Investing in Human Capital During Wartime: Experimental Evidence from Ukraine*

Lelys Dinarte-Diaz James Gresham Renata Lemos
Harry A. Patrinos Rony Rodriguez-Ramirez

May 1, 2025

Abstract

This paper provides insights into human capital investments during wartime by presenting evidence from three experiments of an online tutoring program for Ukrainian students amid Russia's full-scale invasion. The program provides three hours of weekly math and Ukrainian language tutoring in small groups over six weeks, and uses academic and psychosocial tools to address student challenges at different phases of the invasion. Results from all three experiments show significant improvements in math and Ukrainian language scores by as much as 0.49 SD and 0.40 SD –equivalent to 2 to 2.5 years of learning, respectively,– and a reduction in stress levels by up to 0.12 SD. The program improves peer support, fosters positive learning attitudes, and helps students develop social-emotional skills, contributing to improved learning and mental health. The benefit-cost ratios show that benefits exceed costs in each experiment.

Keywords— Ukraine, war, tutoring, achievement, mental health

JEL Codes— I21, I24

*Corresponding authors: Lelys Dinarte-Diaz at ldinartediaz@worldbank.org and Renata Lemos at rlemos@worldbank.org. Dinarte-Diaz: The World Bank, IZA, HiCN, CESifo; Gresham: The World Bank; Lemos: The World Bank, CEP-LSE, IGC, and CEPR; Patrinos: University of Arkansas, IZA, and GLO; Rodriguez-Ramirez: Harvard University. We thank Noam Angrist, Alejandro Ganimian, Lawrence Katz, Jostin Kitmang, Rachael Meager, Laura McDonald, Vesall Nourani, Daniel Rodriguez-Segura, and Eric Taylor for helpful discussions and feedback, as well as participants at seminars at the World Bank, Harvard Economics of Education, 2024 Evidence in Governance and Politics Conference, and the 2024 Northeastern Universities Development Consortium. We are deeply grateful for the incredible support of Teach for Ukraine, led by Oksana Matiiash, for implementing the intervention, and to Olga Anisimova, Yulia Bilyk, Eduardo Cascallar, Sylvia Cesar, Juan Miguel Jimenez, Antonella Novali, Gabriele Sabino, Maria Montoya Aguirre, Mariel Musso, Zezhen Wu, Tracy Xu for their invaluable assistance with research and project management. We also appreciate the collaboration with the Harvard Program in Refugee Trauma for the third experiment. We acknowledge financial support from the World Bank (Strategic Impact Evaluation Fund and Research Support Budget) for the evaluation and the UBS Optimus Foundation for the implementation of the program. AER RCT Registry ID AEARCTR-0010634. We received IRB approval from Innovations for Poverty Action IRB (Protocol 16680) on March 30, 2023. We include a structured ethics appendix in Appendix A. The findings, interpretations, and conclusions expressed in this report belong entirely to the authors and do not necessarily represent the views of the World Bank, its affiliated organizations, Executive Directors, or the governments they represent.

> "There is as much reason for the inclusion in the cost of the war of the loss of human capital sustained by the nation as there is for the loss of material capital."
>
> Harold Boag, *Journal of the Royal Statistical Society, 1916*

1 Introduction

Failing to invest in human capital during wartime can significantly undermine long-term economic development (Blattman and Miguel, 2010; Collier, 1999). However, during conflicts, governments may divert investments in human capital, such as education, to other priorities that are perceived to provide immediate, measurable economic benefits. The risk of destruction to infrastructure, disruption to services, and population displacement makes the delivery of government social services challenging and the return on these investments uncertain. To date, there is limited experimental evidence on interventions implemented during conflicts that can mitigate the adverse effects of war on human capital.

This paper contributes to filling this gap by providing evidence on the feasibility, effectiveness, and scalability of an online tutoring program that was offered to Ukrainian students during Russia's full-scale invasion through three consecutive experiments.[1] The experiments took place between early 2023 to mid-2024 –a period marked by frequent armed attacks and substantial instability in Ukraine (Figure A1). In this context, we expect online tutoring programs to address learning gaps caused by disruptions to regular schooling through targeted support, while also supporting student wellbeing by providing stability and a sense of normalcy via service continuity. Taken together, these benefits could positively impact student learning and mental health, contributing to human capital accumulation. However, there are at least three reasons why such programs might not yield positive results during wartime. First, participation may be hindered by logistical issues including power outages, internet disruptions, frequent evacuations, and temporary accommodations. Second, the psychological toll of war may affect students' ability to focus, tutors' ability to teach, and parents' level of engagement. Third, families may

[1] Russia's full-scale invasion of Ukraine began on February 24, 2022, marking a significant escalation of the ongoing Russo-Ukrainian War that started in 2014 with the annexation of Crimea and the conflict in Donbas.

prioritize immediate needs over education, reducing parental buy-in or student effort, especially if short-term benefits are not clear.

To evaluate the potential of online tutoring to contribute to human capital accumulation during Russia's full-scale invasion of Ukraine, we partnered with Teach for Ukraine[2] to design a program for students in grades 5 to 10. The program provided three hours of weekly tutoring for six weeks in two core subjects–math and Ukrainian language–through an online platform, with students learning in small groups of three. We conducted three experiments, reaching a sample of approximately 10,000 students. In each experiment, the program was adapted to address the evolving challenges students and tutors faced at the time.

The first experiment began in February 2023, 11 months after the invasion was launched, during a harsh winter with frequent power outages. Schools —many of which were damaged or destroyed— operated in various in-person and online formats. The tutoring program, which was aligned with the Ukrainian curriculum, aimed to help students catch up in core subjects by adapting sessions to students' needs. The second experiment began 14 months into the conflict as students were completing their first full academic year under war conditions. The program maintained the same structure as the first experiment but introduced initial diagnostic assessments to sort students into groups by ability and provided diagnostic reports and formative assessments to tutors. The third experiment began 23 months after the invasion began. During this period, displacement had declined, many refugees had returned, and violent attacks had lessened, however, trauma among students had intensified. To address this, the program was similar to the one offered in the second experiment, but also incorporated psychosocial support in the form of trauma-informed care exercises.

In each experiment, the program was advertised online, with a 30-day enrollment window. All students were invited to join small groups and given access to the online platform for interactions with their group peers. However, due to the high volume of applications exceeding TFU's tutoring capacity, we randomized the allocation of available slots for participation in the tutoring sessions. We conducted a two-stage stratified random assignment at the household level. In the first stage, all eligible *households* with

[2] Teach for Ukraine is a non-governmental organization that is part of the Teach for All Global Network.

at least one registered student in the program were randomly assigned to treatment or control groups. In the second stage, we performed another stratified random assignment of students from both control and treatment groups to their specific tutoring groups. The procedure slightly differed between the experiments in the second stage of randomization, as grouping were based on ability. The total evaluation sample across all three experiments was 9,832 students (2,518 students in the first experiment, 2,767 in the second, and 4,547 in the third). There was no overlap in the samples of students across experiments.

We collected data before (at the time of registration), during, and at the end of each experiment to measure student take-up and the impact of the program on academic outcomes (math and Ukrainian language scores), mental health (stress and anxiety), and the potential mechanisms driving changes in outcomes (attitudes toward learning, aspirations, socioemotional skills, and investments). The data came from various sources: take-up data was provided by Teach for Ukraine, academic performance data was collected using self-administered assessments on math and Ukrainian language, and data on mental health and mechanisms was collected through self-reported questionnaires. In addition, we collected data from parents (or guardians) on mental health and well-being and sociodemographic characteristics at the time of registration. Finally, data on each student's attendance and engagement was collected by tutors through daily session journals, and information on student interactions was collected in the tutoring platform.

Overall, intent-to-treat estimates show that the tutoring program led to large improvements in students' academic performance and mental health outcomes. Students assigned to the treatment group improved their math scores by 0.49 standard deviations (SD) in the first experiment, 0.22 SD in the second experiment, and 0.22 SD in the third experiment. Similarly, students assigned to the treatment group improved their Ukrainian language scores by 0.40 SD in the first experiment and 0.32 SD in the third experiment (the impact was null for the second experiment). Taken together, these estimated effects on academic scores are equivalent to the learning gains that can be achieved during between approximately 1 year of education (for 0.21 SD on math scores in the second experiment) and 2.5 years of education (for 0.49 SD on math scores in the first experiment).[3] In addition, we show that students in the treatment group experienced similar improvements in

[3] We use the estimates from Avvisati and Givord (2021) to translate SD of learning gains into equivalent years of schooling.

their stress levels between experiments (between 0.10 SD in the first experiment and 0.12 SD in the third experiment). Impacts on anxiety levels were null.

The tutoring program may have improved students' outcomes through various mechanisms. We explore: (i) structured interactions with other peers; (ii) attitudes toward learning and aspirations; (iii) social-emotional skills; and (iv) parental and student investments. First, we show that students in the treatment group increased the probability of enrolling and interacting on the platform in all three experiments. Second, we document improvements in measures of student attitudes toward learning, but without effects on aspirations for their future, in all three experiments. Third, we also find positive effects of the intervention on students' persistence and self-efficacy in the second and third experiments. This analysis in combination with the positive treatment effects suggest that the online tutoring programs provided students with the necessary peer support, attitudes toward learning, and social-emotional skills to improve their academic performance and improve their mental health.

To provide rigorous evidence for the complementary investments mechanism, we measured the impacts of student investments and parental investments, separately. First, we show that treated students are more likely to seek additional tutoring support while they are enrolled in the intervention. Second, we conducted a parallel experiment to specifically test for parental investments. At the end of the first experiment, we randomly assigned a sub-sample of non-treated households to two groups. Both groups had the opportunity to participate in the tutoring program as it was delivered in the first experiment, but parents of students in the second group additionally received text messages with information on how they could support their children's participation. The results show that academic performance declined in the group in which parents received additional information. This finding highlights potential challenges in designing programs that involve family support in a war context.

Our paper contributes to several strands of literature. First, it contributes to the discussion of policy-informed scalability of interventions (Al-Ubaydli et al., 2017; Banerjee et al., 2017; Muralidharan and Niehaus, 2017; Mobarak, 2022; List, 2022; Vivalt, 2020). Effective scaling requires evaluating whether an intervention produces consistent results across different populations and contexts, accounts for spillover effects (such as general equilibrium effects) that arise at scale, and does not lead to disproportionately higher costs as it expands (List, 2022). Our study contributes to the understanding of scalability by replicating an online tutoring experiment across diverse student populations in all regions of Ukraine as well as Ukrainian refugees outside Ukraine, and across three distinct phases of the war – each characterized by various levels of intensity, disruptions, and population displacement. The consistent positive impacts of the program on academic learning and mental health reinforce its potential scalability while reducing concerns about false positives and population and situation representativeness (List, 2022). In addition, the fact that randomization is conducted at the household level ensures estimates capture potential spillover effects between students in the same household, as would occur if the program were scaled up where all students in the same household would have access to online tutoring. Finally, while we nearly doubled the sample size in our third experiment, the average cost per participant remained similar to that of the first two experiments, suggesting the program maintains cost-efficiency at scale. To our knowledge, this is the first study to address scalability in this context.

Second, this paper adds to the growing body of research on effective strategies that can be used to mitigate the impact of school closures on both learning losses and mental health, particularly in the context of emergencies. Research has shown that school disruptions and closures caused by emergencies—such as wars, pandemics, or natural disasters—can have short- and long-term negative effects on human capital.[4] However, while pandemics primarily disrupt schooling through health-related restrictions, wars not only force school closures but also lead to large scale destruction of physical capital, creating severe resource constraints for governments and making access to education even more difficult. Despite the significant disruptions caused by different types of emergencies, most studies have focused on the potential of remote learning interventions during pandemics. For instance, blended learning has shown promise, and its implementation during school closures has yielded valuable insights (Bettinger et al., 2020; Angrist et al., 2020a,b, 2022, 2023; Hassan et al., 2023). Similarly, free online tutoring programs have proven effective in addressing learning losses during pandemic-induced school clo-

[4] Several papers estimate the impact of wars on human capital accumulation. For example, Austrian and German children who were ten years old during WWII received less education than comparable individuals from non-war countries, leading to negative effects on earnings approximately 40 years after the war (Ichino and Winter-Ebmer, 2004). Similar findings are documented from other wars in the former Yugoslavia (Lai and Thyne, 2007; Eder, 2014), Greece (Patrinos et al., 2022), Spain (Arrazola and de Hevia, 2008; M. Arrazola and Sanz, 2003), Peru (Leon, 2012), El Salvador (Acosta et al., 2020), and Tajikistan (Shemyakina, 2011).

sures in countries such as Italy (Carlana and La Ferrara, 2021) and Spain (Gortazar et al., 2023). Our paper contributes to this literature by providing experimental evidence of the positive impact of online tutoring programs on students' learning outcomes and mental health during wars, where traditional educational recovery strategies may be challenging to implement.

Finally, this paper contributes to research on remedial education as a tool for improving outcomes for vulnerable students. Research has shown that remedial (prevention) programs for primary and secondary students enhance short-term academic performance (Jacob and Lefgren, 2004; Lavy and Schlosser, 2005; Banerjee et al., 2007, 2010; Battistin and Schizzerotto, 2019) and positively impact long-term school progression and labor market outcomes (Lavy et al., 2022). Over the past decades, the leading remedial education program has been "Teaching at the Right Level" (TARL), which has been evaluated by various studies in low- and middle-income countries (Banerjee et al., 2017, 2016, 2015; Banerji and Chavan, 2016; Gutiérrez and Rodrigo, 2014; Lakshminarayana et al., 2013). The tutoring programs in the second and third experiments of our study incorporate some TARL elements, including: (i) assessing student learning levels with a simple tool, (ii) grouping them by learning levels, (iii) using a variety of engaging teaching and learning activities, and (iv) tracking student progress over time. By integrating these elements into the tutoring intervention, we provide further evidence of the relationship between participation in a remedial educational program and short-term academic achievement in a wartime context.

2 Context and Intervention

During the first two years of Russia's full-scale invasion, the Ukrainian government significantly reduced its investment in the education sector. In 2021, education spending, adjusted for 2015 prices, was 178.6 billion Ukrainian hryvnias, representing 17% of total expenditure. However, by 2022, this had decreased to 138.1 billion Ukrainian hryvnias (9.6% of total spending), and further dropped to 110.5 billion Ukrainian hryvnias in 2023, accounting for only 7% of the total expenditure.[5] At the same time, the Ministry of Edu-

[5] Open Budget data from Ukraine's Ministry of Finance

cation and Science estimated that between 2,900 and 3,500 schools (10-13% of all schools) were either partially damaged or completely destroyed, exacerbating the already strained conditions caused by reduced funding.[6]

To ensure student safety and enable in-person education, schools were required to have bomb shelters for air raids. While some schools had shelters, many did not meet safety standards, limiting in-person learning. As a result, only around 30% of secondary schools offered fully in-person classes, while 34% operated exclusively online, and 36% used a blended approach.[7] Online learning tools, such as the Ministry's All-Ukrainian School Online platform, offered education aligned with the Ukrainian state curriculum goals through video tutorials, tests, and assignments in 18 core subjects for grades 5-11. The shift to online education was not new to Ukrainian students, who had experienced remote learning during COVID-19.[8] During this time, the Ministry of Education and Science was interested in working with local institutions to find feasible solutions to support students' education amid these challenges.

In early 2023, we partnered with Teach for Ukraine (TFU) to adapt and evaluate an online tutoring program designed to support students in the public education system, supplementing the education they were receiving through these various learning formats. The program was implemented across three separate, consecutive experiments. It targeted students in grades 5 to 10, and offered three hours of tutoring per week, which were divided into two 1.5 hour sessions. Each session consisted of 45 minutes of math and 45 minutes of Ukrainian language tutoring, and were conducted in groups of three students over a six-week period. To be eligible for the program, students had to be between 10 to 17 years old, with parental consent and student assent required for participation.

In each experiment, the program was adjusted to incorporate features that addressed the evolving challenges faced by Ukrainian students and tutors at different phases of the war. Figure A2 summarizes the intensity of war-related events at different phases of the war, as well as the geographic distribution of students across Ukrainian regions during

[6] Information extracted from Second and Third Rapid Damage and Needs Assessments Reports led by the Government of Ukraine, the World Bank Group, the European Commission, and the United Nations (World Bank et al., 2023; World Bank, 2024).

[7] Information provided by the Ministry of Education and Science

[8] However, sustained attacks on critical infrastructure, particularly the energy grid, frequently disrupted access to online education. The telecommunications sector worked to maintain services despite missile strikes and power outages, relying on generators to stay operational.

the periods in which each of the experiments described below was conducted.

First experiment. The first experiment ran from February to March 2023. By the 2022-23 school year, which began 6 months after the launch of the invasion, in-person education had resumed only in schools equipped with adequate shelters. Students across the country faced a mix of in-person, online, and hybrid learning modalities.

The tutoring program aimed to help students recover from learning losses caused by school closures in the first year of the invasion. Students were randomly assigned to small groups, and tutors followed the Ukrainian curriculum learning goals. The program was intentionally designed to provide tutors with flexibility, allowing them to assess the needs of their students and adapt their approach accordingly. The intervention was open to students residing in Ukraine and Ukrainian refugee students abroad.

Second experiment. The second experiment ran from late April to mid-June 2023, which coincided with the end of the school year. To ease the burden on students and teachers, the Ministry of Education and Science had announced at the time that the state final examinations would be canceled for the second consecutive year.

The tutoring program continued to focus on helping students catch up academically, while also supporting them in completing their first, full school year under wartime conditions. We introduced three modifications to the program to support tutors in their work. First, a short multiple-choice diagnostic assessment was administered at registration to assess students' prior knowledge, which allowed us to group them by ability level.[9] Second, tutors received diagnostic reports before the first week of the tutoring program, summarizing how their assigned groups performed in the short multiple-choice assessments relative to the overall results of all of the students participating in the second experiment.[10] Third, tutors received short, curriculum-based formative assessments consisting of five open-ended, short-essay questions for each grade-subject in the second week of the program, which were to be used as needed to help them gauge student progress on key learning outcomes. The program remained open to both students in Ukraine and Ukrainian refugees living abroad.

[9] The short multiple-choice assessment included 12 questions, 5 math questions and 7 Ukrainian language questions. Students were ranked by their total number of correct answers within their strata and assigned to groups of three in rank order.

[10] Figure A3 shows an example of the diagnostic report sent to the tutors for one of the math groups.

Third experiment. The third experiment ran from February to March 2024. Despite the ongoing conflict, the number of internally displaced people fell from 5.4 million in December 2022 to 3.7 million by September 2023, while the number of returnees had reached 4.6 million. Although 80% of schools had shelters by December 2023, online learning remained a key modality due to capacity constraints in shelters and the continued closure of schools in high-risk areas. In addition, the mental health burden on Ukrainian children and adolescents had become substantial. A 2023 study of over 8,000 Ukrainian adolescents, both within the country and abroad, found that those exposed to the war were significantly more likely to screen positive for psychiatric conditions (Goto et al., 2024).

While the tutoring program remained focused on academic catch-up, tutors were trained to integrate psychosocial support tools into their sessions. In collaboration with the Harvard Program on Refugee Trauma (HPRT), we designed and incorporated four trauma-informed care exercises. The first exercise, a storytelling activity based on the Stone Soup fable, was introduced in the initial session to foster resilience and community solidarity. The second, guided deep breathing, aimed to promote emotional calmness and was reinforced with instructional videos for consistent use. The third combined icebreakers with tapping exercises, where students selected an animal, identified a positive trait, and tapped an acupuncture point while repeating an affirmation, for example, as tapping the hand while expressing the qualities of an eagle. The fourth exercise, Happy Faces, allowed students to privately indicate their mood at the start of each session, enabling tutors to offer emotional support as needed. Tutors were trained by the HPRT team and provided with a manual and lesson plans on integrating these activities into their sessions. While the tools from the second experiment remained available, the third experiment prioritized psychosocial interventions over academic support tools. A detailed description of the tutoring sessions in presented in Appendix B.

3 Experimental Design

3.1 Recruitment of Students and Tutors

For all three experiments, registration for the tutoring program was actively promoted through TFU's social media channels for approximately up to 30 days before the start of each experiment. Parents (or guardians) could enroll their children by clicking the registration link provided in the invitation. After clicking the link, they were first asked to provide consent for their child's participation. If they agreed, they were prompted to complete a 10-minute self-reported survey. At the end of the survey, instructions directed parents to hand their internet-enabled device to each child they wished to enroll. Students who assented to participate in both the program and the study were then asked to complete a 35-minute self-reported survey, which collected information to assess their eligibility for the study. As shown in Figure A4, the number of households registered ranged from approximately 2,500 in the first two experiments to 4,800 in the third. Demand for the tutoring program exceeded TFU's implementation capacity in each experiment. Due to oversubscription, we randomly assigned eligible households to treatment and control groups, as described in section 3.2.

Tutor recruitment for all experiments was conducted exclusively by TFU. Tutors were required to have prior teaching experience, experience working with student groups online, proficiency in using various digital teaching tools (platforms, websites, etc.), and flexibility in their schedules. In addition, candidates were expected to be highly motivated and have prior experience in similar educational initiatives. During the enrollment phase, tutors were required to provide consent to participate in both the program and the study, and to complete a 20-minute questionnaire. Shortlisted candidates were then invited for an interview with TFU staff. Finally, selected tutors participated in a two-day training program,[11] conducted by the International Tutoring Academy (Kyiv-based), before beginning their tutoring activities.

3.2 Randomization

We conducted the randomization at the household level to mimic real-world conditions, where entire households—rather than students within them—would typically have access to the tutoring program if the program were scaled up. This household-level ran-

[11] The training program focused on two areas: (i) socio-emotional support, helping tutors learn strategies to support students in managing their emotions, maintaining motivation, and creating a positive learning environment, and (ii) neurodidactics, helping tutors explore how the brain learns and processes information, gain insights into brain functions related to memory, attention, and emotion, and apply this knowledge during tutoring sessions.

domization approach helped capture spillover effects that would occur at scale, providing more accurate estimates of the program's potential impact. A summary of the number of eligible households and students enrolled and assigned to the treatment and control groups is presented in Figure A5.

First experiment. The registration process took place in December 2022 and January 2023, with a total of 2,322 guardians and 2,518 students completing the baseline survey. We randomly assigned 1,161 households (1,259 students) to the treatment group and 1,161 households (1,259 students) to the control group. Random assignment was stratified based on two variables: (i) whether the student's parent had completed higher education and (ii) whether the student was residing in Ukraine at the time of registration.

Following random assignment to either the treatment or control group, students were further randomized into tutoring groups of three students.[12] This assignment was stratified by treatment status, grade level, and preferred schedule for the online tutoring sessions. Students in the treatment group were assigned to two specialized tutors—one for math and one for Ukrainian language—and remained in the same peer group for both subjects.

Second experiment. The second experiment's registration process took place between March and April 2023 and followed the same procedure as the first experiment, with the addition of a short 12-question diagnostic assessment in math and Ukrainian language. An additional eligibility criteria in this experiment was that students could not have participated in the previous experiment.

A total of 2,573 eligible guardians and 2,767 eligible students completed the baseline survey. We randomly assigned 1,286 households (1,379 students) to the treatment group and 1,287 households (1,388 students) to the control group. The stratification was based on (i) whether the student's parent had completed higher education and (ii) the student's region of residence, categorized into five regions: central, eastern, southern, western, and outside Ukraine.

Following random assignment to either the treatment or control group, students were stratified by treatment status, grade level, and preferred tutoring schedule. Within each

[12] Group size was to some extent dependent on students' schedule preferences and tutor availability. As a result, across experiments, 79.2% of groups had 3 students, 19.7% had 1 or 2 students, and 1.1% had 4 or 5 students.

stratum, students were ranked by their total number of correct answers on the baseline short diagnostic assessment and assigned to groups of three students in rank order. As in the first experiment, students in the treatment group were assigned specialized tutors for math and Ukrainian language.

Third experiment. The third experiment's registration process took place between December 2023 and January 2024, following the same procedure as the first two experiments. This experiment introduced two eligibility criteria: only students residing in Ukraine and those who had not participated in the previous two experiments were eligible to enroll. A total of 4,299 eligible guardians and 4,547 eligible students completed the baseline survey, with 2,148 eligible households (2,273 students) randomly assigned to the treatment group and 2,151 eligible households (2,274 students) to the control group. The randomization and grouping of students were identical to the second experiment.

Across all three experiments, the study design included three key features. First, each experiment was conducted independently, with no overlap in student samples. For example, in the second and third experiments, we verified that the national IDs and names of the enrollees did not match those from previous experiments. We also verified that parental information, such as contact details and identifiers, did not overlap across waves. Second, students in the control group were guaranteed a place in the tutoring program at the end of each experiment, after all follow-up data had been collected. Third, all students, regardless of treatment status, were invited to join the Discord online platform, where they could communicate with peers in the same assigned group.[13]

4 Data

4.1 Data Collection

In each experiment, we conducted three rounds of data collection. First, as discussed in Section 3.1, baseline data was collected from parents, students, and tutors during the enrollment stage, before the intervention started in each of the experiments.

[13] We use this feature to test whether simply providing an online space for interaction is sufficient to observe the impacts of the intervention or if structured interaction led by tutors is also necessary.

Second, we administered self-reported online follow-up surveys to students shortly after each experiment ended. The timing of these follow-ups was designed to capture the short-term effects of the tutoring program while minimizing attrition. For the first experiment, which concluded in March 2023, follow-up data from 1,563 students (62% response rate) was collected between March and April 2023. Follow-up data for the second experiment, which ended in June 2023, was collected from 1,368 students (49.4% response rate) between June and July 2023. Follow-up data for the third experiment, which ended in March 2024, was collected in April 2024 from 2,500 students (54% response rate) in April 2024.

To reduce survey fatigue and improve data quality, we sent two links to each student. One link included the Ukrainian language assessment and the other link included the math assessment, mental health module, and the modules that collected data on other secondary outcomes (see more information below). Across all experiments, out of those students who completed at least one assessment, 84.2% completed both assessments while 8.3% completed the math assessment only and 7.5% completed the Ukrainian language assessment only.

Third, to assess implementation fidelity and take up, we collected data on student attendance and engagement during tutoring sessions. This information was gathered through a journal instrument that we developed as well as through student self-reports, the latter included in the follow-up survey.

4.2 Outcomes, Instruments, and Other Variables

The selection of all outcomes was informed by the theory of change of the intervention, as registered in the American Economic Association RCT registry—AEARCTR-0010634. A summary of the outcomes collected in each survey round and experiment is presented in Table A1 and a description of each instrument is presented in Appendix C.

4.2.1 Main outcomes

Academic achievement: We measure academic achievement using students' scores on math and Ukrainian language assessments. In collaboration with McGraw Hill and an assessment specialist, we designed a math assessment instrument consisting of 30 items selected from a validated pool developed by McGraw Hill, ensuring their alignment with the Ukrainian math curriculum for each grade. The Ukrainian language assessment was developed in collaboration with a psychometrician, an assessment expert, and local assessment specialists, also consisting of 30 items based on the Ukrainian curriculum. To estimate academic achievement outcomes, we apply item response theory (IRT) scoring separately for math and Ukrainian language assessments and standardize the scores relative to the control group within each experiment.

Student's mental health: We measure mental distress among students using the Youth Depression, Anxiety, and Stress Scale (DASS-Y) (Szabo and Lovibond, 2022).[14] However, only the two DASS-Y subscales measuring anxiety and stress were used as a means of reducing survey burden and as the program's scope with regard to impacts on depression were limited. To estimate the impact on participants' mental health, we standardize scores separately for stress and anxiety relative to the control group within each experiment. The higher standardized score, the more severe level of stress or anxiety.

4.2.2 Implementation fidelity measures

We collected student-level data through journals kept by tutors, which tracked the attendance and engagement of students during tutoring sessions. For each session, tutors documented whether students attended, if they had their cameras on, participated actively, came prepared, and were paying attention. Additionally, we gathered self-reports from students in the follow-up survey, which provided information on their attendance, tutoring experience, satisfaction, and interactions with peers.

4.2.3 Mechanisms

Interactions with other peers: We collected data on whether students interacted with other peers from different sources. Data on student enrollment in the Discord platform and the number of interactions with others in the platform was obtained from administrative records provided by TFU. We also collected data from an endline survey asking

[14] A key advantage of using DASS-Y is that in addition to assessing perceptions of mental distress, the scale includes items on physiological responses to anxiety and stress, which are less likely to be affected by response bias.

students if at least one of their friends were enrolled in the TFU tutoring program.

Persistence: We measure this trait using eight items of the grit scale from Duckworth and Quinn (2009), which captures students' capacity to maintain effort in the face of obstacles. The survey was administered in each data collection round, with responses recorded on a 1 to 5 scale. Higher scores indicate greater perseverance of effort. The outcome variable is standardized relative to the control group.[15]

Self-efficacy: Self-efficacy refers to students' belief in their ability to achieve specific goals, which has been found to predict better performance in both academic and non-academic domains (Bandura and Wessels, 1997; Schunk and DiBenedetto, 2016, 2022). We assess students' self-efficacy using the General Self-Efficacy Scale (GSE) (Schwarzer and Jerusalem, 1995). This scale measures a general sense of perceived self-efficacy in handling daily stressors and adapting to stressful life events. It consists of 10 items, yielding a final score ranging from 10 to 40, with higher scores indicating greater self-efficacy. The outcome is standardized relative to the control group.

Future educational aspirations: To examine whether tutoring influenced students' higher education aspirations, we asked the students "Thinking about your future, how long do you think you will continue to study?" Response options included "I think I will start working as soon as I complete this school"; "I think I will continue studying and enroll in high school, then start working after obtaining a diploma"; "I think I will continue studying and enroll in a professional and vocational school (e.g., cosmetologist, mechanic), then start working after obtaining a diploma"; "I think I will continue studying and attend university." The outcome is a binary indicator equal to 1 if the student selected either of the last two options, which indicated educational aspirations beyond high school.

Positive attitudes toward academic subjects: To measure students' enthusiasm for math and Ukrainian language, we asked "How much do you like the following subjects?" Responses were recorded on a 1 to 5 Likert scale. The outcome variable is standardized relative to the control group, with higher scores indicating more positive attitudes toward the subject.

[15] We also collected data on locus of control using the instrument from Carlana and La Ferrara (2021), but a validation analysis indicated low construct reliability. Therefore, we do not include estimates using this measure.

4.2.4 Other student, household, and tutors characteristics

We collected sociodemographic information from students and parents, including age, sex, education level, area of residence, and digital skills. Additionally, we measured parents' baseline mental health (stress) using the DASS-21 instrument (Lovibond and Lovibond, 1996). From tutors, we collected sociodemographic characteristics at baseline, prosocial behavior, empathy, and perceptions at both baseline and endline, and attitudes toward tutoring and their self-efficacy related to tutoring at endline.

4.3 Sample Characteristics, Balance Checks, and Survey Attrition

Table 1 reports the mean baseline characteristics of students and households in the control group, disaggregated by experiment (columns [1], [3], and [5]). In all three experiments, 54% of students in the control groups are female, with an average age of 12.6 and enrollment in the 7th grade. Before the tutoring program, most students in the control group were enrolled in a Ukrainian school with either a virtual modality (ranging from 38% in the third experiment to 48% in the second) or an in-person modality (ranging from 32% in the first experiment to 55% in the third). Fewer than 4% of students had previously used TFU services.

Regarding household characteristics, the majority of parents who registered students in the control group were women (92%), with an average age of 39. Parental displacement due to the war was more common in the first two experiments (39% and 43%, respectively) than in the third, where only 26% reported a change in residence. Access to digital devices for remote learning was high across all experiments, with 99% of control group students reporting access to internet-enabled devices (e.g., mobile phones or laptops) at home.

The prevalence of anxiety among students in the control group remained stable across experiments, with up to 43% reporting normal anxiety levels. However, stress levels varied across experiments. While 56% of control group students in the first and second experiments reported stress levels at or below the normal threshold, only 34% in the third experiment exhibited normal stress levels.

Table 1 also presents the differences in mean baseline characteristics between the treatment and control groups for each experiment (columns [2], [4], and [6]). These differences are estimated using a regression of each student or household characteristic on the treatment indicator, including stratification block-fixed effects. Statistically significant differences are observed in some of the variables in the first and third experiments.[16] However, given the p-values from the overall F-test for joint orthogonality, we do not reject the null hypothesis that the mean characteristics of the treatment and control groups are statistically indistinguishable within each experiment. This suggests that the randomization successfully produced comparable treatment and control groups.

We also collected data on tutors characteristics and present these results in Table A2. On average, tutors were 41 years old and 95% of them were women. Half of tutors had completed a bachelor degree and the other half had completed graduate studies (either masters or PhD). Notably, 97% of them have a teaching training. In terms of working experience, they have 19 years of experience in general, and 17 years of teaching experience in particular. Despite 56% of them have done volunteer work in the past, only 18% of them have worked as tutors before. We also find that most of the tutors (81%) have normal levels of stress.

5 Results

5.1 Assessing Feasibility

The feasibility of implementing an online tutoring program amid ongoing instability, population displacement, and frequent power outages was a key concern throughout the three experiments. We assess the feasibility of implementing the tutoring program using measures of take-up, attendance, and engagement.

As shown in Figure 1, both take-up and attendance exceeded our expectations. Between 68% and 72% of students assigned to the treatment group attended at least one math session, while up to 71% attended at least one Ukrainian language session. On av-

[16] In the first experiment, guardians of students in the treatment group are 2 percentage points (pp) more likely to be female than those in the control group. In the third experiment, treatment group students (i) have guardians who are 3 pp less likely to have changed residence during the war and (ii) score 0.13 standard deviations higher on baseline academic assessments relative to the control group.

erage, students participated in more than six sessions (out of 12) per subject across all experiments. Notably, session-level attendance rates showed only a slight decline over the course of the tutoring program in all three experiments (Figure A6). In addition, student engagement during tutoring sessions remained consistently high across all three experiments. Data from tutor journals indicate that between 54% and 64% of students in attendance had their cameras on, 97% to 98% responded to questions, and 95% to 97% appeared attentive. Only 3% to 4% arrived unprepared (Table A3).

Another potential concern regarding the feasibility of the online tutoring program is whether students miss sessions due to the program's structure or perceptions of its quality. To explore this, we asked students in the second and third experiments to report the reasons for missing at least one session. Overall, most absences were attributed to personal or external factors rather than program-related issues. The most common reasons included lack of electricity or internet access (39% in the second experiment and 32% in the third experiment), feeling unwell (24% and 37% in the second and third experiments, respectively), and school responsibilities (30% in the second experiment and 31% in the third experiment) (Figure A7).

5.2 Main Results

Given our randomized experimental design, our analysis follows a straightforward approach. We estimate the intent-to-treat (ITT) effects of the tutoring program within each experiment using the following specification for student i in stratum s in experiment w:

$$Y_{iswt} = \beta_0 + \beta_1 T_i^w + \beta_2 \mathbf{X}_{isw(t-1)} + \gamma_s + \varepsilon_{iswt} \tag{1}$$

where Y_{iswt} represents the outcome variable of interest (for example, math scores) for student i in stratum s in experiment w, measured after the tutoring program ended, i.e., in period t. T_i^w is an indicator variable denoting the assignment of student i's household to the treatment group in the experiment w. To improve precision, we include a vector of baseline control variables, $\mathbf{X}_{isw(t-1)}$, which we select using a Least Absolute Shrinkage and Selection Operator (LASSO) analysis to identify variables with strong predictive relationships with Y_{iswt} (Bruhn and McKenzie, 2009).

We also include stratum fixed effects, γ_s. As described in Section 3.2, in the first experiment, strata are defined by the interaction between whether the student's parent or guardian has completed higher education and whether the student was residing in Ukraine at baseline. In the second and third experiments, strata are defined by the interaction between parental education and region of residence (central, eastern, southern, western, or outside Ukraine). The term $\varepsilon_{i,t}$ represents the idiosyncratic error. Standard errors are clustered at the tutoring group level. The coefficient β_1 provides the ITT estimate of the treatment effect within each experiment. To address concerns related to multiple hypothesis testing, we compute Westfall-Young stepdown adjusted p-values, which control for the Familywise Error Rate (FWER) while allowing for dependence across outcomes (Westfall and Young, 1993).

First experiment. Using Equation (1), we find that the online tutoring program in the first experiment led to substantial improvements in student learning outcomes. As shown in Figure 2 and columns (1) and (2) of Table A4, the program increased students' math scores by 0.49 SD ($p < 0.001$) and Ukrainian language scores by 0.40 SD ($p < 0.001$).[17] These effects translate into learning gains equivalent to 2.5 and 2 years of education, respectively (OECD, 2019). We also examined the online tutoring program's effect on student mental health (stress and anxiety). As shown in Figure 3 and columns (3) and (4) of Table A4, students in the treatment group experienced a reduction in stress levels of nearly 0.10 SD ($p = 0.079$). The estimated effect on anxiety is not statistically significant at conventional levels ($p = 0.384$).

Second experiment. We present the impact of the online tutoring program on learning in the second experiment in Figure 2 and columns (1) and (2) of Table A5. Treated students improved their math scores by 0.22 SD ($p = 0.001$) compared to the control group, equivalent to learning gains of 1 year of education. The effect on Ukrainian language is null. In terms of mental health outcomes, Figure 3 and columns (3) and (4) of Table A5 show that the tutoring program improved students' stress levels by 0.106 SD. Although this estimate is not statistically significant after adjusting for multiple hypotheses ($p = 0.207$), the magnitude of the effect is economically large and similar to the estimated impacts from the other experiments. The estimated effect on anxiety is small and not statistically

[17] In this and the following sections, we present in parentheses the p-values adjusted for multiple hypothesis testing.

significant.

Third experiment. The results on learning from the third experiment are presented in Figure 2 and columns (1) and (2) of Table A6. The findings indicate that treated students improved their math scores by 0.22 SD ($p < 0.001$) and their Ukrainian language scores by 0.32 SD ($p < 0.001$) relative to the control group. These effects correspond to learning gains equivalent to approximately 1.35 and 1.7 years of education, respectively (OECD, 2019). In addition, we also observe improvements in students' mental health. As shown in Figure 3 and columns (3) and (4) of Table A6, treated students experienced a reduction in stress levels of 0.12 SD ($p = 0.004$). Again, we observe a null impact on student anxiety levels.

Discussion of the results across experiments. We compare our results to recent research on the effectiveness of in-person and online tutoring programs in various contexts. While the average effect of in-person tutoring programs is estimated at 0.28 SD (Nickow et al., 2023), two online tutoring programs implemented in Spain and Italy during the COVID-19 pandemic show estimated effects of 0.26 SD (Gortazar et al., 2023) and between 0.20 and 0.23 SD (Carlana and La Ferrara, 2021), respectively.

In our first and third experiments, we observe sizable effects compared to the existing literature on tutoring programs. In the first experiment, the impacts are 1.7 to 2.4 times greater on math, and 1.4 to 2 times greater on Ukrainian language in comparison with the average effects of in-person and online tutoring programs in Spain and Italy implemented during the pandemic. In the third experiment, our impacts are comparable in magnitude to the average effects of in-person tutoring programs and those observed in the online tutoring programs in Spain and Italy. Specifically, the impacts are between 1 to 1.3 times greater than the average effect on math and 1.2 to 1.7 times greater than the average effect on Ukrainian language.

In contrast, the second experiment shows smaller impacts. While the effect size in math is comparable to recent evidence (75% the size of the average effect estimated by Nickow et al. (2023)), the impacts on the Ukrainian language are null. These findings align with existing research indicating that math tutoring tends to yield larger effect sizes in later grades compared to language-related tutoring programs (Nickow et al., 2023).

What explains the smaller estimated effects on learning outcomes in the second experiment relative to the first and third experiments? One possibility is that tutors used the formative assessments as a teaching tool, thereby reducing time for tailored tutoring activities. However, we find this unlikely, as tutors were explicitly instructed that these assessments were optional and intended solely for diagnostic purposes. A more plausible explanation for the lower estimated outcomes in the second experiment is the Ministry of Education and Science's announcement that the state final examinations would not be held that year. This policy change may have lowered treated students' motivation levels as they were entering their final weeks of the school year. This hypothesis is supported by the observed decline in session attendance from the first to the second experiment (Table A3).

To interpret the impacts on standardized measures of stress, we conduct a back-of-the-envelope calculation. A reduction of 0.10 standard deviations in stress levels, as observed in the first experiment, corresponds to a 1-point decrease in the DASS-Y score. Under some assumptions, such as homogeneous effects of the program across all participants, this 1-point reduction would result in approximately 137 students shifting from mild to normal levels of stress, representing 8.8% of all students who completed the endline survey (137 out of 1,562). Applying the same approach to the other two experiments, the estimated share of students moving from mild to normal stress levels would be 7.6% (104/1,368) in the second experiment and 6.7% (165/2,456) in the third experiment.

Finally, we observe no significant impact on anxiety across experiments. One potential explanation is that the support activities focused primarily on stress-coping techniques. According to existing evidence (Edge et al., 2009), stress-coping techniques may affect the implicit dimensions of anxiety but not the explicit symptoms measured by the DASS-Y scale.

Taken together, these findings suggest that online tutoring programs implemented during conflict could be a promising solution to sustaining investments in human capital. When pooling data from all three experiments and estimating the average effect of the tutoring program using Equation (1) with experiment fixed effects, we find that the program improved math scores by 0.32 SD ($p < 0.001$) and Ukrainian language scores by 0.28 SD ($p < 0.001$), corresponding to learning gains of approximately 1.6 and 1.4 years, respectively, at the end of 6 weeks (see column (2) in Table A7). Similarly, the program reduced stress levels by 0.12 SD ($p < 0.001$). These results highlight the effectiveness of a short, targeted education intervention in mitigating both learning losses and mental health challenges for students in conflict-affected settings.

Treatment on the treated (TOT) results: To estimate TOT effects, we define "participation" in the program as having attended at least one tutoring session. These estimated effects are shown in column (3) in Table A7 (pooled) and in Table A8 (by wave). As expected given the high participation in the program, the effects are only slightly higher than the average intent-to-treat estimates for all main outcomes of interest.

In addition, Figure A8 shows the relationship between attendance to subject-specific tutoring sessions and standardized outcomes. Overall, higher student attendance in the tutoring program is associated with greater improvements in math scores across all experiments, which may reflect the more structured and cumulative nature of math instruction that benefits more directly from repeated practice and tutor guidance, and with improved Ukrainian language performance in the first and third experiments.

In terms of mental health outcomes, as shown in Figure A8, we observe that the reductions on stress are associated with attending at least 6 sessions of the program for the first two experiments, likely because stress-coping strategies require repeated exposure to become effective. Students who attended fewer sessions may not have had sufficient exposure to these techniques. Notably, the improvements observed in the third experiment are unconditional to the number of sessions the students attended. The integration of psychosocial elements may have had an immediate, broad-based effect, benefiting even students with lower attendance.

6 Heterogeneity

We test for heterogeneity in the impacts of the intervention by including an interaction term between the treatment indicator and specific student characteristics, baseline outcomes, parental well-being, and conflict intensity in Equation (1). Specifically, we examine the following student characteristics: gender (=1 if the student is female), age (=1 if the student is older than the median age in the sample, i.e., 12 years), academic performance (=1 if the student's baseline scores in math and Ukrainian language are above the median), and mental health (=1 if the student's mental health — stress and anxiety — was at normal levels). For parental well-being, we use an indicator of whether the parent's stress level was at normal levels. For conflict intensity, we include an indicator for high intensity if the number of war incidents in the student's region during the three months preceding the experiments exceeded the median.

The heterogeneity results by student characteristics and baseline outcomes are presented in Figure 4, while those by parental well-being and conflict intensity are shown in Figure 5. These figures display the estimated impacts (relative to the control group) and the associated 95% confidence intervals for our four main outcome indices.

Two caveats are important to note. First, we are underpowered to conduct the heterogeneity analysis separately by experiment. However, since the effects of the tutoring program are fairly consistent across experiments, we pooled the data from all three experiments for this analysis. Second, this heterogeneity analysis was not pre-registered. Nevertheless, given its policy relevance, we include these exploratory findings to inform decisions regarding potential future program roll-out.

Our results suggest that the tutoring program had larger impacts on the learning outcomes of students older than 12 (0.40 SD in math and 0.37 SD in Ukrainian language) and for those who were underperforming at baseline (0.36 SD in math and 0.34 SD in Ukrainian language) (Figure 4). Older students may have benefited more from the program as they are more independent and may require less support for learning activities, which enables them to engage more effectively with the tutoring program's format. In addition, low-performing students who may have struggled with traditional instructional methods and materials during regular schooling may have benefited more from the tutoring activities, which likely helped to reinforce concepts and topics taught by their teachers. The differential impacts by baseline performance combined with the absence of differences by gender align with findings from other online tutoring programs, including Italy's Tutoring Online Program by Carlana and La Ferrara (2021). In contrast, we find no evidence that these student characteristics drive differences in mental health outcomes observed in Section 5.

In addition, our findings show that parental stress and conflict intensity influence both academic performance and mental health outcomes. As illustrated in Figure 5, children of parents with high stress levels achieved higher scores in math (0.36 SD) and Ukrainian language (0.33 SD) compared to other treated children whose parents had normal stress levels. In addition, children exposed to low levels of conflict demonstrated greater gains in math (0.38 SD) and Ukrainian language performance (0.37 SD) compared to those exposed to high conflict. A potential explanation for this is that high conflict exposure might distract participants or exacerbate connectivity issues during tutoring sessions.

These findings suggest that the tutoring program's impacts are influenced not only by student characteristics, such as age and baseline performance, but also by external factors, including parental stress and conflict exposure. This underscores the importance of designing and implementing tutoring programs that are responsive to individual and contextual differences, particularly in conflict settings, to maximize both academic and mental health benefits.

7 Mechanisms

In this section, we examine the mechanisms through which the online tutoring program may have improved learning and mental health outcomes. First, we assess its impact on non-academic outcomes—including peer interactions, attitudes and aspirations, and social-emotional skills and well-being—to explore whether these factors drive improvements in academic performance and mental health outcomes. Second, we leverage a fourth experiment conducted at the end of the first experiment to provide causal evidence on the role of parental investment.

7.1 Structured Peer Interactions

A potential mechanism through which tutoring improved academic and mental health outcomes is by fostering interactions among students facing similar challenges, such as prolonged disruptions to their education due to the war. Social interactions are known to play a critical role in the emotional well-being and learning of students, particularly in crisis settings, where peer support can mitigate the adverse effects of stress and isolation (Bohren et al., 2019; Paluck et al., 2016; Boursier et al., 2023). Prior research suggests that structured peer interactions, particularly when facilitated by an instructor, enhance learning outcomes by promoting engagement, collaborative problem-solving, and motivation (Roseth et al., 2008; Slavich and Zimbardo, 2012).

A key feature of our design is that all students, including those in the control group, had the opportunity to enroll in the Discord platform. Thus, any increase in interactions was not solely driven by providing an online space for engagement but also by the program's inclusion of tutors who actively fostered peer interaction.

To formally test this mechanism, we measure students' interactions with peers assigned to their tutoring groups. The results from all three experiments, presented in Table 2, indicate that although both treatment and control students had access to the program's online platform for peer interaction, those in the treatment group—who were also assigned to tutors—experienced significantly higher engagement. Specifically, across experiments, they were 14 to 42 pp more likely to enroll in the platform, 22 to 48 pp more likely to interact with others outside of tutoring sessions, and 9 to 11 pp more likely to have more than ten interactions during the program. Moreover, the tutoring program strengthened peer support for academic activities, with treatment group students being 8 to 13 pp more likely to report that their friends were also enrolled in the program.

These findings suggest that structured, tutor-facilitated engagement was a key driver of peer interaction, beyond mere access to an online space. The results highlight the potential of integrating social interaction components into remote learning interventions to foster both academic and psychosocial benefits, particularly in crisis settings.

7.2 Attitudes Toward Learning and Educational Aspirations

A key determinant of academic success is students' attitudes toward learning and their educational aspirations. Prior research suggests that fostering positive attitudes toward academic subjects can enhance student engagement, motivation, and long-term educational attainment (Khattab, 2015). Similarly, students with higher educational aspirations tend to exert greater effort in their studies, leading to improved academic performance (Jacob and Wilder, 2010; Carlana and La Ferrara, 2021; Golan and You, 2021). Structured learning environments, particularly those incorporating small-group attention and mentorship—such as tutoring—have been shown to positively influence both attitudes toward learning and students' expectations for their educational futures (Carlana and La Ferrara, 2021). Given these findings, we explore whether the tutoring program influenced student outcomes by shaping their enthusiasm for learning and educational goals.

We present the results in Table 3. Across experiments, we find positive and statistically significant effects of the tutoring program on improving students' positive attitudes for both academic subjects. This suggests that the intervention may have strengthened students' intrinsic motivation to engage with the material, a factor that has been linked to improved learning outcomes. However, in terms of future aspirations, we find only small and mixed effects. In the first experiment, there is a small, statistically significant increase in the aspiration to work, while in the second experiment, we observe a reduction in the probability of students wishing to pursue higher education. However, neither of these results remain statistically significant after adjusting for multiple hypothesis testing.

These findings suggest that the tutoring program's impact on academic performance is more likely to operate through increased enthusiasm for learning rather than shifts in long-term aspirations. In the context of war, students may be more focused on short-term survival and adapting to immediate disruptions in their education rather than planning for their long-term educational trajectories.

7.3 Social-Emotional Skills

Another potential channel through which the tutoring program may influence educational outcomes is by improving students' social-emotional skills—specifically their persistence and self-efficacy— as well as their general well-being. Social-emotional skills including the ability to persevere in the face of challenges and believing in one's ability to succeed have been linked to both academic achievement and mental health improvements, particularly in high-stress environments like conflict zones. Research in both fields of psychology and economics has shown that students who develop higher levels of social-emotional skills tend to have better academic performance and greater resilience in the face of adversity. For instance, social-emotional skills such as perseverance and emotional regulation can help students cope with stress and persist through challenges (Dinarte-Diaz and Egana-delSol, 2024), ultimately enhancing their educational outcomes.

As shown in Table 4, we find a positive effect on both persistence and self-efficacy measures in the second experiment, with an estimated improvement of 0.11 SD. In the third experiment, the impact on social-emotional skills was more pronounced. We found larger effects, with grit improving by 0.34 SD and self-efficacy by 0.31 SD (see Table 4).

These results suggest that the tutoring program may have contributed to students' ability to persevere and develop a stronger belief in their own capabilities—-key components of social-emotional well-being. These changes in social-emotional skills may have translated into better academic performance and mental health outcomes.

7.4 Complementary Student and Parental Investments

We also test for whether complementary student and parental investments are channels through which academic achievement and mental health improved.

Student investments: Using data from all three experiments, we examine whether students independently increased their learning investments as a result of the tutoring program. Specifically, we asked students from both the treatment and control groups whether they had received any additional tutoring or subject-specific support outside of the online tutoring program offered by TFU in the previous six weeks.[18] As shown in Table 5, students in the treatment group were 21 to 33 pp more likely to seek additional tutoring support outside the program, suggesting that the intervention promoted additional self-driven learning investments.

Parental investments: To provide rigorous evidence for the effects of parental investments, we conducted a parallel experiment at the same time as the second experiment. We designed an information intervention aimed at encouraging parents to support their children's participation in the tutoring program, with the goal of evaluating whether this intervention could positively impact take-up, academic performance, and mental health outcomes. We conducted a stratified randomization and randomly assigned a subsample of 743 households (797 students) who were not treated in the first experiment (either because they were randomly assigned to the control group or to a waitlist) to two treatment groups with equal probability.[19] The strata were defined using geographic region and an indicator of whether the student was in the control group in the first wave (=0 if student was in the waitlist).The first group was offered the opportunity to participate in the

[18] The online tutoring program was referred as "Education Soup" by Teach for Ukraine and consisted of the full tutoring program for the treatment group and access to the Discord platform for the control group.

[19] We invited all students in the control group from the first experiment (1,259 students) as well as 235 students from the waitlist to participate in this additional experiment. Only 681 from the control group and 116 from the waitlist consented to participate.

tutoring program as delivered during the first experiment. The second group was also offered the opportunity to participate in the tutoring program. In addition, parents of students in this group received one text message per week for six weeks, each containing motivational messages for parents to engage and support their children's participation in the tutoring program.

The text messages, based on successful behavioral interventions (Robinson et al., 2022; Calzolari and Nardotto, 2017; Karlan et al., 2016; Allcott and Rogers, 2014; Gerber et al., 2008), focused on three main strategies: reminders (for example, "How is [student's name] doing? Remind [student's name] to log-in to the tutoring session for help with his/her school work and wellbeing."), social norms (for example, "Many students are getting tutoring support this week! Encourage [student's name] to join his/her classmates in the next session."), and accountability (for example: "Make sure [student's name] attends the next tutoring session. The tutor will cover content that will help [student's name] at school!").

We collected baseline data during student registration for the first experiment (7 to 8 weeks before the tutoring program started for this fourth experiment) and follow-up data after the six-week tutoring program (along with students participating in the second experiment). As shown in Table A9, there were no significant differences in follow-up survey completion rates between the two experimental groups, confirming that attrition is unlikely to bias our estimates.

To estimate the impact of parental engagement through text messages, we modify equation (1) as follows:

$$Y_{iswt} = \beta_0 + \beta_1 \text{Text}_i + \beta_2 X_{isw(t-1)} + \gamma_s + \varepsilon_{iswt} \tag{2}$$

where Text_i is an indicator for students whose parents received text messages in addition to the tutoring program. Control variables in vector $X_{isw(t-1))}$ were selected using a double LASSO procedure. We also include strata fixed effects using three variables: whether the guardian has higher education, region of residency, and whether the student was randomly assigned to the control or waitlist in the first experiment. All other variables remain as previously defined. The coefficient β_1 captures the ITT effect of parental engagement through text messages.

Table 6 presents the results. We find that students whose parents received text messages performed worse in both math (-0.15 SD) and Ukrainian language (-0.23 SD) and presented worse mental health outcomes (although not statistically significant).[20] These results are consistent with Robinson et al. (2022), who found that a similar parental communication strategy did not improve academic performance (despite increasing tutoring take-up in their study). Anecdotal evidence from conversations with parents and the TFU team suggests that parents in our study did not welcome the additional reminders to support their children, which made them feel pressured and led to dissatisfaction with the program.

These findings highlight the complexities of complementary investments, such as student investments and parental engagement, as mechanisms for the observed impacts, particularly in war-affected settings. While behavioral nudges to parents may work in some contexts, in crisis situations, they may inadvertently lead to disengagement and negatively impact children's academic achievement and well-being.

8 Robustness Checks

8.1 Exploring Potential Bias Due to Differential Attrition

As we mentioned in the 4.1 section, we collected endline data in two consecutive rounds to reduce survey fatigue. In one survey, we included the math assessment and the survey modules with the other outcomes, and in the second survey, we collected Ukrainian language scores. We assess differential attrition between the treatment and control groups in each of the follow-up survey rounds, with results presented in Table A10. We find no statistically significant differences in survey completion rates between the treatment and control groups in the first and third experiments after accounting for multiple hypothesis. However, in the second experiment, students in the treatment group were 8 pp more likely to complete the first round and 10.8 pp more likely to complete the second round. A potential concern is that this differential attrition could bias our estimates. Specifically,

[20] In terms of average attendance to the tutoring sessions, we find no significant differences between the two groups. Students in the tutoring only treatment attended 7.1 math sessions and 7.2 Ukrainian language sessions. Similarly, students in the tutoring and text messages treatment attended 7.2 math sessions and 7.3 Ukrainian language sessions.

if the attrition rate in the treatment group is correlated with worse academic performance or higher levels of mental distress, the treatment effect estimates may be upward biased.

To address this concern, we conduct a bounds analysis following the approach of Fairlie et al. (2015), using various assumptions about the treatment effects for attritors. Columns (1), (4), and (7) of Table A11 reproduce the relevant average treatment effect estimates for each experiment from Tables A4 to A6. For each experiment, we impute to the lower (upper) bound the mean minus (plus) 5% of the SD of the observed treatment group distribution of attritors in the treatment group, and the mean plus (minus) the same SD multiple of the observed control group distribution for attritors in the control group.[21]

We find that all of the estimated effects presented in Figures 2 and 3 (and Tables A4 to A6) are robust with respect to the imposition of these bounds, showing that the treatment effects are not sensitive to adding or subtracting 5% of the standard deviation from the means.

8.2 Assessing Experimenter Demand due to Self-Reports

Given the war-affected context and the need to collect data efficiently, we chose to measure mental health outcomes using self-report questionnaires administered through the online survey. This method was the most feasible and cost-effective approach given the large sample size and the constraints of the situation. While self-reports can be prone to experimenter demand effects, we believe this concern is mitigated for at least three reasons. First, the DASS-Y includes items that assess the physiological manifestations of stress and anxiety (e.g., "I was easily annoyed" or "my hands felt shaky"), rather than directly asking subjective questions like "Do you feel stressed?" or "Do you feel anxious?" This indirect approach makes the responses less susceptible to experimenter demand effects, as students are not explicitly asked to self-diagnose their emotional state.

Second, the primary content of the tutoring program was academic, and while mental health activities were included in the third wave, these activities were designed to complement the academic focus rather than dominate the program's structure.

[21] Given the large magnitude of the standard deviations of the observed distributions of our main outcomes, we argue that estimating bounds by adding (subtracting) 5% of the standard deviation represents a conservative scenario. For example, subtracting 5% of the standard deviation from the math outcome in the second experiment creates a lower bound that is nearly 60% of the original estimate.

In addition, tutors were able to report their assessment on their students' well-being through the tutor journal, providing an additional, external measure of the students' mental health and emotional state. This triangulation between self-reports and tutor reports strengthens the validity of our mental health outcomes. Indeed, using pooled data across experiments, we find negative correlations of 0.26 ($p < 0.1$) and 0.43 ($p < 0.01$) between students' endline stress and anxiety measures (self-reported using DASS-Y) and the tutors' observations of whether the students "seemed happy, relaxed, or calmed" during the final week of the program, respectively. This suggests that the self-reported data are consistent with the external observations of well-being.

9 Cost-Effectiveness

We conduct a cost-effectiveness analysis by comparing the costs and economic benefits of the online tutoring program. The economic benefits are estimated under the assumption that improvements in learning and mental health from program participation will translate into increased productivity and higher earnings in the labor market. However, this approach has its limitations, as it relies on strong assumptions about future earnings, which are particularly uncertain in the context of an ongoing war. Ukraine's labor market is currently experiencing significant disruptions, including labor shortages due to internal and external migration, skill mismatches, and elevated unemployment compared to pre-war levels. However, the tight labor market is driving wage growth and boosting consumption, which is expected to continue over the forecast horizon despite economic uncertainties (Ministry of Finance, 2024; National Bank of Ukraine, 2024).

We adopt a conservative approach following projections made by the Ukrainian government and present our assumptions for the cost-benefit analysis of the online tutoring program in Table 7. We assume a labor force participation rate of 56%,[22] average annual earning of $6,124 for 2024,[23] and annual real wage growth at 2.9%.[24] The working age is

[22] Labor force participation rates are based on World Development Indicators data (2010-2021, modeled ILO estimate). Surveys by Info Sapiens (2021-2024), cited in National Bank of Ukraine (2024), indicate that participation rates in 2024 remain close to the 2010-2021 average.

[23] The Ministry of Economy's "Recovery Growth Scenarios Report 2024-2027" estimates the average annual earnings at $6,124 for 2024, based on the forecasted exchange rate of 40.33 Ukrainian hryvnia/$)

[24] National Bank of Ukraine's forecast for 2026.

defined as 22 to 65 years, which translates into a 43-year period of labor market participation. A discount rate of 5% is assumed in the baseline scenario. While most education projects use a 3% discount rate, this higher discount rate reflects the uncertainty of future benefits of the program given the context of war. Together, these parameters suggest that the average present value (PV) of total future earnings for participants of the tutoring program (across all treatments) is $147,060 (Panel A in Table 7).

The analysis of expected increases in future earnings associated with the tutoring program relies on existing literature linking earnings to learning and mental health.[25] For learning, we rely on the study conducted by Hanushek et al. (2015) where the the OECD's Programme for the International Assessment of Adult Competencies (PIAAC) survey of adult skills in 23 countries was used to show that, on average, a 1 SD increase in numeracy skills is associated with a 18% wage increase among prime-age workers, and a 10% wage increase after controlling for years of schooling. In Eastern European countries, such as Poland, Slovakia, and the Czech Republic, returns are shown to be around 7.1% to 8.6% after controlling for years of schooling. Thus, we assume a 8% return on earnings for a 1 SD increase in learning in the baseline scenario in Ukraine. Similarly, for mental health, we rely on two meta-analysis (Cabus et al., 2021; Vella, 2024) looking at social-emotional skills in 15 and 52 studies, respectively, including the relationship between emotional stability (fear, worry, paranoia, and stress) and earnings. These studies conclude that the earnings returns to improved emotional stability are approximately 1.6 to 1.8%. Thus, we assume a 1.7% return on earnings for a 1 SD improvement in mental health in the baseline scenario.

We use a costing tool developed by World Bank's Strategic Impact Evaluation Fund to calculate the costs of the program. This tool is designed to capture detailed (disaggregated) listing and valuing of all resources and efforts required to implement a remote instruction program. The cost of the interventions totals $523,202.4. The average cost per targeted participant in the treatment group is relatively similar across experiments: $89.12 in the first experiment, $92.74 in the second experiment, and $93.41 in the third

[25] Multiple studies show that a 1 SD increase in test scores is associated with higher annual earnings, though the effect varies by country and career stage. Hanushek et al. (2015) notes that in the US, a 1 SD increase in high school math performance correlates with a 10-15% rise in early-career earnings, potentially reaching 20% over a lifetime. This effect is likely higher in the US than in lower- and middle-income countries (Hanushek, 2015; Mulligan, 1999; Murnane et al., 2000; Hanushek and Zhang, 2009).

experiment. For the control group, the average cost per participant in all experiments is $7.04. The incremental cost associated with the delivery of the treatment ranges from $82.09 (experiment 1) to $86.37 (experiment 3). The cost per standard deviation of improvement in learning is equivalent to $204.19 for the first experiment, $384.66 for the second experiment, and $394.38 for the third experiment (Panel C in Table 7).

In the baseline scenario, the comparison between the costs and the expected increase in salaries expressed in 2024 prices) from participation in the program and the costs, shows that the tutoring program is highly cost-effective. Using a 5% discount rate, the base case scenario shows substantial benefits, with a benefit-to-cost ratio of 56.4 for the first experiment, 31.4 for the second experiment, and 32.9 for the third experiment (Panel D in Table 7). Thus, the benefits of the programs significantly outweigh its costs in each experiment.

We assess the sensitivity of the estimated benefit-cost ratio by varying the unknown parameters in Table 7, following the approach adopted by Ganimian et al. (2024). The unknown parameters correspond to wage growth, the discount rate, and the earnings premium associated with learning and mental health. Each unknown parameter is extracted from a distribution defined over a range of possible values, with the values from Table 7 positioned at the center. It is assumed that discount rates range from 3% to 7%, that wage growth rates ranged from 1% to 5%, that earnings gains from improved learning from 5% to 11%, and earnings gains from improved mental health ranged from 0.5% to 3%. The parameters were randomly selected from two types of distributions: a uniform distribution, where all parameters within the range have an equal probability, and a truncated normal distribution, where values near the midpoint have a higher probability. The sensitivity analysis reveals that all interventions are highly cost-effective (Figure A9).

Finally, we estimate the marginal value of public funds (MVPF), following Hendren and Sprung-Keyser (2020). The MVPF measures the after-tax benefits that participants receive for each dollar spent by the government. Although the online tutoring program was implemented with donor funding, had the program been funded by the government, the MVPF of the first experiment would be 1.77% (112.1/4,380.0 x 100), 3.18% for the second experiment, and 3.04% for the third experiment. That is, government costs could be fully offset by future tax revenue, as long as program participants pay a net tax rate of at least 1.77% to 3.18% on their future earnings. This suggests that each program offers a significant return on investment in future tax revenue.

10 Conclusion

In this paper, we present robust evidence from three experiments that demonstrated the benefits of investing in education during wartime is important. In each experiment, demand for the online tutoring program was high, and take-up and attendance exceeded expectations. The intervention in each experiment was tailored to address the unique challenges faced by students, teachers, and the education system during distinct phases of the full-scale Russian invasion. While the effects differ across outcomes, overall, we show that the positive impacts of the online tutoring program on learning and mental health are sizable and mostly consistent across all three experiments. In addition, the interventions enhance peer support, foster positive learning attitudes and investments, and develop social-emotional skills, contributing to both academic success and improved mental health.

From a policy point of view, the ratios of benefits to costs from these interventions are considerable and support the argument that human capital should not be discounted as an important priority for investment during wartime. One important takeaway is that dedicated non-governmental organizations can effectively supplement public education, leveraging technology to facilitate peer interactions, improve attitudes toward learning, and develop socio-emotional skills among students. As Ukraine rebuilds its education system, such education programs can support children at a reasonable cost, ensuring a high level of accessibility despite of the challenging external conditions.

Future research should continue to engage with organizations on the ground to evaluate promising interventions, learn from their implementation, and assess their long-term impact, ensuring that efforts to build human capital persist even amidst conflict.

References

Acosta, Pablo, Javier E Baez, German Caruso, and Carlos Carcach, *The Scars of Civil War: The Long-Term Welfare Effects of the Salvadoran Armed Conflict*, World Bank Group, Poverty and Equity Global Practice, 2020. (Cited on 6)

Al-Ubaydli, Omar, John A List, and Dana L Suskind, "What can we learn from experiments? Understanding the threats to the scalability of experimental results," *American Economic Review*, 2017, *107* (5), 282–286. (Cited on 5)

Allcott, Hunt and Todd Rogers, "The short-run and long-run effects of behavioral interventions: Experimental evidence from energy conservation," *American Economic Review*, 2014, *104* (10), 3003–3037. (Cited on 29)

Angrist, Noam, Micheal Ainomugisha, Sai Pramod Bathena, Peter Bergman, Colin Crossley, Claire Cullen, Thato Letsomo, Moitshepi Matsheng, Rene Marlon Panti, Shwetlena Sabarwal et al., "Building Resilient Education Systems: Evidence from Large-Scale Randomized Trials in Five Countries," Technical Report, National Bureau of Economic Research 2023. (Cited on 6)

_ , Peter Bergman, and Moitshepi Matsheng, "School's out: Experimental evidence on limiting learning loss using "low-tech" in a pandemic," Technical Report, National Bureau of Economic Research 2020. (Cited on 6)

_ , _ , Caton Brewster, and Moitshepi Matsheng, "Stemming learning loss during the pandemic: A rapid randomized trial of a low-tech intervention in Botswana," *Available at SSRN 3663098*, 2020. (Cited on 6)

_ , Simeon Djankov, Pinelopi Goldberg, and Harry Patrinos, "The loss of human capital in Ukraine," *Global Economic Consequences of the War in Ukraine Sanctions, Supply Chains and Sustainability*, 2022, *26.* (Cited on 6)

Arrazola, María and Jose de Hevia, "Three measures of returns to education: An illustration for the case of Spain," *Economics of Education Review*, 2008, *27* (3), 266–275. (Cited on 6)

Asiedu, Edward, Dean Karlan, Monica Lambon-Quayefio, and Christopher Udry, "A call for structured ethics appendices in social science papers," *Proceedings of the National Academy of Sciences*, 2021, *118* (29), e2024570118. (Cited on 56)

Avvisati, Francesco and Pauline Givord, "How much do 15-year-olds learn over one year of schooling? An international comparison based on PISA," *OECD Education Working Papers*, 2021, (257). (Cited on 4)

Bandura, Albert and Sebastian Wessels, *Self-efficacy*, Cambridge University Press Cambridge, 1997. (Cited on 16)

Banerjee, Abhijit, Rukmini Banerji, James Berry, Esther Duflo, Harini Kannan, S Mukherji, and Michael Walton, "Teaching at the right level: Evidence from randomized evaluations in India," *NBER Working Paper*, 2015, *22746*, 2369–2429. (Cited on 7)

—, —, —, —, —, Shobhini Mukerji, Marc Shotland, and Michael Walton, "From proof of concept to scalable policies: Challenges and solutions, with an application," *Journal of Economic Perspectives*, 2017, *31* (4), 73–102. (Cited on 5, 7)

—, —, —, —, —, Shobhini Mukherji, Marc Shotland, and Michael Walton, "Mainstreaming an effective intervention: Evidence from randomized evaluations of "Teaching at the Right Level" in India," Technical Report, National Bureau of Economic Research 2016. (Cited on 7)

Banerjee, Abhijit V, Rukmini Banerji, Esther Duflo, Rachel Glennerster, and Stuti Khemani, "Pitfalls of participatory programs: Evidence from a randomized evaluation in education in India," *American Economic Journal: Economic Policy*, 2010, *2* (1), 1–30. (Cited on 7)

—, Shawn Cole, Esther Duflo, and Leigh Linden, "Remedying education: Evidence from two randomized experiments in India," *The quarterly journal of economics*, 2007, *122* (3), 1235–1264. (Cited on 7)

Banerji, Rukmini and Madhav Chavan, "Improving literacy and math instruction at scale in India's primary schools: The case of Pratham's Read India program," *Journal of Educational Change*, 2016, *17* (4), 453–475. (Cited on 7)

Battistin, Erich and Antonio Schizzerotto, "Threat of grade retention, remedial education and student achievement: evidence from upper secondary schools in Italy," *Empirical Economics*, 2019, *56*, 651–678. (Cited on 7)

Bettinger, Eric, Robert W Fairlie, Anastasia Kapuza, Elena Kardanova, Prashant Loyalka, and Andrey Zakharov, "Does EdTech Substitute for Traditional Learning? Experimental Estimates of the Educational Production Function. NBER Working Paper No. 26967.," *National Bureau of Economic Research*, 2020. (Cited on 6)

Blattman, Christopher and Edward Miguel, "Civil war," *Journal of Economic literature*, 2010, *48* (1), 3–57. (Cited on 2)

Bohren, J Aislinn, Alex Imas, and Michael Rosenberg, "The dynamics of discrimination: Theory and evidence," *American economic review*, 2019, *109* (10), 3395–3436. (Cited on 25)

Boursier, Valentina, Francesca Gioia, Alessandro Musetti, and Adriano Schimmenti, "COVID-19-related fears, stress and depression in adolescents: The role of loneliness and relational closeness to online friends," *Journal of Human Behavior in the Social Environment*, 2023, *33* (2), 296–318. (Cited on 25)

Bruhn, Miriam and David McKenzie, "In Pursuit of Balance: Randomization in Practice in Development Field Experiments," *American Economic Journal: Applied Economics*, 2009, *1* (4), 200–232. (Cited on 19)

Cabus, Sofie, Joanna Napierala, and Stephanie Carretero, "The returns to non-cognitive skills: A meta-analysis," Technical Report, JRC Working Papers Series on Labour, Education and Technology 2021. (Cited on 33)

Calzolari, Giacomo and Mattia Nardotto, "Effective reminders," *Management Science*, 2017, *63* (9), 2915–2932. (Cited on 29)

Carlana, Michela and Eliana La Ferrara, "Apart but connected: Online tutoring and student outcomes during the COVID-19 pandemic," *CEPR Discussion Paper No. DP15761*, 2021. (Cited on 7, 16, 21, 24, 26)

Collier, Paul, "On the economic consequences of civil war," *Oxford economic papers*, 1999, *51* (1), 168–183. (Cited on 2)

Crowne, Douglas P and David Marlowe, "A new scale of social desirability independent of psychopathology.," *Journal of consulting psychology*, 1960, *24* (4), 349. (Cited on 75)

Dinarte-Diaz, Lelys and Pablo Egana-delSol, "Preventing violence in the most violent contexts: Behavioral and neurophysiological evidence from el salvador," *Journal of the European Economic Association*, 2024, *22* (3), 1367–1406. (Cited on 27)

Duckworth, Angela Lee and Patrick D. Quinn, "Development and Validation of the Short Grit Scale (Grit–S)," *Journal of Personality Assessment*, feb 2009, *91* (2), 166–174. (Cited on 16, 52, 60)

Eder, Christoph, "Displacement and education of the next generation: evidence from Bosnia and Herzegovina," *IZA Journal of Labor & Development*, 2014, *3* (1), 1–24. (Cited on 6)

Edge, Michael D, Wiveka Ramel, Emily M Drabant, Janice R Kuo, Karen J Parker, and James J Gross, "For better or worse? Stress inoculation effects for implicit but not explicit anxiety," *Depression and anxiety*, 2009, *26* (9), 831–837. (Cited on 22)

Fairlie, Robert W, Dean Karlan, and Jonathan Zinman, "Behind the GATE experiment: Evidence on effects of and rationales for subsidized entrepreneurship training," *American Economic Journal: Economic Policy*, 2015, *7* (2), 125–161. (Cited on 31, 84)

Ganimian, Alejandro J., Karthik Muralidharan, and Christopher R. Walters, "Augmenting State Capacity for Child Development: Experimental Evidence from India," *Journal of Political Economy*, 2024, *132* (5), 1565–1602. (Cited on 34, 55, 73)

Gerber, Alan S, Donald P Green, and Christopher W Larimer, "Social pressure and voter turnout: Evidence from a large-scale field experiment," *American political Science review*, 2008, *102* (1), 33–48. (Cited on 29)

Golan, Jennifer and Jing You, "Raising aspirations of boys and girls through role models: Evidence from a field experiment," *The Journal of Development Studies*, 2021, *57* (6), 949–979. (Cited on 26)

Gortazar, Lucas, Claudia Hupkau, and Antonio Roldán, "Online tutoring works: Experimental evidence from a program with vulnerable children," *Available at SSRN 4390248*, 2023. (Cited on 7, 21)

Goto, Ryunosuke, Irina Pinchuk, Oleksiy Kolodezhny, Nataliia Pimenova, Yukiko Kano, and Norbert Skokauskas, "Mental Health of Adolescents Exposed to the War in Ukraine," *JAMA pediatrics*, 2024, *178* (5), 480–488. (Cited on 10)

Gutiérrez, Emilio and Rodimiro Rodrigo, "Closing the achievement gap in mathematics: evidence from a remedial program in Mexico City," *Latin American Economic Review*, 2014, *23*, 1–30. (Cited on 7)

Hanushek, Eric A, "Economics of Education," Technical Report, Hoover Institution 2015. (Cited on 33)

_ **and Lei Zhang**, "Quality-consistent estimates of international schooling and skill gradients," *Journal of Human Capital*, 2009, *3* (2), 107–143. (Cited on 33)

_ **, Guido Schwerdt, Simon Wiederhold, and Ludger Woessmann**, "Returns to skills around the world: Evidence from PIAAC," *European Economic Review*, 2015, *73*, 103–130. (Cited on 33)

Hassan, Hashibul, Asad Islam, Abu Siddique, and Liang Choon Wang, "Emotional and Behavioral Impacts of Telementoring and Homeschooling Support on Children," in "AEA Papers and Proceedings," Vol. 113 American Economic Association 2023, pp. 498–502. (Cited on 6)

Hendren, Nathaniel and Ben Sprung-Keyser, "A Unified Welfare Analysis of Government Policies*," *The Quarterly Journal of Economics*, 03 2020, *135* (3), 1209–1318. (Cited on 34)

Hevia, M. Risueño M. Arrazola J. De and J. F. Sanz, "Returns to education in Spain: Some evidence on the endogeneity of schooling," *Education Economics*, 2003, *11* (3), 293–304. (Cited on 6)

Holla, Alaka, Magdalena Bendini, Lelys Dinarte-Diaz, and Iva Trako, "Is Investment in Preprimary Education Too Low? Lessons from (Quasi) Experimental Evidence across Countries," *World Bank Policy Research Working Paper Series*, 2021, *Working Paper No. 9723.* (Cited on 55)

Ichino, Andrea and Rudolf Winter-Ebmer, "The long-run educational cost of World War II," *Journal of Labor Economics*, 2004, *22* (1), 57–87. (Cited on 6)

Jacob, Brian A and Lars Lefgren, "Remedial education and student achievement: A regression-discontinuity analysis," *Review of economics and statistics*, 2004, *86* (1), 226–244. (Cited on 7)

_ **and Tamara Wilder**, "Educational expectations and attainment," Technical Report, National Bureau of Economic Research 2010. (Cited on 26)

Karlan, Dean, Margaret McConnell, Sendhil Mullainathan, and Jonathan Zinman, "Getting to the top of mind: How reminders increase saving," *Management science*, 2016, *62* (12), 3393–3411. (Cited on 29)

Khattab, Nabil, "Students' aspirations, expectations and school achievement: What really matters?," *British educational research journal*, 2015, *41* (5), 731–748. (Cited on 26)

Lai, Brian and Clayton Thyne, "The effect of civil war on education, 1980—97," *Journal of peace research*, 2007, *44* (3), 277–292. (Cited on 6)

Lakshminarayana, Rashmi, Alex Eble, Preetha Bhakta, Chris Frost, Peter Boone, Diana Elbourne, and Vera Mann, "The Support to Rural India's Public Education System (STRIPES) trial: A cluster randomised controlled trial of supplementary teaching, learning material and material support," *PloS one*, 2013, *8* (7), e65775. (Cited on 7)

Lavy, Victor and Analia Schlosser, "Targeted remedial education for underperforming teenagers: Costs and benefits," *Journal of Labor Economics*, 2005, *23* (4), 839–874. (Cited on 7)

_, Assaf Kott, and Genia Rachkovski, "Does remedial education in late childhood pay off after all? Long-run consequences for university schooling, labor market outcomes, and intergenerational mobility," *Journal of Labor Economics*, 2022, *40* (1), 239–282. (Cited on 7)

Leon, Gianmarco, "Civil conflict and human capital accumulation: The long-term effects of political violence in Perú," *Journal of Human Resources*, 2012, *47* (4), 991–1022. (Cited on 6)

List, John A, *The Voltage Effect: How to Make Good Ideas Great and Great Ideas Scale*, Currency, 2022. (Cited on 5, 6)

Lovibond, Sydney H and Peter F Lovibond, *Manual for the depression anxiety stress scales*, Psychology Foundation of Australia, 1996. (Cited on 17, 48)

Ministry of Finance, "Ukraine: Recovery Growth Scenarios - Consensus Forecast," Technical Report, Department of Strategic Planning and Macroeconomic Forecasting 2024. (Cited on 32)

Mobarak, Ahmed Mushfiq, "Assessing social aid: the scale-up process needs evidence, too," *Nature*, 2022, *609* (7929), 892–894. (Cited on 5)

Mulligan, Casey B, "Galton versus the human capital approach to inheritance," *Journal of political Economy*, 1999, *107* (S6), S184–S224. (Cited on 33)

Muralidharan, Karthik and Paul Niehaus, "Experimentation at Scale," *Journal of Economic Perspectives*, November 2017, *31* (4), 103–24. (Cited on 5)

Murnane, Richard J, John B Willett, Yves Duhaldeborde, and John H Tyler, "How important are the cognitive skills of teenagers in predicting subsequent earnings?," *Journal of Policy Analysis and Management*, 2000, *19* (4), 547–568. (Cited on 33)

National Bank of Ukraine, "Inflation Report: July 2024," Technical Report, National Bank of Ukraine 2024. (Cited on 32)

Nickow, Andre, Philip Oreopoulos, and Vincent Quan, "The Promise of Tutoring for PreK—12 Learning: A Systematic Review and Meta-Analysis of the Experimental Evidence," *American Educational Research Journal*, 2023, *61* (1), 74–107. (Cited on 21)

OECD, "What Students Know and Can Do," Technical Report, OECD 2019. (Cited on 20, 21)

Paluck, Elizabeth Levy, Hana Shepherd, and Peter M Aronow, "Changing climates of conflict: A social network experiment in 56 schools," *Proceedings of the National Academy of Sciences*, 2016, *113* (3), 566–571. (Cited on 25)

Patrinos, Harry Anthony, Emiliana Vegas, and Rohan Carter-Rau, "An analysis of COVID-19 student learning loss," Technical Report 2022. (Cited on 6)

Robinson, Carly D, Biraj Bisht, and Susanna Loeb, "The inequity of opt-in educational resources and an intervention to increase equitable access," *Annenberg Institute at Brown University EdWorkingPaper, 22*, 2022, *654.* (Cited on 29, 30)

Roseth, Cary J, David W Johnson, and Roger T Johnson, "Promoting early adolescents' achievement and peer relationships: the effects of cooperative, competitive, and individualistic goal structures.," *Psychological bulletin*, 2008, *134* (2), 223. (Cited on 25)

Schunk, Dale H and Maria K DiBenedetto, "Self-efficacy theory in education," in "Handbook of motivation at school," Routledge, 2016, pp. 34–54. (Cited on 16)

_ **and** _ , "Academic self-efficacy," in "Handbook of positive psychology in schools," Routledge, 2022, pp. 268–282. (Cited on 16)

Schwarzer, R. and Matthias Jerusalem, "General Self-Efficacy Scale," 1995. (Cited on 16, 52, 61)

Shemyakina, Olga, "The effect of armed conflict on accumulation of schooling: Results from Tajikistan," *Journal of Development Economics*, 2011, *95* (2), 186–200. (Cited on 6)

Slavich, George M and Philip G Zimbardo, "Transformational teaching: Theoretical underpinnings, basic principles, and core methods," *Educational psychology review*, 2012, *24*, 569–608. (Cited on 25)

Szabo, Marianna and Peter F Lovibond, "Development and psychometric properties of the DASS-youth (DASS-Y): an extension of the depression anxiety stress scales (DASS) to adolescents and children," *Frontiers in Psychology*, 2022, *13*, 766890. (Cited on 15, 44, 46, 54)

The Economist and Sondre Solstad, "The Economist war-fire model," *The Economist*, 2023. First published in "A hail of destruction", February 25th issue. (Cited on 48, 65)

Vella, Melchior, "The relationship between the Big Five personality traits and earnings: Evidence from a meta-analysis," *Bulletin of Economic Research*, 2024, *76* (3), 685–712. (Cited on 33)

Vivalt, Eva, "How Much Can We Generalize From Impact Evaluations?," *Journal of the European Economic Association*, 2020, *18* (6), 3045–3089. (Cited on 5)

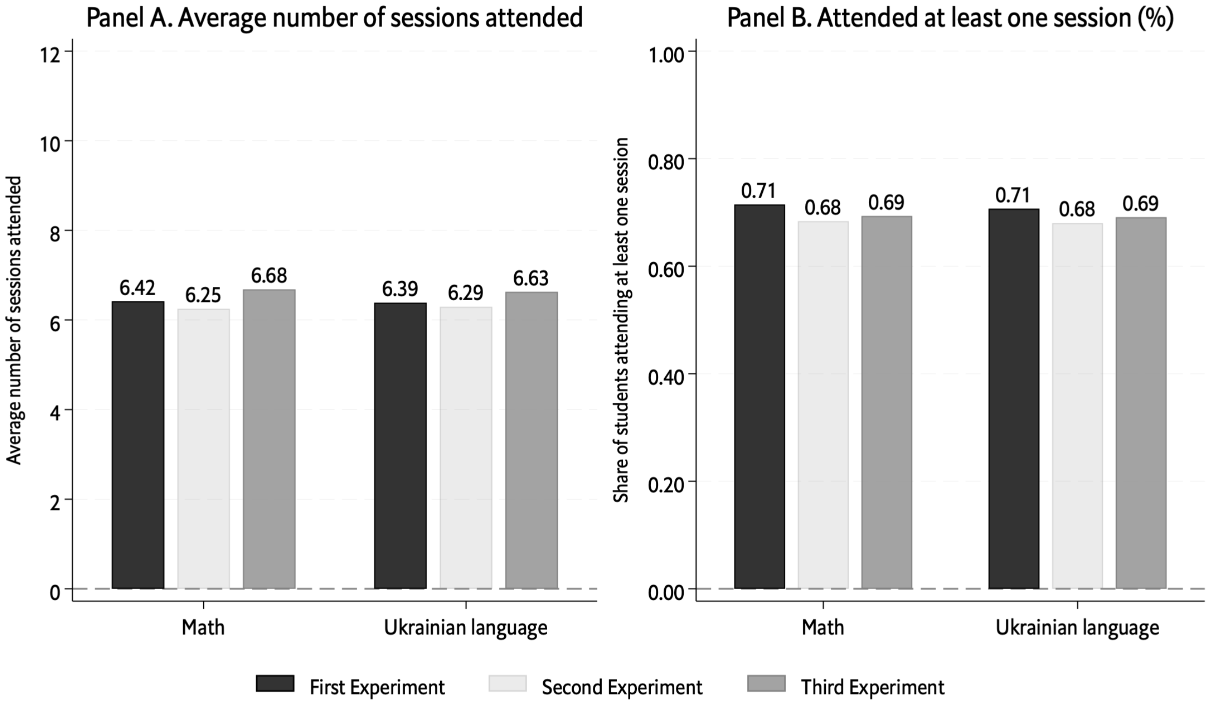
Westfall, Peter H and S Stanley Young, *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Vol. 279, John Wiley & Sons, 1993. (Cited on 20, 50, 51, 52, 54, 77, 78, 79, 80, 81, 82, 83)

World Bank, "Ukraine - Third Rapid Damage and Needs Assessment (RDNA3): February 2022 - December 2023," 2024. English. Available at: http://documents.worldbank.org/curated/en/099021324115085807. (Cited on 8)

_ **, Government of Ukraine, European Union, and United Nations**, "Second Ukraine Rapid Damage and Needs Assessment (RDNA2): February 2022 - February 2023," 2023. English. Available at: http://documents.worldbank.org/curated/en/099184503212328877. (Cited on 8)

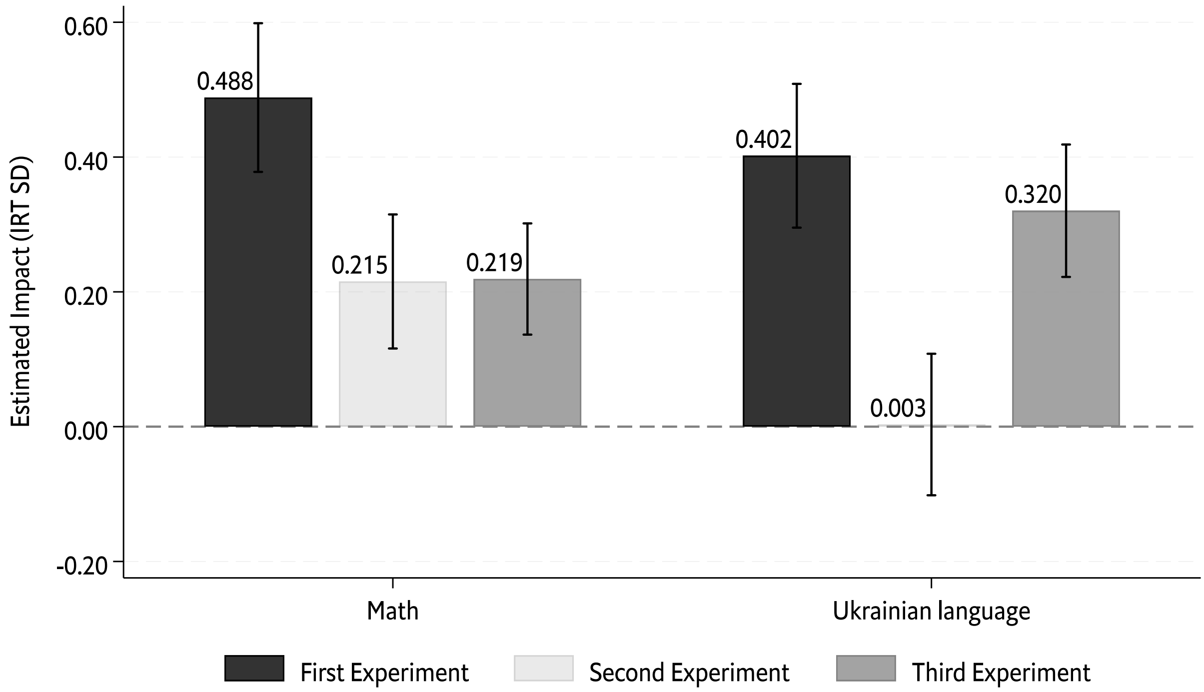
Tables and Figures

Figure 1: Attendance to Tutoring Sessions



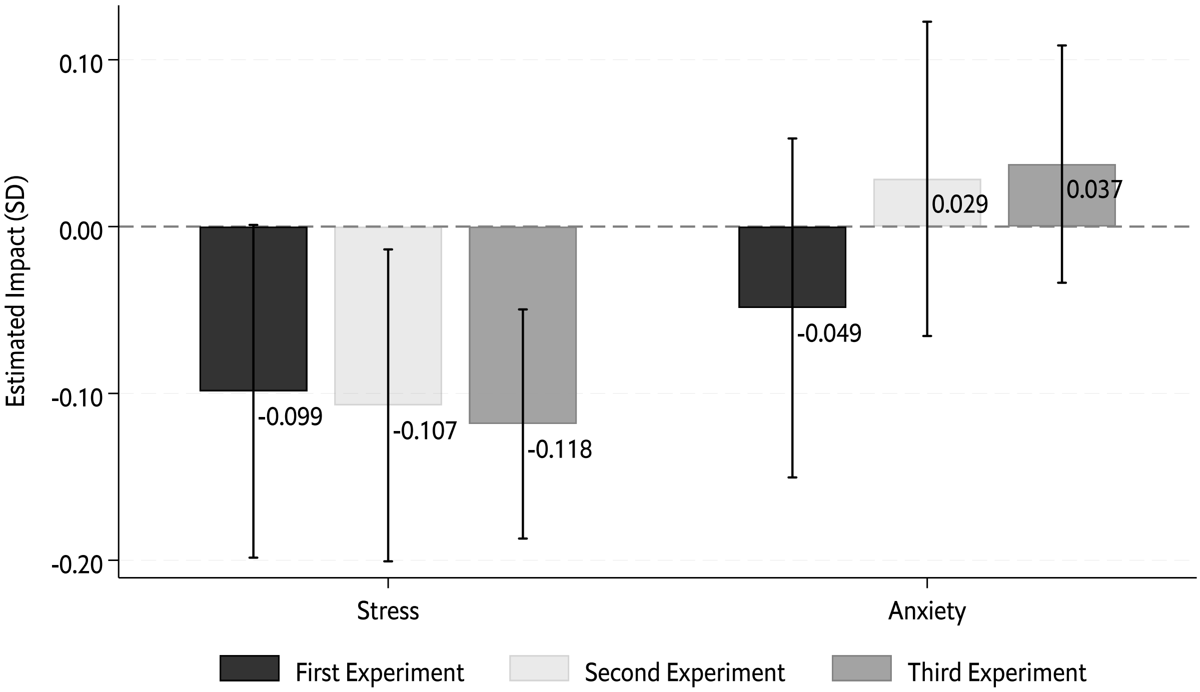
Notes: This figure shows the average attendance to the tutoring program (panel A) and take up of the intervention (panel B), by experiment. The average attendance is estimated as the average number of sessions of attended, separated by subject (math or Ukrainian language). As a reference, the total number of session by subject was 12 across all experiments. The take up is defined as the share of students who attended at least one session of math or Ukrainian language.

Figure 2: Impacts of the Online Tutoring Program on Academic Outcomes



Notes: This figure presents estimates of β_1 from equation (1) on math and Ukrainian language test scores. These estimated coefficients, along with standard errors, adjusted p-values for multiple hypothesis testing, outcome mean for the control group, and number of observations, are presented in Tables A4 to A6. The solid lines represent the 95% confidence intervals. Each outcome has been estimated using item response theory (IRT) scores and then standardized relative to the control group within each experiment. All specifications include controls selected using LASSO and strata fixed effects. Standard deviation (SD) units of IRT scores are used for the y-axis.

Figure 3: Impacts of the Online Tutoring Program on Mental Health Outcomes



Notes: This figure presents estimates of β_1 from equation (1) on standardized scores of stress and anxiety. Outcomes have been standardized relative to the control group within each experiment. These estimated coefficients, along with standard errors, adjusted p-values for multiple hypothesis testing, outcome mean for the control group, and number of observations, are presented in Tables A4 to A6. The solid lines represent the 95% confidence intervals. Each outcome has been estimated using the scoring templates from Szabo and Lovibond (2022). All specifications include controls selected using LASSO and strata fixed effects. Standard deviation (SD) units are used for the y-axis.

Figure 4: Heterogeneity of Results by Student Characteristics

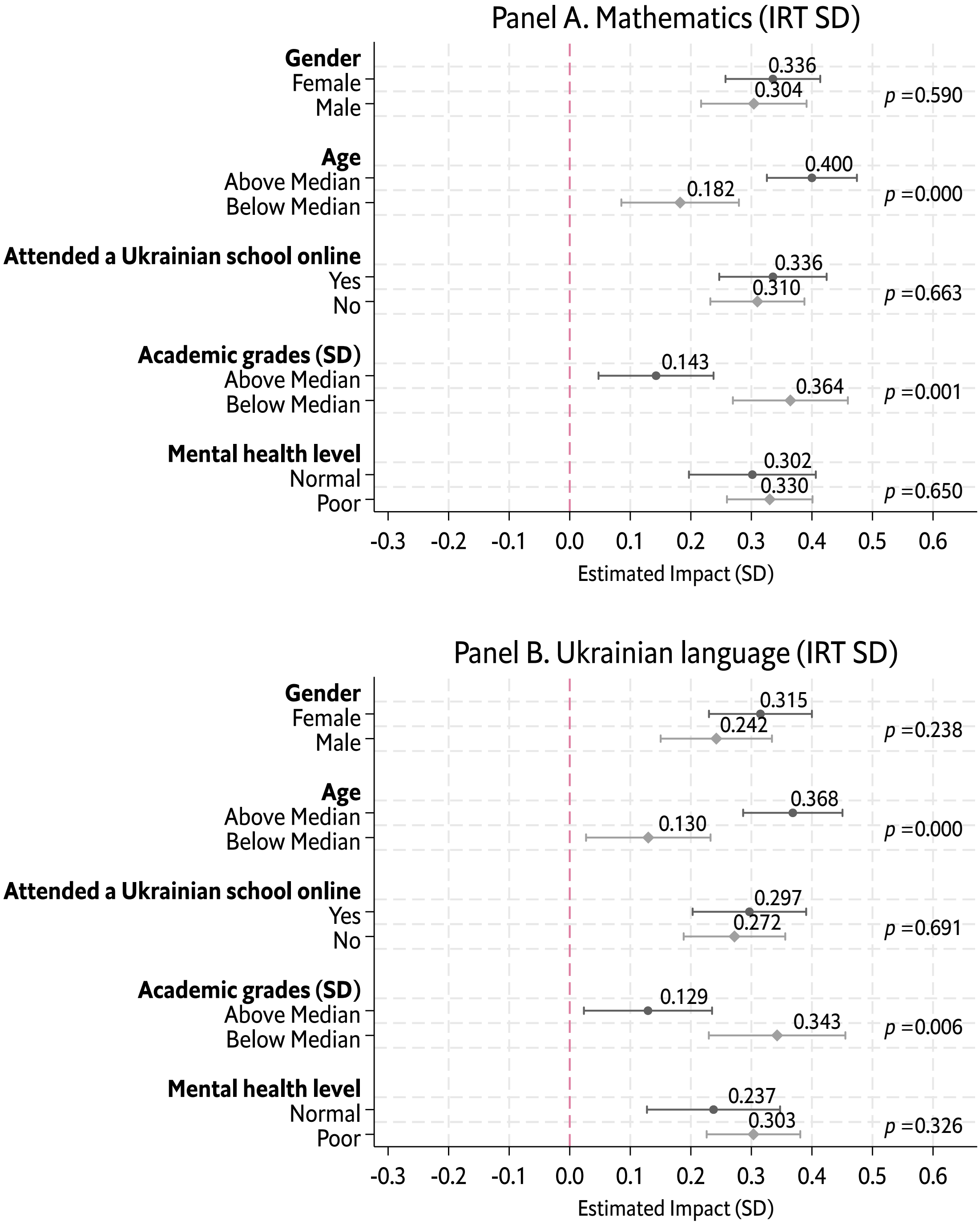
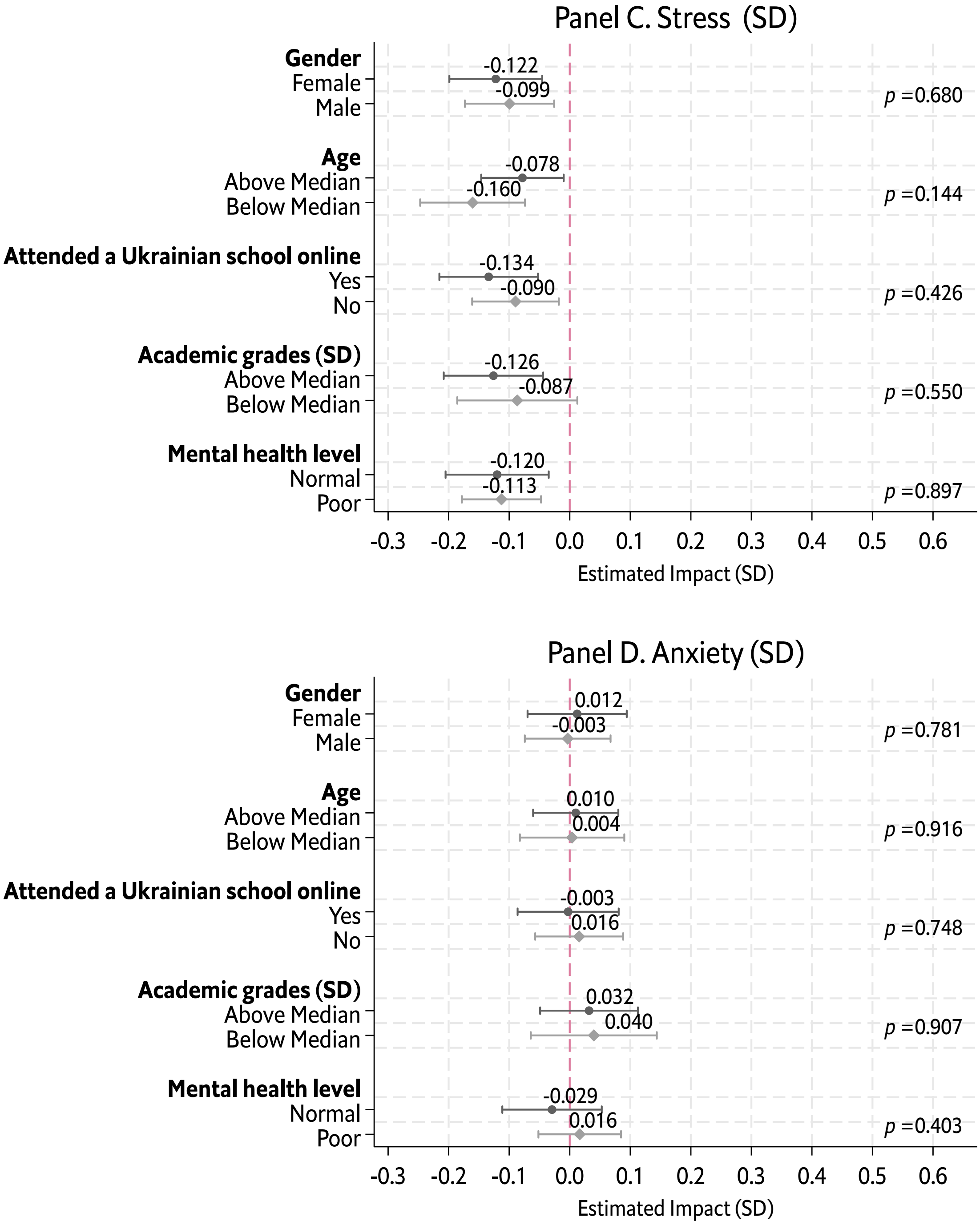


Figure 4: *Continued*–Heterogeneity of results by Student Characteristics



Notes: This Figure reports OLS estimates and 95% confidence intervals of the heterogeneous impact of the online tutoring program by baseline student characteristics on the main outcomes using pooled data across the three experiments. These results are estimated by including an interaction term between the treatment indicator and specific student characteristics measured at baseline in Specification (1), as well as experiment fixed-effects. The "academic grades (SD)" measure consists of the average between math and Ukrainian language scores (in SD) measured at baseline using students' assessments. As reference, the median of the measure of academic grades is 0.056 and the median students' age is 12 years. The measure of mental health was estimated using data from the baseline scores for stress and anxiety and the severity cut-offs from Szabo and Lovibond (2022). The mental health level is "normal" if the student's DASS-Y score was equal or less to 11 and anxiety DASS-Y raw score was equal to 5 or less and is "poor" otherwise. As a reference, the percentage of students with a normal level of mental health is 32%.

Figure 5: Heterogeneity of Results by External Factors

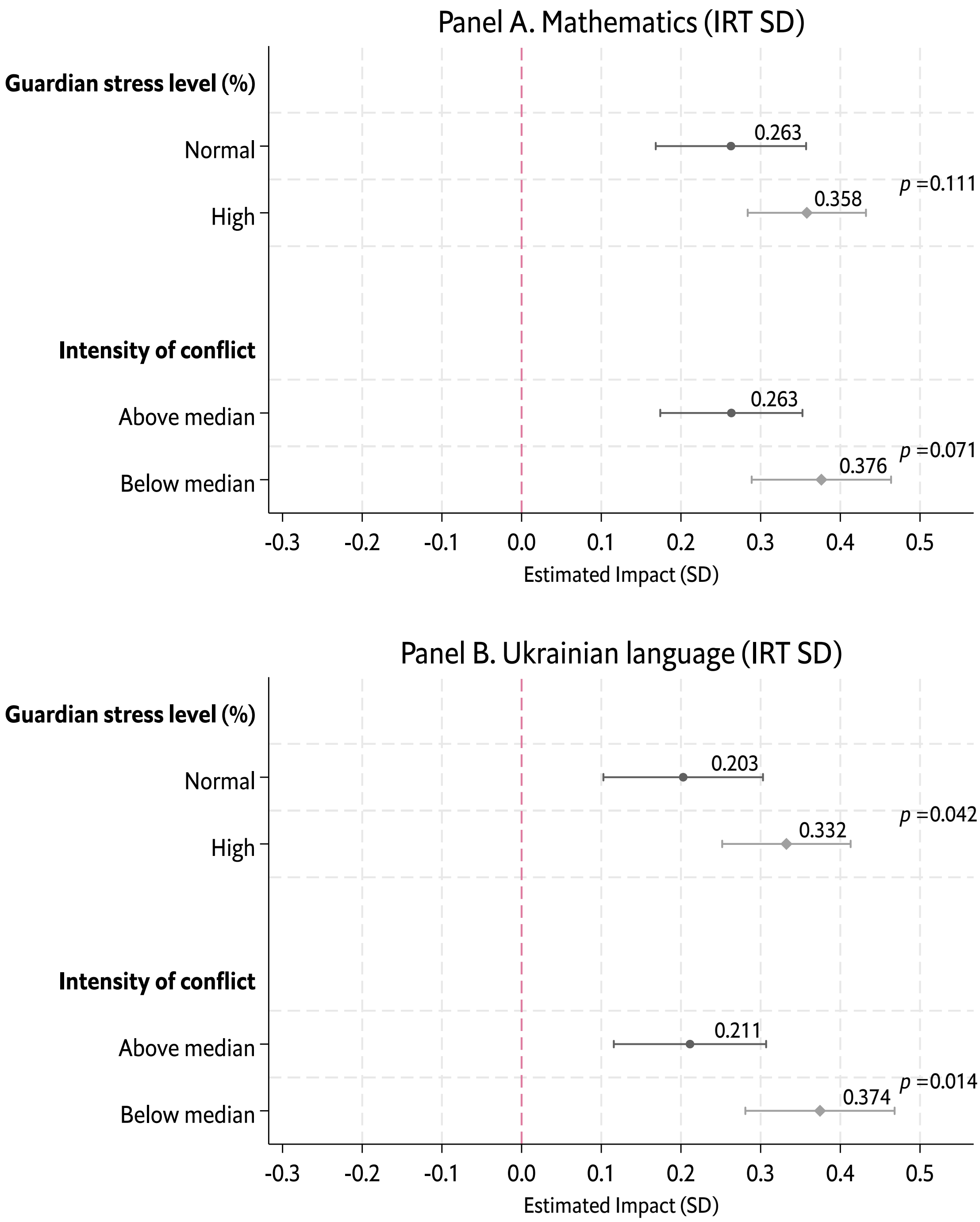
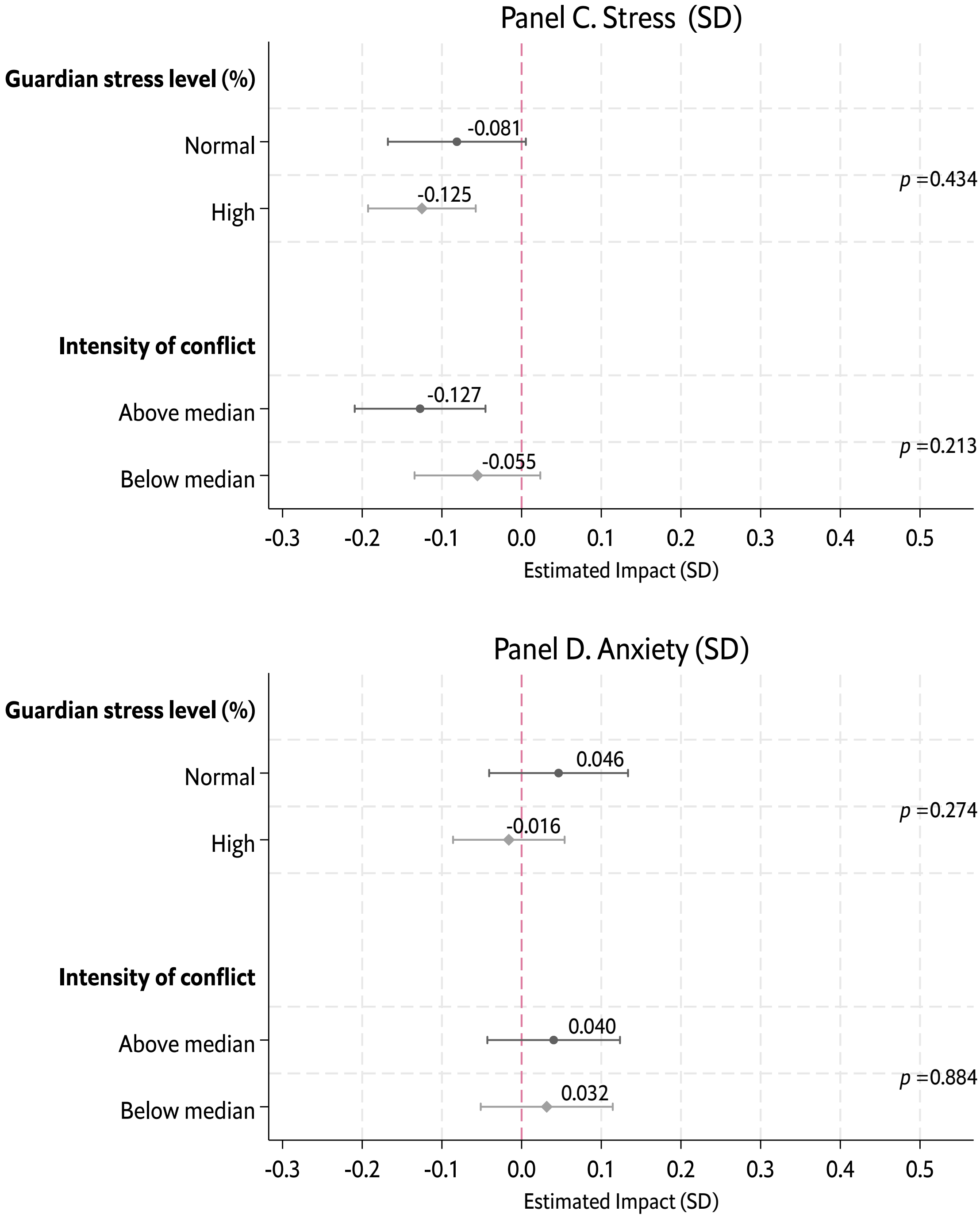


Figure 5: *Continued*–Heterogeneity of Results by External Factors



Notes: This Figure reports OLS estimates and 95% confidence intervals of the heterogeneous impacts of the online tutoring program by external factors measured at baseline on the main outcomes using pooled data across the three experiments. The results are estimated by including an interaction term between the treatment indicator and specific student characteristics measured at baseline in Specification (1), as well as experiment fixed-effects. The "Guardian stress level" measure was estimated using data from the baseline score for stress and the severity cut-off from Lovibond and Lovibond (1996). The stress level is "normal" if the guardian's score was 11 or less and is "high" if the score was higher than 11. As a reference, the percentage of guardians with normal stress levels is 41%. Intensity of conflict was measured using war-fire data from The Economist and Solstad (2023). We construct an indicator of the total number of war-fire events in each third-level administrative unit during the period of each experiment.

Table 1: Balance Between Treatment and Control Groups, by Experiment

	First Experiment		Second Experiment		Third Experiment	
	Control	Diff (F.E.)	Control	Diff (F.E.)	Control	Diff (F.E.)
	(1)	(2)	(3)	(4)	(5)	(6)
Panel A. Student characteristics						
Female	0.55	0.02	0.54	0.02	0.54	0.00
	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)
Age	12.65	–0.07	12.66	0.11	12.47	–0.01
	(0.05)	(0.09)	(0.09)	(0.13)	(0.08)	(0.10)
Current grade	7.19	–0.11	7.07	–0.04	7.03	0.03
	(0.05)	(0.07)	(0.08)	(0.11)	(0.06)	(0.08)
Student's type of school enrollment						
Online in a Ukrainian school	0.44	0.01	0.48	0.00	0.38	–0.01
	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)
Offline (in-person) in a Ukrainian school	0.32	–0.01	0.34	–0.01	0.55	0.00
	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)
Offline abroad + online in a Ukrainian school	0.09	0.00	0.07	–0.00	0.01	–0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.00)	(0.00)
Student has used TFU services	0.04	0.00	0.02	0.00	0.02	–0.00
	(0.01)	(0.01)	(0.00)	(0.01)	(0.00)	(0.00)
Panel B. Household characteristics						
Student has access to electronics at home	0.99	–0.00	0.99	–0.00	0.99	–0.00
	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)	(0.00)
Guardian is female	0.93	0.02*	0.92	0.01	0.92	0.00
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)
Guardian's age	39.57	0.10	39.77	0.26	39.41	–0.08
	(0.14)	(0.20)	(0.14)	(0.20)	(0.12)	(0.16)
Guardian's residency has changed during the war	0.39	0.02	0.43	0.02	0.26	–0.03**
	(0.01)	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)
Panel C. Outomces at baseline						
Math (score)			2.69	0.00	2.73	0.09
			(0.06)	(0.08)	(0.05)	(0.07)
Ukrainian language (score)			3.30	0.02	3.13	0.13*
			(0.07)	(0.09)	(0.05)	(0.08)
Anxiety (SD)	–0.00	0.03	0.00	0.02	–0.00	–0.01
	(0.03)	(0.04)	(0.03)	(0.04)	(0.02)	(0.03)
Stress (SD)	0.00	0.05	–0.00	0.06	–0.00	–0.04
	(0.03)	(0.04)	(0.03)	(0.04)	(0.02)	(0.03)
Normal anxiety level (%)	0.42	–0.00	0.40	–0.01	0.43	0.02
	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)
Normal stress level (%)	0.56	–0.01	0.56	–0.01	0.34	0.01
	(0.01)	(0.02)	(0.01)	(0.02)	(0.01)	(0.01)
Guardian stress level (SD)	–0.00	–0.02	–0.00	0.04	0.00	–0.04
	(0.03)	(0.04)	(0.03)	(0.04)	(0.02)	(0.03)
F-test of joint significance (P-value)		0.58		0.35		0.68
Number of observations	1259	2518	1388	2767	2274	4547

Notes: This table presents the mean for the control group (columns 1, 3, and 5) as well as the difference between the treatment and the control groups (columns 2, 4, and 6) in each experiment. These differences correspond to $\hat{\beta_1}$ in the following specification: $X_{isw} = \beta_0 + \beta_1 T_i^w + \gamma_s + \varepsilon_{iswt}$, where X_{isw} represents the characteristic or outcome of student i in stratum s in experiment w at baseline, T_i^w is the treatment indicator in each experiment w, and γ_s and ε_{iswt} are indicators for the strata fixed effects and the error term. Standard errors are clustered at the tutoring group level and their estimations are in parentheses. The "F-test of joint significance p-value" refers to the null hypothesis that the differences across all observable student characteristics within each experiment are jointly not statistically significant. Statistical significance at the 5% and 10% levels is indicated by ** and *, respectively.

Table 2: Impact of the Online Tutoring Program on Structured Peer Interactions

	Enrolled in online platform (1)	Interacted in online platform (2)	+10 interactions (3)	Friends enrolled in the program (4)
Panel A. First Experiment				
Treatment	0.141***	0.215***	0.105***	0.133***
	(0.015)	(0.022)	(0.017)	(0.021)
	[0.000]	[0.000]	[0.000]	[0.000]
Control group outcome mean	0.823	0.643	0.070	0.134
# of control variables selected	1	0	0	0
Obs.	1,562	1,562	1,562	1,562
Panel B. Second Experiment				
Treatment	0.422***	0.482***	0.087***	0.077***
	(0.022)	(0.024)	(0.015)	(0.021)
	[0.000]	[0.000]	[0.000]	[0.001]
Control group outcome mean	0.541	0.350	0.038	0.146
# of control variables selected	0	0	0	0
Obs.	1,368	1,368	1,368	1,368
Panel C. Third Experiment				
Treatment	0.342***	0.359***	0.108***	0.113***
	(0.014)	(0.018)	(0.013)	(0.018)
	[0.000]	[0.000]	[0.000]	[0.000]
Control group outcome mean	0.637	0.474	0.047	0.224
# of control variables selected	0	0	1	2
Obs.	2,456	2,456	2,456	2,456

Notes: This table presents estimates of β_1 from equation (1) on measures of structure peer interactions. Column (1) presents the results on an indicator of whether the student enrolled in the Discord platform. Columns (2) and (3) show the impacts of the program on an indicator of whether the student interacted in the platform or if the number of interactions was 10 or more, respectively. This measure is equal to zero for the students who did not enroll to the platform. These three measures were collected using data from the platform. Column (4) presents the result on an indicator on whether the student reported that any of their friends enrolled the tutoring program. This outcome was measured using data from the self-reported survey. All estimations include controls variables selected using LASSO and strata fixed effects. The number of control variables selected by LASSO is presented in row "# of control variables selected." Clustered standard errors at the tutoring group level are shown in parentheses. Family-wise p-values are shown in brackets, adjusted for the number of outcome variables, and are estimated using 2,000 bootstraps and the free step-down resampling method of Westfall and Young (1993). In each experiment, the number of outcomes within each family of outcomes consists of the total number of dependent variables shown in the table (i.e., four hypothesis tests). Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Table 3: Impact of the Online Tutoring Program on Attitudes Toward Learning and Educational Aspirations

	Subject Enthusiasm		Future Aspirations	
	Math	Ukrainian language	Working	Pursue Higher Education
	(1)	(2)	(3)	(4)
Panel A. First Experiment				
Treatment	0.073***	0.097***	0.021*	0.003
	(0.025)	(0.022)	(0.012)	(0.024)
	[0.010]	[0.000]	[0.163]	[0.905]
Control group outcome mean	0.558	0.661	0.054	0.658
# of control variables selected	0	0	1	0
Obs.	1,562	1,562	1,562	1,562
Panel B. Second Experiment				
Treatment	0.094***	0.174***	0.008	-0.048*
	(0.027)	(0.025)	(0.014)	(0.026)
	[0.004]	[0.000]	[0.591]	[0.152]
Control group outcome mean	0.514	0.576	0.072	0.618
# of control variables selected	3	1	1	2
Obs.	1,368	1,368	1,368	1,368
Panel C. Third Experiment				
Treatment	0.209***	0.217***	0.008	-0.018
	(0.019)	(0.018)	(0.011)	(0.020)
	[0.000]	[0.000]	[0.775]	[0.775]
Control group outcome mean	0.494	0.556	0.072	0.596
# of control variables selected	3	2	1	1
Obs.	2,456	2,456	2,456	2,456

Notes: This table presents estimates of β_1 from equation (1) on measures of attitudes toward learning and educational aspirations. Columns (1) and (2) presents the results on indicators to whether the students reported that he/she likes Math or Ukrainian language "much" or "very much." Future aspirations outcomes consist of indicators of whether the student wants to start working or pursue higher education (columns (3) and (4), respectively) after completing high school. All these outcomes were measured using data from the self-reported endline survey. All estimations include controls variables selected using LASSO and strata fixed effects. The number of control variables selected by LASSO is presented in row "# of control variables selected." Clustered standard errors at the tutoring group level are shown in parentheses. Family-wise p-values are shown in brackets, adjusted for the number of outcome variables, and are estimated using 2,000 bootstraps and the free step-down resampling method of Westfall and Young (1993). In each experiment, the number of outcomes within each family of outcomes consists of the total number of dependent variables shown in the table (i.e., four hypothesis tests). Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Table 4: Impact of the Online Tutoring Program on Social-Emotional Skills

	Grit (1)	Self-Efficacy (2)
Panel A. Second Experiment		
Treatment	0.106*	0.104*
	(0.054)	(0.054)
	[0.110]	[0.110]
Control group outcome mean	-0.000	0.000
# of control variables selected	2	2
Obs.	1,368	1,368
Panel B. Third Experiment		
Treatment	0.331***	0.305***
	(0.040)	(0.041)
	[0.000]	[0.000]
Control group outcome mean	0.000	-0.000
# of control variables selected	3	0
Obs.	2,456	2,456

Notes: This table presents estimates of β_1 from equation (1) on measures of social-emotional skills. Column (1) shows the results on grit, which was measured using the Short Grit (8 items) scale from Duckworth and Quinn (2009) and column (2) presents the results on self-efficacy, which was measured using the 10-item scale from Schwarzer and Jerusalem (1995). Both outcomes were measured using data from the self-reported endline survey and are standardized using the outcome mean and standard deviation of the control group. All estimations include controls variables selected using LASSO and strata fixed effects. The number of control variables selected by LASSO is presented in row "# of control variables selected." Clustered standard errors at the tutoring group level are shown in parentheses. Family-wise p-values are shown in brackets, adjusted for the number of outcome variables, and are estimated using 2,000 bootstraps and the free step-down resampling method of Westfall and Young (1993). In each experiment, the number of outcomes within each family of outcomes consists of the total number of dependent variables shown in the table (i.e., two hypothesis tests). Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Table 5: Complementary Student Investments:
Participation in Additional Tutoring-related Activities

	First Experiment (1)	Second Experiment (2)	Third Experiment (3)
Treatment	0.211***	0.306***	0.330***
	(0.025)	(0.024)	(0.019)
	[0.000]	[0.000]	[0.000]
Control group outcome mean	0.302	0.229	0.287
# of control variables selected	0	0	0
Obs.	1,562	1,368	2,456

Notes: This table shows estimates of β_1 from equation (1) on measures of whether the student sought and received additional tutoring or subject-specific support during the past weeks. The outcome was measured using data from the self-reported endline survey. All estimations include controls variables selected using LASSO and strata fixed effects. The number of control variables selected by LASSO is presented in row "# of control variables selected." Clustered standard errors at the tutoring group level are shown in parentheses. Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Table 6: Impact of Text Messages on Academic and Mental Health Outcomes in the Parental Investment Experiment

	Academic Outcomes		Mental Health Outcomes	
	Math	Ukrainian language	Stress	Anxiety
	IRT	IRT	SD	SD
	(1)	(2)	(3)	(4)
Tutoring + Text	-0.151	-0.230**	0.064	0.145
	(0.095)	(0.095)	(0.090)	(0.096)
	[0.223]	[0.063]	[0.435]	[0.223]
Control group outcome mean	0.000	0.000	0.000	0.000
# of control variables selected	1	1	1	1
Obs.	466	456	466	466

Notes: This table presents estimates of β_1 from equation (2) on academic performance (math and Ukrainian language) and mental health outcomes (intensive and extensive margins of stress and anxiety) in the parental investment experiment. All outcomes have been standardized using the outcome mean and standard deviation for the group that received only tutoring (no text messages) within each experiment. Outcomes in columns (4) and (6) consist of indicators taking the value of 1 if the DASS-Y raw score for anxiety is 5 or less or if it is 11 or less for stress (Szabo and Lovibond, 2022). All specifications include control variables selected using LASSO and strata fixed effects. The number of control variables selected by LASSO is presented in row "# of control variables selected." Clustered standard errors at the tutoring group level are shown in parentheses. Family-wise p-values are shown in brackets, adjusted for the number of outcome variables, and are estimated using 2,000 bootstraps and the free step-down resampling method of Westfall and Young (1993). In each experiment, the number of outcomes within each family of outcomes consists of the total number of dependent variables shown in the table (i.e., six hypothesis tests). Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Table 7: Cost-Benefit Analysis

Parameter		Value		Source
A. Projection of Future Earnings				
Earnings per year (2024) (USD)		6,124		Ministry of Economy (2024)
Discount Rate (%) [a]		5		Assumption
Working Age [b]		22-65		Assumption
Labor-Force Participation Rates (%) [c]		56.1		WDI and UNB
Real growth in salaries (%) [d]		2.9		National Bank of Ukraine
Average PV of lifetime earnings for a beneficiary		147,060		Calculated
B. Impact of the Interventions				
Earnings gain per SD of learning (%)		8.0		Literature
Earnings gain per SD of mental health (%)		1.7		Literature
	Treatment 1	**Treatment 2**	**Treatment 3**	
Program Effect on learning (ITT) (SD) [e]	0.402	0.215	0.219	Program
Program Effect on mental health (ITT) (SD) [f]	0.099	0.106	0.118	Program
Number of participants offered treatment	1,259	1,379	2,273	Program
C. Cost of the interventions				
Average cost per treatment participant (USD)	89.12	89.74	93.41	Program
Average cost per control participant (USD)	7.04	7.04	7.04	Program
Incremental cost per treatment participant (USD)	82.09	82.70	86.37	Program
Cost-effectiveness ratio (learning) (USD)	204.19	384.66	394.38	Program
D. Benefit Cost Ratio				
Present Value of Benefits (thousand USD)	6,327.7	3,887.5	6,456.4	Calculated
Learning (thousand USD)	6,013.1	3,518.8	5,793.10	Calculated
Mental Health (thousand USD)	314.7	368.7	663.3	Calculated
Present Value of Costs (thousand USD) [g]	112.1	123.7	196.3	Calculated
Benefit Cost Ratio (USD)	56.4	31.4	32.9	Calculated

Notes: This table presents parameters fort the cost-benefit analysis of the tutoring program based on projected impacts on earnings and well-being. Panel A lists parameters and assumptions used to project the future earnings of individuals, considering factors such as participation rates, salary growth, and working age. Panel B outlines the estimated impacts of the interventions on cognitive skills and social-emotional well-being, including treatment effects and participant numbers.

[a]Calculated based on the projected average monthly salary of employees for 2024, as outlined in the Ministry of Economy's Recovery Growth Scenarios Report (April 2024), and the anticipated average exchange rate for 2024.

[b]Used in the cost-effectiveness analysis papers by Ganimian et al. (2024) and Holla et al. (2021).

[c]2010–2021 average extracted from World Development Indicators provide measures for the labor force participation rate (modeled ILO estimate). Surveys conducted by Info Sapiens and cited by the National Bank of Ukraine in the Inflation Report (July 2024) show that labor force participation rates in 2024 is close to 56%.

[d]Based on projections from the National Bank of Ukraine, which forecast a real growth in wages by 2.9 in 2026 (assuming gradual normalization of the economy).

[e]The least significant effect is considered between math and Ukrainian language.

[f]On social-emotional skills, the effect corresponds to stress, as anxiety was not significant.

[g]Treatments 1 and 2 were carried out in 2023; therefore, costs are projected to 2024 using the 2023-2024 inflation rate, as projected by the **?**.

Appendix

For Online Publication Only

A Structured Ethics Appendix

For more explanation of each question, see Asiedu et al. (2021).

1. Policy Equipoise.

In each experiment, the treatment arm provides online, 3-1 group tutoring sessions to students in grades 5 to 9 in an unstable, conflict-affected environment. Despite recent literature on the impacts of individual-based tutoring programs, there was no consensus among experts regarding the impact of small group tutoring on students affected by wars on learning and mental health given the different challenges they face, so the control and treatment arms were in policy equipoise. Furthermore, for those performing poorly on learning and presenting mental health concerns through self-assessments at baseline for experiments 2 and 3 (for which data was collected), we believe that there is equipoise given limited evidence of effectiveness in this setting.

In addition, while financial resources were available since the start of the program, it was agreed that the program would be delivered in a staggered manner because Teach for Ukraine expected to face human resources challenges while attempting to substantially increase the number of tutoring sessions, including challenges in recruiting and training a large number of tutors and assigning and managing the tutoring sessions. Once effectiveness of the program was proven, students in the control group were offered spots in later waves of the the online tutoring program. A total of 2,843 spots were filled by students from the control group.

2. Role of researchers with respect to implementation.

There was no direct interaction between participants and the research team. The core program was designed by Teach for Ukraine. The research team played an active role in suggesting modifications to the design and monitoring of the program (initially intended for large groups of 8-10 children over 6 hours per week for three subjects - math, Ukrainian language, and Ukrainian history) of the program, and designing the tools for monitoring and evaluation. While the research team supported Teach for Ukraine in securing funding for the implementation of the interventions, funding was provided from UBS Optimus Foundation directly to Teach for Ukraine. The research team secured separate funding from the World Bank for the evaluation of the interventions.

The program was implemented by Teach for Ukraine, from hiring tutors and recruiting participants to implementing the tutoring sessions. IRB approval was received from Innovations for Poverty Lab for the implementation of the programs and their evaluation. Informed consent from guardians and assent from their children included taking part in the tutoring program if selected via lottery and in both the baseline and endline surveys.

3. Potential harms to participants or nonparticipants from the interventions or policies.

The IRB reviewed protocols for the online tutoring program, participation in which was free and voluntary and from which participants were always free to withdraw. Protocols were in place for responding to sensitive issues and distress that emerged during or as a result of the sessions. In particular, tutors were requested to fill out journals for each tutoring session delivered, responding to questions related to attitudes and engagement during the sessions for each student. Any student identified in the journals as in distress was directed to a psychologist hired under Teach for Ukraine. The sessions did require participation, effort, and time, but were limited to three hours a week and the participants ultimately decided how much to engage. Participants were not required to attend sessions, and there was no consequence to them for non-attendance. Participants' access to future programs was not reduced by access to this program. All participants assigned to the control group in all three experiments were offered placement in tutoring programs run by Teach for Ukraine at a later date.

4. Potential harms to research participants or research staff from data collection (e.g., surveying, privacy, data management) or research protocols (e.g., random assignment).

Data collection and management procedures were in adherence with human subjects protocols around privacy and confidentiality and respectful of cultural norms. Baseline and endline data collection with students, parents and tutors was entirely self-reported and without the participation of enumerators. Questions considered more sensitive in our context (such as mental health questions) come from well-tested and validated instruments. We also performed psychometric validation of these questions for each experiment before using in the following experiment. In addition, session-based tutor journals were also self-reported. There were no special risks to research staff.

5. Financial and reputational conflicts of interest.

None of the researchers have financial or reputational conflicts of interest with regards to the research results.

6. Intellectual freedom.

There were no contractual restrictions. A Memorandum of Understanding was signed between Teach for Ukraine and the Ministry of Education and Science of Ukraine establishing a partnership for the implementation of education programs during the Russian invasion, including the intervention evaluated by this paper.

7. Feedback to participants or communities.

The intervention and its outcomes were presented to a public working group organized by the Ukraine Education Cluster, which includes representatives of the Ministry of Education and Science of Ukraine and donor partners involved in organizing catch-up learning programs across the country. The results of this intervention have been used to inform design of other donor-funded catch-up learning programs in Ukraine in 2024. In addition, Teach for Ukraine has continuously disseminated the results of the intervention to students, parents, and communities through social media. However, no activity for sharing results to individual participants directly by the research team is planned due to resource constraints.

8. Foreseeable misuse of research results.

There is no foreseeable and plausible risk that the results of the research will be misused or deliberately misinterpreted by interested parties.

9. Other Ethics Issues to Discuss

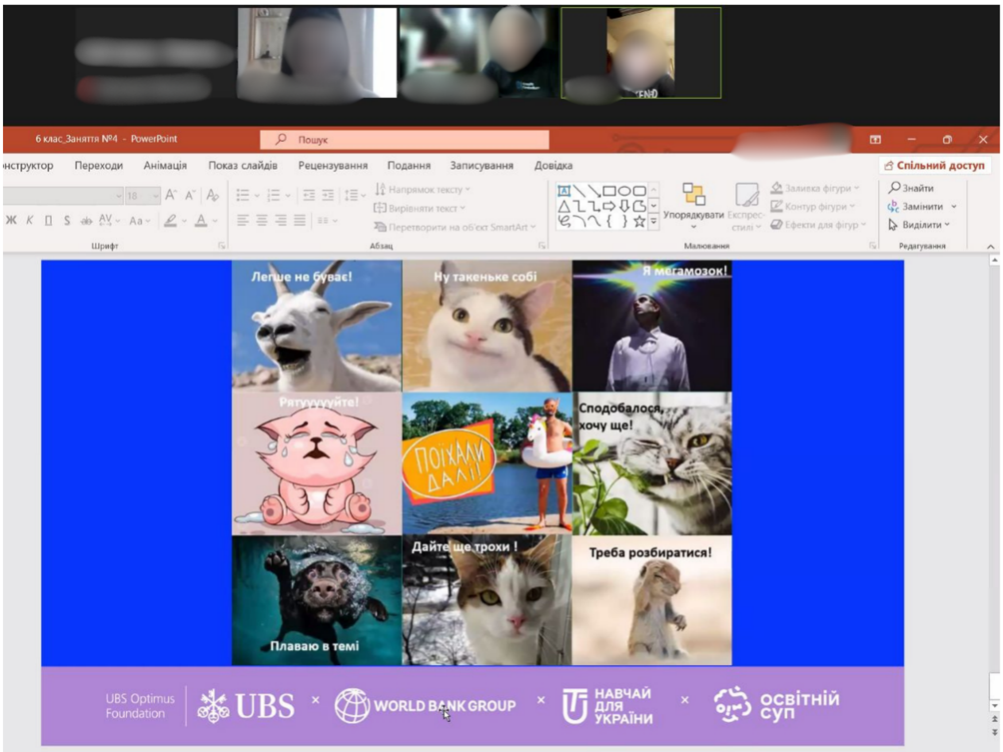
None.

B Description of the Tutoring Sessions

In this section, we provide an overview of the tutoring sessions conducted across all experiments. The sessions combined structured academic instruction with psychosocial support to offer a holistic educational experience. Tutors employed a range of teaching methods and technological tools to engage students, adapt to their individual needs, and support both cognitive and emotional development.

Tutoring sessions typically began with psychosocial activities aimed at creating a positive and welcoming learning environment. Tutors used simple mood check-ins, such as "happy face" or animal image assessments, and brief breathing exercises to help students relax and prepare for learning. These activities set a supportive tone and encouraged participation from the outset. Figure B.1 shows an example of a tutor initiating a session by asking students how they feel.

Figure B.1: Ice-breaker activities during tutoring sessions.



During math sessions, tutors covered topics such as rational numbers, decimals, and geometry concepts like perpendicularity in space, as well as operations with numbers. Interactive tools—including JAM boards, liveworksheets.com, worldwall.net, and GeoGebra—were used to make abstract content more accessible. Students often collaborated to solve problems, participated in quizzes, and engaged in educational games to reinforce key concepts.

Ukrainian language sessions focused on grammar, morphology, and parts of speech. Tutors used presentations, videos, and interactive quizzes, along with collaborative writing tasks. For example, students might complete poems, edit texts, or analyze fables in real time using shared platforms. These activities aimed to strengthen both analytical and creative language skills.

Across subjects, tutors emphasized active listening and created respectful, supportive spaces. Positive reinforcement was common, with tutors regularly offering affirmations such as "well done" or "great job." Differentiated instruction was also observed: tutors adjusted activities based on student proficiency, offering more support or additional challenges as needed. When students struggled, tutors provided extra explanations, guided practice, or real-life examples to aid comprehension.

In the third experiment, psychosocial support was more intentionally integrated into the academic sessions. Exercises such as breathing techniques, mood assessments, and light-hearted games were woven throughout to sustain engagement and alleviate stress.

Students generally participated actively—answering questions, engaging in discussion, and completing interactive tasks. Tutors encouraged participation by addressing students by name and creating inclusive opportunities for them to share their ideas. When students were hesitant or disengaged, tutors adapted their approach by adjusting task difficulty or providing additional support.

Finally, the effective use of digital tools helped sustain interest and participation. These platforms enabled dynamic presentations and collaborative learning, even in a remote setting. Tutors effectively managed technical issues and ensured that all students, regardless of device capabilities, could access and benefit from the sessions.

C Instruments

Academic Assessments for Math and Ukrainian language

The academic assessments were designed to evaluate students' proficiency in math and Ukrainian Language across grades 5 to 10. Each assessment was tailored to align with the curriculum and learning objectives specific to each grade level.

Math Assessment. Students from grades 5 to 10 completed an assessment consisting of 30 items appropriate for their respective grade levels.

Ukrainian language Assessment. The Ukrainian Language assessments varied slightly between experiments to accommodate specific instructional content. In Experiments 1 and 2, students in Grades 5 and 6 completed assessments consisting of 35 items, while students in Grades 7 through 10 had assessments with 40 items each. For Experiment 3, the number of items was adjusted per grade level: Grade 5 students completed 33 items, Grade 6 had 34 items, Grades 7 and 8 each had 39 items, Grade 9 had 38 items, and Grade 10 students completed 39 items.

Mental Health

Description

The DASS-21 Youth Edition is a set of self-report scales designed to measure the negative emotional states of anxiety and stress in adolescents. Each scale contains seven items, providing a quantitative measure of distress along these dimensions. In this study, only the **Stress** and **Anxiety** scales were administered to participants.

Participants responded to each item using a 4-point Likert scale ranging from 0 (Not true) to 3 (Very true). The scores for each scale were summed to produce a total score for Stress and Anxiety, respectively.

For the Anxiety scale, individuals scoring between 0 and 5 were categorized as having normal levels of anxiety. For the Stress scale, scores between 0 and 11 indicated normal levels of stress.

Additionally, for both constructs, the total scores were standardized to the control group to have a mean of 0 and a standard deviation of 1 for each of the experiments. In this standardization, a lower score indicates better outcomes, reflecting lower levels of stress or anxiety. Below, we show the items we used for both categories.

Stress Scale Items

1) I got upset about little things.
2) I found myself over-reacting to situations.
3) I was stressing about lots of things.
4) I was easily irritated.
5) I found it difficult to relax.
6) I got annoyed when people interrupted me.
7) I was easily annoyed.

Anxiety Scale Items

1) I felt dizzy, like I was about to faint.
2) I had trouble breathing (e.g., fast breathing), even though I wasn't exercising and I was not sick.
3) My hands felt shaky.
4) I felt terrified.
5) I felt like I was about to panic.
6) I could feel my heart beating really fast, even though I hadn't done any hard exercise.
7) I felt scared for no good reason.

Grit Scale. The Grit Scale is a measure of perseverance and passion for long-term goals, developed by Duckworth and Quinn (2009). It assesses an individual's consistency of interests and perseverance of effort over time, which are key components of grit. The following items are used in this measure:

1) New ideas and projects sometimes distract me from previous ones.

2) Setbacks don't discourage me.

3) I have been obsessed with a certain idea or project for a short time but later lost interest.

4) I am a hard worker.

5) I often set a goal but later choose to pursue a different one.

6) I have difficulty maintaining my focus on projects that take more than a few months to complete.

7) I finish whatever I begin.

8) I am diligent.

General Self-Efficacy Scale. The General Self-Efficacy Scale measures an individual's belief in their ability to handle various situations and to achieve desired outcomes through their actions. It reflects optimism and confidence in one's competence. Following Schwarzer and Jerusalem (1995), we used the following items:

1) I can always manage to solve difficult problems if I try hard enough.

2) If someone opposes me, I can find the means and ways to get what I want.

3) It is easy for me to stick to my aims and accomplish my goals.

4) I am confident that I could deal efficiently with unexpected events.

5) Thanks to my resourcefulness, I know how to handle unforeseen situations.

6) I can solve most problems if I invest the necessary effort.

7) I can remain calm when facing difficulties because I can rely on my coping abilities.

8) When I am confronted with a problem, I can usually find several solutions.

9) If I am in trouble, I can usually think of a solution.

10) I can usually handle whatever comes my way.

Attitudes Towards Math and Ukrainian language. These items assess students' attitudes toward math and Ukrainian Language using a Likert scale to provide insight into their interest and motivation in these subjects.

1) How do you feel about Math?

2) How do you feel about Ukrainian Language?

Interaction with the Online Platform. These questions evaluate the student's engagement with the digital platforms (Discord) and their interactions with peers, reflecting the social component of the learning experience. During endline, we asked students the following questions:

1) Did you enroll in/join the digital platform of EduSoup (Discord or Prosvita)?

2) Were you able to interact with other children through the platform?

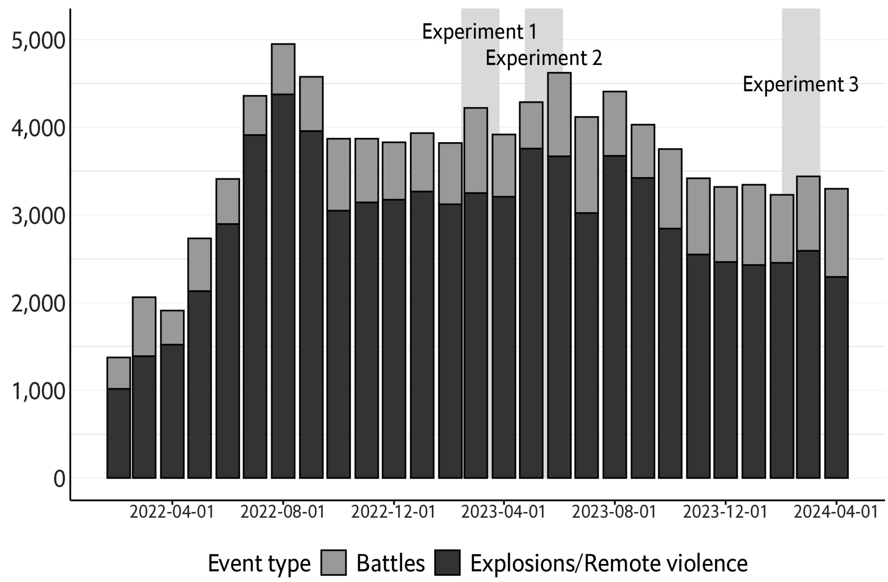
3) How many interactions did you have with them during the past 6 weeks?

4) Think about your friends with whom you interact/chat regularly. Did any of them enroll in the program as well?

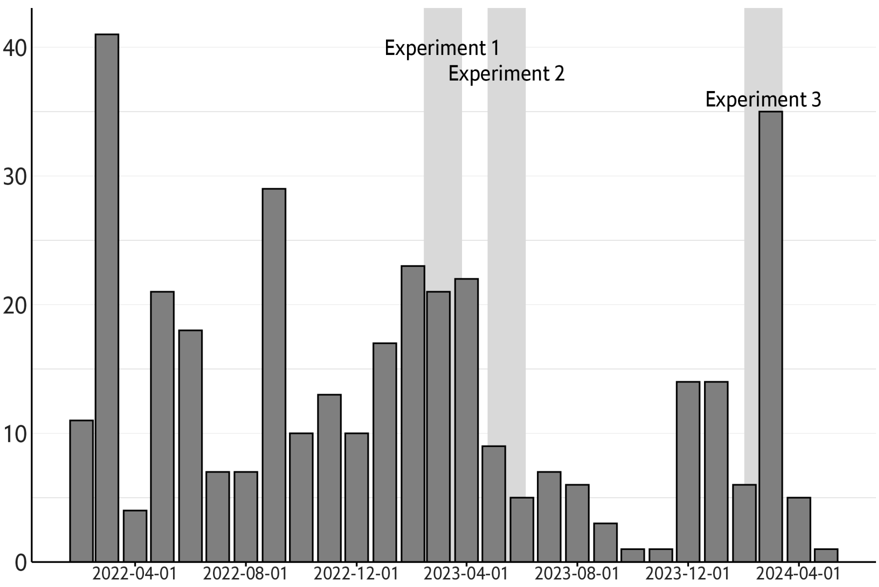
Appendix Figures

Figure A1: Attacks and Violent Events in Ukraine during Russia's full-scale invasion

(a) Violent Events in Ukraine



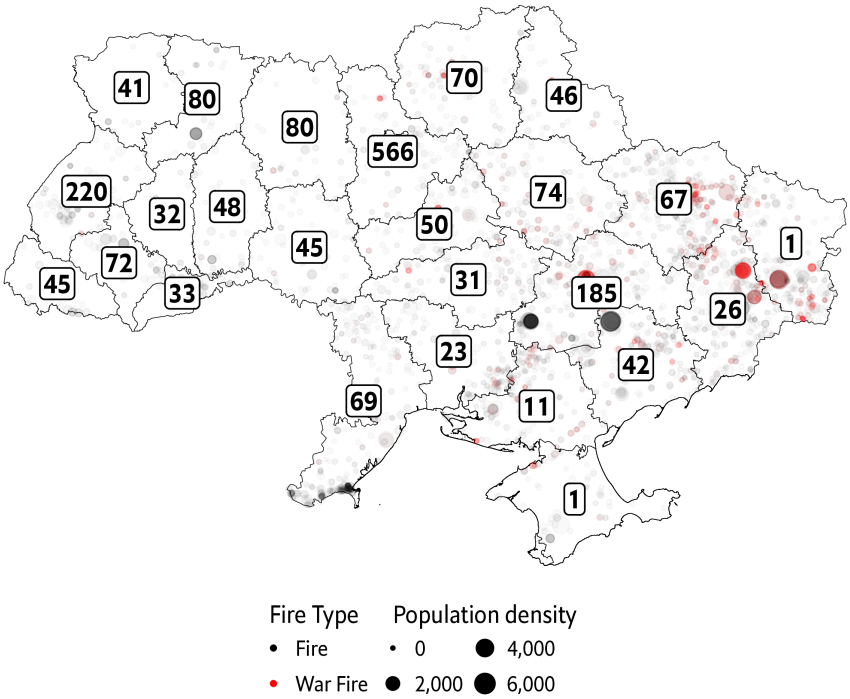
(b) Attacks on the Educational System in Ukraine



Notes: The top figure shows the total number of violent events in Ukraine from November 1, 2021, to May 1, 2024. The gray areas represent the periods when the tutoring experiments took place. "Battles" refers to armed clashes involving both the Ukrainian Armed Forces and the Novorossiya Armed Forces (NAF). "Explosions/Remote violence" includes shelling, artillery, or missile attacks in which only one side was involved. Bars represent monthly accumulated events. The bottom figure shows the total number of school attacks in Ukraine from November 1, 2021, to May 1, 2024. The gray areas represent the periods when the tutoring experiments took place. Five categories of attacks are included: Direct attacks on schools; attacks on students, teachers, and other education personnel; military use of schools or universities; child recruitment at, or en route to/from, school; and attacks on higher education infrastructure. Bars represent monthly accumulated events.
Source: Top figure: Armed Conflict Location & Event Data Project (ACLED). Bottom figure: Global Coalition to Protect Education from Attack, Insecurity Insight.

Figure A2: War Fire Activity in Ukraine Across Different Experiments

Panel A. First Experiment
Feb 13, 2023 - Mar 27, 2023



Panel B. Second Experiment
Apr 24, 2023 - Jun 05, 2023

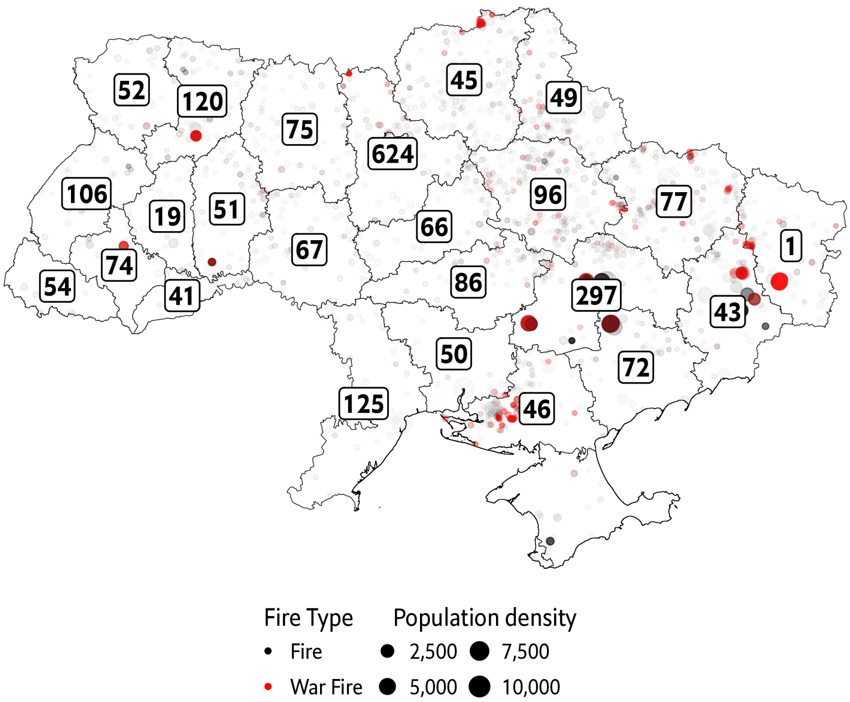
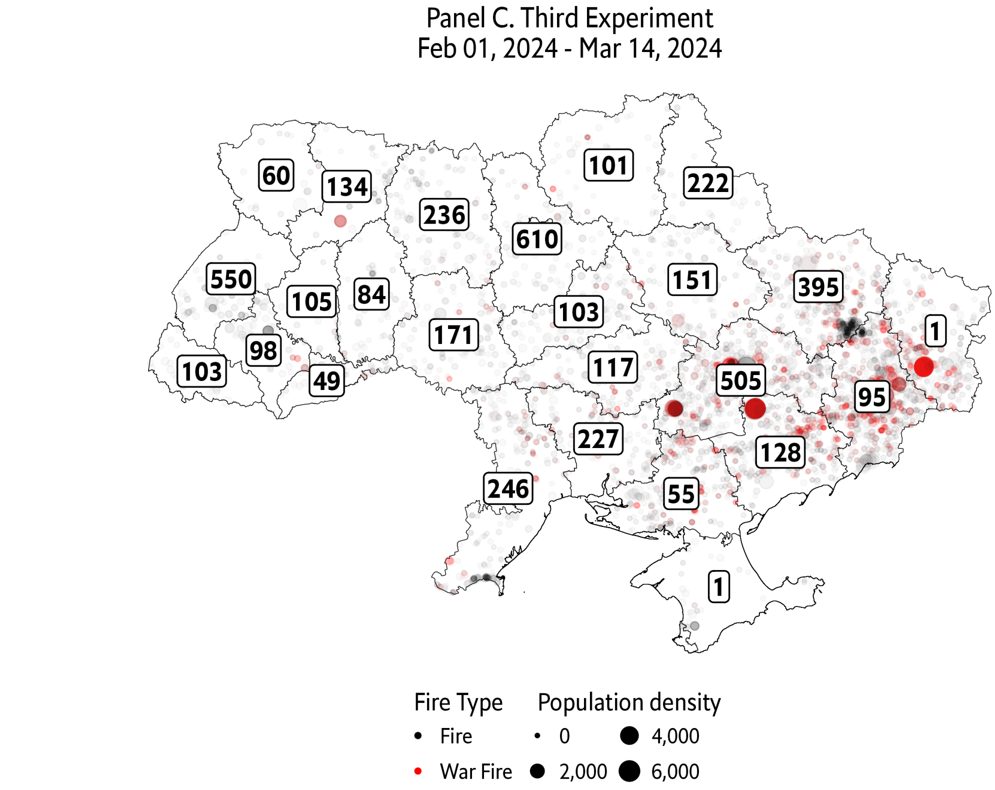


Figure A2: *Continued*–War Fire Activity in Ukraine Across Different Experiments



Notes: This figure shows the distribution of students participating in each experiment (treatment and control) by region (outlined in boxes). Black dots represent fire events not classified as war-related, while red dots represent fire events classified as war-related. War-related fires are defined as excess fire activity in a given 0.1° latitude by 0.1° longitude cell in Ukraine on a given day that is so high it has less than a 5% probability of occurring in a normal year. All dots are scaled by the average population density of the cell in which the fire was detected. *Source:* The Economist and Solstad (2023).The Economist war-fire model. First published in the article "A hail of destruction," *The Economist*, February 25th issue, 2023.

Figure A3: Example of Diagnostic Reports Sent to the Tutors

Математика. 5 клас

Вступне оцінювання учасників та учасницьt

Група 5_106

Шановний тьюторе/ шановна тьюторко!

Усі учасники поточної хвилі під час реєстрації на програму «Освітній Суп» пройшли вступне оцінювання у вигляді 5 питань з математики. Учні 5 класу в середньому набрали **81.0%** правильних відповідей.**Учні групи 5_106 набрали 81.0% правильних відповідей у цьому тестуванні**. Це **значно вище** середнього балу для учнів цього класу.

Це чудова новина! Студенти групи 5_106 можуть продовжувати досягати успіху завдяки зусиллям, практиці та **Вашій підтримці!**

Ми віримо в силу тьюторства! Для нас це означає, що **Ви можете допомогти** своїм підопічним **розкрити весь свій потенціал і подолати труднощі в навчанні**, такі як: російське вторгнення, домашнє середовище та фінансовий стан – **та вплинути на їхню майбутню академічну успішність та добробут.**

Ви маєте **унікальну можливість прищепити** цим учням пристрасть до вивчення математики. Ваші заняття можуть стати безпечним простором для їх відкриттів і саморозвитку, що, зрештою, **сприятиме зміцненню життєстійкості українських дітей та молоді.**

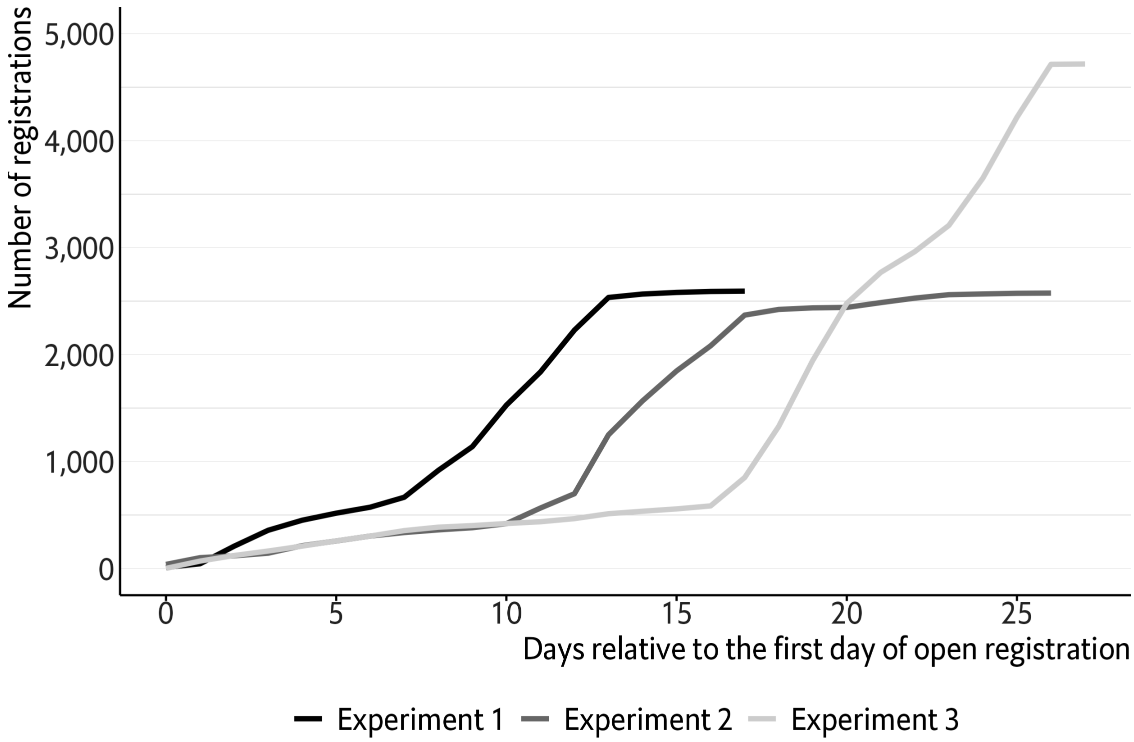
У таблиці нижче наведено середній бал групи **5_106** і всіх учнів цього класу за темами та очікуваними навчальними результатами, що, сподіваємось, допоможе Вам у індивідуалізації роботи з цими учасниками і учасницями.

Наприкінці другого тижня ви отримаєте завдання для короткого поточного оцінювання. Рекомендуємо Вам використати їх для супроводу та корегування процесу підтримки Ваших підопічних.

Очікуваний результат	Конкретизація теми/ очікуваного результату	Group	Grade
ДРОБОВІ ЧИСЛА І ДІЇ З НИМИ			
Здобувач освіти розв'язує вправи, що передбачають: порівняння, додавання і віднімання звичайних дробів з однаковими знаменниками; порівняння, округлення, додавання, множення і ділення десяткових дробів; перетворення мішаного числа у неправильний дріб; перетворення неправильного дробу в мішане число або натуральне число; знаходження відсотка від числа та числа за його відсотком; знаходження середнього арифметичного кількох чисел, середнього значення величини	перетворення неправильного дробу в мішане число або натуральне число	**100.0**	98.3
Здобувач освіти розв'язує вправи, що передбачають: порівняння, додавання і віднімання звичайних дробів з однаковими знаменниками; порівняння, округлення, додавання, множення і ділення десяткових дробів; перетворення мішаного числа у неправильний дріб; перетворення неправильного дробу в мішане число або натуральне число; знаходження відсотка від числа та числа за його відсотком; знаходження середнього арифметичного кількох чисел, середнього значення величини	порівняння, додавання і віднімання звичайних дробів	**100.0**	92.3
Здобувач освіти читає і записує: звичайні та десяткові дроби; мішані числа;	читає і записує десяткові дроби	**100.0**	91.6
Здобувач освіти розв'язує вправи, що передбачають: порівняння, додавання і віднімання звичайних дробів з однаковими знаменниками; порівняння, округлення, додавання, множення і ділення десяткових дробів; перетворення мішаного числа у неправильний дріб; перетворення неправильного дробу в мішане число або натуральне число; знаходження відсотка від числа та числа за його відсотком; знаходження середнього арифметичного кількох чисел, середнього значення величини	додавання звичайних дробів з однаковими знаменниками	**100.0**	63.8

Notes: This figure shows examples of diagnostic reports sent to the tutors of the group 5-106 using baseline math scores. Reports for Ukrainian language had the same structure and format.

Figure A4: Cumulative Registrations of Eligible Households Over Time



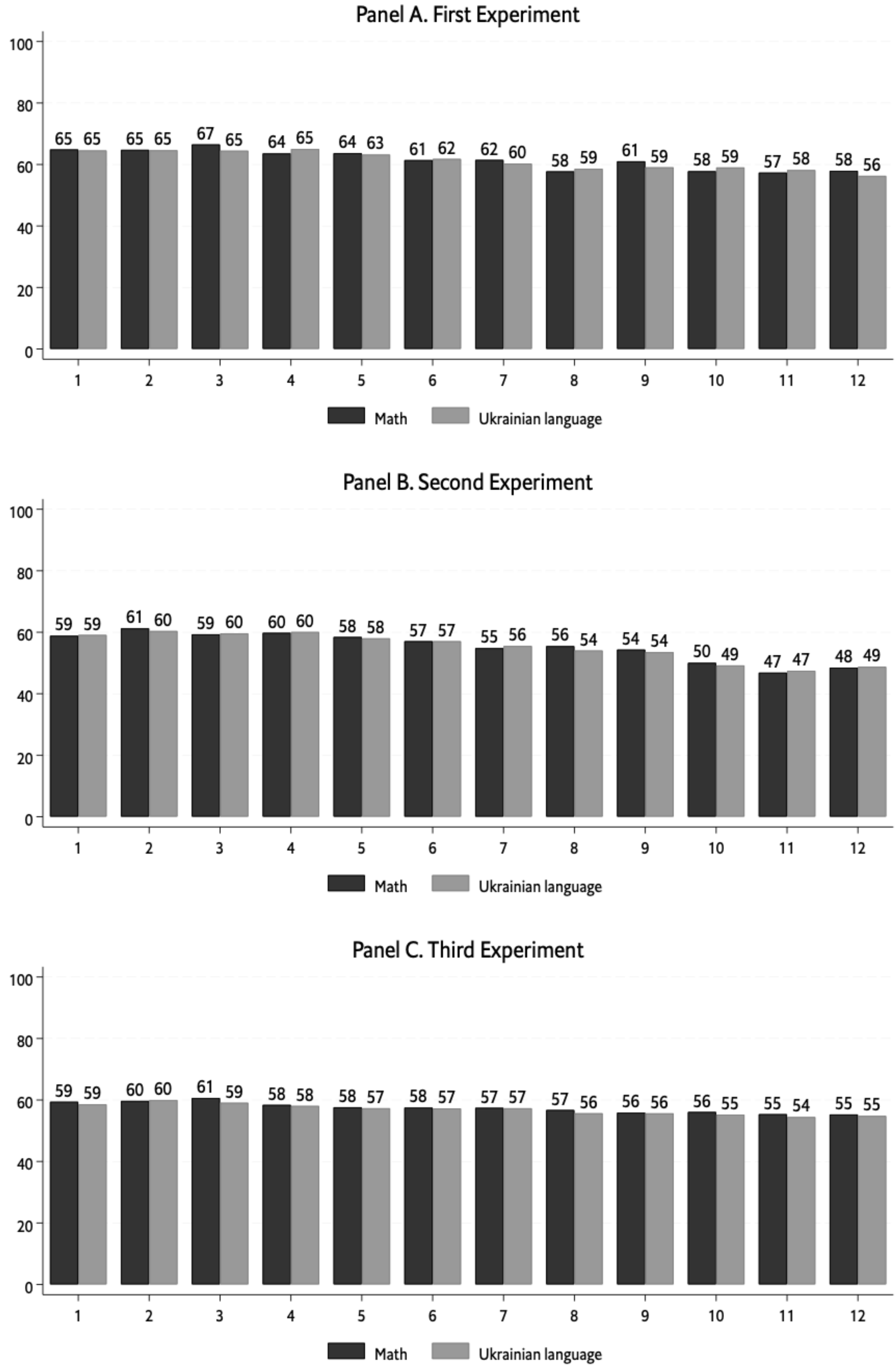
Notes: The figure shows the cumulative number of *eligible* households registered for the study in each experiment. Eligibility criteria varied across experiments. In the first experiment, there was no eligibility criteria. In the second experiment, only households that had not participated in the first experiment were eligible. In the third experiment, eligibility was restricted to households with members living in Ukraine who had not participated in either of the previous experiments. The x-axis represents the number of days since the start of the registration campaigns.

Figure A5: Sample Sizes, Treatment Assignments, and Endline Survey Completion Rates

	First Experiment		Second Experiment		Third Experiment	
Panel A	**First Experiment** Guardians (Students) 2,322 (2,518)		**Second Experiment** Guardians (Students) 2,573 (2,767)		**Third Experiment** Guardians (Students) 4,299 (4,547)	
Panel B	**Treatment** 1,161 (1,259)	**Control** 1,161 (1,259)	**Treatment** 1,286 (1,379)	**Control** 1,287 (1,388)	**Treatment** 2,148 (2,273)	**Control** 2,151 (2,274)
Panel C	*Completed*:	*Completed*:	*Completed*:	*Completed*:	*Completed*:	*Completed*:
Math:	767 (61%)	796 (63%)	740 (54%)	628 (45%)	1,230 (54%)	1,226 (54%)
Ukrainian language:	758 (60%)	802 (63%)	717 (52%)	566 (41%)	1,276 (56%)	1,224 (54%)

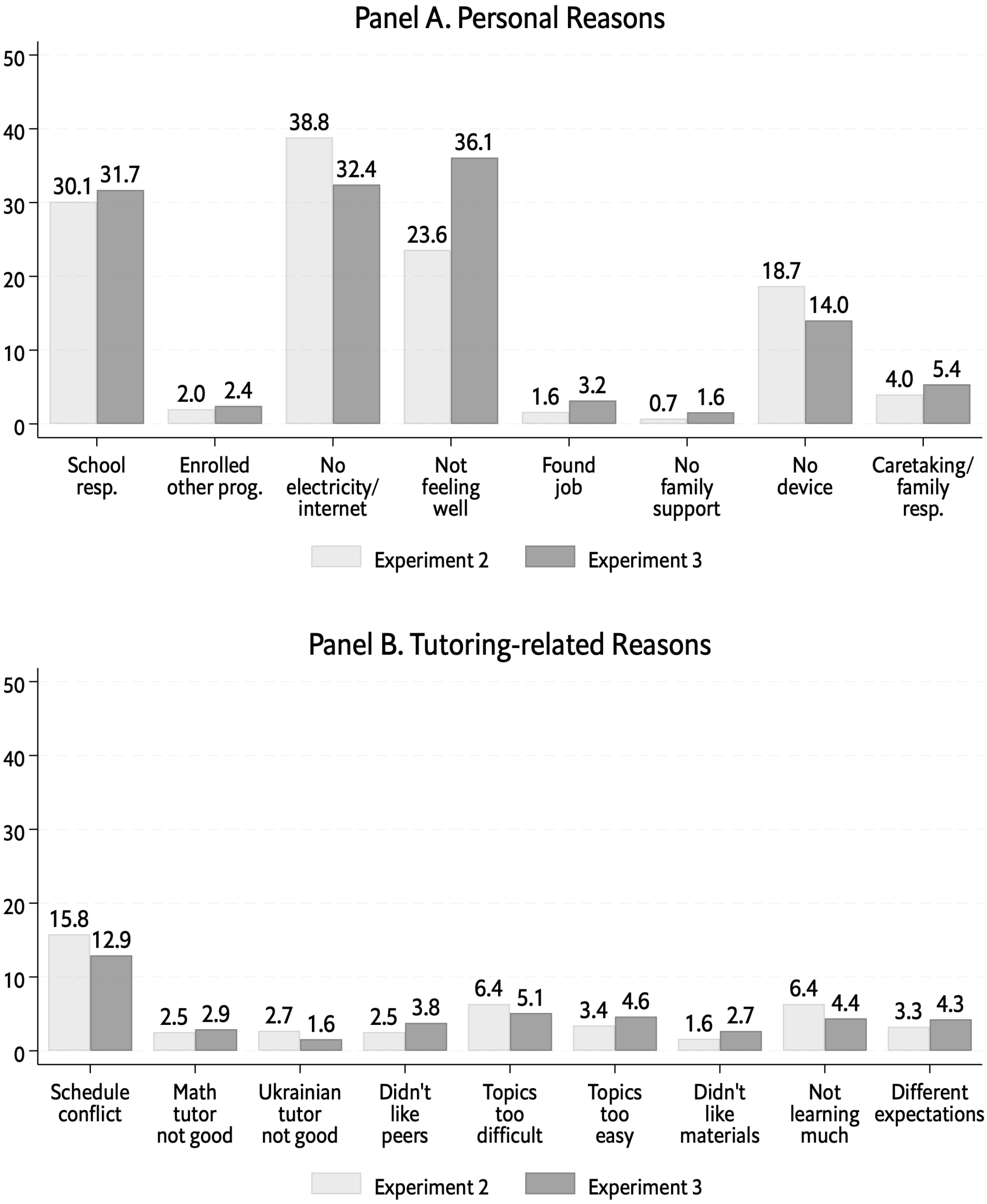
Notes: Panels A and B in the figure show the total number of eligible guardians (households) and students enrolled in each experiment, along with their assignment to the treatment or control group. Across all experiments, 9,194 eligible households and 9,741 students registered for the study. The third panel shows the number of students who completed each of the two endline survey founds (Math and self-reported survey and Ukrainian language), disaggregated by treatment and control groups. Response rates within each group are provided in parentheses.

Figure A6: Attendance Rates by Session and Subject



Notes: This figure presents attendance rates (in %) by session and academic subject, with separate panels for each experiment. Average session attendance is calculated as the percentage of enrolled students who attended each math or Ukrainian language session, disaggregated by experiment.

Figure A7: Reasons for Missing Tutoring Sessions



Notes: This figure shows the percentage of students who reported reasons for missing tutoring sessions in Experiments 2 and 3. Reasons are categorized into two types: Personal reasons (Panel A) and Tutoring-related reasons (Panel B). The y-axis represents the percentage of students citing each reason, while the x-axis lists the specific reasons. This question was asked only to students who missed at least one tutoring session.

Figure A8: Relationship between Tutoring Attendance and Endline Test Scores in Math and Ukrainian language

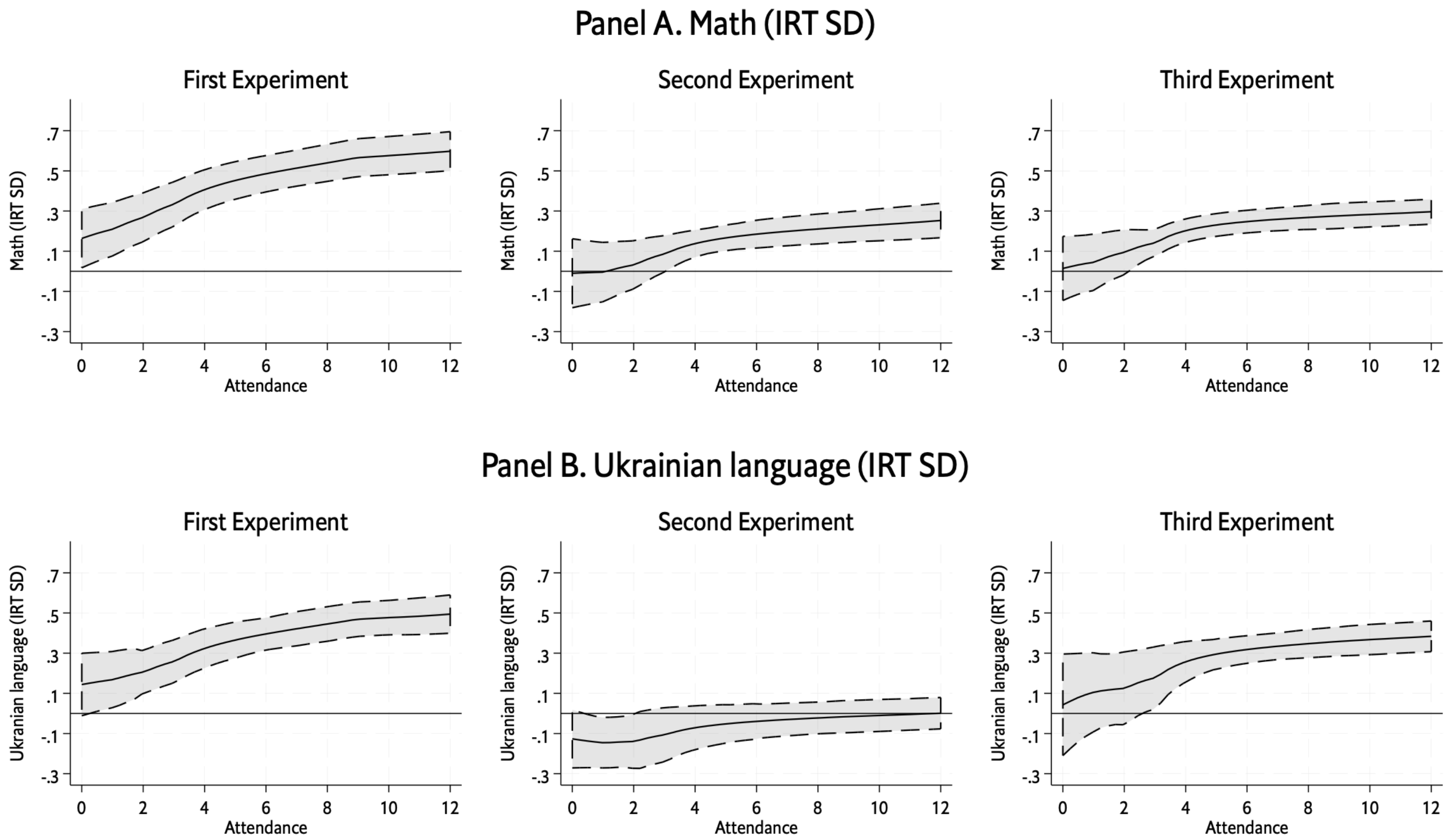
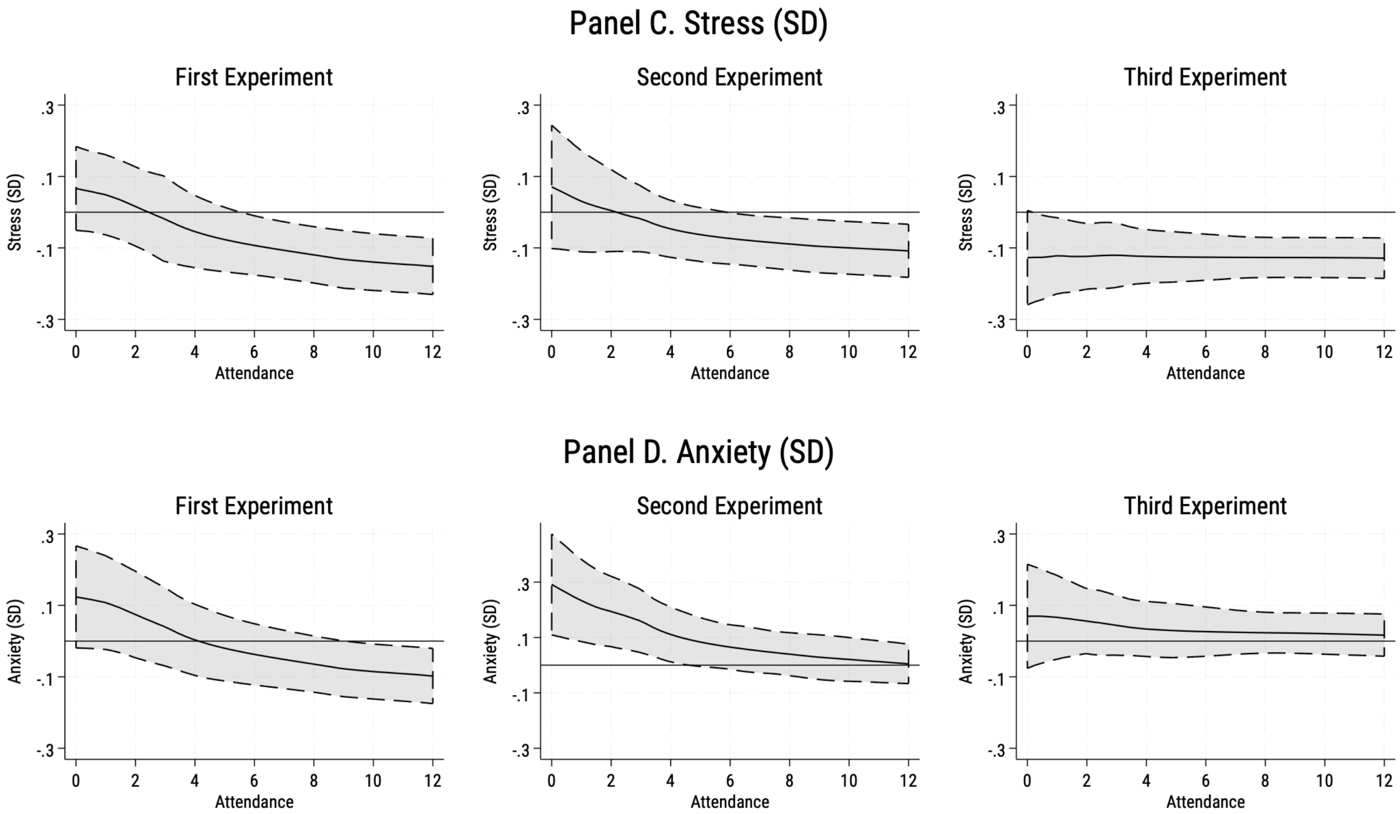
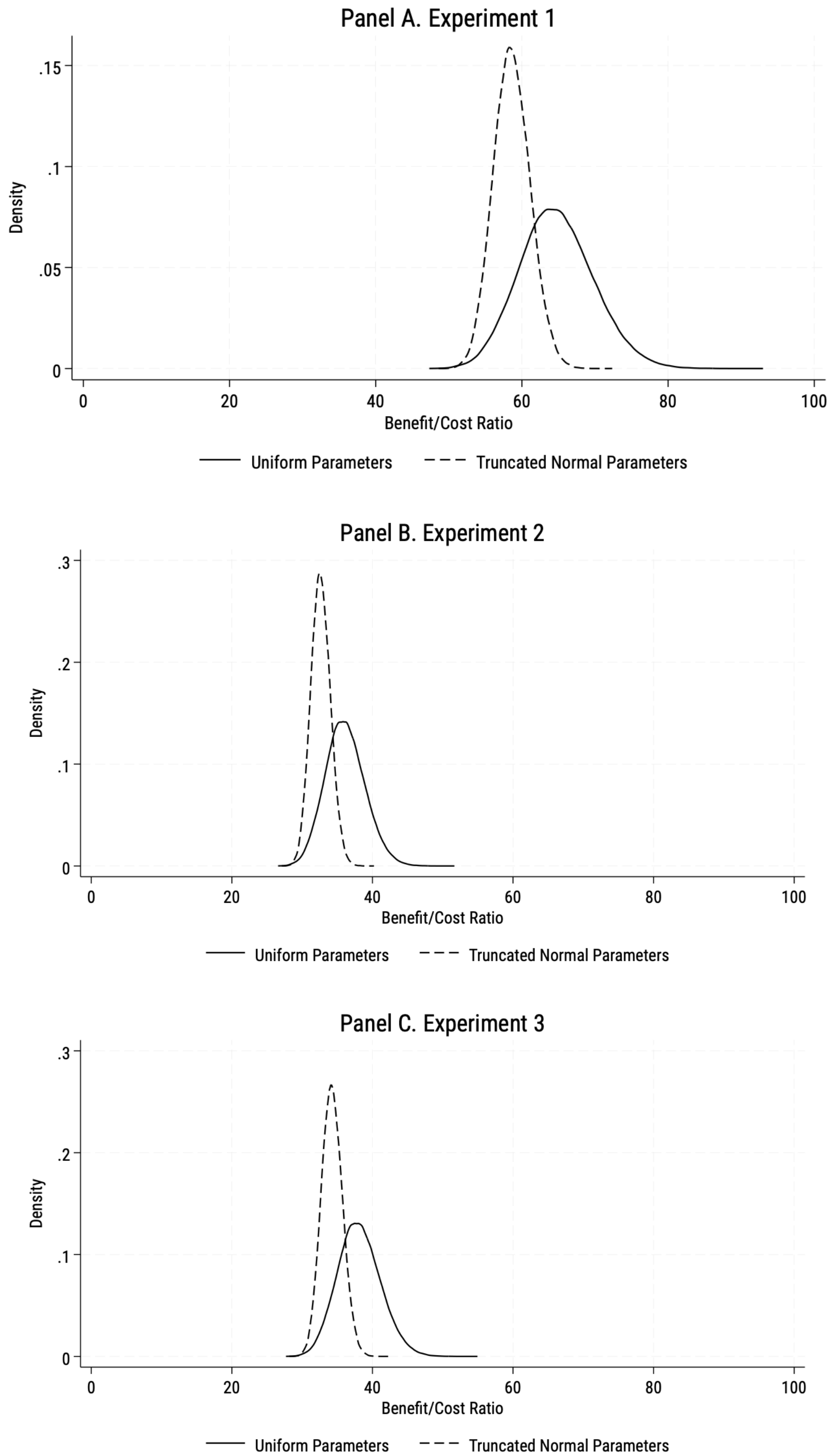


Figure A8: *Continued*–Relationship between Tutoring Attendance and Stress and Anxiety



Notes: These figures present kernel-weighted local polynomial regressions showing the estimated conditional mean of standardized endline outcomes as a function of the number of sessions attended, separated by experiment. Panel A shows the association between attendance at math tutoring sessions and math scores, while Panel B presents the association between attendance at Ukrainian language sessions and Ukrainian language scores. Panels C and D present the relationship between average attendance across both subjects (Math and Ukrainian Language) and stress (Panel C) and anxiety (Panel D), measured in standard deviations. Dashed lines consists of 95% confidence intervals.

Figure A9: Benefit-to-Cost Sensitivity Analysis



Notes: The sensitivity analysis evaluates how changes in critical assumptions, such as wage growth, labor force participation rates, and discount rates, impact the benefit-cost ratio. This involves assessing how variations in parameters like discount rates (ranging from 3% to 7%), wage growth rates (1% to 5%), test-score gains (5% to 11% increase in earnings per standard deviation), and returns of non-cognitive skills (0.5% to 3% increase in earnings per standard deviation) affect outcomes. Two types of distributions are used: a uniform distribution, where each value within a defined range has an equal probability of selection, and a truncated normal distribution, which tends to concentrate values around a central point, with fewer instances of extreme values. This analysis follows the methodology in Ganimian et al. (2024).

Appendix Tables

Table A1: Timing of the Collection of Outcomes and Mechanisms Measures

	Experiment 1		Experiment 2		Experiment 3	
Outcomes	Baseline	Endline	Baseline	Endline	Baseline	Endline
Assessments						
Math		✓	✓	✓	✓	✓
Ukrainian language		✓	✓	✓	✓	✓
Mental Health						
Anxiety (DASS-Y)	✓	✓	✓	✓	✓	✓
Stress (DASS-Y)	✓	✓	✓	✓	✓	✓
Short Grit Scale (Grit-S)			✓	✓	✓	✓
Self-Efficacy				✓		✓
Engagement						
Enrolled in online platform		✓		✓		✓
Interacted in online platform		✓		✓		✓
Number of Interactions		✓		✓		✓
Friends enrolled in program		✓		✓		✓
Subject Enthusiasm						
Math	✓	✓	✓	✓	✓	✓
Ukrainian language	✓	✓	✓	✓	✓	✓
Future Aspirations						
Education Goals	✓	✓	✓	✓	✓	✓

Notes: This table indicates the period (baseline or endline) during which each outcome and mechanism measure was collected, disaggregated by experiment. For example, math scores were collected only at endline (using the endline survey) in the first experiment, but at both baseline and endline in the second and third experiments. Note that measures of student engagement were collected continuously during program implementation by the implementing partner, although the authors only received access to this data after the program concluded.

Table A2: Tutors Characteristics

Variable	All experiments (1) Mean/(SD)	First Experiment (2) Mean/(SD)	Second Experiment (3) Mean/(SD)	Third Experiment (4) Mean/(SD)
Socio-demographics				
Age (years)	41.21(8.89)	41.31(8.57)	41.72(8.83)	41.51(8.84)
Female (%)	0.95	0.94	0.94	0.95
Bachelor/ Specialist	0.51	0.46	0.52	0.54
Masters	0.43	0.48	0.41	0.39
PhD / doctoral	0.05	0.05	0.06	0.05
Professional experience				
Are you trained as a teacher?	0.97	0.98	0.98	0.97
Work Experience (Years)	19.10(9.97)	18.10(10.07)	19.61(10.02)	19.61(9.67)
Teaching Experience (Years)	17.09(10.09)	15.85(9.87)	17.53(10.20)	17.65(9.89)
Have you worked as a tutor before?	0.18	0.25	0.18	0.18
Have you done volunteer work before?	0.56	0.63	0.58	0.57
Taught: Ukrainian language/literature	0.50	0.49	0.49	0.50
Taught: Math	0.48	0.50	0.47	0.49
Mental health				
DASS–Stress				
Normal	0.81	0.82	0.80	0.80
Mild	0.10	0.11	0.10	0.10
Moderate	0.08	0.07	0.08	0.09
Extremely Severe	0.01	0.00	0.01	0.00
Extremely Severe	0.01	0.00	0.01	0.00
Marlow-Crowne Social Desirability	22.92(2.19)	22.89(2.11)	22.95(2.17)	22.98(2.21)
Obs.	326	203	193	238

Notes: This table presents descriptive statistics for tutors characteristics and working experience. Data was collected from a subsample of tutors who responded the tutors survey (76.5%). Tutors who participated in multiple experiments are included in each experiment. For example, 106 tutors from Experiment 1 also participated in Experiment 2, and 139 in Experiment 3. Similarly, 150 tutors from Experiment 2 also participated in Experiment 3. The Social desirability score was collected using the instrument from Crowne and Marlowe (1960).

Table A3: Students Engagement During the Tutoring Sessions

	Mean (1)	SD (2)	Obs. (5)
Panel A. First Experiment			
Was present during this class?	0.61	0.49	27,373
Present for half the class	0.98	0.14	16,821
Had the camera on	0.54	0.50	16,821
Responded to questions	0.97	0.17	16,821
Seemed happy, relaxed, calmed	0.91	0.28	16,821
Seemed tuned-in/paid attention	0.95	0.21	16,821
Came to class unprepared	0.04	0.20	16,821
Did more than required in session	0.35	0.48	16,821
Panel B. Second Experiment			
Was present during this class?	0.55	0.50	30,988
Present for half the class	0.99	0.11	17,179
Had the camera on	0.63	0.48	17,179
Responded to questions	0.98	0.15	17,179
Seemed happy, relaxed, calmed	0.92	0.27	17,179
Seemed tuned-in/paid attention	0.96	0.20	17,179
Came to class unprepared	0.04	0.20	17,179
Did more than required in session	0.32	0.47	17,179
Logged in within 5 minutes	0.94	0.25	17,179
Panel C. Third Experiment			
Was present during this class?	0.57	0.49	52,345
Present for half the class	0.99	0.11	29,981
Had the camera on	0.64	0.48	29,981
Responded to questions	0.98	0.13	29,981
Seemed happy, relaxed, calmed	0.95	0.22	29,981
Seemed tuned-in/paid attention	0.97	0.17	29,981
Came to class unprepared	0.03	0.18	29,981
Did more than required in session	0.38	0.48	29,981
Logged in within 5 minutes	0.95	0.22	29,981

Notes: This table presents descriptive statistics on student engagement during tutoring sessions, as reported by tutors in the tutor journal. Column (1) presents variable mean and column (2) presents the standard deviation (SD). All the variables are measured as indicators, i.e., we recoded the variables: responded to questions, seemed happy, relaxed, calmed, seemed tuned-in/paid attention, came to class unprepared, and did more than required in session to be equal to one if the tutor reported very true or sort of true. Tutors recorded attendance for all students in each session for each subject and provided engagement assessments for those who attended. Consequently, the values presented in column (3) are measured at the student-session-subject level.

Table A4: Impacts of the Online Tutoring Program on Academic and Mental Health Outcomes in the First Experiment

	Academic Outcomes		Mental Health Outcomes	
	Math	Ukrainian language	Stress	Anxiety
	IRT	IRT	SD	SD
	(1)	(2)	(3)	(4)
Treatment	0.488***	0.402***	-0.099*	-0.046
	(0.056)	(0.054)	(0.051)	(0.052)
	[0.000]	[0.000]	[0.079]	[0.384]
Control group outcome mean	0.000	-0.000	0.000	0.000
# of control variables selected	0	0	1	1
Obs.	1,563	1,560	1,562	1,562

Notes: This table presents estimates of β_1 from equation (1) on academic performance (math and Ukrainian language) and mental health outcomes (intensive and extensive margins of stress and anxiety) in the first experiment. All outcomes have been standardized relative to the control group within each experiment. All specifications include controls selected using LASSO and strata fixed effects. The number of selected control variables is presented in row "# of control variables selected." Clustered standard errors at the group level are shown in parentheses. Family-wise p-values are shown in brackets, adjusted for the number of outcome variables, and are estimated using 2,000 bootstraps and the free step-down resampling method of Westfall and Young (1993). In each experiment, the number of outcomes within each family of outcomes consists of the total number of dependent variables shown in the table per academic/mental health domains (i.e., two for academic and two for mental health). Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Table A5: Impacts of the Online Tutoring Program on Academic and Mental Health Outcomes in the Second Experiment

	Academic Outcomes		Mental Health Outcomes	
	Math	Ukrainian language	Stress	Anxiety
	IRT	IRT	SD	SD
	(1)	(2)	(3)	(4)
Treatment	0.215***	0.003	-0.106**	0.026
	(0.051)	(0.053)	(0.048)	(0.048)
	[0.001]	[0.984]	[0.207]	[0.379]
Control group outcome mean	-0.000	-0.000	-0.000	-0.000
# of control variables selected	3	3	3	2
Obs.	1,368	1,283	1,368	1,368

Notes: This table presents estimates of β_1 from equation (1) on academic performance (math and Ukrainian language) and mental health outcomes (intensive and extensive margins of stress and anxiety) in the second experiment. All outcomes have been standardized relative to the control group within each experiment. All specifications include controls selected using LASSO and strata fixed effects. The number of selected control variables is presented in row "# of control variables selected." Clustered standard errors at the group level are shown in parentheses. Family-wise p-values are shown in brackets, adjusted for the number of outcome variables, and are estimated using 2,000 bootstraps and the free step-down resampling method of Westfall and Young (1993). In each experiment, the number of outcomes within each family of outcomes consists of the total number of dependent variables shown in the table per academic/mental health domains (i.e., two for academic and two for mental health). Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Table A6: Impacts of the Online Tutoring Program on Academic and Mental Health Outcomes in the Third Experiment

	Academic Outcomes		Mental Health Outcomes	
	Math	Ukrainian language	Stress	Anxiety
	IRT	IRT	SD	SD
	(1)	(2)	(3)	(4)
Treatment	0.219***	0.320***	-0.118***	0.037
	(0.042)	(0.050)	(0.035)	(0.036)
	[0.000]	[0.000]	[0.004]	[0.635]
Control group outcome mean	0.000	0.000	0.000	0.000
# of control variables selected	5	3	4	5
Obs.	2,456	2,500	2,456	2,456

Notes: This table presents estimates of β_1 from equation (1) on academic performance (math and Ukrainian language) and mental health outcomes (intensive and extensive margins of stress and anxiety) in the third experiment. All outcomes have been standardized relative to the control group within each experiment. All specifications include controls selected using LASSO and strata fixed effects. The number of selected control variables is presented in row "# of control variables selected." Clustered standard errors at the group level are shown in parentheses. Family-wise p-values are shown in brackets, adjusted for the number of outcome variables, and are estimated using 2,000 bootstraps and the free step-down resampling method of Westfall and Young (1993). In each experiment, the number of outcomes within each family of outcomes consists of the total number of dependent variables shown in the table per academic/mental health domains (i.e., two for academic and two for mental health). Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Table A7: Impacts of the Online Tutoring Program on Academic and Mental Health Outcomes
Pooled Experiments

	Control Mean (1)	ITT (2)	TOT (3)	# of control variables selected (4)
Math IRT (SD)	-0.000	0.320***	0.337***	2
		(0.030)	(0.032)	
		[0.000]	[0.000]	
Language IRT (SD)	-0.000	0.281***	0.298***	3
		(0.033)	(0.035)	
		[0.000]	[0.000]	
Stress (SD)	0.000	-0.108***	-0.114***	3
		(0.027)	(0.029)	
		[0.000]	[0.000]	
Anxiety (SD)	0.000	0.008	0.009	4
		(0.028)	(0.029)	
		[0.985]	[0.985]	
Obs.	5,386			

Notes: This table presents intent-to-treat (ITT) and treatment on the treated (TOT) estimates on the main outcomes using pooled data across the three experiments. Column (1) indicates the mean outcome for the control group. Column (2) presents the ITT effects obtained by estimating specification (1). Column (3) presents the treatment on the treated effects estimated using instrumental variables and instrumenting "participation" as having joined at least one session of the program. The outcomes are defined as before. All estimations include control variables selected with LASSO and strata fixed effects, the number of selected variables are shown in Column (4) and are used in both Column (2) and Column (3). Family-wise p-values are shown in brackets, adjusted for the number of outcome variables, and are estimated using 2,000 bootstraps and the free step-down resampling method of Westfall and Young (1993). In each experiment, the number of outcomes within each family of outcomes consists of the total number of dependent variables shown in the table per academic/mental health domains (i.e., two hypothesis tests for academic and two for mental health). Clustered standard errors at the group level are shown in parentheses. Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Table A8: Treatment on the Treated Estimates of the Online Tutoring Program on Academic and Mental Health Outcomes

	Academic Outcomes		Mental Health Outcomes	
	Math	Ukrainian language	Stress	Anxiety
	IRT	IRT	SD	SD
	(1)	(2)	(3)	(4)
Panel A. Experiment 1				
Attended at least 1 session	0.552***	0.460***	-0.112*	-0.052
	(0.063)	(0.062)	(0.057)	(0.059)
	[0.000]	[0.000]	[0.079]	[0.377]
Obs.	1,563	1,560	1,562	1,562
First-stage F-statistic	5,038	4,521	5,023	5,023
# of control variables selected	0	0	1	1
Panel B. Experiment 2				
Attended at least 1 session	0.225***	0.003	-0.111**	0.027
	(0.053)	(0.056)	(0.050)	(0.050)
	[0.001]	[0.959]	[0.052]	[0.576]
Obs.	1,368	1,283	1,368	1,368
First-stage F-statistic	17,773	16,295	17,518	17,606
# of control variables selected	3	3	3	2
Panel C. Experiment 3				
Attended at least 1 session	0.269***	0.346***	-0.120***	0.031
	(0.045)	(0.053)	(0.036)	(0.037)
	[0.000]	[0.000]	[0.002]	[0.393]
Obs.	2,456	2,500	2,456	2,456
First-stage F-statistic	63,313	67,115	65,245	65,066
# of control variables selected	2	2	4	4

Notes: This table shows the second stage estimates of the treatment on the treated using treatment assignment as an instrument for participation in the tutoring program, which is defined as an indicator to whether the student attended at least one session. The outcomes are defined as before. All specifications include control variables selected by LASSO and strata fixed effects. Family-wise p-values are shown in brackets, adjusted for the number of outcome variables, and are estimated using 2,000 bootstraps and the free step-down resampling method of Westfall and Young (1993). In each experiment, the number of outcomes within each family of outcomes consists of the total number of dependent variables shown in the table per academic/mental health domains (i.e., two hypothesis tests for academic and two for mental health). Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Table A9: Endline Completion in the Parental Investment Experiment

	Completed Endline	
	Math (1)	Ukrainian language (2)
Tutoring + Text	0.004	0.009
	(0.036)	(0.036)
	[0.940]	[0.940]
Control group outcome mean	0.583	0.568
# of control variables selected	1	1
Obs.	797	797

Notes: This table presents estimates of the differences on probability of completing the endline survey between the two experimental groups during the parental investment experiment. All estimations include controls selected using LASSO and strata fixed effects. The number of selected control variables is presented in row "# of control variables selected." Clustered standard errors at the group level are shown in parentheses. Family-wise p-values are shown in brackets, adjusted for the number of outcome variables, and are estimated using 2,000 bootstraps and the free step-down resampling method of Westfall and Young (1993). The number of outcomes within each family of outcomes consists of the total number of dependent variables shown in the table (i.e., completed endline math/mental health and endline Ukrainian language). Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Table A10: Endline Survey Completion, by Experiment

	Completed Endline Academic/Mental Health Assessment		
	First Experiment	Second Experiment	Third Experiment
	(1)	(2)	(3)
Panel A. Math/Mental Health			
Treatment	-0.023	0.080***	-0.004
	(0.019)	(0.018)	(0.014)
	[0.233]	[0.000]	[0.888]
Control group outcome mean	0.632	0.452	0.539
# of control variables selected	0	5	4
Obs.	2,518	2,767	4,547
Panel B. Ukrainian language			
Treatment	-0.035*	0.108***	0.018
	(0.020)	(0.018)	(0.014)
	[0.113]	[0.000]	[0.170]
Control group outcome mean	0.637	0.408	0.538
# of control variables selected	0	5	3
Obs.	2,518	2,767	4,547

Notes: This table presents estimates of the differences on probability of completing the endline survey between treatment and control groups during each of the experiments. As explained in section 4.1, the endline data collection was conducted in two rounds: one round that included the math assessment and the survey and another round that included the Ukrainian language assessment only. Considering this, we assess endline completion by round and present the results in panels A and B. All estimations include controls variables selected using LASSO and strata fixed effects. The number of control variables selected by LASSO is presented in row "# of control variables selected." Clustered standard errors at the tutoring group level are shown in parentheses. Family-wise p-values are shown in brackets, adjusted for the number of outcome variables using 2,000 bootstraps and the free step-down resampling method of Westfall and Young (1993). In each experiment, the number of outcomes within each family of outcomes consists of the total number of dependent variables shown in the table (i.e., endline completed math/mental health and endline Ukrainian language). Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

Table A11: Fairlie Bounds

	Experiment 1			Experiment 2			Experiment 3		
	Main	Fairlie 5%		Main	Fairlie 5%		Main	Fairlie 5%	
	Estimate	Lower	Upper	Estimate	Lower	Upper	Estimate	Lower	Upper
Outcome	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)
Math (IRT)	0.488***	0.449	0.534	0.215***	0.160	0.265	0.219***	0.209	0.304
Ukrainian language (IRT)	0.402***	0.366	0.449	0.003	-0.078	0.031	0.320***	0.286	0.391
Stress (SD)	-0.099*	-0.139	-0.061	-0.106**	-0.145	-0.045	-0.118***	-0.171	-0.080
Anxiety (SD)	-0.046	-0.086	-0.008	0.026	-0.014	0.089	0.037	-0.018	0.075

Notes: This table presents the results from a bounds analysis performed following Fairlie et al. (2015). For each experiment, the lower (upper) bound the mean minus (plus) 5% of the standard deviation (SD) of the observed treatment group distribution for attritors was imputed to the treatment group. In addition, the mean plus (minus) the 5% of the SD of the observed control group distribution for attritors was imputed to the control group. The outcomes are defined as before. Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.