice options: (1) Yes, everytime there was an online class (2) Yes, but not always (3) Sometimes (4) No .
- **Tutoring experience and satisfaction:** we included few questions only for treated students.

B.4.1 Teacher Questionnaire

- **Beliefs on academic outcomes:**
  - **Beliefs on academic outcomes.** As part of the final questionnaire for the project, we will ask 7 (7/5) questions in math (Italian/English). These are multiple choice questions prepared by middle school teachers that collaborate with us. How many correct answers do you expect student X to get? We will not share your answers with your students.
  - **Grade.** Overall, considering the performance of your students in all assignments (homework, oral tests, written tests) in the month of May, how would you rate student X? Consider a scale from 1 to 10, where 10 are the best-performing

students (top 2-3 students) in the class and 1 are the least-performing students in the class (bottom 2-3 students).

- **Aspirations:**
  - **Education Goals.** Thinking about the future of the student, how long do you think he/she should continue to study? Multiple choice options: (1) I think he/she should start working as soon as he/she completes compulsory schooling (2)I think he/she should continue studying and enroll in high school, and start working after obtaining a diploma (3) I think he/she should continue studying and enroll in a technical institute, and start working after obtaining a diploma (4) I think he/she should continue studying and enroll in a vocational high-school (such as cosmetology, auto mechanic, etc.) and then start working (5) I think he/she should continue studying and reach university.
- **Additional outcomes:**
  - **Homework.** Did the student X do his/her homework during the month of May 2020? Multiple choice options: (1)Yes, regularly did all assigned homework (2) Yes, did the assigned homework most of the times, but not always (3) Sometimes/rarely (4) No
  - **Issue Behavior.** Especially thinking about the month of May, do you think that the student X had behaviour problems? Multiple choice options: (1) No (2) Yes - minor difficulties (3) Yes - some difficulties (4) Yes - serious difficulties
- **Tutoring experience and satisfaction:** we included few questions only for treated students.

B.4.2 Tutor Questionnaire

- **Empathy.** Below is a list of statements. Please read each statement carefully and rate how strongly you agree or disagree with it. There are no right or wrong answers. (4-points likert scale)
  1. I find it easy to put myself in somebody else's shoes.
  2. I am able to make decisions without being influenced by people's feelings.
- **Hard work.**

1. We would like to start by asking your views on a few issues. How would you place your views on this 1-10 scale? 1 means you agree completely with the statement on the left; 10 means you agree completely with the statement on the right; and if your views fall somewhere in between, you can choose any number in between.
   - Incomes should be made more equal vs. We need larger income differences as incentives for individual effort
   - In the long run, hard work usually brings a better life vs. Hard work doesn't generally bring success – it's more a matter of luck and connections
2. How much do you agree with the following statement? If students put effort in studying, they can get a well-paid job, independent of their family background. (4-points likert scale)

- **Tutoring experience and satisfaction:** we included few questions only for treated tutors.

C Heterogeneous Treatment Effects

C.1 Devices and Internet Connection

The key feature of TOP is the virtual nature of the interaction between tutor and student. By definition, the program requires a minimum technological input, namely an internet connection and a device that the two can use to have a video call. When we recruited middle school students, we told school principals that the beneficiaries should have access to a tablet or PC and to an internet connection for at least 3 hours per week. We have information on these aspects in the tutor endline survey and, for TOP2020, through the registries, where tutors recorded for each meeting the type of device used by the student and the occurrence of technical issues. Based on these sources, 22 percent of the students mainly used a smartphone to connect, and 77 percent of the students had technical issues during at least one of the meetings.

In Panel A of Table D.I we test whether the impact of the program was different for students who connected using a smartphone, compared to those who used a PC or a tablet. We find that it was not, except for aspirations, where the effect on students who used a smartphone is zero. Importantly, column 1 shows that, compared to an increase in test score of 0.27 SD for the students who connected with better devices, the impact for students who connected through a smartphone was 0.22 SD, significant at the 5 percent level. Considering that this may be a more disadvantaged group of students, the sizeable effect is particularly encouraging. It also leaves room for optimism in considering online tutoring as a tool that could be applied in contexts where the population may have lower income and/or only have access to smartphones.

In Panel B of Table D.I we consider the occurrence of technical issues during the tutoring sessions. While the coefficient of the variable 'Technical issues' is never statistically significant, its sign points to a lower effectiveness of the treatment in the presence of technical problems. This may be due to disruptions in the learning process as well as to shortening of the duration of the sessions.

Table D.I. Heterogeneity by whether student used phone for tutoring

	(1)	(2)	(3)	(4)
	Performance	Aspirations Index	Socio-emotional Index	Wellbeing Index
Panel A: Heterogeneous effects based on using phone reported by parents				
Treatment	0.271**	0.241**	0.122	0.049
	(0.105)	(0.105)	(0.100)	(0.107)
Tutored on phone only	-0.202	-0.136	0.021	0.037
	(0.159)	(0.150)	(0.158)	(0.190)
Treat+Tutored on phone==0	0.664	0.499	0.367	0.651
	[]	[]	[]	[]
Sharpened q-value	[1.000]	[1.000]	[1.000]	[1.000]
Obs	383	278	355	346
R^2	0.187	0.430	0.249	0.077
Panel B: Heterogeneous effects by technical isssues reported by tutors				
Treatment	0.342***	0.310***	0.120	0.070
	(0.109)	(0.115)	(0.106)	(0.117)
Technical issues during tutoring	-0.329**	-0.249*	0.016	-0.049
	(0.151)	(0.145)	(0.139)	(0.154)
Treat+Technical Issues==0	0.927	0.654	0.314	0.879
wyoung FWER p-value	[]	[]	[]	[]
Sharpened q-value	[0.543]	[0.543]	[0.836]	[0.836]
Obs	383	278	355	346
R^2	0.198	0.436	0.249	0.078

Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values in square brackets, based on 10,000 replications. The dependent variable for each regression is listed in the column heading. The mean of technical issues for students in the treatment group is 0.336, while the mean of phone usage is 0.22.

C.2 Students' and Parents' Characteristics

An important dimension of heterogeneity pertains to student demographics and socioeconomic background.

In Figures D.1 and D.2 we show the impact of treatment separately for different subgroups of students, split according to predetermined characteristics: gender, immigrant status, and learning disorders, SES status, parental education and occupation. The figure shows the estimated impact (relative to the control group) and associated 95 percent confidence interval, for our three main academic outcomes. Finally, Figures D.3 and D.4 show the effect on the three indexes related to non-academic outcomes. All indexes are standardized so that the coefficients represent the effect of treatment in units of standard deviations.

We find that boys and girls benefited from the program to the same extent, as did

native and immigrant students. One category that benefited slightly more was that of students with learning disorders, especially during the school closure. This is a group of students that may have faced particular difficulties with the learning methods and materials that schools provided during distance learning, in most cases not tailored to the needs of students with dyslexia, dyscalculia, or other disorders. Our program helped alleviate such difficulties. This is noteworthy because only 20 out of 523 tutors had specific training on learning disorders, so the vast majority of the 149 students in the treatment group who had a learning disorder got a tutor without specific training.[26]

When we look at aspirations and socio-emotional skills, the treatment effect appears to be similar across all these subgroups – the only pattern worth mentioning is that aspirations increase for natives but not for immigrants, possibly because the latter face different types of barriers when planning their future education (Carlana et al., 2022).

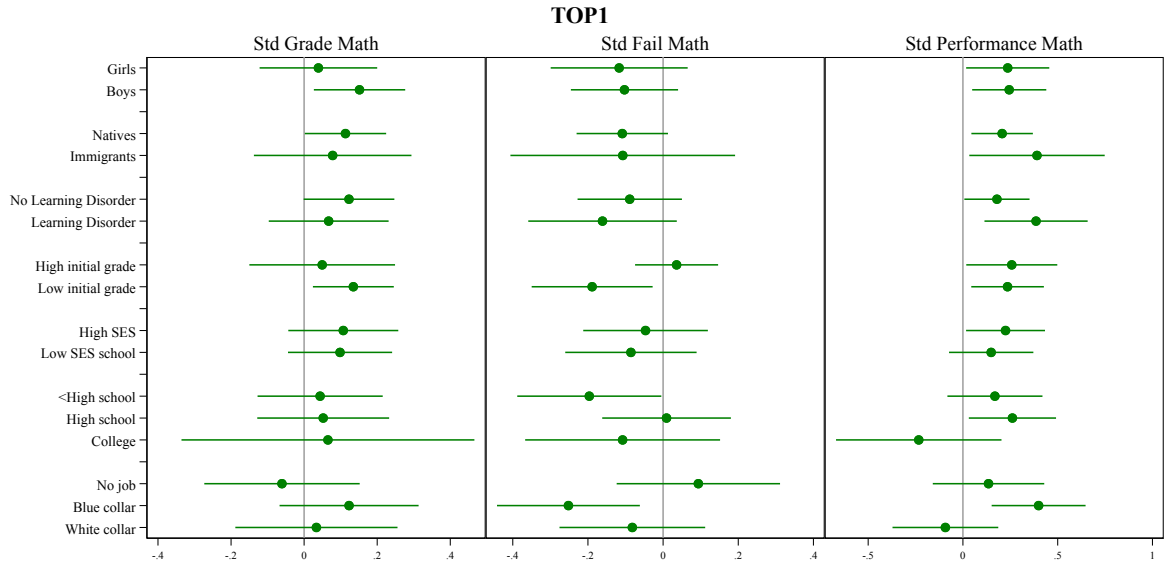
The outcome for which heterogeneity in treatment effects is most striking is psychological well being. TOP worked equally well for boys and girls and for students with and without learning disorders. But when we compare native and immigrant students, it is clear that the increased happiness and reduced depression we detected in TOP2020 is entirely driven by immigrant students. The magnitude of the effect for this group is a striking 0.77 SD increase in well-being. One possible interpretation is that immigrant students have a less dense network of friendships, hence felt more isolated during the lockdown: they were prevented from meeting classmates in school and they may not have been included in conversations that were happening online through WhatsApp or other groups. In fact, among students in the control group, immigrants have on average a 0.52 SD lower well-being compared to natives. Meeting regularly with a tutor proved particularly beneficial for the psychological well-being of these students.

Next, we examine impact heterogeneity by socioeconomic status of the family, as measured in the parents' questionnaire. We focus on three characteristics: SES status: education (less than high school, high school diploma, or higher); type of employment (none, blue collar job, or white collar job).

While the effects are not very precisely estimated, it appears that the gains in academic performance are slightly more concentrated among students from a disadvantaged background.

[26]We did prepare a module on how to teach students with learning disorders as part of our online support for tutors, so this was probably a useful resource for them. Indeed, we know from the endline survey that 62 percent of tutors assigned to students with learning disorders watched the videos compared to 47 percent of other tutors. They were also 12 percent more likely to participate in the group meetings for tutors organized by the pedagogical experts and 20 percent more likely to ask for a one-on-one meeting with an expert to get recommendations on how to effectively help their student.

Figure D.1. Heterogeneity on performance by student characteristics for TOP2020



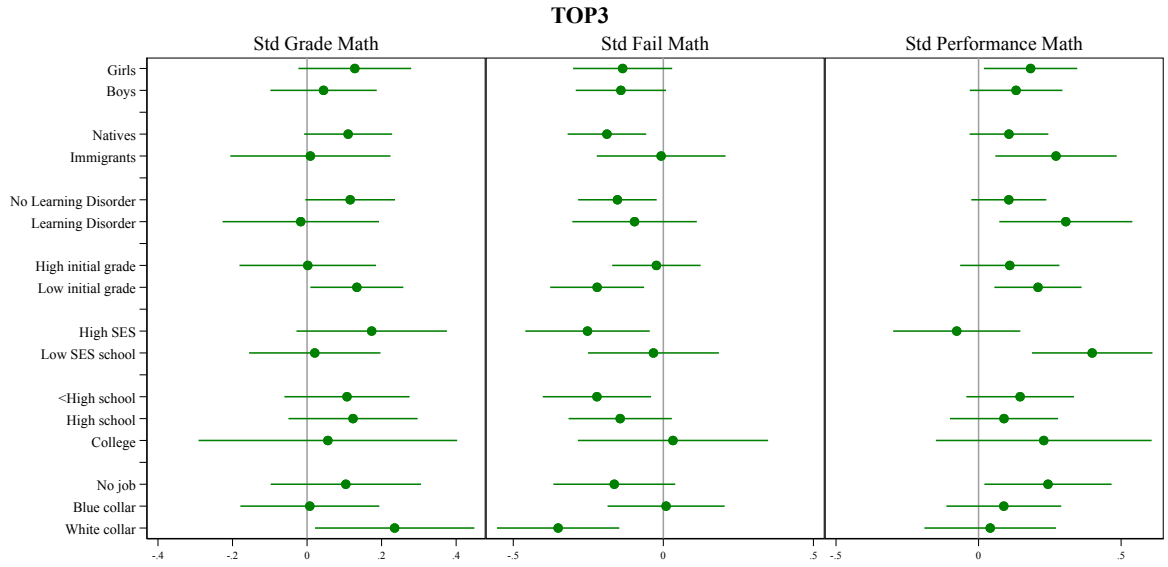
Notes: This Figure reports OLS estimates of the assignment to the TOP tutoring treatment by student characteristics. Randomization round fixed effects included in all regressions. Ex-ante student baseline controls include gender, immigrant, grade, parental education for each parent, employment type for each parent, learning disability, interest for the different subjects, perseverance, importance of luck, and familiarity with computers. The mean of the control group for each index is 0 and the standard deviation is 1. The bar shows 95% confidence intervals.

C.3 Tutor characteristics

After discussing impact heterogeneity in terms of students' and parents' characteristics, we investigate whether tutors' characteristics played a significant role in explaining the effects of the program.

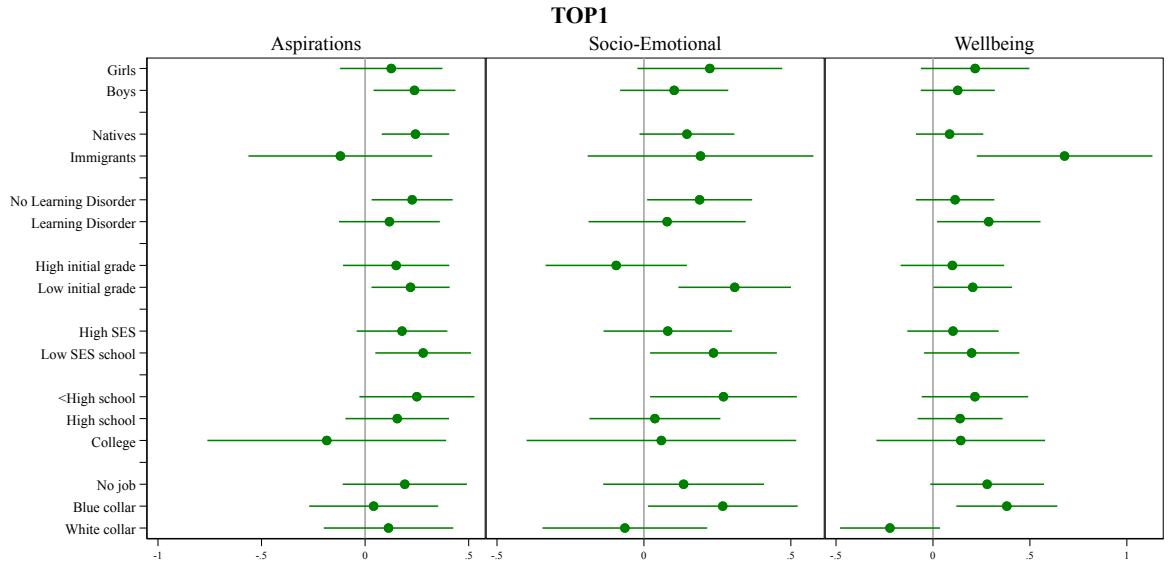
In Figures D.5 and D.6 we explore three sets of tutor baseline characteristics: gender, academic performance, and pro-social attitudes. For each of these characteristics, we report the coefficient of the treatment dummy interacted with the relevant subgroup, and a p-value for the null hypothesis that the two coefficients are the same.

Figure D.2. Heterogeneity on performance by student characteristics for TOP2022



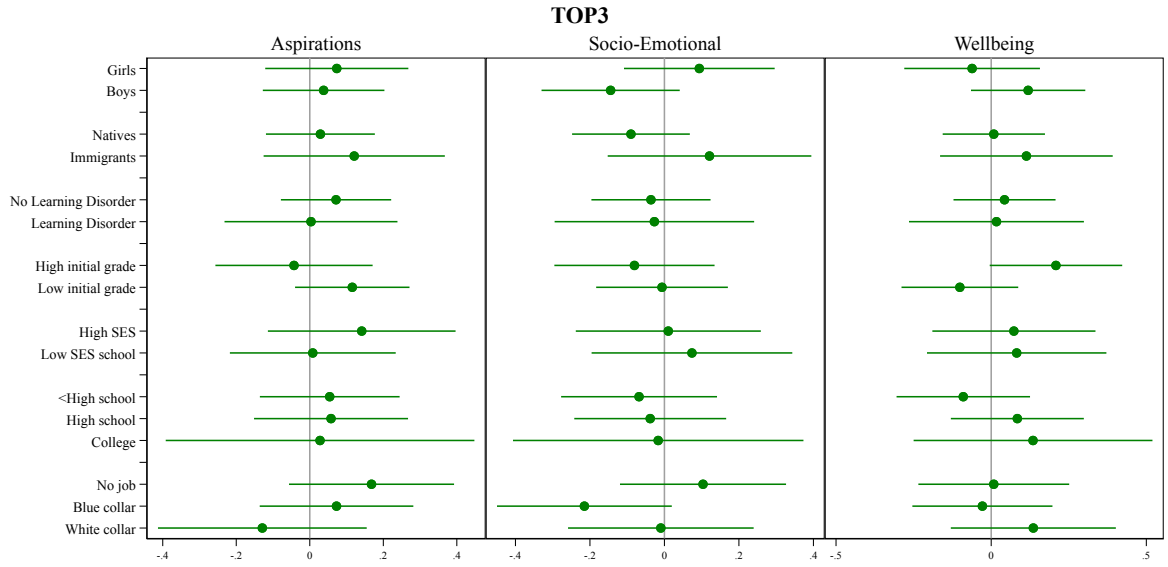
Notes: This Figure reports OLS estimates of the assignment to the TOP tutoring treatment by student characteristics. Randomization round fixed effects included in all regressions. Ex-ante student baseline controls include gender, immigrant, grade, parental education for each parent, employment type for each parent, learning disability, interest for the different subjects, perseverance, importance of luck, and familiarity with computers. The mean of the control group for each index is 0 and the standard deviation is 1. The bar shows 95% confidence intervals.

Figure D.3. Heterogeneity on other indexes by student characteristics for TOP2020



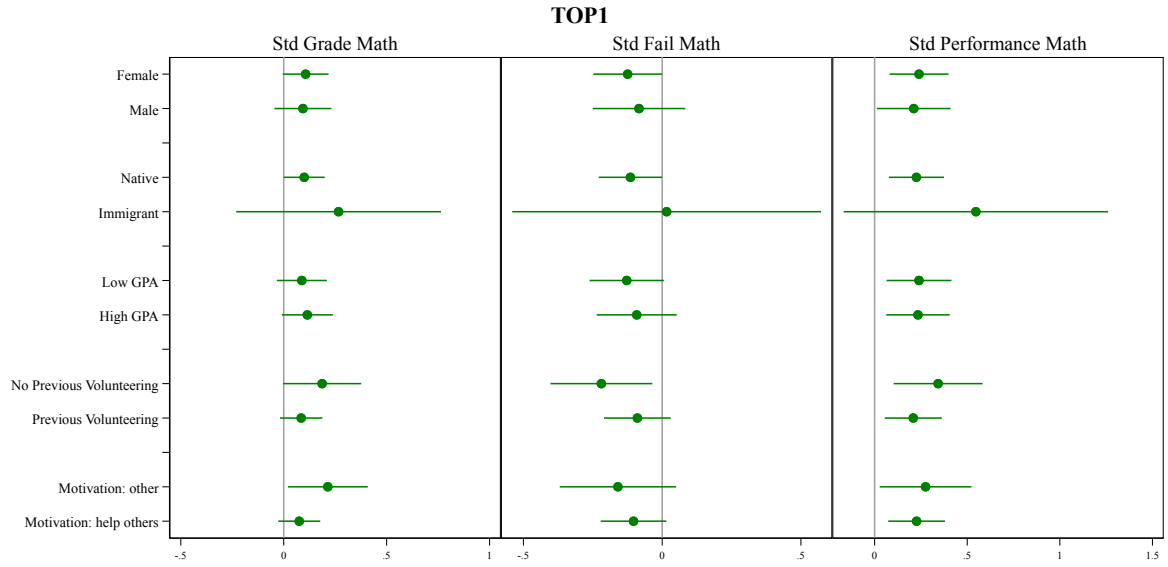
Notes: This Figure reports OLS estimates of the assignment to the TOP tutoring treatment by student characteristics. Randomization round fixed effects included in all regressions. Ex-ante student baseline controls include gender, immigrant, grade, parental education for each parent, employment type for each parent, learning disability, interest for the different subjects, perseverance, importance of luck, and familiarity with computers. The mean of the control group for each index is 0 and the standard deviation is 1. The bar shows 95% confidence intervals.

Figure D.4. Heterogeneity on other indexes by student characteristics for TOP2022



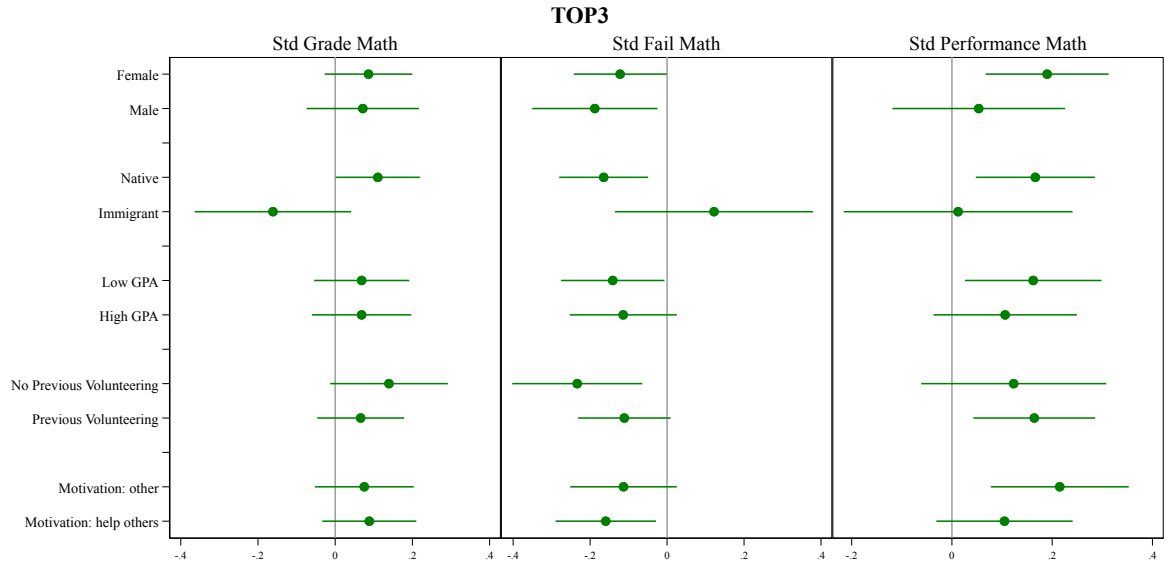
Notes: This Figure reports OLS estimates of the assignment to the TOP tutoring treatment by student characteristics. Randomization round fixed effects included in all regressions. Ex-ante student baseline controls include gender, immigrant, grade, parental education for each parent, employment type for each parent, learning disability, interest for the different subjects, perseverance, importance of luck, and familiarity with computers. The mean of the control group for each index is 0 and the standard deviation is 1. The bar shows 95% confidence intervals.

Figure D.5. Heterogeneity on performance by tutor characteristics for TOP2020



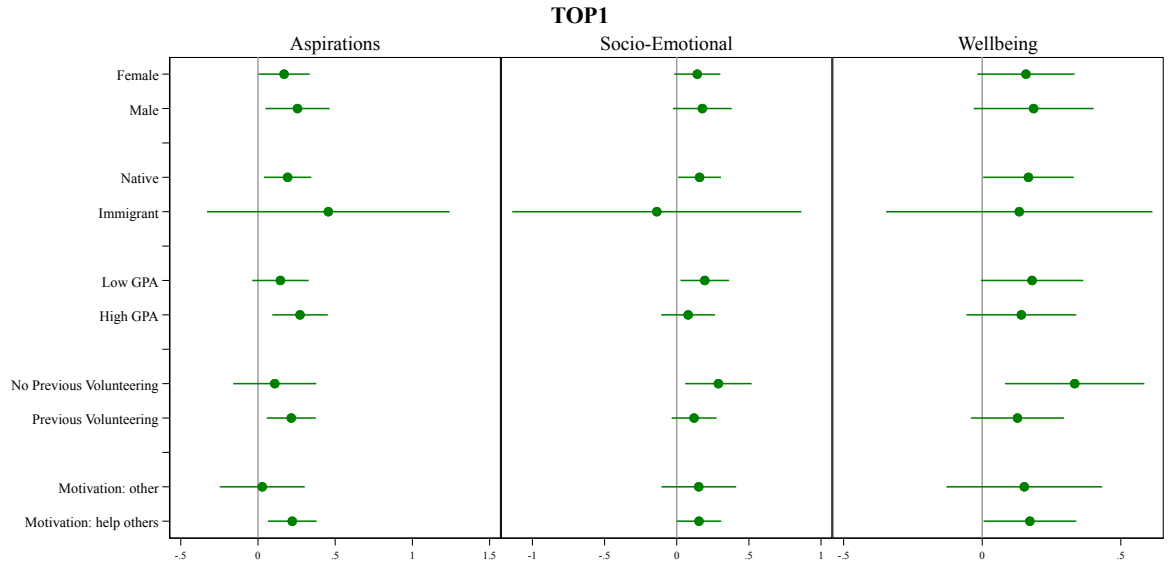
Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values in square brackets, based on 10,000 replications. The dependent variable for each regression is listed in the column heading. GPA scores have been standardized for each faculty: education, humanities, STEM, economics, and other faculties. Additionally, we control for missing GPA values and faculty in the regression. The p-value reported is the p-value of this post-estimation test of the differences between the two groups (e.g., male vs. female tutors).

Figure D.6. Heterogeneity on performance by tutor characteristics for TOP2020



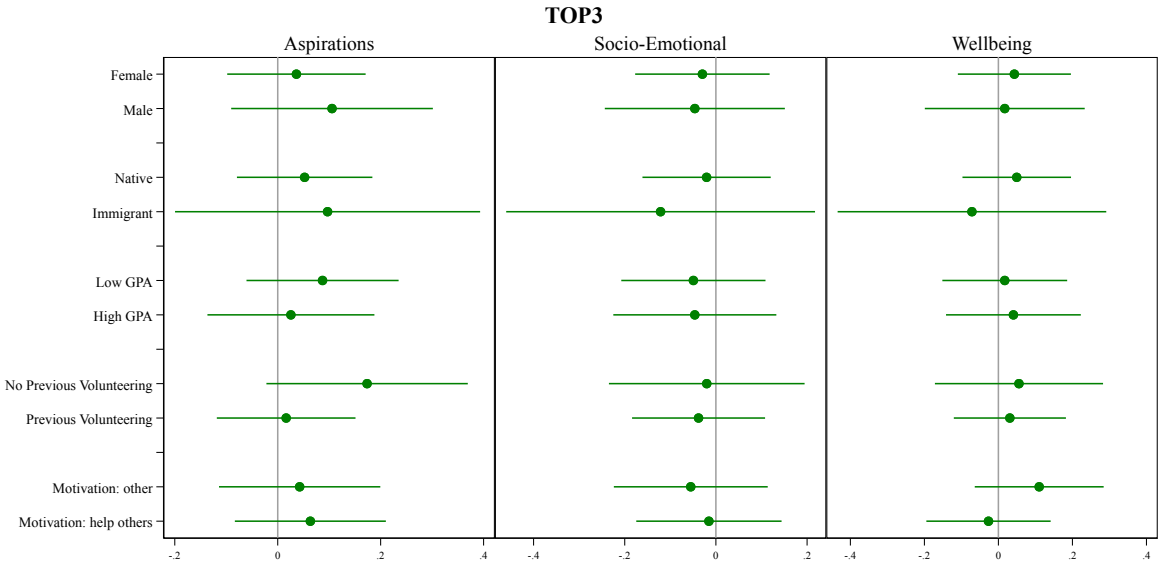
Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values in square brackets, based on 10,000 replications. The dependent variable for each regression is listed in the column heading. GPA scores have been standardized for each faculty: education, humanities, STEM, economics, and other faculties. Additionally, we control for missing GPA values and faculty in the regression. The p-value reported is the p-value of this post-estimation test of the differences between the two groups (e.g., male vs. female tutors).

Figure D.7. Heterogeneity on other indexes by tutor characteristics for TOP2020



Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values in square brackets, based on 10,000 replications. The dependent variable for each regression is listed in the column heading. GPA scores have been standardized for each faculty: education, humanities, STEM, economics, and other faculties. Additionally, we control for missing GPA values and faculty in the regression. The p-value reported is the p-value of this post-estimation test of the differences between the two groups (e.g., male vs. female tutors).

Figure D.8. Heterogeneity on other indexes by tutor characteristics for TOP2020



Notes: Robust standard errors in parentheses; ***, ** and * denote significance at the 1, 5 and 10 percent level, respectively. FWER adjusted p-values in square brackets, based on 10,000 replications. The dependent variable for each regression is listed in the column heading. GPA scores have been standardized for each faculty: education, humanities, STEM, economics, and other faculties. Additionally, we control for missing GPA values and faculty in the regression. The p-value reported is the p-value of this post-estimation test of the differences between the two groups (e.g., male vs. female tutors).

C.4 Causal Forest

In different empirical settings, treatment effects may vary widely between subgroups of the population. It is important then to evaluate if there are significant heterogeneities in treatment effects across observable variables. There are a few limitations to the traditional approach in which economists include interactions between different variables to estimate treatment effects. Including too many variables might lead to computational challenges and increase the risk of overfiting. Choosing one particular set of variables to estimate heterogeneous treatment effect can lead to arbitrary decision making and to the loss of information on possibly important heterogeneities. To deal with these challenges, we compute heterogeneous treatment effect using the honest causal forest algorithm of Wager and Athey (2018).

In our estimation procedure, we follow Davis and Heller (2017) and average treatment effects only across predictions made on observations that were neither part of the tree growing subsample nor the subsample used to estimate the conditional treatment effects. Accordingly, before estimating the causal forest we split the data into a subsample that will be used in the implementation of the causal forest algorithm, the *training sample* and a subsample that will be used to compute average treatment effects, the *test sample*. To do so, we implement the following procedure:

1. From our final sample, we obtain a random subsample—without replacement—consisting of 50% of the observations in the original sample. This subsample is our training sample and the remaining data is our test sample.

2. We use the training sample to estimate the causal forest (using R's command *causal_forest* from the *grf* package). We implement this command building a forest with 2500 trees. To build each tree, we divide the training sample in half and then use 70% of this new sample to determine splits. The other 30% is used to estimate the conditional treatment effect.

3. We use the conditional treatment effect obtained in step 2 and the test sample to estimate average treatment effects.

4. We implement 1000 replications of steps 1, 2, and 3.

5. The final estimate of average treatment effect is given by average the result across all simulations.