Apart but Connected: Online Tutoring, Cognitive Outcomes, and Soft Skills *

Michela Carlana †, Eliana La Ferrara ‡

This version: April 2025

Abstract

We study the Tutoring Online Program (TOP), where tutoring is entirely online and tutors are *volunteer* university students, matched with underprivileged middle school students. We leverage random assignment to estimate effects during and after the pandemic (2020 and 2022), investigating channels of impact. Three hours of individual tutoring per week increased math performance by 0.22 SD in 2020 and 0.20 SD in 2022. Higher-dosage yielded stronger effects, while group tutoring smaller effects. TOP enhanced students' aspirations, socio-emotional skills and psychological well-being, but only during school closures. We also estimate the impact of TOP on tutors, finding an increase in empathy.

Keywords: tutoring, COVID-19, education, achievement, aspirations, socio-emotional skills, well-being.

JEL: I24, I21

*We thank seminar participants at several universities and webinars for helpful comments. Micol Morellini, Vrinda Kapoor, Angelica Bozzi, Luca Calianno, Marco Cappelluti, Isabela Duarte, Agnese Gatti, Gaia Gaudenzi, Matteo Fossi, Federica Mezza, Giovanni Paraboschi, Chiara Soriolo, Amy Tan, and Monia Tomasella provided excellent research assistance. We are truly grateful to all the TOP tutors, to the schools that took part in the intervention for their collaboration, to the team of pedagogical experts led 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). We are indebted to Arianna Ghia for inspiring us to launch this project and to Valerio Nispi Landi for suggesting the acronym "TOP". For TOP 2020, we acknowledge support from Fondazione Romeo ed Enrica Invernizzi through LEAP-Bocconi, the Malcolm Wiener Center for Social Policy at Harvard Kennedy School, and JPAL PPE. For TOP 2022, we acknowledge financial support from Cariplo Foundation and the implementation support from Centro Italiano Aiuti all'Infanzia (CIAI). AER RCT Registry ID: AEARCTR-0005607. The project has obtained IRB approval from Bocconi University and Harvard University.

†Harvard Kennedy School and Bocconi University, 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

Tutoring has been proved to be one of the most effective tools for enhancing educational outcomes and mitigating disparities in learning (Nickow et al., 2020; Fryer Jr, 2017). This potential is particularly relevant in the aftermath of the Covid-19 pandemic, which generated significant learning losses across countries, estimated at 0.17 standard deviations in achievement test scores (Patrinos et al., 2022). One of the main challenges of tutoring programs is their scalability, both in terms of costs and in terms of supply of qualified tutors –especially when offered in disadvantaged settings. This paper presents the results from an innovative *online* tutoring program that we designed and evaluated at the outset of the pandemic and subsequently continued implementing during regular school times.

The "Tutoring Online Program" (TOP) is a pioneering policy experiment launched in Italy in the spring of 2020. To the best of our knowledge, TOP is the first online tutoring program created in response to the pandemic and it has become a systematic policy to mitigate learning gaps in Italy. It is offered free of charge and it targets underprivileged students in grades 6 to 8 (middle school). TOP has two distinguishing features. First, tutoring sessions are conducted exclusively *online*. As mentioned above, the circumstances of the lockdown and the limited availability of qualified tutors in underserved areas make remote delivery a particularly interesting feature to test. Second, the tutors in TOP are not professional educators but *volunteer university students*, who undergo training and receive support from pedagogical experts. While teachers and professionals undoubtedly possess the necessary qualifications, volunteer tutors offer advantages from a budgetary perspective and may also enhance the quality of interpersonal interaction. Indeed, TOP harnesses the intrinsic motivation of university students to volunteer and helps them develop valuable soft skills in the process.

Since its creation, TOP has reached over 4,000 disadvantaged students. In this paper, we focus on the editions implemented in 2020 and 2022, which involved a clean evaluation design and are broadly comparable. TOP 2020 was implemented during a period in which all schools were closed, while TOP 2022 during normal school times. In both editions, the demand for tutors by middle schoolers could not be fully met due to budgetary and administrative constraints. As a result, tutors were allocated randomly among eligible students, taking into account the subjects requested by the student (specifically, math, Italian, and English) and each tutor's availability to teach a particular subject. This allows us to causally estimate the impact of being assigned a tutor on the students' performance, as well as on a number of non-academic outcomes. To this end, we collected both administrative data and detailed individual surveys from students, parents, tutors

and teachers before and after each intervention.

We find sizeable and significant improvements in math performance for students who were assigned an online tutor, compared to those who were not.[1] Math grades given by teachers increased by 0.11 SD and 0.09 SD in TOP 2020 and TOP 2022 and the probability of failing the subject fell by 24 and 21 percent, respectively. TOP also improved students' performance in a standardized math test that we administered at endline, by 0.23 SD in 2020 and 0.16 SD in 2022. When these three outcomes are aggregated into an overall math performance index, the effect of the program is a 0.3 SD increase in 2020 and +0.16 SD in 2022. These are remarkable effects, given that the median length of the online tutoring was about 6 weeks in both rounds.

Our experimental design allows us to examine the *channels of impact*, following the conceptual framework by Nickow et al. (2020). First, to investigate the role played by the *dosage* of the intervention, we exploit random variation in the number of hours of tutoring per week. While the majority of students in TOP 2020 received 3 hours per week, a random subset of those who needed help in more than one subject was assigned a tutor for 6 hours per week. We find that math performance gains double and grade retention halves with the more 'intense' tutoring. The positive impact on test scores persists in the longer term, with an improvement of +0.24 SD in the national standardized math test score for students receiving high-dosage tutoring, more than one year after the end of the intervention.

Second, we investigate the impact of *tailored instruction* by randomizing students to individual vs. small group tutoring in TOP 2022. Small group tutoring may be more cost effective, but one-on-one tutoring enables more personalized instruction. We find larger effects on the math performance index for individual tutoring (+0.2 SD) vs. group tutoring (+0.11 SD), but since there were two students per group the overall cost effectiveness of group tutoring is similar to individual tutoring.

Third, the tutor/tutee relationship may extend beyond academic content, involving a *mentorship* aspect that may influence students' aspirations, socio-emotional skills, and psychological well-being. This was particularly valuable during the pandemic, when opportunities for social interactions were reduced and students experienced higher levels of depression and slower development of socio-emotional skills (Orgilés et al., 2020; Golberstein et al., 2020). We find that TOP 2020 had sizeable and significant effects on an index of educational aspirations (+0.17 SD), on a socio-emotional skills index capturing perseverance, grit and locus of control (+0.15 SD) and on an index of psychological well-being

[1]We also discuss impacts on student performance in Italian and English, but since math was the subject covered by the vast majority of students, our main analysis focuses on math.

measuring happiness and depression (+0.16 SD). No significant effect on these outcomes was found in TOP 2022. This aligns with the notion that, during periods when students' aspirations and soft skills are shaped by many other in-person interactions, the role of the online tutor may become less critical.

We test for heterogeneity of treatment effects by student and tutor characteristics. 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 immigrant students in TOP 2020 (who likely had a more sparse social network). Improvements in academic performance tend to be more concentrated among students from disadvantaged backgrounds, though the differences are not always statistically significant. Interestingly, tutor characteristics such as gender, GPA, pro-social attitudes and motivation do not systematically affect the effectiveness of the tutoring.[2]

Finally, we investigate how the experience of being a TOP tutor during the pandemic affected the tutors themselves. We can do so because in 2020 we randomly selected the university students to whom we assigned a tutee from the pool of those who applied to be tutors.[3] Four months after the end of the program, we find that volunteers who were included in TOP have significantly higher empathy than those who were not (a 0.27 SD increase). We 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 considerable body of 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).[4] However, tutoring is much costlier than classroom instruction and it may be challenging to arrange in-person tutoring in the presence of geographical constraints. Also, tutoring may sometimes be attached with the stigma of being identified as in-need and pulled out of class (Coie and Krehbiel, 1984). We contribute to this literature by providing evidence on large-scale *online* tutoring done by volunteers. Our model allows to substantially reduce costs and to efficiently reach students located in disadvantaged areas through virtual learning.[5] Also, online tutoring is less observable from peers than in-person tutoring,

[2]For pro-social attitudes and motivation, it should be observed that our tutors are already positively selected, implying relatively low variation along these dimensions.

[3]As we explain below, we could not accept all tutor applicants because we were constrained in the number of hours of tutor training and support that we could pay for.

[4]The tutor-tutee relationships also involves a mentorship connection that may affect cognitive and social skills (Kosse et al., 2020; Resnjanskij et al., 2024).

[5]Cortes et al. (2025) achieve a reduction in costs by employing part time tutors who need not have a teaching certification.

which may reduce potential stigma.

Our results are directly relevant to the debate on effective strategies to mitigate the effects of Covid-19 on education.[6] To the best of our knowledge, very few policy experiments have used remote tools to improve learning during the pandemic and replicated them after the school closures (Angrist et al., 2023). For example, Angrist et al. (2022) evaluate two low-tech interventions in Botswana that use SMS text messages and phone calls to support parents in the education of their children, finding sizeable effects. Schueler and Rodriguez-Segura (2023), on the other hand, do not find a significant impact of teacher-student phone tutoring during school closures, mainly due to students' low compliance.

Following our early work in 2020, other remote tutoring interventions have been implemented, generally finding positive effects on academic outcomes. Gortazar et al. (2024) evaluate an online group tutoring program in Spain, where the tutors are professional teachers; Hardt et al. (2022) evaluate a peer mentoring intervention at a German university, where peers met online to discuss self-organization; Hassan et al. (2024) study an over-the-phone learning support intervention for primary school students in Bangladesh; Kraft et al. (2022) conduct a pilot in which they assign a (volunteer) college student tutor to middle school students. More recent virtual tutoring interventions implemented outside of the pandemic include Robinson et al. (2024), who find effects ranging from 0.05 SD to 0.15 SD for K-2 students in Texas, and Ready et al. (2024), who finding effects around 0.05 SD for a virtual tutoring platform offered during school hours in California.

Our work contributes to this body of work in various ways. First, we evaluate the effects of the *first* online tutoring intervention implemented in response to Covid-19. TOP is an innovative and low cost online tutoring program targeting teenage students who were adversely affected by school closures. Second, we provide evidence on the effectiveness of the same intervention implemented during school closures (2020) and in-person schooling (2022). Third, we show impacts on learning outcomes as well as on soft skills and psychological well-being. Fourth, within the same intervention, we study the effects on the students and on the tutors themselves.

Finally, a different strand of literature related to our study is the recent work on organizations that highlights the power of intrinsic motivation and social recognition for improving public service delivery (e.g., Ashraf et al., 2014; Gauri et al., 2021). 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 TOP and their intrinsic motivation may have contributed

[6] See, among others, Stantcheva (2022), Carlana et al. (2023), Werner and Woessmann (2023).

to the effectiveness of our intervention.

2 Intervention and experimental design

2.1 COVID-19 and school closures in Italy

Italy was the first country after China hardly hit by the COVID-19 pandemic, with around 80,000 deaths as of January 2021.[7] In response to the emergency, all school buildings closed on March 5th, 2020, to reopen only in September 2020 for the new school year. At the onset of the pandemic, teachers' digital competencies were still limited, with less than 50 percent of the teachers using any digital tool in their daily lectures (Agcom, 2019). Nonetheless, by the end of the Spring 2020 most students in Italy received some synchronous interactions with teachers. The infrastructure for remote learning substantially improved thanks to substantial resources allocated by the Italian Ministry of Education to devices and internet connectivity.[8]

2.2 The Tutoring Online Program (TOP)

TOP 2020

Two weeks after the schools closed in Italy, we started the process of designing and implementing a new program, the "Tutoring Online Program" (TOP), as an attempt to provide immediate response to the emergency situation. Since its launch, TOP has had two defining features. First, tutoring is done entirely *online* through personal computers, tablets or smartphones. Second, the tutors in TOP are *volunteer university students*, trained and supported by pedagogical experts.

Students recruitment. To launch the program, we sent a recruitment email to the principals of all Italian middle schools (grades 6 to 8).[9] Interested schools answered a brief survey and sent us a list of potential beneficiary students including up to three pupils per class. The selection criteria we recommended to the schools included students who "may need TOP the most in terms of their learning level and family environment". Each child could be recommended for one or more subjects among math, Italian, and English. We

[7] One third of these deaths were concentrated in the region of Lombardy, and the North was generally much more affected than the Center-South by the first wave of the pandemic.

[8] In the Spring 2020, the Italian Ministry of Education allocated 70 million euro to buy tablets that students could temporarily borrow and 10 million for schools' connectivity and online platforms of schools. These interventions facilitated access to devices and internet for disadvantaged students.

[9] In Italy, there are approximately 5,000 middle schools.

clarified that we could not guarantee a tutor to all applicants and that –if the number of requests exceeded the number of tutors– we would randomly assign students to the program, in order to give every applicant the same chances. Tutoring was offered for 3 or 6 hours per week. The intense tutoring was randomly allocated to students needing help in more than one subject. A total of 76 schools from 40 different provinces (out of 107 Italian provinces) participated.[10] Out of 1,594 students identified as in need of the program by the schools, 1,059 students and their parents completed the baseline survey by the end of enrollment period.[11] These 1,059 students constitute our experimental sample for evaluating the effects of TOP 2020. Table A.2 shows that students selected for TOP are substantially disadvantaged compared to their schoolmates, as intended by the program.

Tutors recruitment. To recruit volunteer tutors, we sent an email to all the students enrolled in three large universities in Milan.[12] Applicant tutors completed a baseline survey indicating the subjects in which they felt comfortable tutoring, their time availability, and some background information. The number of applications from volunteers neared 2,000, far exceeding our expectations. Unfortunately, due to budget and capacity constraints, we could only train and support 523 tutors in 2020.

Implementation. As our volunteers were not professionals, we hired a team of pedagogical experts from Bicocca University to train and support the tutors. Within four weeks, these experts set up an online learning platform with a self-training program that included slides and videos on the key elements of tutoring based on the pedagogical implications of Bruner's theory (Pastori and Farina, 2022).[13] The training mainly targeted the academic dimensions. The platform also included a supervised forum where tutors could ask questions and share their experiences. The pedagogical team also organized regular group meetings with around 20 tutors at a time, as well as one-on-one meetings to offer

[10]To assess the representativeness of our sample of schools, Appendix Table A.1 compares the provinces where at least one school participated and those with no participation. The main difference is that schools in the North of Italy are over-represented in our sample; this regional imbalance explains the differences in the table. This could be mainly due to two factors: first, the North was the area most severely hit by the pandemic in Spring 2020; second, the principals in wealthier areas in the North may have been more likely to respond to our invitation.

[11]Each parent was informed by the teachers about the project. Then, they received a text message and an email from our research team requiring them to complete the baseline form. Finally, we also called each non-responding parent to ensure the non-response truly meant they were not interested (we made up to three phone call attempts). More than 90% of the parents who were contacted by our research team decided to take-up the opportunity for their children. Parents signing up for the program agreed to have an adult in the house during the tutoring, to minimize safety concerns.

[12]The three universities were Bicocca University, Bocconi University, and University of Milan, which in 2020 enrolled approximately 33,000, 14,000 and 61,000 students respectively.

[13]The topics included: how to approach students; tools and online platforms for effective online tutoring; how to teach students with learning disorders; and tips to engage students in math, Italian and English.

support in specific circumstances. Despite being the most expensive component of TOP (see section 7 for cost estimates), tutor training and support are an important component of TOP, ensuring that our volunteers receive professional advice in case of need.

The implementation of TOP 2020 was entirely done by the research team. Tutoring lasted from mid-April to the end of May 2020 and endline surveys were collected in June 2020 (Figure A.1 reports the timeline). On average, treated students had 14 tutoring meetings over the course of the program, for a total of 17 hours over 34 days (see Figure A.2 for the full distribution).[14] The subject covered by the great majority of students was math: 69 percent of the students, compared to 36 percent for both Italian and English (Figure A.3). Math is also the subject most commonly covered in TOP 2022, and for this reason we will focus our evaluation of academic impact on math scores, though we also report performance in other subjects in the Online Appendix.

TOP 2022

The structure of the program in 2022 was almost identical to that of 2020, with few exceptions. First, the program was implemented by an Italian NGO supported by the research team, with the objective of bringing the program to scale in the medium term.

Second, all students were offered *three* hours of tutoring per week, in up to two subjects between math, Italian, and English (in TOP 2020 some students received six hours per week).

Third, instead of exclusively relying on 1:1 tutoring, we randomized whether students were exposed to individual or group tutoring. The goal was to better understand the trade-off between personalized instruction on the one hand (stronger under individual tutoring), and cost savings and reinforcement through peer learning on the other (stronger under group tutoring).

Fourth, tutors' training and tutoring sessions happened through an online platform adapted to the needs of our program, *WeSchool*. This platform offers the following advantages: (i) attendance to sessions is automatically recorded; (ii) the platform includes a video-conferencing tool that standardizes the format of the meetings (in TOP 2020 students used different video call systems); and (iii) *WeSchool* allows tutors to upload exercises and tests that students can take to assess their understanding of the material.

[14] After each meeting, tutors were required to record some basic information about the session using a management tool, including whether there were issues with the connection. 16.8 percent of the tutees were reported as having significant problems with the connection. Note that about 5 percent of students chose not to start the tutoring (hence have zero meetings). These students will be included in our intention-to-treat estimates.

Recruitment. TOP 2022 involved a different set of students and tutors compared to the previous edition. The recruitment of students was managed by the implementing partner following a protocol similar to TOP 2020, but focusing on schools in the North.[15] Our experimental sample for TOP 2022 is constituted by 1,137 students who completed the baseline survey. Tutors were recruited form a larger number of universities than in TOP 2020, yielding a total of 481 student volunteers.[16] The training of tutors was led by the same pedagogical experts as in the previous edition, in close collaboration with 'tutor supervisors' from the implementing partner, who were tasked with maintaining high frequency communication with tutors.

Implementation. Figure A.1 shows the timeline of TOP 2022. On average, treated students had 15 tutoring meetings, for a total of 19 hours. The subject covered by the vast majority of students (69 percent) was math (see Figure A.3).

2.3 Randomization and baseline characteristics

TOP 2020 involved 523 tutors and 1,059 student applicants; TOP 2022 involved 481 tutors and 1,137 applicants. For both programs, the randomization proceeded in two steps. First, students were divided into two groups: a treatment group that received tutoring and a control group that did not. In order to allow the tutoring to start as soon as possible, we processed applications on a rolling basis by creating 'blocks'. We stratified the randomization at the block level, where each block consisted of about 100 applicants. Blocks were created based on the timing of baseline survey completion.

Second, in each wave of the program we randomized an additional treatment to investigate mechanisms and test cost-effective alternatives. For TOP 2020, among the 530 treated students, teachers indicated 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 randomization will allow us to estimate the impact of higher treatment dosage. For TOP 2022, among the 607 treated students, we randomized 265 to receive tutoring in a two-person group that included another student from the same class and/or school-grade.[17] This randomization will allow us to study the effectiveness of targeted individual instruction.

[15]The geographic focus was dictated by the statute and mandate of the Foundation that funded TOP 2022.

[16]TOP 2022 volunteers were recruited from the following universities: Bicocca University, Bocconi University, Cattolica University, Insubria University, IULM, Polytechnic University of Milan, University Carlo Cattaneo, University of Brescia, University of Milan, and University of Pavia.

[17]We chose not to create groups of three or more students because finding a regular time when both the tutor and the various students would be available would have been very challenging.

Table 1 shows balance in observable characteristics across treatment and control, pooling the sample for TOP 2020 and 2022.[18] No statistically significant differences emerge.

Table 1. Students' Characteristics at baseline (full sample)

	(1) All	(2) Control	(3) Treatment	(4) P-value	(5) Std diff.
Students					
Survey Data					
Female	0.425	0.414	0.436	0.277	0.045
Immigrant	0.226	0.211	0.240	0.106	0.069
Learning Disorder	0.288	0.292	0.285	0.849	-0.015
School grade 6	0.310	0.297	0.323	0.166	0.056
School grade 7	0.354	0.353	0.355	0.996	0.004
School grade 8	0.336	0.350	0.322	0.173	-0.059
Admin data					
Baseline math grade	6.294	6.308	6.281	0.496	-0.025
Baseline Italian grade	6.493	6.463	6.520	0.227	0.059
Baseline English grade	6.578	6.547	6.606	0.444	0.052
Parents					
SES Status	-0.337	-0.336	-0.339	0.897	-0.003
Single parent household	0.242	0.252	0.232	0.267	-0.047
Mother Education: High-school	0.437	0.455	0.420	0.163	-0.071
Mother Education: University	0.122	0.125	0.119	0.549	-0.018
Father Education: High-school	0.377	0.366	0.387	0.315	0.043
Father Education: University	0.075	0.077	0.074	0.813	-0.011
Mother not employed	0.319	0.311	0.327	0.386	0.034
Mother Occupation: white-collar	0.269	0.276	0.263	0.350	-0.029
Mother Occupation: blue-collar	0.379	0.386	0.373	0.648	-0.027
Father not employed	0.159	0.139	0.179	0.006	0.109
Father Occupation: white-collar	0.273	0.276	0.271	0.830	-0.011
Father Occupation: blue-collar	0.528	0.550	0.508	0.031	-0.084
Parental time helping homework	41.113	40.627	41.571	0.464	0.024
Joint test				0.192	
Observations	2,196	1,059	1,137		

Notes: This table reports characteristics of all students (column 1), control students (column 2), and treated students (column 3) in TOP 2020 and TOP 2022. Column 4 reports p-values for difference in means, controlling for randomization round fixed effects (as in our main specification). Column 5 reports the standardized difference between group averages. All variables are measured at baseline. The second to last row shows the P-value of an F-test for the joint significance of all the coefficients in a regression of the treatment on all the covariates in the table and randomization round FE.

Overall, 42% of students in our sample are female, 23% are immigrants and 29% have learning disorders. The average performance of these students is only slightly above

[18] In Tables A.3 and A.4 we provide an additional set of balance checks, separately for the two waves of the program and with the additional comparison of high vs. low dosage for TOP 2020 and group vs. individual tutoring for TOP 2022. These tables show that the samples are overall well-balanced, with a limited number of statistically significant differences consistent with random sampling. We control for these baseline characteristics in all regressions.

the pass grade in all subjects, which in Italy is a score of 6 out of 10. The average student comes from a disadvantaged background, as measured by parental education and occupation levels.

Table A.5 compares the characteristics of students across the two waves: while parental education and occupation are well-balanced across waves, students from TOP 2022 are more likely to be female, come from an immigrant background, have lower SES status and are less likely to have learning disorders.[19]

Attrition. While we do not have attrition for the administrative outcomes, for survey outcomes attrition rates differed by treatment status. In TOP 2020, 88 percent of treated students and 46 percent of control ones completed the endline survey, while in TOP 2022 the respective figures were 95 and 73 percent. The higher attrition in the control group is not surprising, given that students who received a tutor remained engaged with the program, while control students were contacted after not receiving a tutor.[20]

Table A.6 shows the correlates of attrition. Immigrant students in the control group are more likely to participate in the endline. Students in grade 8 were involved in the final middle school exam during the period of the endline and they were less motivated to devote time to the survey compared to younger students. 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 socio-economic background were more likely to complete the endline test, with the effect being driven mainly by students in the control group (column 4).[21]

Given the direction of the imbalance, 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. Furthermore, the estimates from survey data have lower precision due to the smaller sample size compared to administrative data. Note that administrative outcomes do not suffer from attrition, hence our impact estimates on end-of-year grades and grade retention will be unaffected by attrition.

Tutor allocation. We assigned tutors to students following a sequential procedure.

[19]In the recruitment message to schools, in 2022 we emphasized that our tutors were not expert in learning disorders and this may have affected the selection of students. Grades are higher in 2022, but the result cannot be easily interpreted due to grade inflation.

[20]In 2020, the response rate was overall lower compared to 2022: families were facing severe shocks due to COVID-19 that limited their responsiveness to our data collection effort.

[21]Tables A.7 and A.8 report the balance tables restricting the sample only to students who completed the final survey in 2020 and 2022, respectively. Most characteristics are balanced between treatment and control group. If anything, in 2020 the control group is positively selected in terms of parental education.

First, we restricted the sample of tutors to those currently enrolled in a program at their university and fluent in Italian. Second, we divided tutors into groups depending on their expertise in the various subjects (math, Italian, English or combinations of these). Finally, within each group, we randomly assigned tutors to tutees.[22] While for TOP 2022 we assigned all available tutors, for TOP 2020 some volunteers were not matched.

Table A.9 presents summary statistics on the characteristics of tutors in the two editions of the program. The majority of tutors are female (72 percent) and born in Italy (95 percent). Around one third of tutors are STEM and one third are economics majors in both waves. Students recruited during the pandemic were more likely to be moved by a desire to help others when applying to TOP and to have previous experience as volunteers.

For the 2020 edition, Table A.10 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, very few significant differences emerge.[23] For the 2022 edition, Table A.11 compares tutors assigned to group vs. individual tutoring. Since this assignment was totally random, it is not surprising that there are no significant differences.

2.4 Endline outcomes

We build a unique dataset merging the baseline surveys of parents, students, and tutors with endline data coming from the following sources: (i) administrative data from the Italian Ministry of Education, including teacher assigned grades and failure rate; (ii) data from the National Evaluation Center (INVALSI), including long-term standardized test scores; (iii) the results of a standardized test administered by our research team; and (iv) surveys of parents, students, tutors and teachers.

Student outcomes

Table A.12 reports summary statistics of our main student outcomes measured at endline, divided into four categories: academic, aspirations, socio-emotional skills, and well-being.

[22] 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.

[23] Tutors from STEM are over-represented in treatment because math was the subject most in demand by the students. In column 7 of Table A.10, we report the differences between volunteers that offered their availability for 3 vs. 6 hours per week in 2020. While no variable is significantly different on its own, the joint test suggests that the two groups of tutors differ. This implies that, although students are randomly assigned to the 3- vs. 6-hour dosage, they get a 'package' of different tutor characteristics when assigned to more intensive tutoring.

For the last three sets, we build indexes extracting the first principal component from a series of variables in each category, standardized to have mean zero and standard deviation one in the control group. For details, see Online Appendix B.

Academic outcomes. We have three main academic outcomes measured at the end of the school year: teacher-assigned grades, the probability of failing the grade, and a standardized test score. The former two variables are based on administrative data and are available for all students. The standardized test score, instead, was administered by our research team.[24] The advantage of our test is that it is fully comparable across schools and teachers, and that it conveys information on the quality of learning even when teachers may have been lenient with their assessments due to the pandemic.

Aspirations. We collected information from students on the type of high school track they wished to enroll in and on whether they would like to attend university after high school.[25] We also collected a measure of self-efficacy (Bandura et al., 1999), asking students (and their parents) whether –aside from what they would *like* to do– they think that they (their children) would be capable of successfully attending university if they wanted to.

Socio-emotional skills. We collected three measures of socio-emotional skills. First, to measure *perseverance* in a real effort task (Alan et al., 2019), we asked students to answer a logic question. At the end of the question, they could choose to answer a new question with the same level of difficulty, one with a higher level of difficulty or they could give up. Second, we measure '*grit*' following the Short Grit Scale developed by Duckworth and Quinn (2009), which ranges from 0 (not at all gritty) to 1 (extremely gritty).[26] 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, using four items adapted from Rotter (1966). This outcome ranges from 0 (lowest locus of control) to 1 (highest locus of control).

Well-being. We collected two measures of students' psychological well-being. 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. Values of the index range from 0 (no depression) to 1 (high level of depression). The second measure is a proxy for happiness:

[24]Our test included seven multiple-choice questions in math, seven in Italian, and five in English, close in format to the national standardized test score from INVALSI.

[25]In Italy, after grade 8 students need to choose one of the following high school tracks: top-tier academic (i.e., scientific and classical *lyceum*), other academic, technical, and vocational. Among the students in our sample, only 15 percent are interested in a top-tier academic high school, while around one out of four is planning to attend a vocational high school. On the other hand, 39 percent of students say they are considering university education.

[26]We asked the same questions to children and parents, finding a 0.64 correlation among their answers.

we asked whether students were feeling happy or unhappy on a scale from 1 to 10 (10 being the maximum happiness).[27]

Tutor outcomes

For TOP 2020, we asked all the volunteers that had applied to be tutors to complete a short endline survey in September 2020, six months after the start of the program. 92.4% of 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. The short survey included two outcomes that we used to estimate how participation in TOP affected tutors' ability to empathize with others and their beliefs on the importance of effort.

Empathy. We collected two standard questions inspired by Lawrence et al. (2004), 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 obtain a variable ranging from 0 (lowest empathy) to 1 (highest empathy).

Hard work and effort. 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 an equally-weighted index ranging from 0 to 1.

2.5 Empirical Strategy

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

$$Y_{ir} = \alpha_r + \beta Treated_i + \gamma X_i + \varepsilon_{ir} \tag{1}$$

where Y_{ir} is the relevant outcome for student i 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; X_i is a vector of student level controls measured at baseline, including: gender, immigrant status, grade in which the student is enrolled, mother's and father's education, mother's and father's employment type, learn-

[27] We asked these questions to children and parents. The correlation between the depression measure reported by parents (when asked about their children) and the one reported by students themselves is 0.56, while for happiness it is 0.53.

ing disability, baseline grades and test scores, interest in the different subjects; ε_{ir} is an error term. We report robust standard errors and correct for multiple hypothesis testing using Anderson's sharpened False Discovery Rate (FDR) q-values.[28]

Online Appendix D.1 reports the results of our robustness analysis. First, we show the results on all outcomes selecting control variables in a systematic way with double post LASSO procedure, following Belloni et al. (2012). Second, we present inverse probability-weighted estimates, which help address potential bias related to attrition in our survey outcomes. Overall, the different robustness checks presented provide a consistent picture of the impact of TOP.

3 Impact of TOP on students' performance

In Panel A of Table 2 we estimate the impact of TOP 2020 on students' math performance. Panel A1 reports average effects from the program, while panel A2 distinguishes between 3-hour and 6-hour tutoring (we discuss these results in section 4.1).

Column 1 shows that the teacher-assigned grade increased by 0.11 points for treated students (+0.11 SD relative to the control group). The failure rate (i.e. the probability of getting a score below 6) decreased by 4.4 percentage points, corresponding to a 24 percent reduction compared to the control group (column 3, Panel A1). While these results refer to the universe of students in our sample, the point estimates are not statistically different when we restrict the sample to students who completed the endline survey (columns 2 and 4). This is reassuring, as it mitigates our concerns on selective attrition for those outcomes that are only available from the survey data.

Column 5 of Panel A1 shows that TOP improved treated students' standardized test score by 0.23 SD compared to the control group (significant at the 1 percent level). Finally, column 6 reports the impact of TOP 2020 on the math performance index, an aggregate index obtained through principal component analysis to encompass the information from the grades, failure rate, and test score. TOP improved overall performance by 0.30 SD compared to students in the control group, statistically significant at the 1 percent level. This is an impressive result, considering that the median duration of tutoring was five weeks and that tutors did not specifically prepare the students for this type of test, but rather focused on helping students find a method for studying and doing homework. The sharpened q-values reported in square brackets show that our results remain significant when adjusting for multiple-hypothesis testing.

[28] All our results are similar using the Westfall-Young stepdown adjusted p-values, which also control the family-wise error rate (FWER).

In Panel B1 of Table 2, we report symmetric results on the impact of TOP 2022. Overall, the program had qualitatively similar effects: it improved teacher-assigned grades by 0.09 points (+0.09 SD relative to the control group), it decreased failure rates by 7 percentage points (a 21 percent reduction), it improved the standardized test score by 0.16 SD and the overall performance index by 0.17 SD. The effect are statistically significant at the 5 percent level.

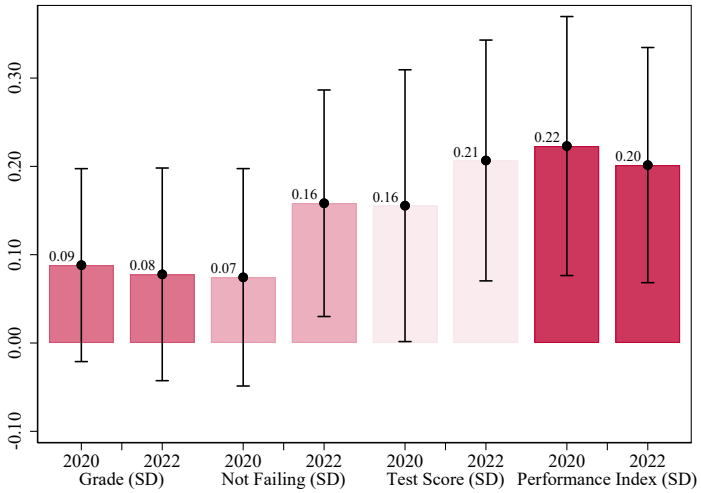
Table 2. Impact of TOP on math academic outcomes

	(1)	(2)	(3)	(4)	(5)	(6)
	Grades		Fail		Test Score	Performance Index
Panel A: TOP2020						
Panel A1: Overall results						
Treatment	0.113	0.113	-0.044	-0.056	0.234	0.303
	(0.056)	(0.056)	(0.022)	(0.029)	(0.075)	(0.071)
		[0.024]		[0.024]	[0.003]	[0.001]
R^2	0.365	0.365	0.124	0.147	0.217	0.306
Panel A2: Results of 3 vs. 6 hours of tutoring						
Treatment	0.098	0.118	-0.028	-0.036	0.156	0.223
	(0.062)	(0.077)	(0.024)	(0.031)	(0.078)	(0.075)
Intense Treatment	0.039	0.093	-0.056	-0.073	0.279	0.280
	(0.086)	(0.092)	(0.034)	(0.036)	(0.103)	(0.100)
R^2	0.382	0.397	0.127	0.152	0.226	0.319
Treat+Intense Treatment==0	0.088	0.025	0.010	0.005	0.000	0.000
Mean Dep:	6.37	6.35	0.18	0.18	0.00	-0.00
Obs	1058	711	1058	711	712	711
Panel B: TOP2022						
Panel B1: Overall results						
Treatment	0.088	0.086	-0.066	-0.070	0.155	0.163
	(0.055)	(0.061)	(0.027)	(0.029)	(0.059)	(0.058)
		[0.029]		[0.019]	[0.019]	[0.019]
R^2	0.241	0.233	0.135	0.132	0.192	0.236
Panel B2: Individual vs. Group tutoring						
Treatment	0.080	0.082	-0.074	-0.080	0.207	0.201
	(0.063)	(0.069)	(0.030)	(0.033)	(0.069)	(0.068)
Group Treatment	0.018	0.009	0.018	0.024	-0.122	-0.091
	(0.075)	(0.080)	(0.037)	(0.039)	(0.084)	(0.082)
R^2	0.241	0.233	0.135	0.132	0.193	0.238
Treat+Group Treatment==0	0.165	0.233	0.114	0.140	0.265	0.138
Mean Dep:	5.97	6.01	0.31	0.30	-0.00	-0.00
Obs	1121	943	1121	943	953	943
Sample	Admin	Admin+Endline	Admin	Admin+Endline	Endline	Admin+Endline

Notes: OLS estimates, robust standard errors in parentheses, and Anderson sharpened q-values in square brackets. The dependent variable is teacher-assigned grades in columns 1-2, failure rate in columns 3-4, standardized test score in column 5 and a performance index including all three academic outcomes in column 6. "Treatment" is an indicator for being assigned a tutor; "Intense treatment" is an indicator for being assigned to 6 hours of tutoring; "Group treatment" is an indicator for being assigned to group tutoring vs. individual tutoring. "Treat + Intense Treatment" is the p-value for testing the null that the sum of the two coefficients in Panel A2 is zero. "Treat + Group Treatment" is the p-value for testing the null that the sum of the two coefficients in Panel B2 is zero. "Mean Dep." is the mean of the dependent variable at endline for students in the control group. "Sample" indicates whether the sample used for the analysis is from administrative data (Admin) or the endline test score (Endline). Controls included in all regressions: parental education and occupation, gender, immigration status, learning disorders, school grade, SES status,teacher-assigned grades, and test scores in math at baseline. Panel A2 also includes a dummy for whether the student was eligible for 6-hour tutoring.

As explained in section 2.2, there are important differences between TOP 2020 and TOP 2022. First, in-person interactions were limited and school was remote in 2020, while students were back to school in 2022. Second, conditional on being assigned to treatment, different features of the program were cross-randomized: the 3 vs. 6 hours of individual tutoring in 2020, and the group vs. individual tutoring (both of 3 hours) in 2022. We next report effects that are fully comparable across the two editions of TOP, to ease the interpretation. Figure 1 shows the impact of *three* hours of *individual* tutoring on the various math outcomes, expressed in standard deviations, in 2020 and 2022. The coefficients plotted are the equivalent of the effect of individual 3h tutoring as in Panel A2 and Panel B2 of Table 2.

Figure 1. 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 math outcomes, i.e. teacher-assigned grades, admission (not failing), standardized test scores, and an overall performance index. Each index is standardized to have mean 0 and standard deviation 1 in the control group. For each outcome, the first bar shows the impact on TOP 2020 and the second bar on TOP 2022, with 95% confidence intervals. Controls included are parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline.

Strikingly, the three hours of individual tutoring have an impact that is statistically indistinguishable during and after the school closure due to the pandemic. The effects on grades and failure rate are slightly smaller in magnitude than the estimates for the overall program impact (between 0.07 and 0.16 SD) and in some cases statistically insignificant. The impact on the standardized test score ranges from 0.16 to 0.21 SD and is significant at 5 percent level for both editions. The aggregate performance index shows increases

between 0.23 and 0.20 SD, significant at the 5 percent level. By splitting the sample and focusing only on the three hours of individual tutoring we lose power, but the results qualitatively suggest a positive impact on all three performance measures. Interestingly, the impact is comparable in magnitude to the average impact of large-scale in-person tutoring (+0.25 SD), as reported in the meta-analysis by Nickow et al. (2020).[29] This is remarkable and suggests significant potential for remote tutoring interventions.

The results discussed so far focus on math because around 70% of students predominantly covered this subject. Between 30 and 55% of students received support in Italian and English. Tables A.13 and A.14 report the effects on these two subjects: impacts are positive across the board, but smaller in magnitude and less precisely estimated. This is not surprising, as tutors devoted considerably less time to these subjects relative to math.

Finally, Online Appendix D.2 presents the medium-term impacts of TOP on academic outcomes. More than one year after the end of the intervention, the intense 6-hour tutoring increased test scores administered by the national agency (INVALSI) by +0.24 SD (p-value = 0.08), while the effect of 3-hours individual tutoring is less than half in size and statistically not significant. TOP also decreased the probability of choosing a vocational track by 8%, i.e., increased educational aspirations, though the effect is imprecisely estimated.

4 Mechanisms

In this section we explore the mechanisms through which online tutoring may improve learning outcomes. First, we exploit specific design features of our program to investigate the causal effect of tutoring dosage and the tailoring of instruction with individual versus group tutoring. Second, we analyze the impact of online tutoring on non-academic outcomes (aspirations, socio-emotional skills, and psychological well-being) to investigate whether these may be driving channels of the observed improvement in performance, e.g., due to the mentorship aspect of tutoring. Finally, we complement this section with an Online Appendix C analyzing the heterogeneity on the main outcomes based on baseline characteristics of students and tutors. The results show limited heterogeneity in treatment effects that could depend also from the selection of students most in need of the tutoring intervention from schools and the self-selection of motivated university students

[29]This figure refers to large-scale tutoring interventions involving more than 400 observations. Note that, different from 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 and show larger effects for younger children.

as tutors.

4.1 Instruction time and tailored instruction

Tutoring dosage: 3 hours vs. 6 hours. One of the most prominent explanations for why tutoring improves performance is the increase in instruction time (Nickow et al., 2020; Fryer Jr, 2017). The experimental design of TOP 2020 allows us to investigate this channel. Among the 427 students who needed help in more than one subject (out of the 530 treated), 27 percent were randomly assigned a tutor who offered 6 hours per week.

We start by assessing how participating in the 3 versus 6 hours version of the program affected the time devoted to homework. Panel A of Table A.15 quantifies the average treatment effect on study time as reported by students and parents for TOP 2020. While TOP led to a modest increase in the amount of time devoted to homework, it potentially substituted 'inefficient' study-alone time with better quality study-with-tutor time. In Table A.16, we complement these data with detailed data from the tutor registry, available only for treated students. Students assigned to intense tutoring increase the total number of sessions by 3.7 (from 13.3 to 17), but also the duration of each session by 16 minutes (from 68 to 84 minutes) and the number of sessions per week by 0.7 (from 2.3 to 3). The span of the tutoring in terms of days and the the share of sessions that took place as planned were not significantly affected.

The effects of tutoring dosage on our main academic outcomes are reported in Panel A2 of Table 2. We regress each outcome on the treatment dummy and on an indicator for whether the student was assigned a 6-hour tutor (the variable '*Intense Treatment*' in the table). Based on the estimates in column 1, 3, 5 and 6, the impact of the 3-hour version of the program is a 0.10 point increase in grades, a 2.8 percentage points reduction in the failure rate, a 0.16 SD increase in the standardized test score and a 0.22 SD increase in the overall performance index, respectively. Having a tutor for 6 hours a week more than doubles the impact on failure rates and test scores, while the (positive) differential effect on teacher-assigned grades is not statistically significant at conventional levels.[30] The magnitude of the effects of the intense treatment is in line with the meta-analysis of Nickow et al. (2020), who show that the average effect of tutoring on learning almost doubles when doubling the dosage of in-person tutoring.

Tailoring instruction: group vs. individual tutoring. Another leading mecha-

[30]The difference in effects may reflect that the intense dosage was particularly effective for students at the margin of failure, and that teacher-assigned grades exhibit less variation compared to our standardized test score.

nism through which tutoring impacts academic outcomes is the customization of learning. A well-established literature underlines the importance of "teaching at the right level" to increase effectiveness of instruction (Banerjee et al., 2016). Teachers face challenges in providing effective support to students when the latter exhibit varying skill levels and require assistance in various subjects. We can investigate this aspect by exploiting the fact that TOP 2022 randomly assigned students to individual vs. group tutoring. Groups were formed in a way that was meant to ease the customization of learning to the extent possible: pairs were created among students from the same class (or if unavailable from the same grade and school) and who needed help *in the same subjects*. Despite this effort, the possibility for the tutor to tailor explanations and allocate time optimally based on the need of each individual student was lower than under one-on-one tutoring.

Panel B of Table A.15 shows that the difference in the time devoted to homework is minimal for group vs. individual tutoring. In Panel B2 of Table 2 we report the impact of group tutoring on our main academic outcomes. In this table, the effect of individual tutoring is given by the coefficient on *'Treatment'*, while the effect of group tutoring is the sum of the coefficients on *'Treatment'* and *'Group Treatment'*. The table also reports the p-value for the null that this sum is zero. For teacher-assigned grades, the impact is imprecisely estimated for both types of tutoring (columns 1-2). When we consider failure rates (columns 3-4), individual tutoring leads to a reduction of 7.3 percentage points (p-value=0.017), while for group tutoring the effect is 5.6 percentage points (p-value=0.112). Finally, the impact on the standardized test score and overall performance index (columns 5 and 6) is large and significant for individual tutoring (0.21 SD and 0.20 SD, with p-value= 0.03 for both), and smaller and insignificant for group tutoring (0.09 SD and 0.11 SD, with p-value= 0.256 and = 0.130 respectively).

Overall, these results suggest that online tutoring is slightly more effective when students receive personalized individual support than in group interactions. Despite a constant 'quantity' of instruction time, the 'quality' of the input received may be lower in a group setting, leading to more muted effects in the absence of individual customization of learning.

4.2 Mentorship: aspirations, socio-emotional skills, well-being

The human connection generated by tutoring can be an additional channel to explain the improvement in academic outcomes: tutors may motivate students to set higher goals for themselves, help them increase their locus of control, and/or their psychological well-being. In-person mentoring relationships have been shown to have positive long-term

implications for children's educational outcomes and life trajectories (Resnjanskij et al., 2024; Kosse et al., 2020; Falk et al., 2024). However, it is unclear whether similar positive impacts may be generated when students are connected with a tutor in an online setting.

Table 3. Impact of TOP 2020 and TOP2022 on non-academic outcomes

	(1)	(2)	(3)
	Aspirations	Socio-Emotional	Well-being
Panel A: TOP2020			
Panel A1: Overall results			
Treatment	0.167	0.146	0.162
	(0.078)	(0.075)	(0.082)
	[0.055]	[0.055]	[0.055]
R^2	0.335	0.179	0.077
Panel A2: Results of 3 vs. 6 hours of tutoring			
Treatment	0.134	0.133	0.136
	(0.083)	(0.081)	(0.086)
Intense Treatment	0.109	0.049	0.092
	(0.121)	(0.106)	(0.116)
R^2	0.343	0.180	0.078
Treat+Intense Treatment==0	0.043	0.089	0.061
Mean Dep:	-0.00	-0.00	-0.00
Obs	523	636	614
Panel B: TOP2022			
Panel B1: Overall results			
Treatment	0.058	-0.039	0.037
	(0.066)	(0.071)	(0.072)
	[1.000]	[1.000]	[1.000]
R^2	0.219	0.059	0.074
Panel B2: Group vs. Individual tutoring			
Treatment	0.022	-0.075	0.004
	(0.075)	(0.081)	(0.082)
Group Tutoring	0.088	0.086	0.077
	(0.092)	(0.100)	(0.104)
R^2	0.220	0.060	0.075
Treat+Group Tutoring==0	0.200	0.898	0.401
Mean Dep:	-0.00	-0.00	-0.00
Obs	889	898	881

Notes: OLS estimates, robust standard errors in parentheses, and Anderson sharpened q-values in square brackets. The dependent variable is the Aspiration index in column 1, Socio-emotional Skills index in column 2, Psychological Well-being index in column 3. "Treatment" is an indicator for being assigned a tutor; "Intense treatment" is an indicator for being assigned to 6 hours of tutoring; "Group treatment" is an indicator for being assigned to group tutoring. "Treat + Intense Treatment" is the p-value for testing the null that the sum of the two coefficients in Panel A2 is zero. "Treat + Group Treatment" is the p-value for testing the null that the sum of the two coefficients in Panel B2 is zero. "Mean Dep." is the mean of the dependent variable at endline for students in the control group. Controls included in all regressions: parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline. Panel A2 also includes a dummy for whether the student was eligible for 6-hour tutoring.

Aspirations. In column 1 of Table 3, we estimate the impact of the program on

students' aspirations and perceived ability to attain educational goals. Panel A1 shows that TOP had a positive and significant impact on aspirations during the pandemic, when students had limited interactions outside their home. The aspiration index improves by 0.17 SD (p-value = 0.055). Panel A2 disaggregates the impacts by the intensity of the tutoring, showing effects of 0.14 SD (p-value = 0.09) for the 3 hours of individual tutoring and 0.25 SD (p-value = 0.038) for intense tutoring. The coefficients for TOP 2022 (panel B of Table 3) are positive but smaller and statistically indistinguishable from zero. It is interesting that, while group tutoring presents a disadvantage in terms of customization of learning, interacting with another student in addition to the tutor reinforced the impact of the program on aspirations with an increase of 0.10 SD (p-value = 0.208).

The first two bars of Figure A.4 compare the effects of the common component in TOP 2020 and 2022, i.e., three hours of individual tutoring. The impact on aspirations is positive and significant at the 10 percent level during the pandemic, while it is virtually zero when the program is implemented during normal school activities. The role that the tutor had in shaping aspirations in 2020 may thus have been heightened by the fact that students did not have many possibilities for in-person interactions.

In Panel A of Tables A.17 and A.18 we estimate the impact of TOP 2020 and TOP 2022, respectively, on the individual components of the aspiration index. The results are in line with those presented above, with positive effects of TOP 2020 and small insignificant effects of TOP 2022.[31] In particular, students who received tutoring during the pandemic are 6.1 percentage points less likely to say that they plan to attend a vocational high school (a 20 percent decrease relative to the control group) and 5.9 percentage points more likely to aspire to enroll in university (an 16 percent increase). Interestingly, the parents of treated students internalize higher educational goals and revise upward the perceived ability of their children to attend university.

Socio-emotional skills and beliefs. We next test whether the program affected students' socio-emotional skills, in particular students' reactions in the face of obstacles and their perceived ability to control what happens in their lives. This is another channel through which the interaction with a tutor may affect educational achievement.

Column 2 of Table 3 reports these results for TOP 2020 and 2022, using as dependent variable an index of socio-emotional skills that comprises perseverance, grit and locus of control (see Appendix B.1 for a detailed description). The results show a positive and significant impact of TOP when implemented during school closures and no effect in 2022, consistent with the results for aspirations. The same pattern emerges in Panel A2, with

[31] Note that when we adjust for multiple hypothesis testing, the effects are not statistically significant at conventional levels for the individual components of the aspiration index.

a slightly more positive impact of intense tutoring in 2020, as well as from Figure A.4 comparing the three hours of individual tutoring across years.

In Panel B of Tables A.17 and A.18 we estimate impacts separately on the different components of the index and find that the positive impact in 2020 is mainly driven by locus of control. A possible interpretation is that, when students who work with a tutor realize positive academic results, they understand that success in school is not a matter of luck. To substantitate this interpretation, in Table A.13 (Panel C) we show that students in TOP 2020 revised upwards their beliefs about how well they would do on the standardized test and about the grades they would receive from their teachers.[32] Parents and teachers also revised their expectations upwards.

Mental health and well-being. An important goal of TOP 2020, in addition to the academic component, was to help students navigate the psychological difficulties that the lockdown and isolation due to Covid-19 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 3 shows the impact of the program on an index of students' well-being that captures depression and happiness (details in Appendix B.1). Consistent with the results on aspirations and socio-emotional skills, we find a positive and significant impact on the well-being of students for TOP 2020 (Panel A1), but no effect for TOP 2022 (Panel B1). The impact in 2020 is larger for the 6 hours of individual tutoring (+0.22 SD, p-value 0.07). The difference among the two editions also emerges when comparing three hours of online tutoring (the rightmost two bars in Figure A.4).[33] Interestingly, psychological well-being is the outcome for which the heterogeneity in treatment effects reported in Online Appendix C is most striking: the increased happiness and reduced depression we detect in TOP 2020 are entirely driven by immigrant students.[34]

The above findings suggest that TOP played an important role in enhancing aspirations, improving socio-emotional skills, and mitigating potential mental health problems associated with the pandemic and with the strict regime of lockdown, which is an important result in and of itself. In terms of mechanisms, however, the asymmetric impact across the two editions suggests that the increase in aspirations, socio-emotional skills,

[32]Students' beliefs were not affected in TOP 2022 (Table A.14). Students in remedial education programs are often exposed to negative stereotypes from teachers and peers, and the isolation during the pandemic may have shielded TOP 2020 students from these negative stereotypes, allowing them to positively update.

[33]Panel C of Tables A.17 and A.18 shows results on the individual outcomes that compose the index.

[34]These are the students who may have suffered the most during the lockdown, having a more sparse social network, and for whom virtual meetings with the tutor may have been particularly important.

and psychological well-being is unlikely to be the driving channel for the improvement of students' learning outcomes, since the positive impact on academic performance emerges also in 2022 (i.e., even in the absence of effects on these dimensions).

5 Results on tutors

The primary purpose of the Tutoring Online Program was to improve outcomes for students who were the direct beneficiaries of the intervention. However, the experience of being a volunteer tutor may have affected tutors as well. To understand tutors' perceptions, in September 2020 (about three months after the end of the program) we administered a short questionnaire to all volunteers who originally *applied* to the TOP program, independent of whether they were assigned a student. The questionnaire focused on two main outcomes: the capacity to empathize and the beliefs on the relative importance of hard work versus luck for achieving success in life.[35]

As mentioned in Section 2.3, the assignment of tutors to students was random, conditional on a set of baseline characteristics used for the allocation of tutors (e.g., subject in which they could tutor and time availability). This allows us to estimate the causal impact of tutoring on the tutors, by comparing the outcomes of volunteers who were assigned to a student and volunteers who were not assigned to any student, conditioning on the allocation criteria. Around half of the respondents who completed our endline questionnaire had been randomly assigned to a student. Table A.10 shows that the characteristics of volunteers who were selected to be TOP tutors (column 2) and of those who were not selected (column 1) are overall balanced, once we take into account the allocation criteria.[36] Furthermore, given that the allocation of students to tutors was random, we can also analyze the causal impact of being assigned to a student with a specific set of characteristics on outcomes.

Table 4 shows the impact of tutoring on our two outcomes of interest: the 'Empathy index' and the 'Hard Work index' (described in Appendix B.6). All specifications control for the factors used in the assignment of tutors to students, i.e., time and subject availability, previous training and tutoring experience, and regular enrollment in university. Columns 2 and 4 additionally control for other baseline characteristics of the volunteers, such as socioeconomic background, academic performance, motivation, etc.

[35]We kept the questionnaire very brief in an attempt to get as many responses as possible, especially from university students who had not been selected to become tutors.

[36]Volunteers who majored in STEM are over-represented in the 'treatment' group (TOP tutor) because most students requested help in math, hence our allocation criteria –which privileged tutors that we could match to the needs of the students– mechanically generate this imbalance.

Table 4. Impact of TOP 2020 on Tutors

	(1)	(2)	(3)	(4)
	Empathy Index		Hardwork Index	
Panel A: overall effects				
Tutor	0.233	0.202	-0.051	-0.073
	(0.087)	(0.092)	(0.087)	(0.088)
	[0.016]	[0.060]	[0.388]	[0.256]
R^2	0.026	0.064	0.017	0.080
Panel B: by SES				
Tutor assigned to low SES student	0.319	0.229	-0.119	-0.139
	(0.106)	(0.105)	(0.102)	(0.096)
Tutor assigned to high SES student	0.138	0.054	-0.003	-0.024
	(0.109)	(0.110)	(0.100)	(0.097)
R^2	0.030	0.063	0.018	0.081
P-value difference	[0.148]	[0.118]	[0.277]	[0.366]
Panel C: by students' immigration status				
Tutor assigned to native student	0.240	0.211	-0.014	-0.035
	(0.093)	(0.097)	(0.090)	(0.091)
Tutor assigned to immigrant student	0.202	0.165	-0.204	-0.240
	(0.130)	(0.133)	(0.126)	(0.130)
R^2	0.026	0.064	0.019	0.083
P-value difference	[0.767]	[0.725]	[0.099]	[0.082]
Panel D: by students' gender				
Tutor assigned to female student	0.376	0.325	-0.039	-0.039
	(0.116)	(0.120)	(0.104)	(0.106)
Tutor assigned to male student	0.162	0.140	-0.089	-0.125
	(0.095)	(0.099)	(0.094)	(0.096)
R^2	0.031	0.067	0.017	0.081
P-value difference	[0.055]	[0.101]	[0.694]	[0.454]
Randomization controls:	Yes	Yes	Yes	Yes
Tutor controls:	No	Yes	No	Yes
Mean Dep:	-0.00	-0.00	-0.00	-0.00
Obs	739	739	734	734

Notes: OLS estimates, robust standard errors in parentheses and p-values in square brackets. The dependent variable is the Empathy index in columns 1-2 and the Hard Work index in columns 3-4. "Tutor" is an indicator for being assigned to a student in the TOP program. "Randomization controls" include: past tutoring experience; past training on learning disorders and immigrant students; expertise in math, Italian, English; availability for 3 vs. 6 hours per week; adhering to the schedule for university exam completion; confirming availability to tutor. Additional "Tutor controls" include: gender; faculty of enrollment; undergraduate vs. master student; GPA; previous volunteering activities; motivation to help others; parental education; familiarity with the computer. "Mean Dep." is the mean of the dependent variable at endline for respondents in the control group (i.e., who did not become tutors).

In columns (1) and (2) of Panel A, we find that participating in TOP increased tutors' empathy by 0.2 SD, relative to volunteers who were not assigned a student. The effect is statistically significant and sizeable.[37] The other Panels of Table 4 show the effect of

[37]In Table A.19 we present the ordered logit results for the individual questions used to build the

participating in TOP depending on the characteristics of the student to whom each tutor was randomly assigned. We find suggestive evidence that the positive effect on empathy is driven by exposure to students with lower socio-economic status (Panel B) and to girls (Panel D). These results nicely complement the work by Kosse et al. (2020), who show an improvement in prosociality for low-SES children exposed to a mentor. Furthermore, while a large literature in psychology and behavioural science (e.g., Kamas and Preston, 2021) has documented that girls are on average more empathic than boys, our results also show that exposure to girls can activate a higher level of empathy in others.

When we examine respondents' perceptions of the role of hard work to achieve success in life (columns 3-4 of Panel A), we do not find any economically or statistically significant effect of being a tutor, on average. [38] However, the aggregate results mask some interesting heterogeneity: qualitatively, tutors assigned to low-SES students and to immigrant students are less likely to believe that hard work can lead to success in life (the effect is significant at the 10 percent level for assignment to immigrant tutees). These tutors may have experienced the struggle faced by disadvantaged students, thus updating their beliefs on the importance of luck in life.

Unfortunately, we do not have any way to estimate the impact of tutoring on TOP 2022 tutors, because in that edition all volunteers who applied to become tutors were assigned a student, in an attempt to give the benefit of tutoring to as many tutees as possible. Nonetheless, the results for TOP 2020 suggest that the experience of tutoring can be beneficial to the tutors themselves, insofar as it improves their soft skills and their ability to empathize. This is a novel finding which deserves further research.

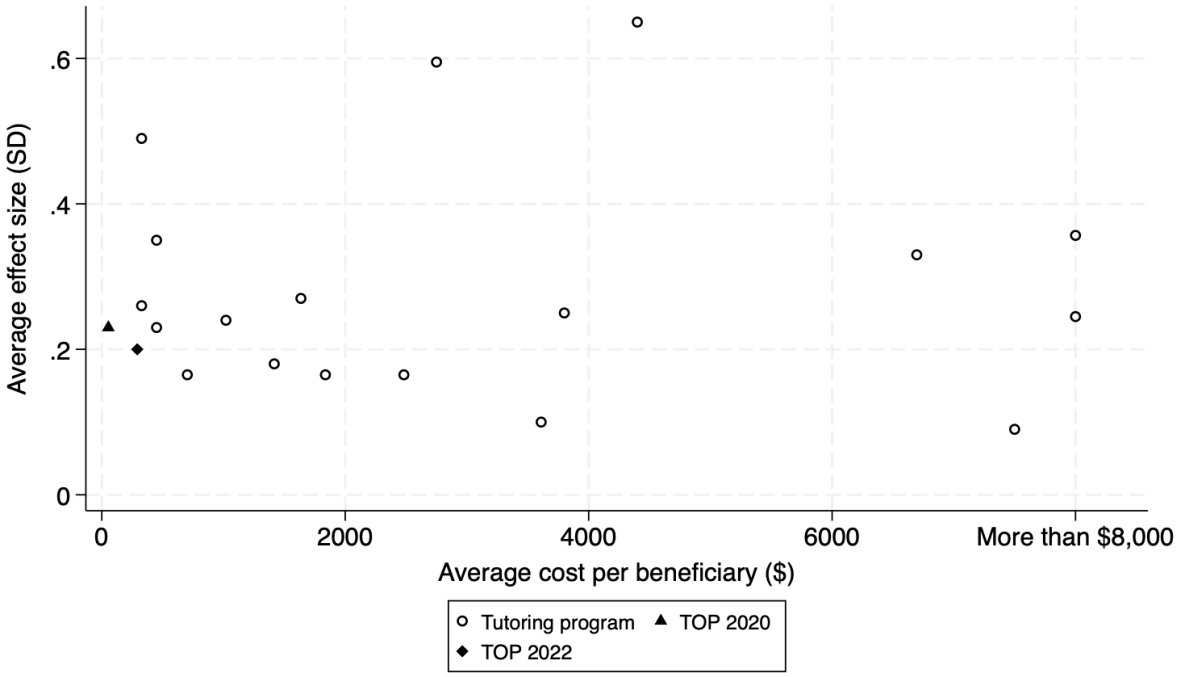
6 Scalability

The evidence discussed so far shows consistent positive effects on learning outcomes of the TOP program under various scenarios: replicating the experiment out-of-sample in different rounds of the program, implemented by different actors (the research team vs. an NGO), and in different environmental contexts (school closures vs. normal in-person interactions). However, could an online tutoring program actually scale? In this section, we emphasize two key aspects for the effective implementation at scale of an intervention of this type: costs and quality of implementation.

Empathy and Hard Work indexes. Most of the effect on empathy is driven by increased ability to put oneself in someone else's shoes.

[38] This is possibly consistent with the interpretation that society-wide factors not affected by the tutoring may play a more important role in shaping these beliefs than the experience of effort (or lack of effort) exerted by the tutee.

Figure 2. Effect on learning outcomes and average cost per beneficiary of tutoring interventions in high income countries



Notes: This figure plots the effect size on learning outcomes against the cost per beneficiaries (in USD) for a sample of tutoring interventions in high-income countries. Detailed information on the selection criteria for the tutoring programs included in the analysis is provided in Appendix E.1.

Costs. First, one of the main concerns for the scalability of tutoring programs is cost (Nickow et al., 2020; Ander et al., 2016). TOP leverages two key features to address this issue: (i) it is implemented fully *online* and (ii) tutors are *volunteer* university students, under the supervision of paid educational experts. Therefore, the main costs are related to training and supervision of tutors, monitoring implementation, recruitment and interactions with schools. The average cost for these activities was 50 euros in 2020, when TOP was fully managed by the research team, and 270 euros in 2022, when it was implemented by an NGO.[39] From 2022 onwards, we improved the program by offering access to an online tutoring platform at an average additional cost of around 75 euros per tutoring.[40]

[39]In 2022, the monitoring was done by paid staff while in 2020 we involved in part volunteers for the supervision of tutors.

[40]Section D.3 of the Appendix provides some evidence that it may be necessary to have a PC or tablet to take full advantage of online tutoring, although in the TOP 2020 edition the impact on performance was not differential for students connected through smartphones. Hence, starting 2022, we also partnered with another NGO (TechSoup) to ensure all children have access to a device and a stable internet connection. This entailed an average additional cost of 150 euros per student receiving the PC/tablet, including all

We conjecture that bringing the program to national scale maintaining the implementation by the same NGO could lead to marginal costs of approximately 300 euros per tutee, exploiting some economies of scale in the overall management of the program.

How do these costs compare to those of other tutoring interventions? Figure 2 plots the impact on learning outcomes (in SD) against the average cost per beneficiary for a number of tutoring interventions in high-income countries.[41] TOP 2020 and TOP 2022 are represented by the black triangle and diamond, respectively. The TOP model appears to be substantially cheaper than most options, while still leading to a sizeable impact on learning outcomes. Online tutoring implemented by volunteer university students trained by educational experts thus represents a viable and cost-effective solution to reach at least 10,000 students in need of support every year and to improve the learning by approximately 0.2 SD.

Quality of implementation. To ensure a sizeable positive impact on learning outcomes while scaling the program, a key ingredient is the recruitment of volunteers who can deliver high quality instruction. The number of TOP volunteers who applied and who successfully completed the training and tutoring sessions, as well as their characteristics, changed between editions. In 2022, tutors were more likely to be young undergraduate students (58% in 2022 vs. 47% in 2020), with less tutoring and volunteering experience and less likely to be solely moved by the desire to help others (see Appendix Table A.9). The supply and commitment of volunteers during the first months of COVID in 2020 was undoubtedly exceptional. However, the program was as effective in 2022 as in 2020, and we did not detect significant heterogeneity in treatment effects by tutor characteristics, as detailed in the Appendix C. This could be due to the fact that (i) the participating universities are all located in relatively wealthy areas in the North, and (ii) the tutors who self-select to participate in TOP may be particularly motivated, thus constituting a relatively homogeneous pool. To the extent that the observable characteristics of the tutor do not appear to be a driving force behind the positive impact of TOP, this can help scale up the policy.[42] The recruitment of tutors could be expanded by partnering with more universities (the universities involved in the current format enroll only 14% of the students enrolled in Italian universities) and by increasing the involvement of the university leadership.[43] We believe that by expanding the recruitment of tutors by increasing

administrative costs of regenerating computers that are donated to the NGO, licensing, and shipping. 47% of the students in TOP 2022 received a PC from TechSoup.

[41] Appendix E.1 details the criteria we used to select the programs included in the graph.

[42] Note that the GPA of TOP tutors is representative of that of students enrolled in the same major and university, with around 53% in TOP 2020 and 51% in TOP 2022 having a GPA higher than the average in their university/major.

[43] There are large differences in the participation rates of tutors across universities, ranging from 0.4%

the number of universities involved, the scale of 10,000 tutors per year in the country may be achieved while keeping the quality high enough to generate learning gains in the ballpark of 0.2 SD.

7 Conclusions

School closures due to the COVID-19 outbreak have created massive learning losses and adverse psychological effects for children, especially for the most vulnerable and those from low socioeconomic background (Agostinelli et al., 2022; Carlana et al., 2023), exacerbating pre-existing inequalities. This paper shows that online tutoring is an effective tool to improve students' academic performance. We exploit over-subscription to an innovative online tutoring program in Italy, the "Tutoring Online Program"(TOP), to evaluate its impact using a randomized control trial. We find that three hours of one-on-one support provided virtually by volunteer university students increased an index of math performance (encompassing teacher grades, failure rate, and a standardized test score) by 0.22 SD in 2020 and 0.20 SD in 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 compared to individual one-on-one tutoring. Importantly, the academic gains from TOP were realized not only during the pandemic but also during normal school times. Furthermore, we find that online tutoring was effective in improving aspirations, psychological well-being, and development of socio-emotional skills, but only in 2020, when students were in distress and socially isolated.

In-person tutoring, especially when implemented by professionals and/or for several days per week, has proved to be highly effective but it is generally viewed as "too costly to be undertaken on a large scale" (Ander et al., 2016). A program like TOP allows to achieve sizeable results on learning and other life outcomes, keeping the costs extremely contained. Several countries have launched national tutoring plans to accelerate learning and mitigate inequalities. Among others, Chile has strongly built on the early evidence on TOP to launch a national plan for tutoring with volunteer university students.[44] Even when schools are open, virtual tutoring implemented by volunteer university students provides an effective tool to help vulnerable children and prevent inequalities to emerge, in a cost-effective way.

to 2%. These differences are correlated with the engagement of the universities in sharing the information with students, with the most successful ones sending multiple emails and pop-up messages to the students' online profiles.

[44]Decreto 2169 Exento, Ministerio de Educacion. Available here: https://shorturl.at/SEPqj.

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-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*, 6(7):941–950.

Angrist, N., Cullen, C., Ainomugisha, M., Bathena, S. P., Bergman, P., Crossley, C., Letsomo, T., Matsheng, M., Panti, R. M., Sabarwal, S., and Sullivan, T. (2023). Learning curve: Progress in the replication crisis. *AEA Papers and Proceedings*, 113:482–88.

Angrist, N., Evans, D. K., Filmer, D., Glennerster, R., Rogers, F. H., and Sabarwal, S. (2020). How to improve education outcomes most efficiently. *A Comparison of*, 150.

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.

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., Shotland, M., and Walton, M. (2016). Mainstreaming an effective intervention: Evidence from randomized evaluations of teaching at the right level in india. Working Paper 22746, National Bureau of Economic Research.

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. (2019). Implicit stereotypes: Evidence from teachers' gender bias. *The Quarterly Journal of Economics*, 134(3):1163–1224.

Carlana, M., La Ferrara, E., and Lopez, C. (2023). Exacerbated inequalities: The learning loss from covid-19 in italy. *AEA Papers and Proceedings*, 113:489–93.

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

Carlana, M., La Ferrara, E., and Pinotti, P. (2022b). Implicit stereotypes in teachers' track recommendations. *AEA Papers and Proceedings*, 112:409–14.

Chernozhukov, V., Demirer, M., Duflo, E., and Fernandez-Val, I. (2018). Generic machine learning inference on heterogeneous treatment effects in randomized experiments, with an application to immunization in india. Working Paper 24678, National Bureau of Economic Research.

Coie, J. D. and Krehbiel, G. (1984). Effects of academic tutoring on the social status of low-achieving, socially rejected children. *Child Development*, 55(4):1465–1478.

Cortes, K. E., Kortecamp, K., Loeb, S., and Robinson, C. D. (2025). A scalable approach to high-impact tutoring for young readers. *Learning and Instruction*, 95:102021.

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.

Falk, A., Kosse, F., and Pinger, P. (2024). 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. (2021). Motivating bureaucrats through social recognition: External validity - a tale of two states. *Organizational Behavior and Human Decision Processes*, 163:117–131.

Golberstein, E., Wen, H., and Miller, B. F. (2020). Coronavirus disease 2019 (covid-19) and mental health for children and adolescents. *JAMA pediatrics*, 174(9):819–820.

Gortazar, L., Hupkau, C., and Roldán, A. (2024). Online tutoring works: Experimental evidence from a program with vulnerable children. *Journal of Public Economics*, (232):3–22.

Hardt, D., Nagler, M., and Rincke, J. (2022). Can peer mentoring improve online teaching effectiveness? an rct during the covid-19 pandemic. *Labour Economics*, 78:102220.

Hassan, H., Islam, A., Siddique, A., and Wang, L. C. (2024). Telementoring and homeschooling during school closures: a randomised experiment in rural bangladesh. *The Economic Journal*, 134(662):2418–2438.

Kamas, L. and Preston, A. (2021). Empathy, gender, and prosocial behavior. *Journal of Behavioral and Experimental Economics*, 92:101654.

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.

Lawrence, E. J., Shaw, P., Baker, D., Baron-Cohen, S., and David, A. S. (2004). Measuring empathy: reliability and validity of the empathy quotient. *Psychological medicine*, 34(5):911–920.

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.

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

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. Working Paper 27476, National Bureau of Economic Research.

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.

Pastori, G. and Farina, E. (2022). What worked? a qualitative study on the tutor-tutee relationship in the tutoring online program (top). *EXCELLENCE AND INNOVATION IN LEARNING AND TEACHING*, (2022/2).

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*.

Ready, D. D., McCormick, S. G., and Shmoys, R. J. (2024). The effects of in-school virtual tutoring on student reading development: Evidence from a short-cycle randomized controlled trial. *EdWorkingPaper: 24–942*.

Resnjanskij, S., Ruhose, J., Wiederhold, S., Woessmann, L., and Wedel, K. (2024). Can mentoring alleviate family disadvantage in adolescence? a field experiment to improve labor market prospects. *Journal of Political Economy*, 132(3):1013–1062.

Robinson, C. D., Pollard, C., Novicoff, S., White, S., and Loeb, S. (2024). The effects of virtual tutoring on young readers: Results from a randomized controlled trial. *Educational Evaluation and Policy Analysis*.

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

Schueler, B. E. and Rodriguez-Segura, D. (2023). A cautionary tale of tutoring hard-to-reach students in kenya. *Journal of Research on Educational Effectiveness*, 16(3):442–472.

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

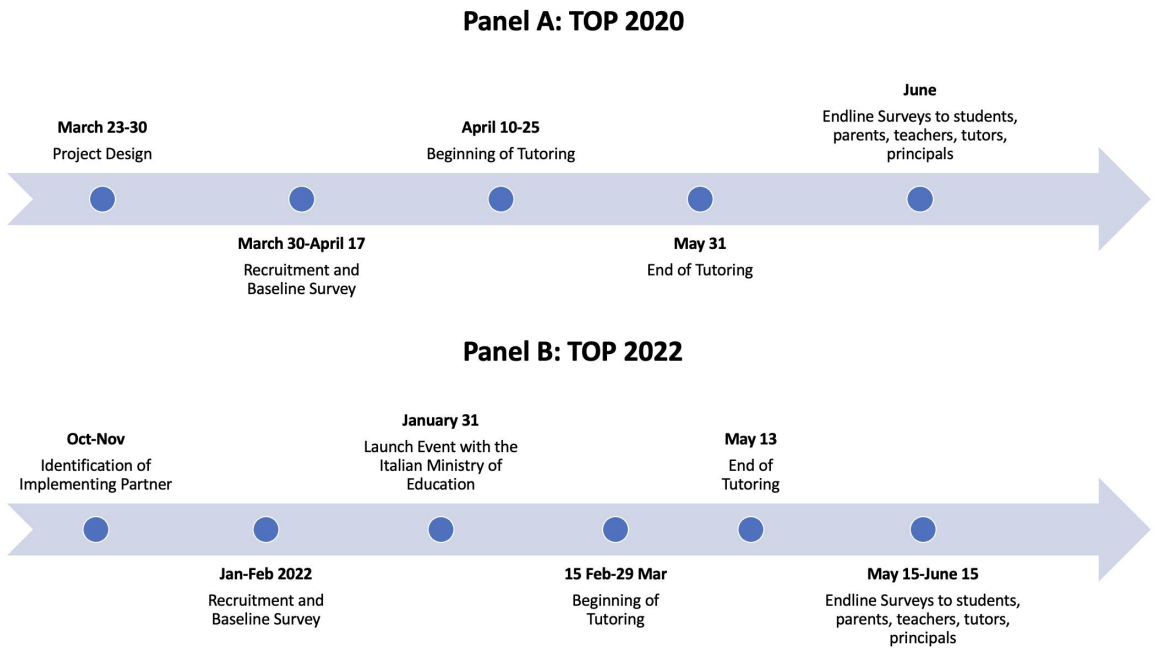
Wager, S. and Athey, S. (2018). Estimation and inference of heterogeneous treatment effects using random forests. *Journal of the American Statistical Association*, 113(523):1228–1242.

Werner, K. and Woessmann, L. (2023). The legacy of covid-19 in education. *Economic Policy*.

Supplemental Appendix

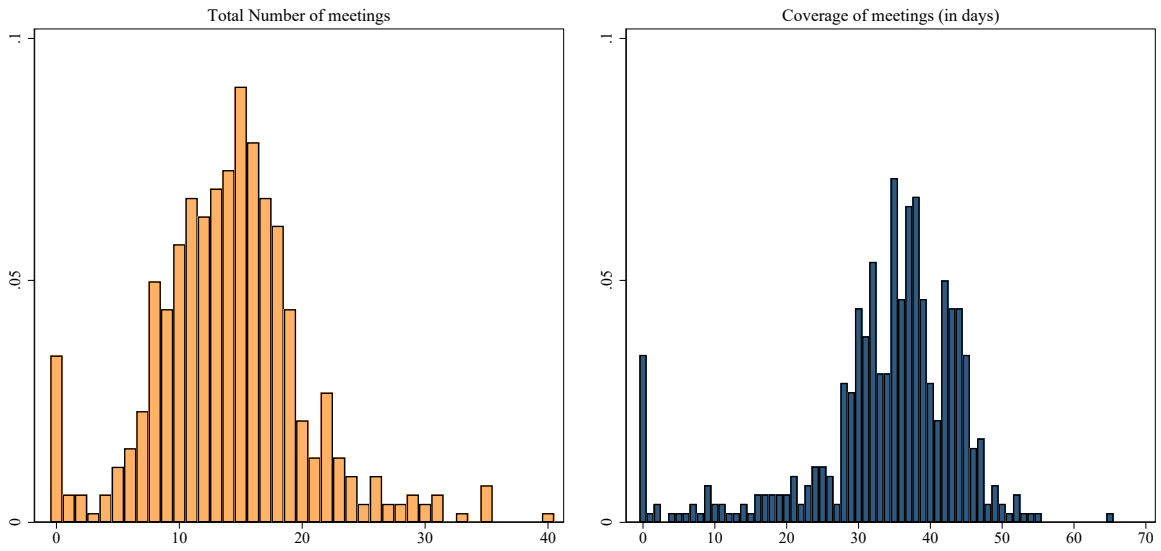
A Online Appendix: Additional Tables and Figures

Figure A.1. Timeline of TOP



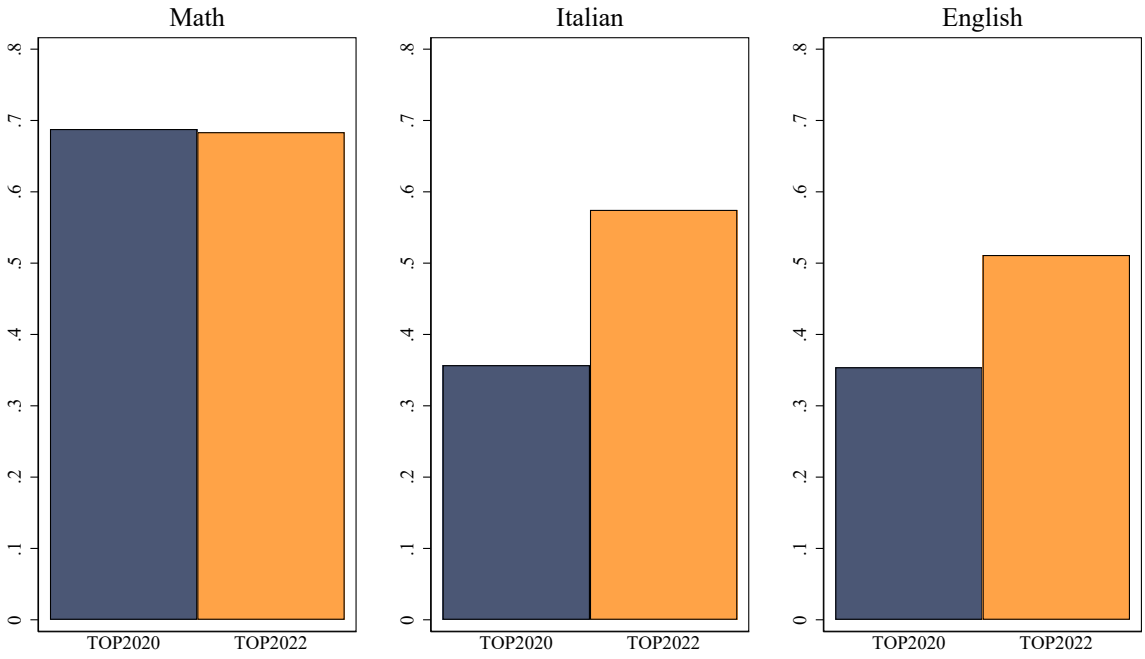
Notes: This Figure shows the timeline for the implementation of TOP 2020 in Panel A and TOP 2022 in Panel B.

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



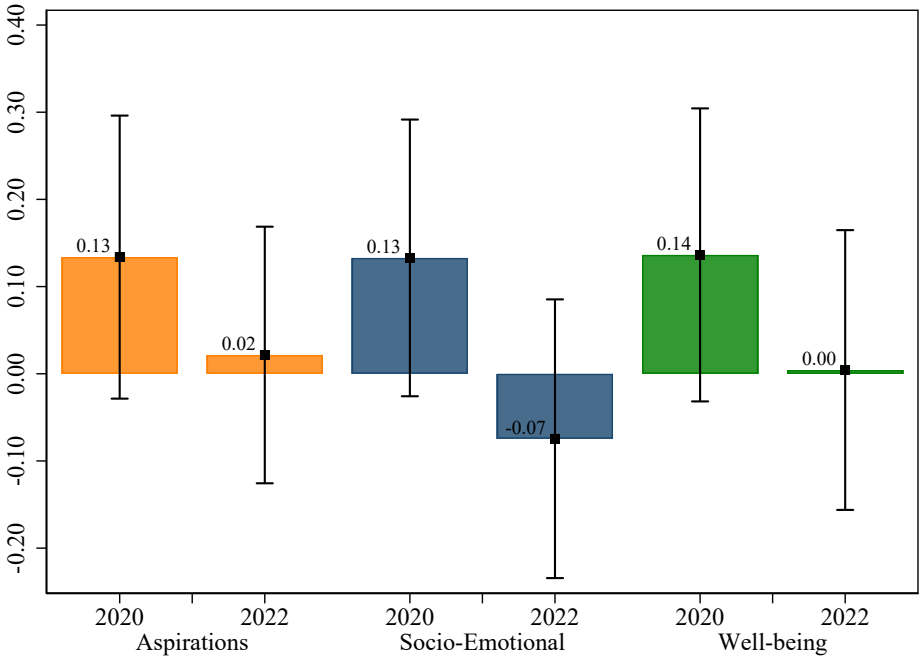
Notes: This Figure shows the total number of tutoring meetings (left panel) and the number of days from the beginning to the end of the tutoring (right panel) for TOP 2020. The data used in these graphs are reported by tutors in the registry. The Figure shows data from 522 treated students: 8 students did the tutoring but we do not have precise information from the tutors on the number of meetings.

Figure A.3. Main Subjects During Tutoring



Notes: This Figure shows the main subjects done during the tutoring meetings for TOP 2020 (blue bar) and for TOP 2022 (orange bar), as reported by tutors at endline.

Figure A.4. The impact of three hours of individual tutoring on non-academic outcomes



Notes: This Figure reports OLS estimates of the impact of 3 hours per week of individual tutoring on "aspirations", "socio-emotional skills", and "psychological well-being". Each index is standardized to have mean 0 and the standard deviation is 1 in the control group. For each outcome, the first bar shows the impact on TOP 2020 and the second bar on TOP 2022, with 95% confidence intervals. Controls included are parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline.

Table A.1. Characteristics of provinces with schools participating/not participating in TOP 2020

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	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)	
Joint test				0.000
Observations	67	40	107	

Notes: This table shows the characteristics of provinces that had at least one treated school (column 2) compared to provinces with no treated schools (column 1) for TOP 2020. Column (3) shows the difference in means between the two groups and column (4) 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. The second to last row shows the P-value of an F-test for the joint significance of all the coefficients in a regression of the a binary indicator equal to one if the province in participating in TOP 2020 on all the covariates in the table.

Table A.2. Balance Table TOP students vs. schoolmates

Variable	(1) TOP	(2) No TOP	(3) Diff.	(4) Std. Diff.
Immigrant	0.170	0.132	-0.038	-0.075
	(0.376)	(0.339)	(0.011)	
Student is male	0.602	0.514	-0.088	-0.126
	(0.490)	(0.500)	(0.017)	
Father was born in Italy	0.814	0.841	0.027	0.050
	(0.389)	(0.366)	(0.013)	
Mother was born in Italy	0.754	0.809	0.055	0.094
	(0.431)	(0.393)	(0.014)	
Mother edu: High-school	0.333	0.393	0.060	0.089
	(0.472)	(0.489)	(0.018)	
Mother edu: Degree	0.101	0.200	0.099	0.197
	(0.302)	(0.400)	(0.015)	
Father edu: high-school	0.279	0.351	0.072	0.110
	(0.449)	(0.477)	(0.018)	
Father edu: Degree	0.079	0.146	0.068	0.152
	(0.269)	(0.353)	(0.013)	
Low SES	0.602	0.445	-0.157	-0.225
	(0.490)	(0.497)	(0.018)	
Std Invalsi score maths, 5th grade	-0.562	0.025	0.587	0.434
	(0.915)	(0.996)	(0.036)	
Std Invalsi score Italian, 5th grade	-0.559	0.024	0.583	0.435
	(0.897)	(0.997)	(0.036)	
Baseline Grade Math	6.154	7.214	1.060	0.595
	(1.132)	(1.375)	(0.056)	
Baseline Grade Italian	6.475	7.273	0.799	0.550
	(0.920)	(1.123)	(0.046)	
Baseline Grade in English	6.466	7.374	0.908	0.545
	(1.054)	(1.292)	(0.053)	
Joint test				0.000
Observations	929	20,931	21,860	

Notes: This table reports characteristics of TOP 2020 students (column 1) and their schoolmates (column 2). Column (3) reports the difference in means between column (2) and columns (1). Column (4) reports the standardized difference between group averages. All variables are measured at baseline in administrative data. The second to last row shows the P-value of an F-test for the joint significance of all the coefficients in a regression of the a binary indicator equal to one if the student participated in TOP 2020 on all the covariates in the table.

Table A.3. Balance Table TOP 2020 (baseline 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	0.381	0.412	0.403	0.437	0.298	0.481	0.547
Immigrant	0.169	0.178	0.161	0.225	0.690	0.962	0.403
Learning Disorder	0.337	0.312	0.297	0.355	0.410	0.307	0.486
School grade 6	0.325	0.313	0.313	0.315	0.692	0.823	0.767
School grade 7	0.338	0.342	0.351	0.315	0.934	0.773	0.571
School grade 8	0.336	0.345	0.336	0.371	0.760	0.945	0.401
Admin data							
Baseline math grade	6.270	6.152	6.190	6.049	0.066	0.121	0.692
Baseline Italian grade	6.330	6.397	6.424	6.324	0.250	0.357	0.880
Baseline English grade	6.456	6.436	6.510	6.234	0.759	0.788	0.099
Parents							
SES Status	-0.273	-0.230	-0.202	-0.306	0.400	0.274	0.401
Single parent household	0.258	0.219	0.221	0.211	0.144	0.326	0.407
Mother Education: High-school	0.469	0.439	0.412	0.514	0.344	0.071	0.026
Mother Education: University	0.126	0.113	0.123	0.087	0.526	0.699	0.495
Father Education: High-school	0.358	0.380	0.384	0.370	0.452	0.472	0.892
Father Education: University	0.084	0.063	0.065	0.058	0.201	0.245	0.945
Mother not employed	0.312	0.317	0.320	0.308	0.880	0.695	0.554
Mother Occupation: white-collar	0.270	0.268	0.274	0.252	0.937	0.983	0.789
Mother Occupation: blue-collar	0.391	0.379	0.372	0.399	0.698	0.543	0.542
Father not employed	0.132	0.185	0.186	0.182	0.019	0.026	0.805
Father Occupation: white-collar	0.282	0.270	0.274	0.259	0.679	0.675	0.958
Father Occupation: blue-collar	0.558	0.511	0.506	0.524	0.124	0.148	0.894
Parental time helping homework	49.062	48.553	49.003	47.319	0.846	0.947	0.777
Joint test					0.117	0.244	0.659
Observations	529	530	387	143			

Notes: This table reports characteristics of control students (column 1) and treated students (column 2) in TOP 2020. Columns (3) and (4) divide the treated students in those assigned to 3 hours vs. 6 hours of tutoring. 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 "Treatment" and "Treatment Intense" are reported in columns (6) and (7), respectively. As in our main specification, we control for whether the student was identified for intense tutoring for the last two columns. All variables are measured at baseline. The second to last row shows in Column 5 the P-value of an F-test for the joint significance of all the coefficients in a regression of the treatment and randomization round FE. Next are shown the P-values of a Chi^2-test for the joint significance of all the coefficients in a multinomial logit regression on the 3 hours treatment against the control (Column 6) and the six hours treatment against the 3 hours treatment (Column 7).

Table A.4. Balance Table TOP 2022 (baseline 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
Learning Disorder	0.247	0.262	0.275	0.245	0.546	0.426	0.585
School grade 6	0.268	0.331	0.310	0.358	0.019	0.220	0.161
School grade 7	0.368	0.367	0.363	0.374	0.941	0.785	0.694
School grade 8	0.364	0.301	0.327	0.268	0.026	0.363	0.066
Admin data							
Baseline math grade	6.347	6.393	6.442	6.330	0.444	0.467	0.881
Baseline Italian grade	6.596	6.628	6.630	6.625	0.569	0.596	0.943
Baseline English grade	6.637	6.755	6.761	6.746	0.189	0.113	0.395
Parents							
SES Status	-0.375	-0.442	-0.445	-0.439	0.291	0.173	0.332
Single parent household	0.246	0.244	0.228	0.264	0.898	0.553	0.377
Mother Education: High-school	0.441	0.403	0.423	0.378	0.304	0.818	0.227
Mother Education: University	0.124	0.124	0.128	0.120	0.826	0.713	0.720
Father Education: High-school	0.375	0.392	0.435	0.337	0.504	0.119	0.056
Father Education: University	0.069	0.082	0.083	0.081	0.364	0.505	0.859
Mother not employed	0.309	0.336	0.354	0.313	0.287	0.122	0.219
Mother Occupation: white-collar	0.281	0.259	0.278	0.234	0.217	0.476	0.525
Mother Occupation: blue-collar	0.381	0.367	0.327	0.419	0.796	0.204	0.055
Father not employed	0.145	0.173	0.152	0.200	0.117	0.411	0.338
Father Occupation: white-collar	0.270	0.272	0.287	0.253	0.913	0.895	0.670
Father Occupation: blue-collar	0.542	0.506	0.497	0.517	0.132	0.300	0.658
Parental time helping homework	32.272	35.572	36.121	34.868	0.146	0.227	0.982
Joint test					0.336	0.318	0.074
Observations	530	607	342	265			

Notes: This table reports characteristics of control students (column 1) and treated students (column 2) in TOP 2022. Columns (3) and (4) divide the treated students in those assigned to individual vs. group tutoring. 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 "Individual Tutoring" and "Group Tutoring" are reported in columns (6) and (7), respectively. All variables are measured at baseline. The second to last row shows in Column 5 the P-value of an F-test for the joint significance of all the coefficients in a regression of the treatment and randomization round FE. Next are shown the P-values of a Chi^2-test for the joint significance of all the coefficients in a multinomial logit regression on the individual group treatment against the control (Column 6) and the group treatment against the individual treatment (Column 7).

Table A.5. Characteristics of students: TOP 2020 vs. TOP 2022

	(1) All	(2) TOP2020	(3) TOP2022	(4) P-value	(5) Std diff.
Students					
Survey Data					
Female	0.425	0.396	0.452	0.008	0.113
Immigrant	0.226	0.173	0.274	0.000	0.242
Learning Disorder	0.288	0.324	0.255	0.000	-0.152
School grade 6	0.310	0.319	0.302	0.376	-0.037
School grade 7	0.354	0.340	0.368	0.175	0.059
School grade 8	0.336	0.341	0.331	0.614	-0.021
Admin data					
Baseline math grade	6.294	6.211	6.371	0.001	0.147
Baseline Italian grade	6.493	6.363	6.613	0.000	0.259
Baseline English grade	6.578	6.446	6.700	0.000	0.225
Parents					
SES Status	-0.337	-0.287	-0.411	0.006	-0.141
Single parent household	0.242	0.238	0.245	0.708	0.016
Mother Education: High-school	0.437	0.454	0.421	0.121	-0.067
Mother Education: University	0.122	0.120	0.124	0.759	0.012
Father Education: High-school	0.377	0.369	0.384	0.475	0.031
Father Education: University	0.075	0.074	0.076	0.804	0.008
Mother not employed	0.319	0.314	0.324	0.644	0.021
Mother Occupation: white-collar	0.269	0.269	0.269	1.000	0.000
Mother Occupation: blue-collar	0.379	0.385	0.374	0.580	-0.023
Father not employed	0.159	0.159	0.160	0.927	0.003
Father Occupation: white-collar	0.273	0.276	0.271	0.799	-0.011
Father Occupation: blue-collar	0.528	0.534	0.522	0.572	-0.024
Parental time helping homework	41.113	48.810	34.028	0.000	-0.381
P-value of F test				0.000	
Observations	2,196	1,059	1,137		

Notes: This table reports characteristics of all students (column 1), students in TOP 2020 (column 2), and students in TOP 2022 (column 3). Column 4 reports p-values for difference in means. Column 5 reports the standardized difference between group averages. All variables are measured at baseline. The SES Status is the variable of "Economic, Social and Cultural Status" created by INVALSI including three components: parental occupation, parental education, and the ownership of specific goods (quite space to study, computer, desk, internet at home, own bedroom, encyclopedia). Parents are not employed when they are homemakers, unemployed or retired. The second to last row shows the P-value of an F-test for the joint significance of all the coefficients in a regression of a binary indicator equal to one if the student participated in TOP 2022 on all the covariates in the table.

Table A.6. Attrition between baseline and endline sample

	(1)	(2)	(3)	(4)	(5)	(6)
Sample	All	All	TOP2020		TOP2022	
	Full	Full	Treatment	Control	Treatment	Control
Dependent variable: Dummy for endline completion						
Treatment	0.305	0.300				
	(0.017)	(0.017)				
Female		0.011	0.017	-0.015	0.007	-0.009
		(0.017)	(0.028)	(0.046)	(0.023)	(0.038)
Immigrant		0.047	-0.026	0.097	0.023	0.123
		(0.021)	(0.042)	(0.062)	(0.027)	(0.045)
Learning Disorder		-0.024	-0.041	-0.038	-0.005	-0.043
		(0.019)	(0.029)	(0.048)	(0.027)	(0.046)
School grade 6						
School grade 7						
Baseline math grade		0.005	0.014	-0.015	0.026	0.005
		(0.008)	(0.015)	(0.023)	(0.011)	(0.020)
Single parent household		-0.028	-0.023	0.000	-0.061	-0.003
		(0.022)	(0.037)	(0.053)	(0.030)	(0.049)
Mother Education: High-school		0.015	0.022	0.033	0.031	-0.038
		(0.020)	(0.032)	(0.055)	(0.029)	(0.046)
Mother Education: University		0.005	-0.025	-0.030	0.052	-0.028
		(0.035)	(0.061)	(0.095)	(0.039)	(0.075)
Father Education: High-school		0.017	0.024	0.030	0.033	-0.010
		(0.020)	(0.032)	(0.053)	(0.025)	(0.045)
Father Education: University		0.036	0.103	0.139	-0.022	-0.005
		(0.040)	(0.068)	(0.114)	(0.050)	(0.083)
Mother Occupation: White-collar		-0.035	0.015	0.357	-0.022	-0.196
		(0.051)	(0.229)	(0.263)	(0.049)	(0.122)
Mother Occupation: Blue-collar		-0.082	-0.060	0.226	-0.009	-0.206
		(0.050)	(0.227)	(0.266)	(0.051)	(0.122)
Father Occupation: White-collar		0.002	-0.162	-0.387	0.056	0.037
		(0.052)	(0.157)	(0.198)	(0.065)	(0.104)
Father Occupation: Blue-collar		0.064	-0.120	-0.278	0.101	0.067
		(0.051)	(0.155)	(0.196)	(0.064)	(0.100)
High treatment intensity			-0.024			
			(0.032)			
Higher than average num. meetings			0.089			
			(0.028)			
Group tutoring					-0.027	
					(0.030)	
Mean Dep	.753	.753	.879	.463	.923	.723
Obs.	2196	2196	513	529	607	530
R^2	0.164	0.188	0.084	0.090	0.085	0.122

Notes: OLS estimates, robust standard errors in parentheses. The dependent variable is a dummy which assumes value 1 if the student completed the endline survey. Columns (1) and (2) include all students who completed the baseline surveys in TOP 2020 or 2022. Columns (3) and (4) restrict the sample to students in TOP 2020 assigned to the treatment and control group, respectively. Columns (5) and (6) restrict the sample to students in TOP 2022 assigned to the treatment and control group, respectively.

Table A.7. Balance Table TOP 2020 (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	0.419	0.418	0.424	0.213	0.230	0.924
Immigrant	0.197	0.175	0.162	0.208	0.469	0.434	0.737
Learning Disorder	0.317	0.312	0.303	0.339	0.912	0.853	0.769
School grade 6	0.367	0.315	0.318	0.310	0.196	0.308	0.609
School grade 7	0.335	0.337	0.335	0.341	0.986	0.953	0.834
School grade 8	0.298	0.348	0.347	0.349	0.193	0.272	0.765
Admin data							
Baseline math grade	6.281	6.158	6.198	6.048	0.103	0.179	0.524
Baseline Italian grade	6.421	6.403	6.421	6.352	0.710	0.602	0.648
Baseline English grade	6.533	6.434	6.510	6.226	0.204	0.447	0.125
Parents							
SES Status	-0.267	-0.278	-0.253	-0.343	0.853	0.967	0.703
Single parent household	0.250	0.214	0.220	0.200	0.330	0.604	0.269
Mother Education: High-school	0.500	0.444	0.410	0.537	0.142	0.023	0.012
Mother Education: University	0.142	0.111	0.122	0.083	0.190	0.280	0.507
Father Education: High-school	0.368	0.381	0.378	0.388	0.853	0.942	0.787
Father Education: University	0.105	0.065	0.065	0.066	0.080	0.065	0.629
Mother not employed	0.310	0.318	0.326	0.294	0.856	0.569	0.269
Mother Occupation: white-collar	0.310	0.273	0.271	0.278	0.262	0.193	0.519
Mother Occupation: blue-collar	0.363	0.373	0.368	0.389	0.683	0.808	0.689
Father not employed	0.118	0.180	0.179	0.183	0.027	0.038	0.953
Father Occupation: white-collar	0.273	0.264	0.265	0.262	0.760	0.660	0.697
Father Occupation: blue-collar	0.584	0.519	0.521	0.516	0.120	0.184	0.739
Parental time helping homework	47.591	49.660	50.775	46.689	0.621	0.438	0.351
Joint test					0.090	0.203	0.756
Observations	245	466	340	126			

Notes: This table reports characteristics of control students (column 1) and treated students (column 2) in TOP 2020, restricting the same to those students who completed the endline test score. Columns (3) and (4) divide the treated students in those assigned to 3 hours vs. 6 hours of tutoring. 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 "Treatment" and "Treatment Intense" are reported in columns (6) and (7), respectively. As in our main specification, we control for whether the student was identified for intense tutoring. All variables are measured at baseline. The second to last row shows in Column 5 the P-value of an F-test for the joint significance of all the coefficients in a regression of the treatment and randomization round FE. Next are shown the P-values of a Chi^2-test for the joint significance of all the coefficients in a multinomial logit regression on the 3 hours treatment against the control (Column 6) and the six hours treatment against the 3 hours treatment (Column 7).

Table A.8. Balance Table TOP 2022 (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	0.457	0.401	0.534	0.867	0.222	0.008
Immigrant	0.287	0.296	0.307	0.282	0.633	0.261	0.158
Learning Disorder	0.236	0.257	0.270	0.239	0.429	0.321	0.536
School grade 6	0.311	0.329	0.311	0.353	0.528	0.910	0.400
School grade 7	0.399	0.368	0.366	0.370	0.335	0.307	0.697
School grade 8	0.290	0.304	0.323	0.277	0.709	0.335	0.204
Admin data							
Baseline math grade	6.396	6.429	6.472	6.371	0.610	0.548	0.757
Baseline Italian grade	6.637	6.646	6.633	6.662	0.878	0.955	0.881
Baseline English grade	6.683	6.768	6.781	6.751	0.399	0.172	0.219
Parents							
SES Status	-0.400	-0.437	-0.442	-0.430	0.638	0.427	0.415
Single parent household	0.235	0.230	0.224	0.239	0.885	0.684	0.581
Mother Education: High-school	0.422	0.414	0.434	0.388	0.985	0.530	0.205
Mother Education: University	0.127	0.126	0.127	0.125	0.777	0.530	0.426
Father Education: High-school	0.366	0.408	0.446	0.355	0.188	0.080	0.208
Father Education: University	0.072	0.080	0.079	0.082	0.543	0.787	0.604
Mother not employed	0.326	0.330	0.348	0.307	0.802	0.351	0.155
Mother Occupation: white-collar	0.266	0.259	0.280	0.231	0.610	0.894	0.542
Mother Occupation: blue-collar	0.373	0.371	0.332	0.424	0.994	0.297	0.044
Father not employed	0.146	0.166	0.146	0.193	0.270	0.574	0.439
Father Occupation: white-collar	0.253	0.270	0.280	0.256	0.541	0.876	0.467
Father Occupation: blue-collar	0.559	0.518	0.509	0.529	0.143	0.297	0.675
Parental time helping homework	32.715	35.678	36.378	34.737	0.256	0.300	0.880
Joint test					0.964	0.713	0.089
Observations	383	560	322	238			

Notes: This table reports characteristics of control students (column 1) and treated students (column 2) in TOP 2022, restricting the same to those students who completed the endline test score. Columns (3) and (4) divide the treated students in those assigned to individual vs. group tutoring. 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 "Individual Tutoring" and "Group Tutoring" are reported in columns (6) and (7), respectively. All variables are measured at baseline. The second to last row shows in Column 5 the P-value of an F-test for the joint significance of all the coefficients in a regression of the treatment and randomization round FE. Next are shown the P-values of a Chi^2-test for the joint significance of all the coefficients in a multinomial logit regression on the individual group treatment against the control (Column 6) and the group treatment against the individual treatment (Column 7).

Table A.9. Characteristics of tutors: TOP 2020 vs. TOP 2022

	(1) All	(2) TOP2020	(3) TOP2022	(4) P-value	(5) Std diff.
Female	0.718	0.701	0.736	0.221	0.078
Immigrant	0.055	0.017	0.097	0.000	0.351
GPA	26.563	26.726	26.380	0.039	-0.135
Volunteering experience	0.773	0.822	0.719	0.000	-0.246
Tutoring experience	0.871	0.958	0.778	0.000	-0.537
Tutor's motivation: wants to help / be useful	0.699	0.831	0.555	0.000	-0.601
Major: Economics	0.288	0.285	0.291	0.845	0.013
Major: Education	0.044	0.065	0.021	0.000	-0.215
Major: STEM	0.303	0.337	0.266	0.014	-0.154
Undergraduate Degree	0.525	0.473	0.582	0.001	0.218
Younger siblings	0.672	0.697	0.644	0.293	-0.067
Father education: University	0.399	0.388	0.411	0.462	0.047
Mother Education: University	0.403	0.387	0.421	0.286	0.069
Joint test				0.000	
Observations	1,003	522	481		

Notes: This table reports characteristics of all tutors (column 1), tutors in TOP 2020 (column 2), and tutors in TOP 2022 (column 3). P-values for difference in means between the two editions are reported in column (4). The standardized difference between group averages is reported in column (5). For TOP 2020, we selected a limited number of tutors among the applicants considering the subject and time availability. For TOP 2022, we assigned all available volunteers. The second to last row shows the P-value of an F-test for the joint significance of all the coefficients in a regression of a binary indicator equal to one if the tutor participated in TOP 2022 on all the covariates in the table.

Table A.10. Balance Table for Tutors in TOP 2020

	(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	0.718	0.701	0.690	0.730	0.277	0.606	0.280
GPA	26.366	26.010	26.400	24.954	0.803	0.833	0.405
Tutoring experience	5.331	5.080	5.155	4.879	0.974	0.964	0.871
Volunteering experience	0.777	0.822	0.816	0.837	0.994	0.998	0.991
Motivation TOP: help others	0.783	0.831	0.845	0.794	0.008	0.021	0.807
Major: Economics	0.308	0.285	0.325	0.177	0.663	0.908	0.494
Major: Education	0.058	0.065	0.055	0.092	0.208	0.355	0.655
Major: STEM	0.158	0.337	0.357	0.284	0.003	0.009	0.939
Undergraduate Degree	0.509	0.473	0.496	0.411	0.407	0.876	0.166
Younger siblings	0.502	0.548	0.554	0.532	0.838	0.551	0.343
Mother education: University	0.420	0.385	0.399	0.348	0.789	0.814	0.991
Father education: University	0.399	0.385	0.383	0.390	0.181	0.413	0.374
Joint test					0.000	0.000	0.000
Observations	1,534	522	381	141			

Notes: This table reports characteristics of tutors in the control group (column 1), and assigned to a tutee in TOP 2020 (column 2). Column (3) and (4) splits the sample between the tutors that self-selected to volunteer for 3 hours per week or 6 hours. 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 "Treatment" and "Treatment Intense" are reported in columns (6) and (7), respectively. The second to last row shows the P-value of an F-test for the joint significance of all the coefficients in a regression respectively of the treatment (5) the 3 hours treatment (6), the six hours treatment (7) on all the covariates in the table and randomization round FE.

Table A.11. Balance Table for Tutors in TOP 2022

	(1) Treatment	(2) Individual Tutoring	(3) Group Tutoring	(4) P-value (2)-(3)
Female	0.736	0.731	0.748	0.696
Immigrant	0.097	0.101	0.087	0.640
GPA	26.380	26.321	26.527	0.470
Volunteering experience	0.719	0.725	0.705	0.660
Tutoring experience	0.778	0.775	0.784	0.823
Tutor's motivation: wants to help / be useful	0.555	0.535	0.604	0.163
Major: Economics	0.291	0.269	0.345	0.105
Major: Education	0.021	0.023	0.014	0.490
Major: STEM	0.266	0.260	0.281	0.651
Undergraduate Degree	0.582	0.594	0.554	0.428
Younger siblings	0.644	0.633	0.672	0.645
Father education: University	0.411	0.418	0.394	0.637
Mother Education: University	0.421	0.408	0.453	0.372
Joint test				0.979
Observations	481	342	139	

Notes: This table reports characteristics of tutors in the treatment group in TOP 2022 (column 1), and among those self-selected for individual and group tutoring (column 2 and 3). P-values for difference in means between tutors in the individual and group tutoring are reported in column (4). The second to last row shows the P-value of an F-test for the joint significance of all the coefficients in a regression of the Group treatment on all the covariates in the table and randomization round FE.

Table A.12. Summary statistics of main outcome variables

	TOP 2020			TOP 2022		
	Obs	Mean	S.D.	Obs	Mean	S.D.
Section I: Academic Outcomes						
Performance Index	711	0.14	1.04	943	0.10	0.96
Grade Math Endline	1058	6.40	1.09	1121	6.03	0.98
Fail Math Endline	1058	0.16	0.37	1121	0.28	0.45
Std Performance Math	712	0.10	1.03	953	0.09	0.96
Section II: Aspirations						
Std Aspirations Index	523	0.07	1.00	889	0.03	1.02
Outcomes reported by Students						
Aspirations University	674	0.39	0.49	961	0.32	0.47
Self-efficacy University	682	0.24	0.42	965	0.26	0.44
High-school: vocational	681	0.28	0.45	966	0.22	0.41
High-school: top tier	681	0.15	0.36	966	0.15	0.36
Outcomes reported by Parents						
Aspirations University	765	0.35	0.48	1000	0.28	0.45
Self-efficacy: university	772	0.33	0.47	1002	0.36	0.48
Outcomes reported by Teachers						
Aspirations 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	1.02
Outcomes reported by Students						
Logic task: difficult	685	0.59	0.49	967	0.56	0.50
Logic task: give-up	685	0.12	0.32	967	0.13	0.34
Grit	673	0.69	0.13	958	0.66	0.13
Locus of control	685	0.72	0.11	961	0.70	0.11
Outcomes reported by Parents						
Grit	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						
Depression	669	0.54	0.12	948	0.58	0.12
Happiness	665	0.63	0.22	955	0.69	0.21
Outcomes reported by Parents						
Depression	731	0.58	0.10	968	0.58	0.10
Happiness	741	0.62	0.21	969	0.68	0.18

Notes: This table shows the summary statistics of all outcome variables for TOP 2020 (columns 1-3) and TOP 2022 (columns 4-6), as reported in the administrative data for grades and failure rate or endline questionnaire from students, parents, and teachers for the other variables. 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.13. Estimation of the impact of TOP 2020 on academic outcomes and beliefs

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Regression Model 1			Regression Model 2					
	Treatment	SE	q-value	Treatment	SE	Intense	SE	Mean Dep	Obs
Outcomes									
Panel A: Italian									
Grade	0.076	(0.050)	[0.078]	0.051	(0.055)	0.076	(0.080)	6.49	1057
Fail	-0.017	(0.017)	[0.171]	-0.013	(0.018)	-0.013	(0.025)	0.08	1057
Std Test Score Endline	0.174	(0.078)	[0.033]	0.142	(0.083)	0.115	(0.118)	0.00	712
Std Test Score Follow-up	0.105	(0.096)	[0.160]	0.051	(0.103)	0.201	(0.149)	-0.00	342
Panel B: English									
Grade	0.004	(0.059)	[0.419]	0.002	(0.063)	-0.010	(0.100)	6.55	1058
Fail	-0.013	(0.019)	[0.237]	-0.008	(0.020)	-0.018	(0.032)	0.12	1058
Std Test Score Endline	0.195	(0.086)	[0.033]	0.093	(0.091)	0.339	(0.128)	0.00	515
Std Test Score Follow-up	0.050	(0.092)	[0.244]	0.009	(0.099)	0.149	(0.152)	-0.00	343
Panel C: Beliefs									
Students									
Beliefs on math test score	0.034	(0.014)	[0.030]	0.028	(0.015)	0.020	(0.021)	0.66	704
Beliefs on all test score	0.033	(0.012)	[0.020]	0.030	(0.012)	0.011	(0.018)	0.65	705
Beliefs on grade	0.292	(0.104)	[0.020]	0.324	(0.110)	-0.105	(0.157)	6.15	680
Parents									
Beliefs on math test score	0.030	(0.013)	[0.033]	0.029	(0.014)	0.003	(0.019)	0.67	746
Beliefs on all test score	0.024	(0.010)	[0.030]	0.194	(0.011)	-0.007	(0.015)	0.69	756
Teachers									
Beliefs on math test score	0.236	(0.097)	[0.030]	0.047	(0.104)	0.140	(0.144)	4.61	704
Beliefs on all test score	0.045	(0.014)	[0.012]	0.047	(0.014)	-0.010	(0.022)	0.48	792
Beliefs on grade	0.045	(0.123)	[0.012]	0.360	(0.129)	0.074	(0.214)	5.49	792

Notes: OLS estimates, robust standard errors in parentheses, and Anderson sharpened q-values in square brackets. The dependent variable is defined in the left column. "Treatment" is an indicator for being assigned a tutor; "Intense" is an indicator for being assigned to 6 hours of tutoring. Columns (1) and (2) report the coefficients and standard errors of our main specification. Columns (4), (5), (6), and (7) report the coefficients and standard errors of the regression model including both "Treatment" and "Intense treatment" as independent variables, controlling for whether the students was identified for intense tutoring (coefficient not shown). Controls included in all regressions: parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline. "Mean Dep" is the mean of the dependent variable at endline for students in the control group.

Table A.14. Estimation of the impact of TOP 2022 on academic outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Regression Model 1			Regression Model 2					
	Treatment	SE	q-value	Treatment	SE	Group	SE	Mean Dep	Obs
Outcomes									
Panel A: Italian									
Grade	0.021	(0.045)	[0.861]	0.013	(0.054)	0.019	(0.066)	6.36	1121
Fail	0.001	(0.019)	[1.000]	0.003	(0.023)	-0.005	(0.028)	0.10	1121
Std Test Score Endline	0.043	(0.065)	[0.807]	0.146	(0.076)	-0.244	(0.094)	-0.00	952
Panel B: English									
Grade	0.090	(0.060)	[0.416]	0.079	(0.069)	0.025	(0.087)	6.44	1121
Fail	-0.060	(0.021)	[0.031]	-0.061	(0.023)	0.003	(0.027)	0.16	1121
Std Test Score Endline	0.083	(0.062)	[0.450]	0.122	(0.074)	-0.092	(0.087)	0.00	951
Panel C: Beliefs									
Students									
Belief on math test score	0.008	(0.013)	[0.807]	0.016	(0.016)	-0.019	(0.019)	0.64	946
Belief on own test score	0.006	(0.010)	[0.807]	0.018	(0.012)	-0.028	(0.015)	0.66	945
Belief on grade	-0.025	(0.110)	[1.000]	-0.090	(0.131)	0.157	(0.164)	5.54	955
Parents									
Belief on math test score	0.013	(0.013)	[0.717]	0.013	(0.015)	-0.001	(0.018)	0.63	998
Belief on test score	0.001	(0.009)	[1.000]	-0.004	(0.010)	0.011	(0.013)	0.69	995
Teachers									
Belief on math test score	0.040	(0.014)	[0.031]	0.046	(0.016)	-0.013	(0.020)	0.48	821
Belief on test score	0.015	(0.009)	[0.322]	0.014	(0.010)	0.003	(0.013)	0.53	1068
Belief on grade	0.179	(0.086)	[0.180]	0.222	(0.101)	-0.102	(0.122)	5.43	1081

Notes: OLS estimates, robust standard errors in parentheses, and Anderson sharpened q-values in square brackets. The dependent variable is defined in the left column. "Treatment" is an indicator for being assigned a tutor; "Group" is an indicator for being assigned to group tutoring. Columns (1) and (2) report the coefficients and standard errors of our main specification. Columns (4), (5), (6), and (7) report the coefficients and standard errors of the regression model including both "Treatment" and "Group" as independent variables. Controls included in all regressions: parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline. "Mean Dep" is the mean of the dependent variable at endline for students in the control group.

Table A.15. Estimation of the impact of TOP2020 and TOP2022 on instruction time

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Regression Model 1			Regression Model 2					
Outcomes									
Panel A: TOP2020	Treatment	SE	q-value	Treatment	SE	Intense	SE	Mean Dep	Obs
Homework time									
Minutes per day reported by students	9.968	(3.491)	[0.004]	6.794	(3.745)	12.190	(4.998)	88.26	690
Minutes per day reported by parents	8.672	(2.782)	[0.003]	5.952	(3.006)	10.941	(4.030)	81.71	778
Always homework reported by teachers	0.106	(0.032)	[0.003]	0.102	(0.034)	-0.003	(0.052)	0.28	851
Online classes									
Always follows reported by students	0.023	(0.029)	[0.203]	0.032	(0.031)	-0.031	(0.043)	0.83	687
Always follows reported by parents	-0.003	(0.024)	[0.432]	0.006	(0.025)	-0.031	(0.038)	0.88	777
Always follows reported by teachers	0.108	(0.032)	[0.003]	0.099	(0.035)	0.030	(0.050)	0.57	859
Panel B: TOP2022	Treatment	SE	q-value	Treatment	SE	Group	SE	Mean Dep	Obs
Homework time									
Minutes per day reported by students	4.213	(2.926)	[0.291]	4.343	(3.398)	-0.313	(4.048)	80.60	967
Minutes per day reported by parents	3.418	(2.988)	[0.291]	3.593	(3.447)	-0.416	(4.123)	83.43	1004
Always homework reported by teachers	-0.041	(0.024)	[0.291]	-0.028	(0.028)	-0.031	(0.032)	0.20	1081

Notes: OLS estimates, robust standard errors in parentheses, and Anderson sharpened q-values in square brackets. The dependent variable is defined in the left column. "Treatment" is an indicator for being assigned a tutor; "Intense" is an indicator for being assigned to 6 hours of tutoring. "Group" is an indicator for being assigned to group tutoring. Columns (1) and (2) report the coefficients and standard errors of our main specification. In Panel A, columns (4), (5), (6), and (7) report the coefficients and standard errors of the regression model including both "Treatment" and "Intense treatment" as independent variables for TOP2020, controlling for whether the students was identified for intense tutoring (coefficient not shown). In Panel B, columns (4), (5), (6), and (7) report the coefficients and standard errors of the regression model including both "Treatment" and 'Group" as independent variables for TOP2022. Controls included in all regressions: parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline. "Mean Dep" is the mean of the dependent variable at endline for students in the control group.

Table A.16. Estimation of the impact of Intense Tutoring in TOP2020 on instruction time

	(1)	(2)	(3)	(4)
	Regression Model			
	Intense	SE	Mean Dep	Obs
Tutoring time (from tutor registry)				
Share of sessions that took place as planned	0.002	(0.016)	0.911	504
Session duration (in minutes)	15.776	(2.097)	67.936	502
Share of sessions planned after the end of school day	-0.029	(0.021)	0.874	502
Share of sessions planned on weekends	0.023	(0.012)	0.067	502
Number of sessions planned	4.317	(0.687)	14.598	504
Number of sessions	3.669	(0.631)	13.321	504
Total time of meetings (in minutes)	511.333	(50.662)	882.451	504
Coverage of meetings (in days)	0.237	(0.895)	36.361	502
Number of sessions per week	0.664	(0.093)	2.345	502

Notes: OLS estimates, robust standard errors in parentheses. The dependent variable is defined in the left column. "Intense" is an indicator for being assigned to 6 hours of tutoring. Columns (1) and (2) report the coefficients and standard errors of the regression model including "Intense treatment" as independent variable for TOP2020, controlling for whether the students was identified for intense tutoring (coefficient not shown). Controls included in all regressions: parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline. "Mean Dep" is the mean of the dependent variable at endline for students in the 3-hours treatment group.

Table A.17. Estimation of the impact of TOP 2020 on non academic outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Regression Model 1			Regression Model 2					
	Treatment	SE	q-value	Treatment	SE	Intense	SE	Mean Dep	Obs
Outcomes									
Panel A: Aspirations									
Students									
Aspirations University	0.062	(0.037)	[0.230]	0.033	(0.039)	0.109	(0.055)	0.36	674
Self-efficacy university	0.040	(0.032)	[0.230]	0.053	(0.036)	-0.047	(0.047)	0.21	682
High-school: vocational	-0.058	(0.034)	[0.230]	-0.047	(0.036)	-0.039	(0.049)	0.30	681
High-school: top tier	-0.000	(0.027)	[0.332]	-0.016	(0.028)	0.059	(0.040)	0.16	681
Parents									
Aspirations University	0.042	(0.033)	[0.230]	0.034	(0.036)	0.033	(0.050)	0.34	765
Self-efficacy university	0.080	(0.032)	[0.106]	0.081	(0.035)	0.000	(0.048)	0.29	772
Teachers									
Aspirations University	0.029	(0.022)	[0.230]	0.020	(0.024)	0.021	(0.037)	0.14	839
Panel B: Socio-emotional skills									
Students									
Logic task: difficult	0.048	(0.037)	[0.230]	0.007	(0.040)	0.153	(0.054)	0.56	685
Logic task: give-up	-0.034	(0.026)	[0.230]	-0.038	(0.027)	0.018	(0.037)	0.14	685
Grit	0.015	(0.010)	[0.230]	0.016	(0.011)	-0.004	(0.015)	0.68	673
Locus of control	0.023	(0.009)	[0.106]	0.022	(0.009)	0.004	(0.012)	0.71	685
Parents									
Grit	-0.006	(0.010)	[0.332]	-0.002	(0.010)	-0.015	(0.015)	0.67	736
Panel C: Psychological well-being									
Students									
Depression	-0.018	(0.009)	[0.182]	-0.016	(0.010)	-0.008	(0.014)	0.55	669
Happiness	0.025	(0.017)	[0.230]	0.023	(0.018)	0.005	(0.026)	0.61	665
Parents									
Depression	-0.011	(0.008)	[0.230]	-0.009	(0.008)	-0.008	(0.012)	0.59	731
Happiness	0.035	(0.016)	[0.156]	0.032	(0.017)	0.011	(0.022)	0.60	741

Notes: OLS estimates, robust standard errors in parentheses, and Anderson sharpened q-values in square brackets. The dependent variable is defined in the left column. "Treatment" is an indicator for being assigned a tutor; "Intense treatment" is an indicator for being assigned to 6 hours of tutoring. Columns (1) and (2) report the coefficients and standard errors of our baseline specification. Columns (4), (5), (6), and (7) report the coefficients and standard errors of the regression model including both "Treatment" and "Intense treatment" as independent variables, controlling for whether the students was identified for intense tutoring (coefficient not shown). Controls included in all regressions: parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline. "Mean Dep" is the mean of the dependent variable at endline for students in the control group.

Table A.18. Estimation of the impact of TOP 2022 on other non academic outcomes

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
	Regression Model 1			Regression Model 2					
	Treatment	SE	q-value	Treatment	SE	Group	SE	Mean Dep	Obs
Outcomes									
Panel A: Aspirations									
Students									
Aspirations university	0.022	(0.030)	[1.000]	0.004	(0.035)	0.042	(0.042)	0.31	961
Self-efficacy university	0.000	(0.029)	[1.000]	-0.001	(0.034)	0.003	(0.040)	0.26	965
High-school: vocational	-0.031	(0.027)	[1.000]	-0.023	(0.031)	-0.018	(0.038)	0.24	966
High-school: top tier	0.038	(0.023)	[1.000]	0.032	(0.027)	0.012	(0.034)	0.13	966
Parents									
Aspirations university	0.031	(0.028)	[1.000]	0.028	(0.033)	0.008	(0.039)	0.27	1000
Self-efficacy university	-0.002	(0.030)	[1.000]	0.002	(0.035)	-0.009	(0.043)	0.36	1002
Teachers									
Aspirations university	-0.007	(0.014)	[1.000]	-0.018	(0.016)	0.025	(0.019)	0.06	1076
Panel B: Socio-emotional skills									
Students									
Logic task: difficult	-0.011	(0.034)	[1.000]	-0.004	(0.039)	-0.017	(0.047)	0.56	967
Logic task: give-up	0.007	(0.023)	[1.000]	0.002	(0.026)	0.011	(0.031)	0.13	967
Grit	-0.009	(0.009)	[1.000]	-0.016	(0.010)	0.018	(0.013)	0.66	958
Locus of control	0.007	(0.008)	[1.000]	0.009	(0.009)	-0.004	(0.011)	0.70	961
Parents									
Grit	-0.008	(0.008)	[1.000]	-0.012	(0.009)	0.008	(0.012)	0.65	955
Panel C: Psychological well-being									
Students									
Depression	-0.006	(0.008)	[1.000]	-0.001	(0.009)	-0.011	(0.011)	0.58	948
Happiness	0.008	(0.014)	[1.000]	0.003	(0.016)	0.013	(0.020)	0.69	955
Parents									
Depression	-0.003	(0.007)	[1.000]	-0.001	(0.008)	-0.005	(0.010)	0.59	968
Happiness	0.001	(0.012)	[1.000]	0.001	(0.014)	0.001	(0.017)	0.68	969

Notes: OLS estimates, robust standard errors in parentheses, and Anderson sharpened q-values in square brackets. The dependent variable is defined in the left column. "Treatment" is an indicator for being assigned a tutor; "Group" is an indicator for being assigned to group tutoring. Columns (1) and (2) report the coefficients and standard errors of our baseline specification. Columns (4), (5), (6), and (7) report the coefficients and standard errors of the regression model including both "Treatment" and 'Group" as independent variables. Controls included in all regressions: parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline. "Mean Dep" is the mean of the dependent variable at endline for students in the control group.

Table A.19. 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
Tutors	0.148	0.122	0.337	-0.259	0.486	-0.194
	(0.196)	(0.200)	(0.232)	(0.222)	(0.239)	(0.229)
Mean Dep:	4.46	3.47	1.81	2.98	3.15	2.79
Obs	738	741	737	737	739	739

Notes: Coefficients from ordered logit regressions. 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" is the mean of the dependent variable for students in the control group.

B Data Appendix

B.1 Description of outcome variables

Test score. 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. We use the same procedure in TOP 2020 and TOP 2022 to collect test scores from all students in the sample in grade 6, 7, and 8.

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. The assessment covered a wide range of competencies and very few students reached a ceiling in terms of correct answers.

The variable used in the paper is constructing by standardizing the average number of correct answers to have mean 0 and standard deviation 1 in the control group for each edition of the program. Some examples of questions are reported in Appendix B.2.

Performance, aspiration, socio-emotional and well-being indexes. For each edition of TOP, we take the first principal component to reduce the dimensionality and preserve the maximum amount of information. For the performance index, we include in polychoric principal component analysis (PCA) the teacher-assigned grade, failure rate, and standardized test score. The list of questions used for the other indexes is available in Appendix B.3, B.4, and B.5. After the PCA, we standardize the outcome to have mean 0 and standard deviation 1 in the control group for each edition of the program.

Empathy and hard work indexes. For each index, we calculate a weighted average

of the answer to the questions reported in Appendix B.6. We standardize the outcome to have mean 0 and standard deviation 1 in the control group.

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.5 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
- **Tutoring experience and satisfaction:** we included few questions only for treated students.

B.6 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)

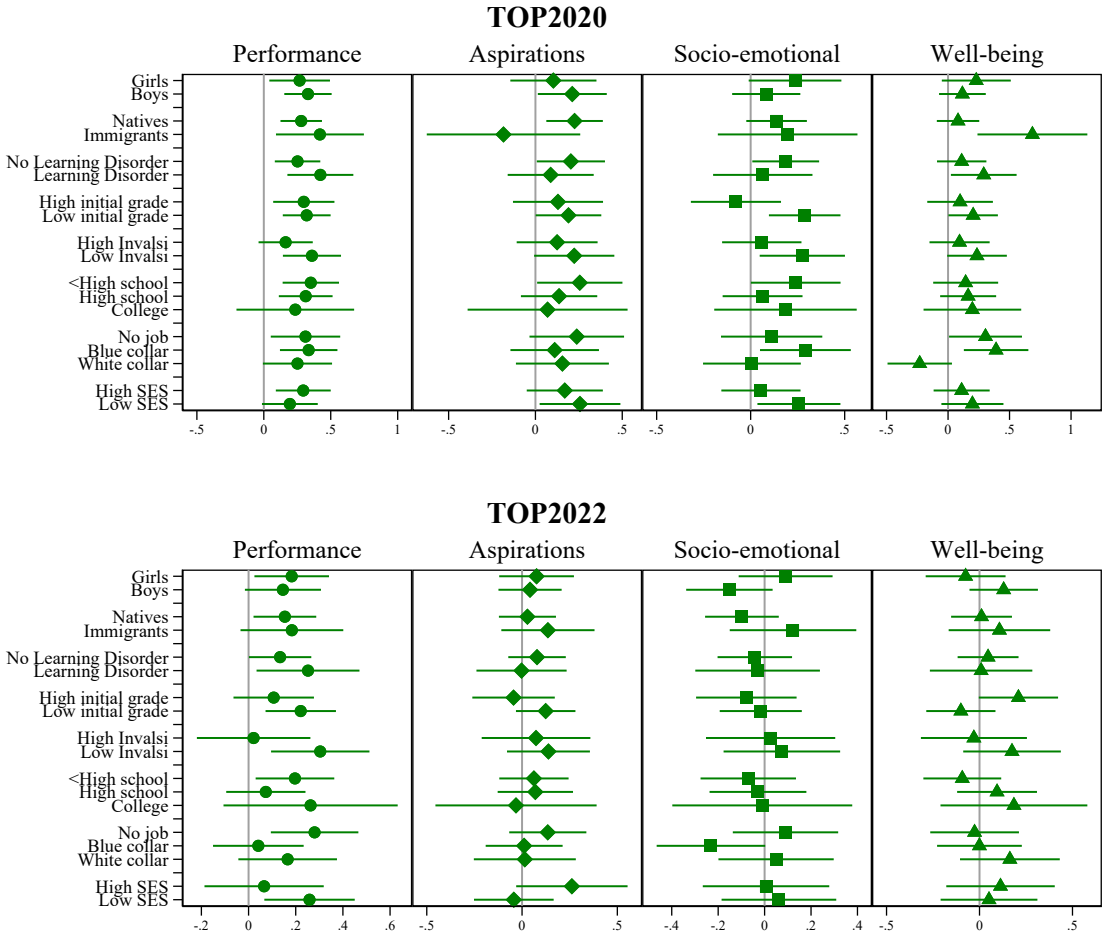
- **Tutoring experience and satisfaction:** we included few questions only for treated tutors.

C Heterogeneous treatment effects

C.1 Results on heterogeneous treatment effects

In this Appendix, we discuss the heterogeneity of impacts depending on the characteristics of students and tutors.

Figure C.1. Heterogeneity by student characteristics



Notes: This Figure reports OLS estimates and 95% confidence intervals of the impact of TOP tutoring by student characteristics for TOP 2020 in the top panel and TOP 2022 in the bottom panel. Randomization round fixed effects included in all regressions. Controls included are parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline. Each index is standardized to have mean 0 and the standard deviation is 1 in the control group.

Student Characteristics. An important dimension of heterogeneity that may help us to understand the mechanisms through which tutoring affects students' outcomes per-

tains to student demographics and socioeconomic background. In Figure C.1 we show the impact of TOP on our aggregate indexes of math performance, aspirations, socio-emotional skills, and psychological well-being for different sub-groups of students, split according to predetermined characteristics: sex, immigrant status, learning disorders, baseline performance in teacher-assigned grades, baseline performance in standardized test scores (Invalsi), mother's education and occupation, and socio-economic status of the household.[45] For each of the four outcomes, the figure shows the estimated impact (relative to the control group) and the associated 95 percent confidence interval.

While the differences across subgroups are not statistically significant in most cases, it appears that the gains in academic performance in both TOP 2020 and TOP 2022 are somewhat more concentrated among students from a disadvantaged background, i.e., immigrants, students with learning disorders, students with lower initial performance, and with lower SES. These are the students who may receive the least support from parents in terms of help with homework, and for whom regular meetings with a tutor may make more of a difference. We observe a similar pattern for socio-emotional skills, which improved relatively more for students with lower initial performance and lower socio-economic status.

When we look at aspirations, the treatment effect appears to be similar across all subgroups – the only result worth mentioning is that in 2020 aspirations increase for natives but not for immigrants, possibly because the latter face different types of barriers when planning their future education, which may have been exacerbated by school closures (Carlana et al., 2022a).

The outcome for which the heterogeneity in treatment effects is most striking is psychological well-being. When we compare native and immigrant students, it is clear that the increased happiness and reduced depression we detect in TOP 2020 is entirely driven by immigrant students. The magnitude of the effect for this group is a striking 0.69 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. In fact, among students in the control group, immigrants have on average a 0.42 SD lower well-being compared to natives in TOP 2020 (p-value =0.03). For TOP 2022, we find a small positive effect on the well-being of immigrant students, although post-pandemic there are no significant differences in the level among immigrant and non-immigrant students in the control group (the difference is +0.11 with a p-value= 0.40).

[45]The socio-economic status (SES) index is constructed by Invalsi considering parental education, occupation, and ownership of relevant resources, such as books, desk, computer, internet connection, and having a quite space to study and a own bedroom.

Table C.I. CLAN of Performance Index

	(1)	(2)	(3)	(4)	(5)	(6)
		TOP 2020			**TOP 2022**	
Variable	Upper quartile	Lower quartile	p-value difference	Upper quartile	Lower quartile	p-value difference
Female	0.385	0.406	0.854	0.479	0.445	0.791
	[0.291,0.484]	[0.302,0.511]		[0.393,0.574]	[0.359,0.539]	
Immigrant	0.191	0.179	0.902	0.39	0.178	0
	[0.112,0.271]	[0.096,0.261]		[0.301,0.478]	[0.109,0.247]	
Learning Disorder	0.32	0.327	0.976	0.251	0.231	0.677
	[0.228,0.415]	[0.227,0.428]		[0.175,0.333]	[0.155,0.307]	
School grade 6	0.335	0.297	0.638	0.318	0.313	0.742
	[0.24,0.431]	[0.202,0.396]		[0.237,0.407]	[0.229,0.397]	
School grade 7	0.361	0.3	0.443	0.364	0.373	0.792
	[0.264,0.46]	[0.199,0.4]		[0.277,0.452]	[0.285,0.46]	
Std INVALSI Math 5	-0.072	0.19	0.042	-0.162	0.004	0.062
	[-0.279,0.117]	[-0.026,0.411]		[-0.256,-0.069]	[-0.124,0.132]	
SES Status	-0.189	-0.329	0.123	-0.454	-0.312	0.064
	[-0.353,-0.042]	[-0.507,-0.136]		[-0.553,-0.345]	[-0.442,-0.191]	
Baseline math grade	6.215	6.081	0.073	6.386	6.449	0.557
	[6.047,6.411]	[5.839,6.302]		[6.203,6.586]	[6.241,6.672]	
Single parent household	0.218	0.236	0.651	0.36	0.145	0
	[0.136,0.301]	[0.143,0.33]		[0.277,0.452]	[0.081,0.209]	
Mother Education: High-school	0.501	0.437	0.381	0.349	0.477	0.056
	[0.403,0.6]	[0.33,0.544]		[0.264,0.435]	[0.39,0.567]	
Mother Education: University	0.127	0.123	0.85	0.144	0.138	0.964
	[0.063,0.195]	[0.053,0.195]		[0.081,0.208]	[0.079,0.203]	
Father Education: High-school	0.412	0.351	0.547	0.306	0.492	0.002
	[0.315,0.51]	[0.252,0.451]		[0.224,0.389]	[0.403,0.582]	
Father Education: University	0.069	0.083	0.78	0.062	0.079	0.553
	[0.017,0.121]	[0.026,0.141]		[0.02,0.105]	[0.031,0.127]	
Mother Occupation: White-collar	0.414	0.421	0.77	0.424	0.376	0.583
	[0.332,0.497]	[0.334,0.509]		[0.353,0.499]	[0.306,0.451]	
Mother Occupation: Blue-collar	0.534	0.538	0.847	0.496	0.578	0.257
	[0.45,0.617]	[0.453,0.626]		[0.42,0.566]	[0.511,0.657]	
Father Occupation: White-collar	0.325	0.334	0.914	0.273	0.377	0.043
	[0.238,0.414]	[0.242,0.424]		[0.208,0.342]	[0.294,0.462]	
Father Occupation: Blue-collar	0.638	0.628	0.827	0.67	0.566	0.098
	[0.548,0.727]	[0.541,0.724]		[0.6,0.74]	[0.481,0.653]	

Notes: The table reports the median CLAN estimates for all covariates over 100 splits. The numbers in columns 1 and 4 (2 and 5) represent the share of individuals with a given characteristic among those belonging to the top (bottom) quartile in terms of impact of the treatment. 90 percent confidence intervals are reported in square brackets. Columns 3 and 6 report the p-value for the hypothesis that the difference between columns (1) and (2) -or between (4) and (5), respectively- is zero.

To complement the above analysis with a more systematic approach, we estimate heterogeneous treatment effects using generic machine learning inference. We follow Chernozhukov et al. (2018) and and apply their method to understand who benefits the most from tutoring. Section C.2 of this online appendix describes our methodology in more detail.

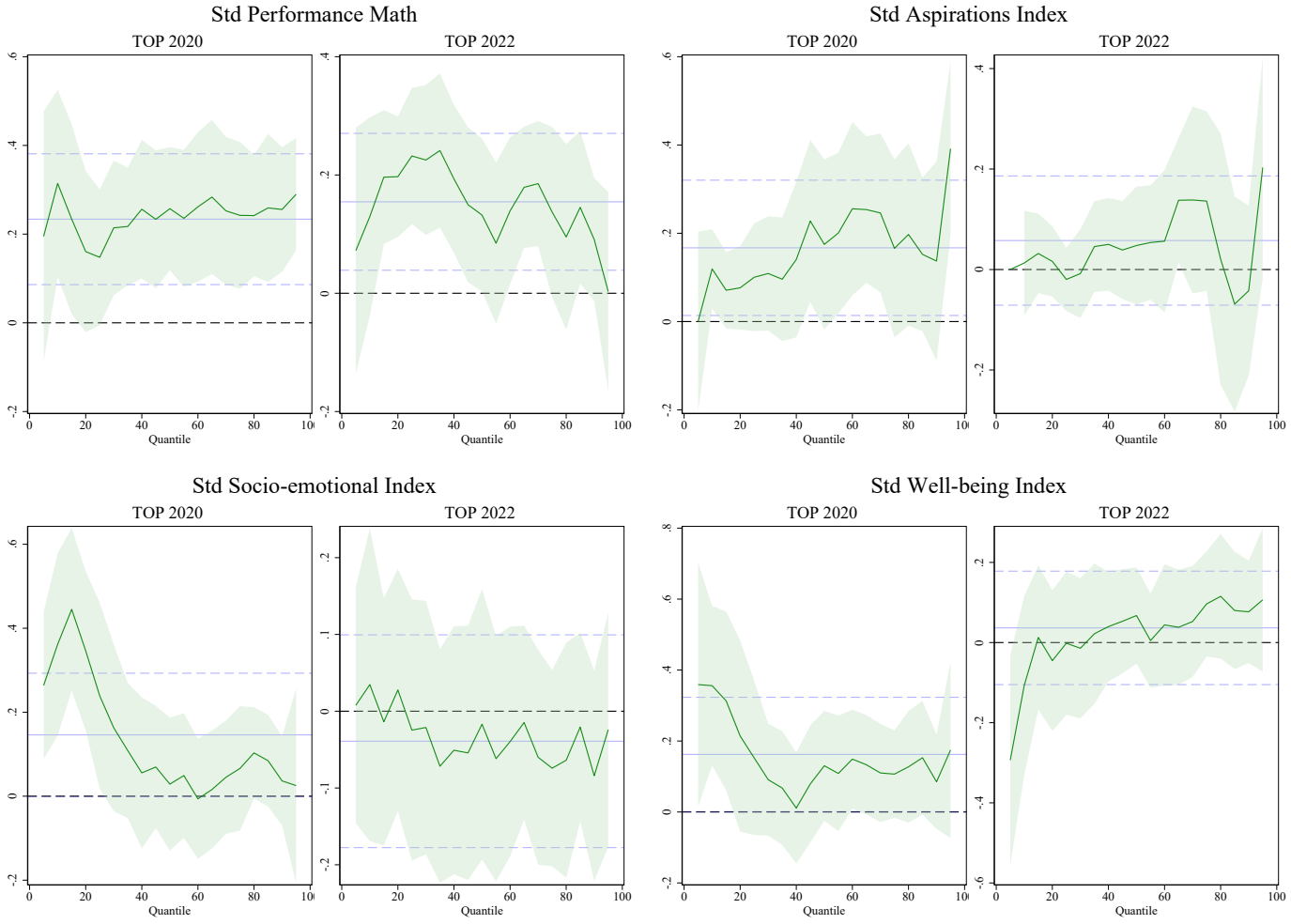
In a nutshell, we do not find strong evidence of heterogeneity in treatment effects. This can be seen by observing the coefficient β_2 in the Best Linear Predictor (BLP) of the Conditional Average Treatment Effect (CATE) in Figure C.4. Nonetheless, to investigate treatment heterogeneity, we also check the predictions on the expected treatment effect for each individual, given the covariates. Table C.I reports the mean of each baseline

characteristic for the students in the top and bottom quartile of predicted impact on math performance. Overall, the results are consistent with our previous discussion of Figure C.1. Students who have lower initial standardized test scores (Invalsi) and are from a more disadvantaged background are over-represented among the students with the highest predicted impact on performance, but most differences are not statistically significant. Tables C.V, C.VI, and C.VII report a similar exercise for the other three outcome indexes. The results show more positive effects for students from lower socio-economic status and with higher initial learning difficulties of all these three 'soft' dimensions in TOP 2020, with particularly striking results for the psychological well-being of immigrants thanks to the support of the tutor.

Finally, Figure C.2 reports the results of the quantile regressions. The pattern is quite interesting: while the impact on academic performance is similar across quantiles in both waves of program implementation, the positive results on socio-emotional and well-being indexes in TOP2020 are driven by the lower quantiles.

Overall, the above results qualitatively suggest that the most disadvantaged children seem to have benefited the most from the tutoring. However, heterogeneity based on parents' or students' characteristics is not stark. To understand why this may be the case, 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 (that is, from a population that may have overall been similarly disadvantaged). This could account for the low degree of heterogeneity in treatment effects.

Figure C.2. Quantile Regression



Notes: These Figures shows the results of the quantile regression (in green) for each of the main outcomes and editions of the program. Results from a least squares regression with 95% confidence intervals are shown in blue.

Tutor characteristics and match with student characteristics. Next, we investigate whether tutors' characteristics played a significant role in explaining the effects of the program. In Figure C.3 we explore a sets of tutor baseline characteristics: sex, education level of the mother (as a proxy for socioeconomic status), academic performance, and pro-social attitudes. For each of these characteristics, we report the treatment effect on students' outcomes, separately by subgroup of tutor characteristic, as well as the associated 95 percent confidence interval.

The figure shows that male and female tutors are on average equally effective in improving student outcomes. Table C.II further investigates this issue by testing for differential effects depending on combinations of sex of the tutor and of the student. We fail to detect statistically significant pairwise differences: mixed-sex pairs perform slightly better than same-sex ones, but the difference is insignificant, as shown by the p-values in square brackets.

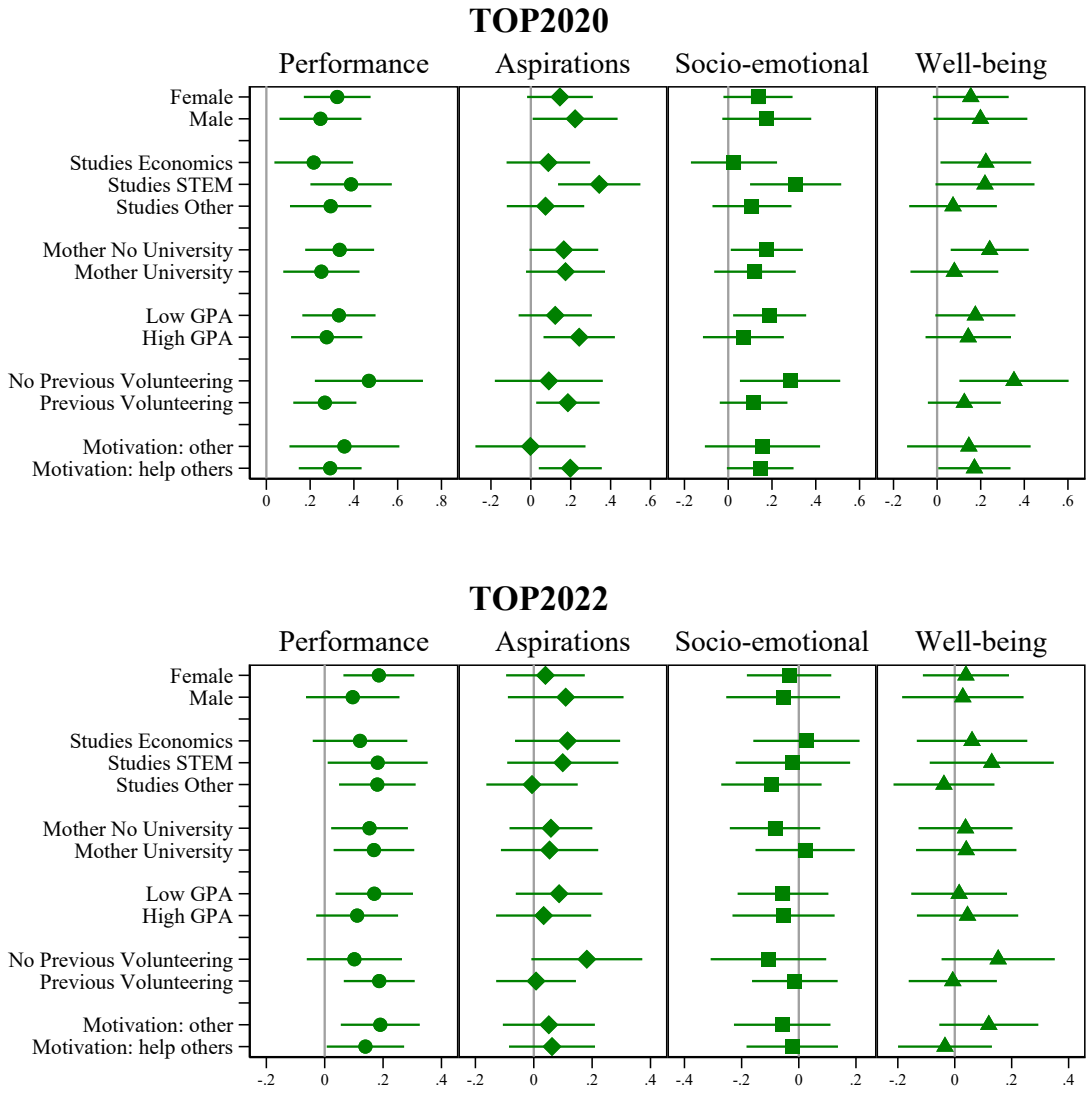
Similarly, in Figure C.3 we do not detect statistically significant differences by tutors' socio-economic status, as proxied by mother's education. Table C.III tests for differential impact by tutor-student socio-economic status match. Overall, the results are not statistically significant with a few notable exceptions. Students with highly educated mothers did not benefit in their socio-emotional development from the interaction with a tutor from low socio-economic background.[46]

In the remaining panels of Figure C.3, we find that tutors' GPA did not significantly affect the impact of the program: treated students benefited equally from interacting with a tutor above or below the median GPA in their faculty.[47] Finally, we capture tutors' pro-social attitudes and motivation. We compare the impact of tutors with and without previous volunteering experience and of tutors who, when asked at baseline what motivated them to participate in TOP, replied that it was to make themselves useful (variable 'Help others'). Note that our tutors are generally highly pro-social: 80 percent had previous experience as a volunteer (82 percent in TOP 2020 and 72 percent 2022), while 70 percent joined TOP to be useful to others (83 percent in TOP 2020 and 56 percent in TOP 2022). For this reason, it is not too surprising that we do not detect significant differences in the outcomes of students who were assigned different types of tutors. The one outcome in which tutors' motivation seems to make a difference is aspirations in TOP 2020, where the positive impact is entirely driven by the more pro-social tutors.

[46]The positive impact on aspirations and well-being in TOP 2020 was driven by low SES students, but not differentially so when matched to a high- vs. a low-SES tutor.

[47]We standardize the GPA within faculty to account for potential differences in grading criteria, number of credits, etc., across programs.

Figure C.3. Heterogeneity by tutor characteristics



Notes: This Figure reports OLS estimates of the assignment to the TOP tutoring treatment by tutor characteristics for TOP 2020 students in the top panel and TOP 2022 students in the bottom panel. Randomization round fixed effects included in all regressions. Student baseline controls include gender, immigrant, grade, education of each parent, employment type of each parent, and learning disorder. The mean of the control group for each index is 0 and the standard deviation is 1. The bar shows 95% confidence intervals.

Table C.II. Heterogeneity by gender match between tutor and tutee

		(1)	(2)	(3)	(4)
		Performance	Aspirations	Socio-emotional	Wellbeing
Panel A: TOP 2020					
Tutor	**Student**				
Female	Female	0.251	0.049	0.179	0.159
		(0.122)	(0.132)	(0.134)	(0.150))
Male	Male	0.209	0.207	0.037	0.058
		(0.119)	(0.134)	(0.130)	(0.126)
Male	Female	0.307	0.240	0.390	0.427
		(0.155)	(0.180)	(0.169)	(0.199)
Female	Male	0.380	0.215	0.106	0.153
		(0.100)	(0.111)	(0.099)	(0.107)
P-value diff (F, F)-(M, M)		[0.806]	[0.408]	[0.442]	[0.602]
P-value diff (F, F)-(M, F)		[0.699]	[0.254]	[0.192]	[0.147]
P-value diff (F, F)-(F, M)		[0.414]	[0.343]	[0.662]	[0.974]
Observations		709	520	633	611
R^2		0.305	0.333	0.180	0.080
Panel B: TOP 2022					
Tutor	**Student**				
Female	Female	0.194	0.079	0.085	-0.080
		(0.088)	(0.107)	(0.111)	(0.118))
Male	Male	0.048	0.142	-0.190	0.104
		(0.122)	(0.139)	(0.131)	(0.137)
Male	Female	0.152	0.070	0.105	-0.061
		(0.104)	(0.142)	(0.155)	(0.174)
Female	Male	0.178	0.009	-0.138	0.139
		(0.087)	(0.088)	(0.102)	(0.100)
P-value diff (F, F)-(M, M)		[0.337]	[0.720]	[0.105]	[0.304]
P-value diff (F, F)-(M, F)		[0.684]	[0.948]	[0.897]	[0.913]
P-value diff (F, F)-(F, M)		[0.902]	[0.609]	[0.138]	[0.152]
Observations		943	889	898	881
R^2		0.238	0.221	0.063	0.077

Notes: OLS estimates, robust standard errors in parentheses, and p-values in square brackets. The dependent variable is the performance index column 1, the aspiration index in column 2, the socio-emotional index in column 3, and the well-being index in column 4. The first column indicates the gender of the tutor, while the second one the gender of the student. The p-value of the difference between groups is presented at the end of each panel: the first letter refers to the gender of the tutor, while the second letter to the gender of the student (M for male and F for female). For example, (F, M) refers to a female tutor assigned to a male student. Panel A presents the results of TOP 2020, while Panel B of TOP 2022. Controls included in all regressions: parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline.

Table C.III. Heterogeneity by SES match between tutor and tutee

		(1)	(2)	(3)	(4)
		Performance	Aspirations	Socio-emotional	Wellbeing
Panel A: TOP 2020					
Mothers' Education					
Tutor	**Student**				
Low	Low	0.379	0.213	0.209	0.167
		(0.089)	(0.106)	(0.097)	(0.106)
High	High	0.164	-0.068	0.281	0.127
		(0.215)	(0.268)	(0.209)	(0.208)
High	Low	0.278	0.203	0.096	0.163
		(0.086)	(0.091)	(0.095)	(0.105)
Low	High	0.161	-0.035	-0.034	0.176
		(0.278)	(0.235)	(0.176)	(0.210)
P-value diff (L, L)-(H, H)		[0.351]	[0.328]	[0.751]	[0.862]
P-value diff (L, L)-(H, L)		[0.255]	[0.927]	[0.267]	[0.970]
P-value diff (L, L)-(L, H)		[0.456]	[0.337]	[0.221]	[0.967]
Observations		711	523	636	614
R^2		0.308	0.337	0.183	0.077
Panel B: TOP 2022					
Mothers' Education					
Tutor	**Student**				
Low	Low	0.196	0.023	-0.068	-0.028
		(0.076)	(0.081)	(0.094)	(0.098)
High	High	0.143	-0.116	0.097	0.337
		(0.190)	(0.212)	(0.190)	(0.195)
High	Low	0.108	0.082	-0.012	0.020
		(0.069)	(0.082)	(0.087)	(0.089)
Low	High	0.377	0.263	-0.231	0.166
		(0.182)	(0.219)	(0.220)	(0.230)
P-value diff (L, L)-(H, H)		[0.796]	[0.537]	[0.431]	[0.090]
P-value diff (L, L)-(H, L)		[0.267]	[0.503]	[0.570]	[0.651]
P-value diff (L, L)-(L, H)		[0.352]	[0.300]	[0.492]	[0.438]
Observations		943	889	898	881
R^2		0.239	0.223	0.062	0.077

Notes: OLS estimates, robust standard errors in parentheses, and p-values in square brackets. The dependent variable is the performance index column 1, the aspiration index in column 2, the socio-emotional index in column 3, and the well-being index in column 4. The first column indicates the soio-economic status of the tutor (high vs. low), while the second one the socio-economic status of the student. The p-value of the difference between groups is presented at the end of each panel: the first letter refers to the SES of the tutor, while the second letter to the SES of the student (H for high and L for low). For example, (H, L) refers to a high-SES tutor assigned to a low-SES student. Panel A presents the results of TOP 2020, while Panel B of TOP 2022. Controls included in all regressions: parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline.

C.2 Methodology

We assess the treatment heterogeneity following the generic machine learning (ML) inference approach by Chernozhukov et al. (2018) aimed at detecting heterogeneity in the treatment effects with an agnostic procedure. This method focuses on the Best Linear Predictor (BLP) of the Conditional Average Treatment Effects (CATEs) and the Classification Analysis (CLAN) instead of the CATEs themselves to overcome the issues posed by the sparsity requirements in the causal forest method by Wager and Athey (2018)

The outcome variable Y in our analysis are: performance, aspirations, socio-emotional skills, well-being indexes. The dummy D represents the random allocation to TOP tutoring. The variables Z used in the heterogeneity analysis are the following: gender, immigration status, learning disorders, school grade attended (Grade 6, 7 or 8), baseline standardized test score Invalsi in math, socio-economic status measured by Invalsi, baseline teacher-assigned grades, whether the child lives in a single parent household, parental education and occupation.

Following the steps of the algorithm in Chernozhukov et al. (2018), we implement the analysis as follows.

- **Step 1:** We set the number of splits (S= 100), the significance level ($\alpha = 0.05$), and the propensity scores $p(Z)$, which in our case are calculated directly from the RCT design.

- **Step 2:** we set the proportion of the splits at 50% (half form the auxiliary sample N and the remaining half forms the main sample M). Each split follows these steps:

    1. Each of the selected ML methods is tuned and trained separately. We consider the following ML methods: RIDGE, decision tree and support vector machine to estimate $B(Z)$ and $S(Z)$, proxy predictors of b_0 and s_0, given the following specification:

$$Y = b_0(Z) + Ds_0(Z) + U$$

    2. Estimate the BLP parameters by weighted OLS in the main sample M:

$$Y_i = \widehat{\alpha}' X_{1i} + \widehat{\beta}_1 \left(D_i - p\left(Z_i\right)\right) + \widehat{\beta}_2 \left(D_i - p\left(Z_i\right)\right) \left(S_i - \mathbb{E}_{N,M} S_i\right) + \widehat{\epsilon}_i, \quad i \in M$$

    such that:

$\mathbb{E}_{N,M}\left[w\left(Z_i\right)\widehat{\epsilon}_i X_i\right]=0$ for $X_i=\left[X_{1i}',D_i-p\left(Z_i\right),\left(D_i-p\left(Z_i\right)\right)\left(S_i-\mathbb{E}_{N,M}S_i\right)\right]'$, where $w\left(Z_i\right)=\left\{p\left(Z_i\right)\left(1-p\left(Z_i\right)\right)\right\}^{-1}$.

3. We estimate the GATES parameters by weighted OLS in the main sample M:

$$Y_i=\widehat{\alpha}'X_{1i}+\sum_{k=1}^{K}\widehat{\gamma}_k\cdot\left(D_i-p\left(Z_i\right)\right)\cdot 1\left(S_i\in I_k\right)+\widehat{\nu}_i,\quad i\in M$$

   Where ℓ_k is the (k/K)-quantile of $\{S_i\}_{i\in M}$. Once again, X_i includes the same controls as the ones described in step 2.

4. We estimate the CLAN parameters in the main sample M:

$$\widehat{\delta}_1=\mathbb{E}_{N,M}\left[g\left(Y_i,Z_i\right)\mid S_i\in I_1\right]\quad\text{ and }\quad\widehat{\delta}_K=\mathbb{E}_{N,M}\left[g\left(Y_i,Z_i\right)\mid S_i\in I_K\right]$$

   where $I_k=[\ell_{k-1},\ell_k)$ and ℓ_k is the (k/K)-quantile of $\{S_i\}_{i\in M}$.

5. We compute the two performance measures for the ML methods:

$$\widehat{\Lambda}=\left|\widehat{\beta}_2\right|^2\widehat{Var}(S(Z))\quad\widehat{\bar{\Lambda}}=\frac{1}{K}\sum_{k=1}^{K}\widehat{\gamma}_k^2$$

- **Step 3:** We choose the best ML methods based on the medians of $\widehat{\Lambda}$ and $\widehat{\bar{\Lambda}}$.
- **Step 4:** We compute the estimates, $(1-\alpha)$-level conditional confidence intervals and conditional p-values for all the parameters of interest.
- **Step 5:** We compute the adjusted $(1-2\alpha)$-confidence intervals and adjusted p-values using Variational Estimation and Inference Methods (VEIN). These methods take into consideration the two different sources of sampling uncertainty, that is, (i) the estimation uncertainty regarding our estimated parameters, conditional on the data split; (ii) the uncertainty or 'variation' induced by the data splitting (our split into the auxiliary N and the main M sample).

Following the steps described above of the algorithm in Chernozhukov et al. (2018) for each of the four main outcomes, we select the following learners:

Table C.IV. Learners performance

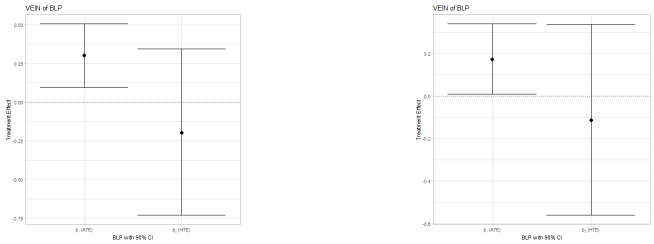
	(1)	(2)	(3)	(4)
	TOP 2020		**TOP 2022**	
Learner	Λ	$\bar{\Lambda}$	Λ	$\bar{\Lambda}$
Performance Index (PPCA)				
RIDGE	0.00236	0.1179	0.00293	0.0509
Decision Tree	0.00534	0.12138	0.00233	0.04753
Support Vector Machine	0.00609	0.12135	0.0035	0.05391
Std Socio-emotional Index				
RIDGE	0.00316	0.04338	0.00239	0.03074
Decision Tree	0.01095	0.05279	0.00607	0.03273
Support Vector Machine	0.00271	0.04652	0.01181	0.03837
Std Well-being Index				
RIDGE	0.01709	0.07202	0.00291	0.03206
Decision Tree	0.0082	0.06032	0.00371	0.03258
Support Vector Machine	0.01254	0.06343	0.00356	0.02916
Std Aspiration Index				
RIDGE	0.00326	0.05903	0.00431	0.0269
Decision Tree	0.00803	0.05655	0.00492	0.02721
Support Vector Machine	0.00813	0.05961	0.01221	0.03998

The BLP results are provided below in Figures C.4. The estimate for the mean prediction $\hat{\beta}_1$ (corresponding to the Average Treatment Effect) match the estimated coefficient of our main analysis, and is statistically significant at the 5 percent level for the performance index in both TOP 2020 and TOP 2022. However, there no clear evidence of heterogeneity in the treatment effects for all outcomes in both waves: none of the coefficients for the differential prediction $\hat{\beta}_2$ is statistically significant.

An analysis of the CLAN results in Tables C.I, C.V, C.VI, and C.VII shows some consistency, with significant differences in the characteristics of the individuals most affected by the intervention as discussed in the main text.

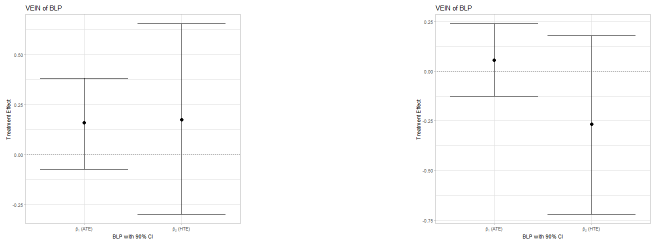
Figure C.4. BLP results for TOP

TOP 2020 TOP 2022



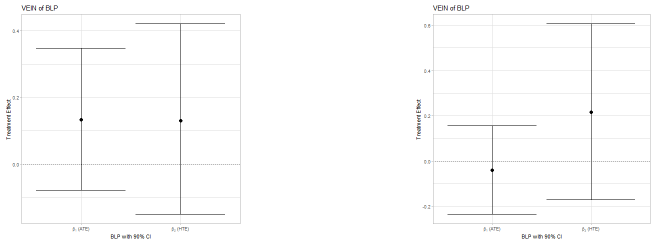
Performance Index (PPCA)

TOP 2020 TOP 2022



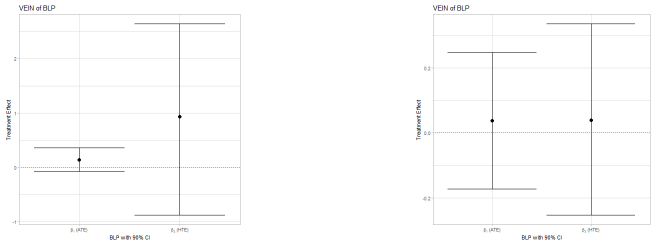
Aspiration Index

TOP 2020 TOP 2022



Socio-emotional Index

TOP 2020 TOP 2022



Well-being Index

Notes: This figure shows the Best Linear Predictor (BLP) of the CATE on the left for TOP 2020 and on the right for TOP 2022 and the GATES analysis of the Generic Machine Learning algorithm on the right for the four main indexes: performance, aspirations, socio-emotional skills, and psychological well-being. In the BLP graphs, $\beta_1(ATE)$ depicts the estimated Average Treatment Effect in the BLP analysis, while $\beta_2(HTE)$ depicts the estimated differential effect. In the GATES graphs, the sample is divided in quartiles based on their heterogeneous treatment effect scores, and the estimated treatment effect is reported for each of the 4 subgroups, including its 95 percent confidence interval.

Table C.V. CLAN of Aspiration Index

	(1)	(2)	(3)	(4)	(5)	(6)
		TOP 2020			**TOP 2022**	
Variable	Upper quartile	Lower quartile	p-value difference	Upper quartile	Lower quartile	p-value difference
Female	0.394	0.47	0.488	0.5	0.415	0.2
	[0.275,0.513]	[0.348,0.591]		[0.407,0.593]	[0.322,0.506]	
Immigrant	0.083	0.176	0.17	0.313	0.261	0.238
	[0.021,0.161]	[0.088,0.276]		[0.226,0.399]	[0.179,0.344]	
Learning Disorder	0.197	0.376	0.018	0.232	0.293	0.208
	[0.1,0.294]	[0.261,0.494]		[0.153,0.311]	[0.212,0.383]	
School grade 6	0.318	0.318	1	0.335	0.342	0.85
	[0.205,0.432]	[0.205,0.432]		[0.251,0.427]	[0.254,0.431]	
School grade 7	0.394	0.288	0.107	0.357	0.356	0.93
	[0.275,0.513]	[0.178,0.398]		[0.268,0.446]	[0.271,0.45]	
Std INVALSI Math 5	0.122	0.07	0.629	-0.15	-0.072	0.366
	[-0.111,0.351]	[-0.14,0.287]		[-0.257,-0.027]	[-0.17,0.042]	
SES Status	-0.178	-0.163	0.885	-0.285	-0.537	0.002
	[-0.404,0.05]	[-0.346,0.027]		[-0.381,-0.17]	[-0.658,-0.416]	
Baseline math grade	6.303	6.228	0.627	6.418	6.577	0.309
	[6.056,6.558]	[5.985,6.479]		[6.214,6.62]	[6.366,6.8]	
Single parent household	0.076	0.394	0	0.241	0.212	0.468
	[0.012,0.14]	[0.275,0.513]		[0.161,0.321]	[0.139,0.293]	
Mother Education: High-school	0.386	0.554	0.031	0.448	0.405	0.511
	[0.275,0.513]	[0.44,0.672]		[0.357,0.54]	[0.314,0.497]	
Mother Education: University	0.155	0.13	0.574	0.136	0.153	0.796
	[0.072,0.245]	[0.053,0.211]		[0.073,0.2]	[0.087,0.22]	
Father Education: High-school	0.496	0.299	0.015	0.483	0.342	0.022
	[0.378,0.62]	[0.191,0.409]		[0.396,0.575]	[0.254,0.431]	
Father Education: University	0.13	0.03	0.033	0.08	0.087	0.938
	[0.053,0.22]	[-0.004,0.072]		[0.03,0.131]	[0.037,0.144]	
Mother Occupation: White-collar	0.437	0.46	0.706	0.423	0.416	0.639
	[0.344,0.529]	[0.358,0.564]		[0.355,0.493]	[0.336,0.496]	
Mother Occupation: Blue-collar	0.542	0.482	0.295	0.494	0.543	0.284
	[0.438,0.632]	[0.374,0.589]		[0.426,0.567]	[0.468,0.621]	
Father Occupation: White-collar	0.474	0.2	0	0.265	0.413	0.001
	[0.359,0.591]	[0.137,0.266]		[0.191,0.34]	[0.328,0.499]	
Father Occupation: Blue-collar	0.505	0.736	0.001	0.675	0.534	0.003
	[0.39,0.622]	[0.658,0.814]		[0.602,0.762]	[0.448,0.621]	

Notes: The table reports the median CLAN estimates for all covariates over 100 splits. The numbers in columns 1 and 4 (2 and 5) represent the share of individuals with a given characteristic among those belonging to the top (bottom) quartile in terms of impact of the treatment. 90 percent confidence intervals are reported in square brackets. Columns 3 and 6 report the p-value for the hypothesis that the difference between columns (1) and (2) -or between (4) and (5), respectively- is zero.

Table C.VI. CLAN of Socio-Emotional Index

	(1)	(2)	(3)	(4)	(5)	(6)
		TOP 2020			**TOP 2022**	
Variable	Upper quartile	Lower quartile	p-value difference	Upper quartile	Lower quartile	p-value difference
Female	0.489	0.359	0.111	0.474	0.42	0.54
	[0.384,0.597]	[0.252,0.469]		[0.385,0.57]	[0.328,0.512]	
Immigrant	0.153	0.125	0.374	0.292	0.237	0.314
	[0.077,0.231]	[0.052,0.198]		[0.208,0.376]	[0.161,0.321]	
Learning Disorder	0.233	0.388	0.03	0.221	0.277	0.291
	[0.144,0.327]	[0.276,0.499]		[0.144,0.298]	[0.194,0.36]	
School grade 6	0.25	0.366	0.117	0.328	0.29	0.621
	[0.157,0.343]	[0.258,0.478]		[0.24,0.414]	[0.21,0.38]	
School grade 7	0.413	0.305	0.142	0.266	0.518	0
	[0.308,0.522]	[0.202,0.411]		[0.184,0.347]	[0.425,0.611]	
Std INVALSI Math 5	-0.135	0.33	0.001	-0.144	-0.109	0.756
	[-0.325,0.054]	[0.105,0.551]		[-0.262,-0.009]	[-0.202,-0.02]	
SES Status	-0.424	0.001	0.001	-0.505	-0.327	0.024
	[-0.606,-0.239]	[-0.182,0.2]		[-0.608,-0.398]	[-0.426,-0.215]	
Baseline math grade	5.881	6.624	0	6.293	6.562	0.097
	[5.708,6.058]	[6.374,6.884]		[6.1,6.515]	[6.364,6.76]	
Single parent household	0.223	0.201	0.842	0.159	0.295	0.017
	[0.135,0.313]	[0.112,0.292]		[0.092,0.227]	[0.21,0.38]	
Mother Education: High-school	0.415	0.541	0.102	0.427	0.393	0.59
	[0.31,0.52]	[0.434,0.652]		[0.337,0.52]	[0.302,0.484]	
Mother Education: University	0.115	0.175	0.296	0.109	0.127	0.437
	[0.051,0.184]	[0.091,0.26]		[0.052,0.167]	[0.067,0.19]	
Father Education: High-school	0.359	0.41	0.56	0.548	0.225	0
	[0.258,0.467]	[0.303,0.518]		[0.457,0.638]	[0.151,0.301]	
Father Education: University	0.072	0.1	0.499	0.091	0.045	0.072
	[0.017,0.125]	[0.034,0.166]		[0.039,0.144]	[0.006,0.083]	
Mother Occupation: White-collar	0.396	0.462	0.238	0.52	0.289	0
	[0.312,0.482]	[0.373,0.563]		[0.449,0.599]	[0.222,0.359]	
Mother Occupation: Blue-collar	0.572	0.494	0.325	0.423	0.669	0
	[0.484,0.659]	[0.401,0.592]		[0.349,0.495]	[0.6,0.748]	
Father Occupation: White-collar	0.333	0.326	0.806	0.282	0.325	0.345
	[0.239,0.429]	[0.226,0.425]		[0.205,0.359]	[0.248,0.406]	
Father Occupation: Blue-collar	0.638	0.646	0.83	0.67	0.633	0.285
	[0.542,0.735]	[0.543,0.748]		[0.592,0.751]	[0.554,0.713]	

Notes: The table reports the median CLAN estimates for all covariates over 100 splits. The numbers in columns 1 and 4 (2 and 5) represent the share of individuals with a given characteristic among those belonging to the top (bottom) quartile in terms of impact of the treatment. 90 percent confidence intervals are reported in square brackets. Columns 3 and 6 report the p-value for the hypothesis that the difference between columns (1) and (2) -or between (4) and (5), respectively- is zero.

Table C.VII. CLAN of Well-being Index

	(1)	(2)	(3)	(4)	(5)	(6)
		TOP 2020			**TOP 2022**	
Variable	Upper quartile	Lower quartile	p-value difference	Upper quartile	Lower quartile	p-value difference
Female	0.52	0.286	0.005	0.407	0.595	0.003
	[0.407,0.632]	[0.184,0.388]		[0.319,0.497]	[0.501,0.691]	
Immigrant	0.318	0.015	0	0.264	0.28	0.815
	[0.219,0.43]	[-0.008,0.041]		[0.185,0.345]	[0.192,0.369]	
Learning Disorder	0.392	0.255	0.128	0.266	0.269	0.82
	[0.285,0.503]	[0.161,0.356]		[0.188,0.351]	[0.183,0.355]	
School grade 6	0.273	0.344	0.316	0.419	0.262	0.008
	[0.173,0.373]	[0.243,0.458]		[0.329,0.512]	[0.176,0.351]	
School grade 7	0.325	0.402	0.244	0.267	0.413	0.023
	[0.219,0.43]	[0.292,0.513]		[0.186,0.35]	[0.315,0.509]	
Std INVALSI Math 5	-0.088	0.355	0.004	-0.085	-0.122	0.405
	[-0.273,0.091]	[0.127,0.578]		[-0.196,0.037]	[-0.239,0.007]	
SES Status	-0.639	0.368	0	-0.47	-0.343	0.123
	[-0.795,-0.465]	[0.199,0.547]		[-0.597,-0.323]	[-0.453,-0.238]	
Baseline math grade	5.89	6.643	0	6.795	6.08	0
	[5.692,6.086]	[6.392,6.895]		[6.577,6.998]	[5.881,6.28]	
Single parent household	0.351	0.094	0	0.195	0.261	0.132
	[0.244,0.458]	[0.029,0.159]		[0.124,0.268]	[0.175,0.348]	
Mother Education: High-school	0.364	0.569	0.004	0.439	0.422	0.795
	[0.256,0.472]	[0.461,0.677]		[0.356,0.528]	[0.329,0.524]	
Mother Education: University	0.078	0.29	0	0.165	0.111	0.193
	[0.018,0.138]	[0.191,0.393]		[0.101,0.232]	[0.052,0.171]	
Father Education: High-school	0.338	0.434	0.171	0.378	0.415	0.458
	[0.231,0.444]	[0.326,0.543]		[0.292,0.467]	[0.323,0.509]	
Father Education: University	0.026	0.172	0.001	0.15	0.038	0.012
	[-0.001,0.062]	[0.089,0.257]		[0.088,0.213]	[0.002,0.073]	
Mother Occupation: White-collar	0.205	0.845	0	0.444	0.377	0.129
	[0.152,0.261]	[0.773,0.919]		[0.373,0.518]	[0.308,0.443]	
Mother Occupation: Blue-collar	0.762	0.093	0	0.505	0.567	0.252
	[0.69,0.824]	[0.035,0.153]		[0.433,0.578]	[0.493,0.638]	
Father Occupation: White-collar	0.238	0.491	0	0.368	0.309	0.308
	[0.17,0.311]	[0.381,0.606]		[0.29,0.448]	[0.226,0.392]	
Father Occupation: Blue-collar	0.75	0.447	0	0.58	0.644	0.266
	[0.675,0.822]	[0.339,0.558]		[0.5,0.663]	[0.563,0.728]	

Notes: The table reports the median CLAN estimates for all covariates over 100 splits. The numbers in columns 1 and 4 (2 and 5) represent the share of individuals with a given characteristic among those belonging to the top (bottom) quartile in terms of impact of the treatment. 90 percent confidence intervals are reported in square brackets. Columns 3 and 6 report the p-value for the hypothesis that the difference between columns (1) and (2) -or between (4) and (5), respectively- is zero.

D Additional results

D.1 Robustness

In Tables D.I and D.II we conduct a robustness analysis of our main results for TOP 2020 and 2022. Columns 1 and 2 display, respectively, OLS estimates and standard errors from our benchmark specification, for comparison purposes.

In columns 3 and 4, we choose 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– and the variables that are sufficiently correlated with control –again, after imposing the LASSO penalty (Ludwig et al., 2017).[48] A comparison of columns 1 and 3 shows that including LASSO-selected controls makes no substantial difference in most results, with the exception of the Aspiration index for TOP 2020 (Table D.I), 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, which help address potential bias related to attrition in our survey outcomes.[49] The estimated effects are almost unchanged and, if anything, slightly larger due to the minor imbalances in our survey data (with the control group being positively selected in terms of parental background). Overall, the different robustness checks presented provide a consistent picture of the positive impact of TOP on student outcomes.

[48]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 a non-zero 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 either the first or the second step. Table D.III lists the controls selected using LASSO for each outcome.

[49]Since there is no attrition in the administrative data, the first two lines in each table (i.e., grade and failure rate obtained from administrative records) do not report any estimates.

Table D.I. Robustness checks: TOP 2020

	(1)	(2)	(3)	(4)	(5)	(6)
	Standard controls		LASSO controls		Inverse Probability Weighting	
	Coeff	Std. Error	Coeff.	Std. Error	Coeff.	Std. Error
Academic Outcomes						
Performance Index	0.303	0.071	0.278	0.072	0.325	0.072
Grade Math Endline	0.113	0.056	0.121	0.054	NA	NA
Fail Math Endline	-0.044	0.022	-0.039	0.022	NA	NA
Std Performance Math	0.234	0.075	0.201	0.075	0.269	0.076
Non academic outcomes						
Aspirations						
Std Aspirations Index	0.167	0.078	0.094	0.065	0.134	0.077
Aspirations University (student)	0.062	0.037	0.035	0.034	0.050	0.037
Self-efficacy University (student)	0.040	0.032	0.027	0.031	0.033	0.032
High-school: vocational	-0.058	0.034	-0.019	0.029	-0.056	0.034
High-school: top tier	-0.000	0.027	-0.000	0.022	0.002	0.026
Aspirations University (parent)	0.042	0.033	0.016	0.031	0.043	0.034
Self-efficacy: university (parent)	0.080	0.032	0.072	0.030	0.079	0.032
Aspirations University (tutor)	0.029	0.022	0.031	0.021	0.029	0.021
Socio-Emotional Skills						
Std Socio-emotional Index	0.146	0.075	0.135	0.070	0.164	0.076
Logic task: difficult	0.048	0.037	0.044	0.037	0.046	0.038
Logic task: give-up	-0.034	0.026	-0.034	0.026	-0.041	0.027
Grit (student)	0.015	0.010	0.011	0.009	0.016	0.010
Grit (parent)	-0.006	0.010	-0.007	0.009	-0.004	0.010
Locus of control	0.023	0.009	0.024	0.009	0.025	0.009
Well-being						
Std Well-being Index	0.162	0.082	0.150	0.077	0.199	0.083
Depression (student)	-0.018	0.009	-0.022	0.009	-0.021	0.009
Happiness (student)	0.025	0.017	0.023	0.017	0.030	0.018
Depression (parent)	-0.011	0.008	-0.010	0.008	-0.014	0.008
Happiness (parent)	0.035	0.016	0.034	0.016	0.033	0.016

Notes: OLS estimates and standard errors in columns (1) and (2), coefficients and standard errors for the regressions with LASSO selected controls in columns (3) and (4) and coefficients and standard errors corrected through inverse probability weighting (IPW) in column (4) and (5). Randomization round fixed effects included in all regressions. The controls included for each regression and selected with LASSO are listed in Table D.III. The results with inverse probability weighting include all standard controls: parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline. Since there is no attrition in the administrative data, IPW estimates are not calculated for teacher-assigned grades and failure rate.

Table D.II. Robustness checks: TOP 2022

	(1)	(2)	(3)	(4)	(5)	(6)
	Standard controls		LASSO controls		Inverse Probability Weighting	
	Coeff	Std. Error	Coeff.	Std. Error	Coeff.	Std. Error
Academic Outcomes						
Performance Index	0.163	0.058	0.170	0.056	0.164	0.058
Grade Math Endline	0.088	0.055	0.075	0.055	NA	NA
Fail Math Endline	-0.066	0.027	-0.059	0.026	NA	NA
Std Performance Math	0.155	0.059	0.160	0.059	0.155	0.060
Non academic outcomes						
Aspirations						
Std Aspirations Index	0.058	0.066	0.043	0.056	0.059	0.066
Aspirations University (student)	0.022	0.030	0.020	0.029	0.018	0.030
Self-efficacy University (student)	0.000	0.029	0.001	0.028	0.000	0.029
High-school: vocational	-0.031	0.027	-0.017	0.023	-0.032	0.028
High-school: top tier	0.038	0.023	0.025	0.020	0.039	0.023
Aspirations University (parent)	0.031	0.028	0.023	0.026	0.032	0.028
Self-efficacy: university (parent)	-0.002	0.030	0.003	0.030	0.000	0.030
Aspirations University (tutor)	-0.007	0.014	-0.005	0.014	-0.009	0.015
Socio-Emotional Skills						
Std Socio-emotional Index	-0.039	0.071	-0.032	0.067	-0.048	0.070
Logic task: difficult	-0.011	0.034	-0.009	0.033	-0.012	0.034
Logic task: give-up	0.007	0.023	0.005	0.023	0.009	0.023
Grit (student)	-0.009	0.009	-0.007	0.008	-0.009	0.009
Grit (parent)	-0.008	0.008	-0.005	0.008	-0.009	0.008
Locus of control	0.007	0.008	0.006	0.007	0.007	0.008
Well-being						
Std Well-being Index	0.037	0.072	0.027	0.067	0.029	0.073
Depression (student)	-0.006	0.008	-0.004	0.007	-0.006	0.008
Happiness (student)	0.008	0.014	0.005	0.013	0.007	0.014
Depression (parent)	-0.003	0.007	-0.004	0.006	-0.003	0.007
Happiness (parent)	0.001	0.012	0.001	0.012	0.002	0.012

Notes: This table shows the main OLS estimates and the standard errors in columns (1) and (2), the coefficients and standard errors for the regressions with LASSO selected controls in columns (3) and (4) and the coefficients and standard errors corrected through inverse probability weighting (IPW) in column (4) and (5). Randomization round fixed effects included in all regressions. The controls included for each regression and selected with LASSO are listed in Table D.IV. The results with inverse probability weighting include all standard controls: parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline. Since there is no attrition in the administrative data, IPW estimates are not calculated for teacher-assigned grades and failure rate.

Table D.III. LASSO selected variables, TOP 2020

(a). Academic Outcomes

	(1)	(2)	(3)
	Grade	Fail	Performance
Baseline math grade	✓	✓	✓
Std INVALSI Math 5	✓		✓
SES Status	✓		
Single parent household	✓		
Math Grade at baseline (Missing)	✓		
Learning disorders			✓
School grade 6			✓
School grade 7			✓

(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
Female	✓	✓				✓		
Immigrant	✓	✓				✓		
Learning disorders	✓			✓		✓	✓	✓
Baseline math grade	✓	✓	✓	✓	✓	✓	✓	✓
Std INVALSI Math 5	✓	✓			✓	✓	✓	
Mother Education: University	✓	✓	✓			✓	✓	✓
Father Education: University	✓	✓			✓	✓	✓	
SES Status			✓					
School grade 6				✓				

(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	
Learning disorders	✓	✓		✓		✓
Baseline math grade	✓				✓	✓
Mother Education: University						
Std INVALSI Math 5		✓				
Female						✓
Mother Occupation: White-collar						✓
Father Education: University						

(d). Well-being

	(1)	(2)	(3)	(4)	(5)
	Well-being Index	Students		Parents	
		Depression	Happiness	Depression	Happiness
Female		✓	✓		
Baseline math grade				✓	

Notes: This table shows the controls selected using LASSO for each outcome variable in TOP 2020.

Table D.IV. LASSO selected variables, TOP 2022

(a). Academic Outcomes

	(1)	(2)	(3)
	Grade	Fail	Performance
Baseline math grade	✓	✓	✓
Std INVALSI Math 5	✓		✓
SES Status	✓		
Single parent household	✓		
Math Grade at baseline (Missing)	✓		
Mother Education: University			
Learning disorders			✓
School grade 6			✓
School grade 7			✓

(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
Female	✓	✓				✓		
Immigrant	✓	✓				✓		
Learning disorders	✓			✓		✓	✓	✓
Baseline math grade	✓	✓	✓	✓	✓	✓	✓	✓
Std INVALSI Math 5	✓	✓			✓	✓	✓	
Mother Education: University	✓	✓	✓			✓	✓	✓
Father Education: University	✓	✓			✓	✓	✓	
SES Status			✓					
School grade 6				✓				

(c). Socio-emotional Skills

	(1)	(2)	(3)	(4)	(5)	(6)
	Socio-Emotional Index	Students				Parents
		Perservance: difficulty	Perseverance: give up	Grit	Locus of control	Grit
Learning disorders	✓	✓		✓		✓
Baseline math grade	✓				✓	✓
Mother Education: University						
Std INVALSI Math 5		✓				
Female						✓
Mother Occupation: White-collar						✓
Father Education: University						

(d). Well-being

	(1)	(2)	(3)	(4)	(5)
	Well-being Index	Students		Parents	
		Depression	Happiness	Depression	Happiness
Female		✓	✓		
Baseline math grade				✓	

Notes: This table shows the controls selected using LASSO for each outcome variable in TOP 2022.

D.2 Medium-term outcomes

Table D.V reports limited evidence on the medium-term impact of the intervention. The limitation stems from the fact that the only administrative data we have available for years other than the school year of the intervention is for students who participated in TOP 2020 while they were in grades 7 or 8. For these students we can investigate the effect of receiving tutoring on standardized test scores *the following year*, as well as on the type of high school track chosen at the end of middle school.

The national agency INVALSI administers a standardized test to all students in Italian schools at the end of grade 8. In column 1 of Table D.V we study how students who participated in TOP 2020 when they were in grade 7 performed in the INVALSI test in 2021, when they were in grade 8. The average effect of participating in the online tutoring is a 0.068 SD increase in test scores. The effect is not statistically significant due to the limited sample for which we observe this outcome, but the qualitative result is encouraging, especially considering that this test is taken one full year after the end of the intervention. Note that the effect is entirely driven by students who participated in the intense 6-hour tutoring: for this group the impact is +0.24 SD (p-value = 0.08).

Column 2 shows the impact of the program on the probability of attending a vocational track – the least academically demanding type of high school in the Italian system. This outcome is available for students who participated in TOP 2020 during grade 7 as well as grade 8 and is measured in 2022 for the former and in 2021 for the latter. Again, the estimates are not statistically significant, but they suggest a reduction in the probability of attending a vocational track by 3.7 percentage points (a 8% relative to the control group) consistent with Falk et al. (2024).

Finally, in column 3 we consider as outcome the probability that, during grade 8, teachers recommend the student for a vocational high school track. Teacher recommendations are a feature of the Italian system and, albeit non-binding, are a useful indicator of teachers' assessment of the future potential of the students.[50] Teacher recommendations are available for *all* students who participated in TOP 2020 and for this outcome the average effect of the program is a precisely estimated zero.

[50]Carlana (2019) and Carlana et al. (2022b) show that teachers' recommendations also reflect implicit bias against female and immigrant students, respectively.

Table D.V. Estimation of the impact of TOP 2022 on medium-term academic outcomes

	(1)	(2)	(3)
	Test Score	Vocational track choice	Vocational track recommendation
Panel A: Overall results			
Treatment	0.068	-0.037	-0.009
	(0.088)	(0.036)	(0.028)
	[1.000]	[1.000]	[1.000]
R^2	0.413	0.153	0.210
Panel B: Results of 3 vs. 6 hours of tutoring			
Treatment	0.009	-0.042	0.005
	(0.098)	(0.039)	(0.030
Intense Treatment	0.226	0.026	-0.044
	(0.146)	(0.059)	(0.047)
Treat+Intense Treatment==0	0.080	0.775	0.384
R^2	0.417	0.156	0.223
Mean Dep:	0.00	0.46	0.62
Sample: TOP2020 in	Grade7	Grade7 and 8	Grade6, 7 and 8
Outcome observed in	2021	2021-2022	2020-2021-2022
Obs	341	721	1010

Notes: OLS estimates, robust standard errors in parentheses, and Anderson sharpened q-values in square brackets. The dependent variable is the INVALSI8 standardized test score of 2021 for students treated in grade 7 in column 1, a dummy variable for whether the student chooses vocational track in column 2, and a dummy variable for whether the student was recommended by teachers to vocational track in column 3. The test score is only available for one of the three treated cohorts while the track choice only for two of the treated cohorts. "Treatment" is an indicator for being assigned a tutor; "Intense treatment" is an indicator for being assigned to 6 hours of tutoring. In Panel B, we control for whether the students was identified for intense tutoring (coefficient not shown). Controls included in all regressions: parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline. "Mean Dep" is the mean of the dependent variable at endline for students in the control group.

D.3 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 enables the tutor and the student 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. In our endline survey, we collected information on whether the student used a phone for the tutoring: 20 percent of the students in TOP 2020 and 28 percent in TOP 2022 mainly used a smartphone to connect.

In Table D.VI 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 in TOP 2020 (Panel A) the impact was not statistically different, except for

aspirations, where the effect on students who used a smartphone is zero. Column 1 shows that, compared to an increase in test score of 0.32 SD for the students who connected with better devices, the impact for students using a smartphone was 0.27 SD, significant at the 1 percent level. However, the positive effect on performance in TOP 2022 (Panel B) is fully driven by students who connected using a PC or tablet, while students who used the phone for tutoring were the only ones showing an improvement in aspirations and socio-emotional skills.

While based on the results during the pandemic one could have concluded that online tutoring was a promising option for relatively low-income settings –provided students had access to smartphones–, the overall picture including TOP 2022 suggests a more nuanced interpretation as far as academic outcomes are concerned. In normal school times, it may be necessary to provide a better device to students in need, in order to improve the effectiveness of online tutoring programs.

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

	(1)	(2)	(3)	(4)
	Performance	Aspirations	Socio-emotional	Wellbeing
Panel A: TOP 2020				
Treatment	0.313	0.175	0.125	0.160
	(0.076)	(0.083)	(0.079)	(0.085)
Phone tutoring	-0.044	-0.040	0.103	0.010
	(0.096)	(0.123)	(0.118)	(0.131)
p-value Treat+Phone	[0.007]	[0.272]	[0.059]	[0.210]
Obs.	711	523	636	614
R^2	0.306	0.335	0.180	0.077
Panel B: TOP 2022				
Treatment	0.194	0.032	-0.088	0.054
	(0.060)	(0.068)	(0.074)	(0.075)
Phone tutoring	-0.190	0.161	0.303	-0.107
	(0.094)	(0.113)	(0.113)	(0.139)
p-value Treat+Phone	[0.969]	[0.097]	[0.061]	[0.706]
Obs.	943	889	898	881
R^2	0.240	0.221	0.066	0.075

Notes: OLS estimates, robust standard errors in parentheses, and p-values in square brackets. The dependent variable is the performance index column 1, the aspiration index in column 2, the socio-emotional index in column 3, and the well-being index in column 4. "Treatment" is an indicator for being assigned a tutor; "Phone tutoring" is an indicator for doing the tutoring using the phone. 20 percent of students in TOP 2020 and 28 percent of students in TOP 2022 used mainly the phone during the tutoring. The p-value of the sum of the coefficients "Treatment" and "Phone tutoring" is presented at the end of each panel. Panel A presents the results of TOP 2020, while Panel B of TOP 2022. Controls included in all regressions: parental education and occupation, gender, immigration status, learning disorders, school grade, teacher-assigned grades, SES status, and test scores in math at baseline.

E Cost and effect size of tutoring interventions

E.1 Data selection process

In the replication package for this study, we include a dataset with the list of all tutoring interventions considered in our analysis, the effect size on the learning outcome, the cost per beneficiary, as well as the location, the sample size, the age range of the children included in the intervention, and the duration of the program.

To determine which studies to include in the dataset, we adopted the following selection criteria. First, to ensure comparability across interventions, we focused on studies reporting learning as an outcome, using standardized test scores (e.g., math, language, overall scores) or comparable performance metrics (e.g., reading fluency scores, comprehension test scores). To maintain consistency in effect size measurement, we report effect sizes in standard deviations (SD). We excluded programs reporting only other educational outcomes, such as enrollment, dropout rate, or attendance. Whenever an intervention reported multiple effect sizes for the same outcome dimension (e.g., effects on math vs. language test scores), we calculated the average effect size.

Second, each study had to clearly report the cost of the intervention per-beneficiary for a given year or provide enough information for this metric to be calculated. If costs were available in currencies other than USD, we converted them using exchange rates from the study's publication date retrieved from https://www.google.com/finance/, to generate Figure 2 in the main text. The reported cost ranges were averaged to obtain a single estimate per intervention.

Finally, we restricted our sample to studies in which the target population was school-aged children in high-income countries and with a minimum sample size of 70 observations.

The selection of educational interventions followed a systematic approach grounded in a thorough review of the literature. The first step of our search was to consider meta-analyses and systematic reviews evaluating or reporting the cost-effectiveness of tutoring programs. We began by reviewing the meta-analyses by Nickow et al. (2020), Fryer Jr (2017) and Angrist et al. (2020), which provided a foundational overview of education interventions and their estimated impacts. Subsequently, we expanded our search by including relevant studies cited in these meta-analyses and using these as a starting point to identify additional pertinent studies through their references and related literature.

We also searched academic databases containing peer-reviewed research articles and scholarly working papers (e.g., EBSCO, JSTOR, NBER, etc.). We conducted a broad search on Google Scholar by inserting combinations of the following terms "education",

"intervention", "cost", "cost-effectiveness", "tutoring", "program", "treatment", "tutor", "student". For the most relevant studies, we applied both backward and forward citation tracking. This involved reviewing the study's references to identify relevant past research and examining subsequent studies that cited it to ensure comprehensive coverage of the literature.

Another approach we used leveraged "Connected Papers" – a tool that maps related academic papers based on conceptual similarity rather than direct citations, creating a visualization of interconnectedness of research articles. This approach facilitated a broader exploration of the relevant literature.

Overall, the above methodology resulted in the inclusion of 21 programs based in high-income countries that included a tutoring component in at least one of their treatment arms. In Section 6 of the paper, we compare the costs and effect size of TOP to those of the other interventions.