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

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

September 23, 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 invasion of Ukraine. Conducted between early 2023 and mid-2024, the experiments reached nearly 10,000 students across all regions of Ukraine. The program offered three hours per week of small-group tutoring in math and Ukrainian language over six weeks, and used academic and psychosocial tools to address student challenges at different intensities of disruption. Results show that the program led to substantial improvements in learning—up to 0.49 standard deviations in math and 0.40 standard deviations in Ukrainian language—and consistent reductions in stress—up to 0.12 standard deviations. High take-up and engagement rates were observed, and four mechanisms were identified as drivers of impact: structured peer interactions, improved attitudes toward learning, enhanced socio-emotional skills, and increased student investments. A complementary experiment using information nudges to increase parental engagement highlights challenges in promoting parental investments in a conflict setting. The program was cost-effective across all experiments, with benefit-to-cost ratios ranging from 31 to 56, and scalable given its reliance on existing infrastructure and teaching capacity.

***Keywords***— Ukraine, wartime, tutoring, student 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, Damien De Walque, Alejandro Ganimian, Lucas Gortazar, Lawrence Katz, Jostin Kitmang, David McKenzie, 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, Insper, CAF, 2025 ESOC Conference, 2024 EGAP Conference, and the 2024 NEUDC conference. We are deeply grateful for the incredible support of Teach for Ukraine, led by Oksana Matiiash, for implementing the intervention, and to Yulia Bilyk, Sylvia Cesar, Juan Miguel Jimenez, Gabriele Sabino, and Maria Montoya Aguirre 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 delivery of public 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 addresses this gap by evaluating an online tutoring program offered to Ukrainian students during Russia's invasion of Ukraine, providing evidence from three consecutive experiments on the program feasibility, impact on learning and mental health, and scalability. While tutoring programs have proven effective in non-war settings (Nickow et al., 2023), it is unclear whether the key components determining their impact hold up in wartime conditions. First, logistical challenges such as power outages and frequent displacement may hinder participation. Yet, students' intrinsic motivation to recover lost learning may lead to high engagement despite adverse conditions. Second, the mechanisms through which tutoring operates may function differently during wartime. For example, structured peer interactions may both provide emotional support and facilitate learning, but also transmit stress or introduce distractions. Third, contextual factors can moderate program impacts. For instance, the psychological toll of conflict may impair students' ability to focus and reduce parents' capacity to support their education.

To evaluate the potential of online tutoring to contribute to human capital accumulation during wartime, we partnered with Teach for Ukraine[1] to design a program for students in grades 5 to 10. The core structure of 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. The experiments took place between early 2023 and mid-2024—a period marked by frequent armed attacks and

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

substantial instability in Ukraine (Figure A1). In each experiment the program content was adapted to address the evolving challenges faced by students and teachers.

The first experiment began in mid-February 2023, 11 months after large-scale disruptions began, 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, aligned with the Ukrainian curriculum, aimed to help students catch up on core subjects by adapting sessions to students' needs. The second experiment began in late March 2023 as students were completing their first full school year amid ongoing instability. The program maintained the same structure as the first experiment but introduced diagnostic assessments to group students by ability and provided diagnostic reports and formative assessments to tutors. The third experiment began in early February 2024. 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.

Across all the three experiments, the sample included 9,194 households (2,322 in the first experiment, 2,573 in the second, and 4,299 in the third) and 9,832 students (2,518 in the first, 2,767 in the second, and 4,547 in the third). In each experiment, the program was advertised online, with a 30-day enrollment window. Due to the high volume of applications exceeding TFU's capacity, we conducted stratified random assignment of eligible households to treatment or control groups.[2] We used parental education and region of residency as stratification variables. To assign students to groups, we created strata using student treatment status, grade, and preferred schedule for sessions. Students were then randomly assigned to groups in the first experiment and based on a ranking by baseline ability in the second and third experiments.

We develop a conceptual framework to guide our empirical analysis. We identify two human capital outcomes that the intervention may primarily impact: academic outcomes and mental health. We further propose that program impact on these outcomes will depend on feasibility, four candidate mechanisms—structured peer interactions, attitudes toward learning and aspirations, social-emotional skills, and student and parental

[2] The eligibility criteria for the first experiment were guardian consent and student assent. For the second and third experiments, we also required that only students who had not enrolled in previous experiment(s) were eligible to participate. This ensured no overlap in the samples of students and households across experiments, mitigating risks of potential contamination. However, since the samples are independent, we are unable to compare effect sizes across experiments.

investments—and relevant contextual and individual characteristics.[3]

We collected data before (at registration), during, and at the end of each experiment. Following our conceptual framework, we measured student take-up, academic outcomes (math and Ukrainian language scores), mental health (stress and anxiety), and potential mechanisms (peer interactions, attitudes, aspirations, socio-emotional skills, and investments). Data from students came from various sources: academic performance data was collected through self-administered assessments in math and Ukrainian language, and data on mental health and mechanisms were collected through self-reported questionnaires. Data on student take-up and interactions was tracked through the online platform, while attendance and engagement was collected by tutors through daily journals. Finally, we also collected data from parents (or guardians) and tutors on mental health and sociodemographic characteristics at registration.

We document five insights into human capital investments during conflict. First, the tutoring program led to sizable improvements in students' academic performance and mental health. Intention-to-treat estimates show that 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). Stress levels improved consistently (between 0.10 to 0.12 SD); anxiety levels did not.

Second, implementing online tutoring during wartime is feasible. Take-up and attendance exceeded expectations: 68–72% of students attended at least one math session and up to 71% attended Ukrainian sessions, with average participation above six sessions (out of twelve) per subject. Attendance remained stable, and engagement—measured by attentiveness, participation, and preparedness—was consistently high. Most absences were due to external constraints such as power outages or illness, rather than dissatisfaction with the program.

Third, we explore mechanisms driving program impact as outlined in the conceptual framework. Treated students were more likely to interact and engage with tutors and peers through the tutoring platform. Their attitudes toward learning, persistence and self-efficacy also improved, though aspirations did not. In terms of student investments,

[3] As discussed in Section 7, there may be additional mechanisms at play. We focus on these four as they have been previously documented in non-war settings (Nickow et al., 2023).

treated students were more likely to seek additional tutoring support and spent more time using online academic resources. However, parental investments did not emerge as a mechanism driving program impact. To test this, we conducted a parallel experiment. At the end of the first experiment, we randomly assigned a sub-sample of non-treated households to two groups. Both groups were offered the tutoring program as previously delivered, but parents of students in the second group also received text messages with tips for support. Academic outcomes declined in this group, highlighting the challenges of engaging parents during wartime.

Fourth, we find that program impacts vary by student characteristics, student baseline ability level, parental well-being, and conflict intensity. Benefits were greater for older students, those underperforming at baseline, and those with more stressed parents. In contrast, students living in regions more severely affected experienced smaller learning gains, possibly due to distractions or disruptions during sessions. Finally, we find that the program is highly cost-effective, with benefit-to-cost ratios ranging from 31 to 56. Under a more conservative scenario where only 20% of the treatment effect persists after 5 years, the benefit-to-cost ratios still range from 6.5 to 11.6.

Our paper contributes to several strands of literature. First, it contributes to the growing body of research on strategies to mitigate the effects of school closures on learning and mental health, particularly in emergencies. Research has shown that school disruptions caused by wars, pandemics, or natural disasters can have short- and long-term negative effects on human capital.[4] In particular, wars can cause prolonged, large scale destruction of both human and physical capital and create severe resource constraints for governments, 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; Carlana and La Ferrara, 2021; Gortazar et al., 2023). Our paper contributes to this literature by providing experimental evidence on the effects of online tutoring during war, where traditional recovery strategies are more challenging to implement.

[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).

Second, this paper adds 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,[5] and across three distinct phases – 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. The program's structure also builds on Ukraine's existing education system by leveraging trained teachers with available capacity, making it operationally feasible and institutionally aligned for broader implementation. 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 of an educational intervention in an active war context.

Finally, this paper contributes to research on remedial education 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 assessing student learning levels with a simple tool, grouping them by learning levels,

[5] The intervention reached students across all regions of Ukraine—including high-conflict areas—and participants closely resembled the national student population on key characteristics.

and using a variety of teaching and learning activities. By embedding these practices into the program, we provide evidence of the academic benefits of remedial education in a wartime context.

## 2 Context and Intervention

Since February 2022, the Ukrainian government has significantly reduced investment in the education sector. In 2021, education spending—adjusted to 2015 prices–was 178.6 billion Ukrainian hryvnias, accounting for 17% of total expenditure. This figure fell to 138.1 billion hryvnias (9.6%) in 2022 and further declined to 110.5 billion hryvnias (7%) in 2023. At the same time, the Ministry of Education and Science estimated that between 2,900 and 3,500 schools (10%–13% of all schools) were either partially damaged or completely destroyed, compounding the challenges posed by reduced funding (World Bank et al., 2023; World Bank, 2024).

To enable safe in-person learning amid air raid threats, schools were required to have bomb shelters. While some schools met this requirement, many did not, limiting in-person instruction. As a result, only 30% of secondary schools offered fully in-person classes; 34% operated exclusively online, and 36% adopted a blended approach.[6] Online or blended instruction was not new to Ukrainian students, who had experienced school closures during the COVID-19 pandemic. The Ministry's All-Ukrainian School Online platform, developed as a response to the pandemic, provided curriculum-aligned content, including video lessons, tests, and assignments, for grades 5–11 in 18 core subjects. However, sustained attacks on critical infrastructure, particularly the energy grid, frequently disrupted access to online education.[7]

Despite these adaptations, education disruptions led to large learning losses. In 2022, Ukrainian students scored 428 in reading and 441 in mathematics on the PISA assessment, well below the OECD averages of 476 and 472, respectively (OECD, 2023). These gaps correspond to approximately 0.48 and 0.31 standard deviations.[8] In response, the Ministry of Education and Science expressed interest in partnering with local organizations to identify feasible strategies to support students during the ongoing crisis.

[6] Information provided by the Ministry of Education and Science.

[7] The telecommunications sector relied on generators to maintain services during power outages and missile strikes.

[8] The 2022 PISA sample included only 18 regions; heavily disrupted areas were excluded, suggesting these are likely upper-bound estimates. In 2018, when all 27 regions were included, Ukraine's scores were close to the OECD average.

Following the MOES demand, we partnered with Teach for Ukraine (TFU) in early 2023 to adapt and evaluate an online tutoring program designed to supplement the education Ukrainian students were receiving through existing modalities. The program was implemented through three consecutive experiments, targeting students in grades 5 to 10. Across all experiments, the program offered three hours of tutoring per week, divided into two 90-minute sessions, each covering 45 minutes of math and 45 minutes of Ukrainian language. Sessions were designed to be delivered in small groups of three students over six weeks. Students aged 10 to 17 were eligible to participate, contingent on parental consent and student assent. Across experiments, the program was adapted to meet the evolving needs of students and tutors in the wartime context.

Tutors recruitment and their assignment to groups of students were conducted exclusively by TFU. Tutors were required to have prior teaching experience, familiarity with online instruction and digital tools, scheduling flexibility, and a demonstrated commitment to education.[9] Shortlisted candidates for the tutoring positions were interviewed by TFU staff. Finally, selected tutors participated in a two-day training program conducted by the International Tutoring Academy (Kyiv), which covered socio-emotional support strategies to help tutors foster motivation and a positive learning environment, and neurodidactics to strengthen their understanding of how students process and retain information.[10]

**First experiment.** The first experiment ran from late January to March 2023. By this time, in-person instruction had resumed only in schools with adequate shelters. Students experienced varied formats of learning—in-person, online, and blended. The tutoring program aimed to mitigate learning losses from prolonged school closures. Students were randomly assigned to small groups, and tutors followed Ukrainian curriculum goals. The program was designed to give tutors flexibility, allowing them to assess the needs of their students and adapt their approach accordingly. The program was open to both students residing in Ukraine and Ukrainian refugees abroad.

**Second experiment.** The second experiment ran from late April to mid-June 2023, coinciding with the end of the school year. The Ministry of Education and Science canceled final state examinations for the second year to reduce burdens on students and teachers.

The tutoring program continued to focus on helping students catch up academically,

[9] These criteria ensured that tutors were at least as qualified as students' regular school teachers. As shown in Section 5.1, 97% of tutors held formal teaching credentials, and nearly all had at least a bachelor's degree; 48% held a master's or doctoral degree.

[10] Tutors provided consent to participate in both the program and study.

while also supporting them in completing their first, full school year amid large-scale disruptions. We introduced three new tools to better support tutors. First, students completed an online short multiple-choice diagnostic test during registration to assess prior knowledge; this allowed us to group them by ability.[11] Second, tutors received diagnostic reports from short, multiple-choice assessments showing their students' results relative to the full participant pool before the first week of the program.[12] Third, tutors received short, curriculum-aligned formative assessments—consisting of five open-ended questions per subject and grade level—to use at their discretion for tracking student progress. The program remained open to students inside Ukraine and Ukrainian refugees abroad

**Third experiment.** The third experiment ran from February to early April 2024. By this time, internal displacement had declined significantly: the number of internally displaced persons fell from 5.4 million in December 2022 to 3.7 million by September 2023, while 4.6 million had returned. Although 80% of schools had bomb shelters by late 2023, online education remained critical due to shelter capacity limits and continued closures in high-risk zones. Meanwhile, the mental health burden on children intensified. A 2023 study of more than 8,000 Ukrainian adolescents found that those exposed to conflict 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 specific psychosocial support tools into their sessions. In partnership with the Harvard Program on Refugee Trauma (HPRT), we developed four trauma-informed care exercises for tutors to integrate into sessions. The first was a storytelling activity based on the fable Stone Soup, designed to promote resilience and a sense of community. The second involved guided deep breathing, which aimed to promote emotional calmness and was reinforced with instructional videos. The third combined icebreakers with tapping techniques: students chose an animal and a trait (e.g., "eagle – strong"), then tapped an acupuncture point while repeating an affirmation. The fourth, called Happy Faces, invited students to indicate their mood at the beginning of each session using emojis, allowing tutors to identify students needing additional support. Tutors were trained on these exercises and provided with a manual on how to integrate these activities into their sessions. Although the tools from the second experiment remained available, the third experiment emphasized psychosocial support over academic support tools. Further details on tutoring session content is presented in Appendix B.

[11] The test consisted of 12 items: 5 in math and 7 in Ukrainian language. Students were ranked by the number of correct responses within their strata and grouped into triplets by ability.

[12] Figure A2 provides an example of a math group's diagnostic report.

# 3 Conceptual Framework

We develop a parsimonious model to explore how the key components that contribute to the success of tutoring interventions in non-war settings—as documented by Nickow et al. (2023) and others—may behave differently during wartime, with implications for the accumulation of human capital.

In the absence of conflict, we assume that student $i$ accumulates human capital $h_{i,t}$ in period $t$ according to the following equation:

$$h_{i,t} = s_{i,t} + m_{i,t} + (1 - \delta_t)h_{i,t-1} + \theta_i T_{i,t} + \epsilon_{i,t} \tag{1}$$

where $s_{i,t}$ and $m_{i,t}$ denote human capital acquired through formal schooling and mental health investments, respectively. The parameter $\delta_t$ represents the depreciation rate of prior human capital ($h_{i,t-1}$). $T_{i,t}$ captures additional investments in human capital beyond $s_{i,t}$ and $m_{i,t}$, and $\theta_i$ captures the student-specific return to these additional investments. $\epsilon_{i,t}$ is an idiosyncratic shock.[13]

For the purpose of our study, $T_{i,t}$ is exogenous due to random assignment and consists of an online tutoring program that provides learning catch-up and mental health support.[14] We define $\theta_i$ as a function of: (i) program participation or take-up, $\gamma_i \in \{0, 1\}$, (ii) mechanisms through which the online tutoring program operates, $m(\text{Mech}_i)$, and (iii) individual or contextual characteristics, $g(C_{i,t-1})$, that moderate the impacts of the program. These characteristics are assumed to be fixed prior to the intervention (in period $t - 1$). We then model the parameter $\theta_i$ according to the following equation:

$$\theta_i = f(\gamma_i, m(\text{Mech}_i), g(C_{i,t-1})) \tag{2}$$

In the presence of conflict, human capital accumulated through $s_{i,t}$, $m_{i,t}$, and $(1 - \delta_t)h_{i,t-1}$ is reduced by a wartime shock $\rho_i \in \{0, 1\}$, such that the student retains only a fraction $(1 - \rho_i)$ of what would have otherwise been accumulated in the absence of conflict. We assume that all these other elements are similar, on average, for those assigned ($T_{i,t} = 1$) or not ($T_{j,t} = 0$) to the online tutoring program. At the same time, the presence

[13] We use a discrete-time model to illustrate the effect of external inputs on human capital accumulation. For simplicity, we assume time allocation decisions are fixed.

[14] We acknowledge that there may be other human capital investments available, as well as other contextual conditions affecting $h_{i,t}$, which are similar, on average, for those assigned or not to the intervention targeting human capital.

of conflict may amplify or diminish the program's impact—by affecting the feasibility of participation, altering the operation of key mechanisms, or shifting the relevance of individual and contextual characteristics. Our objective is to explore how each component of $\theta_i$, in the presence of conflict, behaves when $\rho_i > 0$. We discuss each component below.

**Program participation ($\gamma_i$).** During wartime, logistical challenges such as power outages, internet disruptions, evacuations, and displacement may drive $\gamma_i$ close to zero. However, students' intrinsic motivation to catch up academically when faced with disruptions to formal schooling may sustain high levels of participation, even under adverse conditions.

**Mechanisms ($f(\textbf{Mech}_i)$).** Drawing on evidence from tutoring programs in non-war settings (Nickow et al., 2023), we assume that the function $m(\text{Mech}_i)$ includes *at least* the following mechanisms: structured peer interactions ($\text{Peers}_i$), students' attitudes and aspirations toward learning ($\text{Attitudes}_i$), socio-emotional skills ($\text{SES}_i$), and additional investments complementary to tutoring by students ($\text{SI}_i$) or their parents ($\text{PI}_i$).[15]

First, structured peer interactions ($\text{Peers}_i$) are known to play a critical role in the emotional well-being and learning of students during stable times (Roseth et al., 2008; Slavich and Zimbardo, 2012). In wartime, opportunities to interact with peers in a safe and structured environment may reduce isolation, foster emotional support, and facilitate knowledge-sharing. At the same time, peer interactions can also transmit stress or trauma, introduce distractions or disruptions, and exacerbate exclusion or bullying (Carrell et al., 2018), which can be particularly salient for displaced or highly affected students.

Second, evidence on attitudes towards learning and educational aspirations ($\text{Attitudes}_i$) from non-conflict settings suggests that positive attitudes and higher aspirations are associated with greater effort and improved outcomes (Khattab, 2015; Jacob and Wilder, 2010; Carlana and La Ferrara, 2021; Golan and You, 2021). During wartime, students can improve their attitudes toward school and long-term educational expectations as a way to foster resilience. War, however, may severely lower students' perceived returns to education, shifting aspirations towards short-term survival or work, and therefore tutoring may have muted effects (Justino, 2011).

Third, social-emotional skills ($\text{SES}_i$), including perseverance, emotional regulation, and self-efficacy can help students cope with adversity and persist with learning, even under stress (Dinarte-Diaz and Egana-delSol, 2024). However, these same skills may be impaired by fear, trauma, or displacement, which reduce attention, motivation, and capacity for learning or to cope with mental health issues (Justino, 2011).

[15] We acknowledge that other mechanisms may be at play that we do not observe with our data.

Finally, students motivated to recover lost learning may increase effort ($SI_i$), and parents may respond to crisis by reinforcing educational routines and monitoring participation in the tutoring program ($PI_i$). However, war can suppress both student effort and parental support if household members perceive education as either unsafe or a lower priority (Justino, 2011).

**Contextual factors ($g(C_i)$).** Contextual factors such as exposure to violence and mental health status of the child's parent or guardian can moderate program impacts. For instance, the academic performance and mental health of students can be affected differently depending on the level of violence they are exposed to (Valente, 2014; Rodriguez and Sanchez, 2012). In addition, students who entered the program with higher psychological distress may have more difficulty engaging (Brück et al., 2019). On the other hand, these same students may benefit most from the psychosocial elements of the intervention. Parental trauma (e.g., stress, depression, aggression) can further shape the home environment in ways that either support or hinder student outcomes (Betancourt et al., 2013; Shonkoff et al., 2012).

The salience of conflict may also vary by individual characteristics, such as gender or age. For instance, older boys may face heightened risks of conscription or displacement during war (Blattman and Miguel, 2010; Justino, 2011; Justino et al., 2014; Swee, 2015; Bertoni et al., 2019), potentially altering their engagement with education and responsiveness to tutoring.

This framework guides our interpretation of the empirical findings by exploring how conflict-related shocks ($\rho_i$) may alter program participation, the operation of key mechanisms, and the influence of contextual characteristics. These shocks may lead to different outcomes than those observed in non-war settings, where these components behave under different conditions.

# 4 Experimental Design

## 4.1 Recruitment of Students

For each of the three experiments, student registration was actively promoted through TFU's social media platforms for approximately 30 days prior to the program's start. Parents or guardians could register their children by clicking the invitation link, which first prompted them to provide consent. If consent was granted, they completed a 10-minute self-administered survey. At the end of the survey, instructions directed parents to hand

their internet-enabled device to each child they wished to enroll. Children who assented to participate in both the program and the study then completed a 35-minute survey to assess their eligibility. As shown in Figure A3, demand for the tutoring program exceeded TFU's implementation capacity in all three experiments. Between 2,500 and 4,800 households registered across the waves. Due to oversubscription, eligible households were randomly assigned to treatment or control groups, as described in Section 4.2.

## 4.2 Randomization

Randomization was conducted at the household level to mimic real-world implementation conditions, where all grade-eligible students within a household—not just some grade-eligible individual students—would have access to the program if it were scaled up. This household-level randomization approach helped capture spillover effects that would occur at scale, providing more accurate estimates of the program's potential impact. Sample sizes for enrolled and assigned households and students are shown in Figure A4.

**First experiment.** Registration occurred between December 2022 and January 2023. A total of 2,322 eligible households (2,518 students) completed 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. Randomization was stratified by (i) whether the student's parent had completed higher education and (ii) whether the student was residing in Ukraine at registration.

Following assignment to treatment or control, students were stratified by treatment status, grade, and preferred schedule, and randomly assigned to groups of three.[16] Students in the treatment group were assigned two subject-specific tutors (one for math and one for Ukrainian language) and remained with the same peer group across subjects.

**Second experiment.** Registration occurred between March and April 2023 and followed the same procedure as the first experiment, with one key difference: enrolled students completed a 12-question diagnostic assessment in math and Ukrainian language. Eligibility was restricted to students who had not participated in the first experiment. A total of 2,573 households (2,767 students) were eligible. We randomly assigned 1,286 households (1,379 students) to the treatment group and 1,287 households (1,388 students) to the control group. Stratification variables included (i) parental education and (ii) the student's

[16] Group size depended in part on scheduling preferences and tutor availability. Across experiments, 79.2% of groups had 3 students, 19.7% had 1 or 2 students, and 1.1% had 4 or 5 students.

region of residence, categorized into five regions: central, eastern, southern, western, and outside Ukraine.

As in the first experiment, students were stratified by treatment status, grade, and schedule preference. Within each stratum, students were ranked by their diagnostic test score and assigned to groups of three students in rank order. Students in the treatment group were again assigned subject-specific tutors.

**Third experiment.** Registration occurred between December 2023 and January 2024, with two additional eligibility criteria: students had to reside in Ukraine and could not have participated in the previous two experiments. A total of 4,299 eligible households (4,547 students) completed the baseline survey. We randomly assigned 2,148 households (2,273 students) to the treatment group and 2,151 households (2,274 students) to the control group. The grouping procedure was the same as those in the second experiment.

Three design features were consistent across experiments. First, each wave was conducted independently, with no sample overlap. We verified non-overlap by checking national IDs, student names, and guardian contact information. Second, students in the control group were guaranteed access to the tutoring program after follow-up data collection. Third, all students, regardless of treatment status, were invited to join a dedicated Discord group with their assigned peers.[17] A summary of the experimental design for each experiment is presented in Table A1.

# 5 Data

## 5.1 Data Collection

Each experiment involved three rounds of data collection. First, we collected baseline data from parents, students, and tutors during the enrollment stage, before the start of the intervention. Second, to assess implementation fidelity and take-up, we collected data on student attendance and engagement during tutoring sessions from tutor-reported session journals. Third, we administered follow-up surveys to students shortly after each experiment concluded to capture short-term program impacts while minimizing attrition. To reduce survey fatigue and improve data quality, we split the student surveys into two parts and sent two separate links: one link included the Ukrainian language assessment;

[17] This design feature allows us to test whether simply enabling online peer interaction is sufficient for impact, or whether structured tutor-led engagement is necessary.

the other included the math assessment, mental health module, and other outcomes modules.

For the first experiment, which ended in March 2023, follow-up data were collected from 1,563 students (62% response rate) between March and April. For the second experiment, which ended in June 2023, data from 1,368 students (49.4%) were collected between June and July. For the third experiment, which ended in March 2024, data from 2,500 students (54%) were collected in April 2024. Among students who completed at least one survey, 84.2% completed both, 7.5% completed only the Ukrainian assessment, and 8.3% completed only the math assessment and other modules.[18]

## 5.2 Outcomes, Instruments, and Other Variables

The selection of outcomes was guided by our conceptual framework and pre-specified in the American Economic Association RCT registry (AEARCTR-0010634). Table A2 summarizes the outcomes collected in each experiment. Detailed descriptions of each instrument are provided in Appendix C.

### 5.2.1 Main outcomes

**Academic learning outcomes:** Academic performance was measured through math and Ukrainian language tests. The math test, developed by an assessment expert using donated items from McGraw Hill, and the Ukrainian language test, jointly developed by an expert team with newly created items, each contained 30 items and were aligned with the curriculum for their respective grades. To estimate learning 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 used the stress and anxiety subscales of the Youth Depression, Anxiety, and Stress Scale (DASS-Y) (Szabo and Lovibond, 2022). We excluded the depression subscale to reduce survey burden and because the program's potential impacts on depression were expected to be limited. To estimate the mental health outcomes, we standardize scores for stress and anxiety separately, relative to the control group within each experiment. Higher standardized scores indicate higher levels of stress or anxiety.

[18] We observe differences in response rates by treatment status in the second experiment for both assessments, and a smaller difference in the Ukrainian assessment in the first experiment. We address this using bounds analysis following Fairlie et al. (2015). See Section 8 for details.

### 5.2.2 Feasibility measures

To assess feasibility, we collected data on attendance and engagement from tutor session journals, which recorded whether students attended, turned on their cameras, were prepared, participated, and paid attention. TFU also provided administrative records on platform enrollment, and students self-reported their attendance in the follow-up survey.

### 5.2.3 Mechanisms

**Structured peer interactions:** Across all experiments, we collected data on whether students interacted with other peers from different sources. We used platform administrative data to measure online peer interactions (if they interacted at least once or the total number of interactions) and asked students whether at least one friend had participated in the tutoring program.

**Social-emotional skills:** In the second and third experiments, we measured *persistence* using the eight-item grit scale from Duckworth and Quinn (2009). Responses were on a 1–5 scale and standardized relative to the control group. The higher score, the greater the student's persistence. In addition, we measured *self-efficacy* using the General Self-Efficacy Scale (GSE) (Schwarzer and Jerusalem, 1995),[19] a ten-item scale scored from 10 to 40. Scores were standardized relative to the control group, higher values indicate greater self-efficacy.[20]

**Attitudes toward learning and future aspirations:** To assess attitudes, we asked students how much they liked math and Ukrainian language. Responses were on a 1–5 scale; we constructed an indicator equal to 1 if they reported liking either subject "much" or "very much." For aspirations, students were asked how long they intended to continue studying. We coded as 1 those who indicated intentions to pursue vocational training or university, that is, educational aspirations beyond high school.

**Complementary student investments:** We measured whether students had accessed additional tutoring or subject-specific support in the six weeks preceding the survey, beyond the TFU program.[21] We also asked students to report daily time spent on the All-Ukrainian Online School platform and on homework.

[19] Self-efficacy has been found to predict better performance in both academic and non-academic domains (Bandura and Wessels, 1997; Schunk and DiBenedetto, 2016, 2022).

[20] We also collected data on locus of control using the instrument from Carlana and La Ferrara (2021), but dropped it due to low construct reliability.

[21] The "program" was referred to as "Education Soup", encompassing both the tutoring program (for treated students) and Discord access (for controls).

### 5.2.4 Other variables

At baseline, we collected demographic and background data on students and parents, including age, gender, education, digital access, and geographic location. We also measured parents' stress using the DASS-21 scale (Lovibond and Lovibond, 1996). For tutors, we collected demographic information and tutoring experience, and measures of stress using the DASS-21 scale and growth mindset and bias toward more resourceful students using the instrument developed by Sabarwal et al. (2022).

Moreover, we measured exposure to conflict using data from The Economist's war-fire model (The Economist and Solstad, 2023). We calculated the total number of war-fire events in each student's administrative region during each experiment period.

## 5.3 Sample Characteristics, Balance Checks, and Survey Attrition

Table 1 presents baseline characteristics of control group students and households by experiment (columns 1, 3, and 5). Across all three experiments, 54% of students in the control group were female, with an average age of 12.6 years and most enrolled in grade 7. Prior to the intervention, the majority attended Ukrainian schools using virtual (ranging from 38% to 48%) or in-person (ranging from 32% to 55%) formats. Fewer than 4% of students had prior exposure to TFU services.

Regarding household characteristics, most parents or guardians who registered their children were women (92%) with an average age of 39. Parental displacement at the time of the experiment was more prevalent in the first and second experiments (39% and 43%, respectively) than in the third (26%). Access to digital devices was nearly universal: 99% of students in the control group reported having an internet-enabled device (e.g., phone or laptop) at home.

Baseline mental health indicators showed high but relatively stable levels of anxiety, with up to 43% of students in the control group scoring in the normal range.[22] However, stress levels were high and varied across experiments: while 56% of control group students in the first and second experiments reported normal stress levels, only 34% did so in the third.

Columns 2, 4, and 6 of Table 1 show differences in baseline characteristics between

[22] As a reference, the Global Burden of Disease Study estimates that 4.4% of 10–14-year-old and 5.5% of 15–19-year-old adolescents globally (including war and non-war settings) experienced an anxiety disorder (above normal levels) in 2021 (Ward and Goldie, 2024).

treatment and control groups, estimated using a regression of each variable on the treatment indicator with stratification-block fixed effects. We observe statistically significant differences in a few variables in the first and third experiments.[23] However, joint orthogonality tests do not reject the null hypothesis that the mean characteristics of the treatment and control groups are statistically indistinguishable within each experiment, confirming that the randomization produced comparable treatment and control groups.

We also collected baseline characteristics for tutors, presented in Table A3. On average, tutors were 41 years old and predominantly female (95%). Half of tutors held bachelor's degrees, and the other half had completed graduate studies (master's or PhD). Tutors had, on average, 17 years of teaching experience, and most tutors (81%) reported normal levels of stress. Participants scored an average of 0.25 on the fixed mindset scale and 0.44 on the bias scale. Both scales as normalized from 0 to 1, where higher values indicate a more fixed mindset and stronger bias, respectively.

# 6 Results

Given the randomized experimental design, we estimate intent-to-treat (ITT) effects separately for each experiment using the following specification for student $i$ in stratum $s$ and experiment $w$:

$$Y_{iswt} = \beta_0 + \beta_1 T_i^w + \beta_2 \mathbf{X}_{isw(t-1)} + \gamma_s + \varepsilon_{iswt} \tag{3}$$

where $Y_{iswt}$ represents the outcome of interest (e.g., test score or indicator for normal stress level) measured after the program (in period $t$), $T_i^w$ is an indicator for whether student $i$'s household was assigned to the treatment group in experiment $w$, and $\mathbf{X}_{isw(t-1)}$ is a vector of baseline covariates selected using LASSO (Bruhn and McKenzie, 2009) to improve precision. $\gamma_s$ denotes stratification fixed effects, and standard errors are clustered at the tutoring group level.

Strata are defined based on the stratification variables used during randomization: parental education and whether the student resided in Ukraine at baseline (first experiment), and parental education crossed with region of residence (second and third experiments). The term $\varepsilon_{i,t}$ represents the idiosyncratic error. Standard errors are clustered

[23] In the first experiment, guardians of treatment group students were 2 percentage points more likely to be female. In the third experiment, treatment group students were 3 percentage points less likely to have experienced displacement and had baseline academic scores 0.13 SD higher than those in the control group.

at the tutoring group level. The coefficient $\beta_1$ estimates the ITT effect of the treatment within each experiment. To correct for multiple hypothesis testing, we report Westfall-Young stepdown adjusted $p$-values (Westfall and Young, 1993).

## 6.1 Main results

The results from this analysis are presented in Figure 2 and 3 as well as Tables A4, A5, A6 for each experiment. Overall we find sizable improvements in academic learning and stress levels, but not in anxiety levels.

**First experiment.** Relative to students in the control group, students in the treatment group scored 0.49 standard deviations (SD) higher in math ($p < 0.001$), 0.40 SD higher in Ukrainian language ($p < 0.001$), and experienced a reduction in stress levels of 0.10 SD ($p = 0.079$).[24] We do not find a statistically significant impact on anxiety at conventional levels ($p = 0.384$).

**Second experiment.** Math scores improved by 0.22 SD ($p = 0.001$) due to the tutoring program, while the effect on Ukrainian language scores was null. Stress levels decreased by 0.11 SD ($p = 0.207$), an effect economically large, but not statistically significant after multiple hypothesis correction. The impact on anxiety is null.

**Third experiment.** Math scores improved by 0.22 SD and Ukrainian scores improved by 0.32 SD (both $p < 0.001$). Stress levels decreased by 0.12 SD ($p = 0.004$). We again observe a null impact on anxiety.

We do not compare estimated effects across experiments because the timing of participation was not randomized, introducing potential selection bias such as more motivated or better-informed students enrolling earlier.

## 6.2 Discussion of the results across experiments.

Our results show that, on average, the online tutoring intervention led to substantial improvements in both mathematics and Ukrainian language achievement. We interpret these gains as evidence of learning recovery in a context where baseline performance was already low. Before February 2022, Ukrainian schools had been closed for 36 weeks. Continued disruptions following the COVID-19 pandemic likely exacerbated learning losses. For example, between 2018 and 2022, Ukraine's average PISA scores declined from 466 to

[24] Throughout, we report adjusted $p$-values in parentheses.

428 in reading and from 453 to 441 in mathematics. These 2022 scores likely represent an upper bound, as testing was conducted in only 18 of the country's 27 regions, implying that actual national performance may be even lower.

Next, we compare our results to recent research on the impact of in-person and online tutoring programs in various contexts. A recent meta-analysis reports an average effect size of 0.28 SD for in-person tutoring programs (Nickow et al., 2023). Two online tutoring programs implemented during COVID-19 reported effect sizes between 0.20 and 0.26 SD in Spain and Italy (Carlana and La Ferrara, 2021; Gortazar et al., 2023). However, comparisons should be made cautiously, as Ukrainian students faced more severe and prolonged disruptions, resulting in different initial conditions and baseline achievement levels.[25]

Relative to these benchmarks, our estimated effects from the first and third experiments are large. In the first experiment, the effect on math (0.49 SD) is 1.7 to 2.4 times larger than the average impact reported in Spain and Italy; for Ukrainian language (0.40 SD), the effects are 1.4 to 2 times larger. In the third experiment, effects are comparable to high-performing online programs: the math impact (0.22 SD) are similar to estimates from Italy and Spain, while the Ukrainian impact (0.32 SD) is 1.2 to 1.7 times larger.

In contrast, effects from the second experiment are more modest. The 0.22 SD gain in math is 75% of the average found in in-person programs, while the Ukrainian language effect is null. These patterns echo prior findings that tutoring tends to yield larger effects on math than on language outcomes, particularly in middle and secondary grades (Nickow et al., 2023). One possible explanation for the smaller effects in the second experiment is the Ministry of Education and Science's announcement canceling final exams for the second consecutive year. This may have reduced student motivation, particularly as the experiment occurred during the last weeks of the academic year. In fact, reduced student engagement is also reflected in lower attendance during the last weeks of the second experiment.[26]

We interpret the magnitude of impacts on stress using a back-of-the-envelope calculation. In the first experiment, a 0.10 SD reduction in stress corresponds to a 1-point drop on the DASS-Y scale. Assuming homogeneous effects, this would shift 137 students (8.8%

[25] For example, Italy's average PISA scores declined from 487 to 471 in mathematics but increased from 476 to 482 in reading between 2018 and 2020. Italy's scores remain much closer to the OECD average than scores in Ukraine.

[26] See more details in section 7.1. Math skills may also be perceived as more transferable or economically valuable in the context of war and potential displacement, further explaining the asymmetry in subject-level results.

of survey completers) from mild to normal stress levels. The equivalent shares for the second and third experiments are 7.6% (104 of 1,368) and 6.7% (165 of 2,456), respectively.

Across all experiments, we observe no meaningful impact on anxiety. One explanation is that the psychosocial content emphasized stress-coping techniques—such as breathing and grounding—rather than interventions targeting future-oriented worry. Evidence suggests such techniques may influence the implicit dimensions of anxiety but not the explicit symptoms captured by the DASS-Y anxiety subscale (Edge et al., 2009).

Pooling data across all three experiments, we estimate the average impact of the tutoring program by including experiment fixed effects in Equation (3). As shown in Table A7, column (2), the program improved math scores by 0.32 SD ($p < 0.001$) and Ukrainian language scores by 0.28 SD ($p < 0.001$). Stress levels declined by 0.12 SD ($p < 0.001$). These results underscore the effectiveness of a short, structured intervention in supporting academic learning and mitigating psychological distress in a conflict-affected setting.

### 6.3 Additional results

Because program take-up was less than 100%, we also estimate treatment-on-the-treated (TOT) effects, defining participation as attending at least one tutoring session. Column (3) of Table A7 (pooled) and Table A8 (by experiment) show that TOT estimates are only slightly larger than ITT effects.

In addition, we explore the relationship between the number of tutoring sessions attended and student outcomes in Figure A5. Two patterns emerge. First, improvements in academic learning—especially in math—are positively associated with attendance, consistent with the cumulative nature of math learning. This pattern holds across all experiments and for Ukrainian language in the first and third experiments.

Second, stress reduction appears to require a minimum dose of exposure. In the first and second experiments, meaningful improvements in stress are observed only among students who attended at least six sessions (Figure A5). This suggests that repeated exposure to coping strategies may be necessary for impact. In contrast, in the third experiment, stress reductions are observed regardless of attendance level, suggesting that the integrated psychosocial tools may have had broader reach and faster effects, even among students with lower participation.

# 7 Feasibility, Mechanisms, and Heterogeneity

In this section, we follow the conceptual framework outlined in Section 3 to examine the underlying components that contribute to the program's effectiveness.

## 7.1 Feasibility

We assess feasibility using measures of take-up, attendance, and engagement. As shown in Figure 1, both take-up and attendance were high. Between 68% and 72% of students assigned to treatment attended at least one math session, and up to 71% attended at least one Ukrainian language session. On average, students participated in more than six of the twelve sessions per subject.[27] Session-level attendance declined only slightly over time (Figure A6). Engagement was also high: according to tutor journals, 54% to 64% of students had their cameras on, 97% to 98% responded to questions, and 95% to 97% appeared attentive. Only 3% to 4% arrived unprepared (Table A9).

To assess whether absences were due to the program's structure or quality, students in the second and third experiments were asked why they missed sessions. The most common reasons were external: lack of electricity or internet (39% in the second experiment, 32% in the third), illness (24% and 37%), and school responsibilities (30% and 31%). These patterns suggest that logistical barriers, not dissatisfaction with the program, were the primary drivers of absenteeism (Figure A7).

## 7.2 Mechanisms

**Structured Peer Interactions.** To examine whether peer interaction contributed to program impacts, we compare treatment and control students' engagement with peers. A potential concern is that any observed effects are driven by the availability of an online platform, not by tutor facilitation. To address this, all students—including those in the control group—were invited to join the program's Discord platform, providing them with access to the same online social space.

As shown in Table 2, students in the treatment group—who were also assigned to tutors—were 22 to 48 pp more likely to interact with peers outside of scheduled tutoring sessions, and 9 to 11 pp more likely to have more than ten peer interactions. In addi-

[27] The median attendance rate was 67%. For comparison, the median attendance in an online tutoring program implemented in Spain during the COVID-19 pandemic was 83%.

tion, treated students were 8 to 13 pp more likely to report that their friends were also enrolled in the program. These results suggest that tutor-led group structures facilitated meaningful peer support that extended beyond the platform's availability.

**Attitudes Toward Learning and Educational Aspirations.** We explore whether the program influenced students' enthusiasm for school subjects and future educational goals. Table 3 shows that the program increased the probability that students reported liking math and Ukrainian language by 7 to 22 pp across experiments. However, effects on aspirations were small and inconsistent: while we observe modest shifts in desired education levels in the first and second experiments, none remain statistically significant after adjusting for multiple hypothesis testing. These results suggest that improvements in academic outcomes were more likely driven by increased short-term motivation and enjoyment of learning rather than changes in long-term aspirations.

**Social-Emotional Skills.** To assess the role of social-emotional development, we examine impacts on persistence and self-efficacy—two core constructs linked to learning and well-being. As shown in Table 4, the tutoring program significantly improved both outcomes. In the second experiment, treatment increased persistence and self-efficacy by 0.10 SD each. In the third experiment, the effects were even larger: 0.33 SD for persistence and 0.31 SD for self-efficacy. These results are consistent with previous studies linking tutoring to non-cognitive gains (Carlana and La Ferrara, 2021) and suggest that the program strengthened students' perseverance and confidence in their abilities, contributing to both academic performance and psychological resilience.

**Complementary Student and Parental Investments.** We also examine whether the tutoring program influenced students' and parents' additional educational investments.

*Student investments:* Across all three experiments, treated students were 21 to 33 pp more likely to seek out additional tutoring or academic support outside the program (Table 5). They were also 14 to 27 pp more likely to report using online learning resources (e.g., the All-Ukrainian Online School platform) for at least one hour per day. However, there were no significant changes in time spent on homework. These results suggest that the program fostered greater intrinsic motivation for learning.

*Parental investments:* To isolate the effect of parental engagement, we implemented a parallel experiment at the same time of the second experiment. A subsample of 743 households (797 students) not treated in the first experiment[28] was randomly assigned to one

[28] 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.

of two groups: both were offered the tutoring program, but only one group's parents received weekly motivational text messages for six weeks. 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 messages were drawn from effective behavioral strategies, including reminders, social norms, and accountability framing (Robinson et al., 2022; Calzolari and Nardotto, 2017; Karlan et al., 2016; Allcott and Rogers, 2014; Gerber et al., 2008). Examples included: "How is [student's name] doing? Remind them to log in for help with school and well-being," "Many students are getting tutoring this week! Encourage [student's name] to join," "Make sure [student's name] attends. The tutor will cover content that will help in 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). There were no significant differences in survey attrition between groups (Table A10).

To estimate the impact of parental engagement through text messages, we modify equation (3) as follows:

$$Y_{iswt} = \beta_0 + \beta_1 \text{Text}_i + \beta_2 X_{isw(t-1)} + \gamma_s + \varepsilon_{iswt} \tag{4}$$

where $\text{Text}_i$ 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. All other variables remain as previously defined. The coefficient $\beta_1$ captures the ITT effect of parental engagement through text messages.

As shown in Table 6, students whose parents received messages performed worse in math (-0.15 SD) and Ukrainian language (-0.23 SD), and showed slightly worse mental health (not statistically significant).[29] These results align with Robinson et al. (2022), who found that increased parental outreach did not improve student outcomes, despite boosting take-up. Informal feedback from TFU and families suggests that the messages may have felt intrusive or burdensome in a high-stress setting and made parents feel pressured and led to dissatisfaction with the program. This highlights the complexities of

[29] Attendance did not differ significantly between groups. Students whose parents received messages attended an average of 7.2–7.3 sessions per subject, comparable to those in the tutoring-only group.

applying behavioral nudges in crisis contexts, where well-intended engagement strategies may backfire.

### 7.3 Heterogeneity analysis

We examine whether program impacts vary by student and household characteristics and by exposure to conflict intensity. To do so, we extend Equation (3) by interacting the treatment indicator with key baseline variables: student gender (=1 if student is female), age (=1 if older than median age, i.e., above 12 years), academic performance (=1 if student's baseline scores in math and Ukrainian language are above-median score), and mental health (=1 if baseline stress and anxiety are in the normal range). For parental wellbeing, we use a binary indicator for normal baseline stress levels. For intensity of conflict exposure, we classified students as having high (above-median) or low (at/below-median) exposure based on the distribution of events during the respective experiment period.

Figure 4 presents heterogeneity by student characteristics and baseline outcomes. Figure 5 shows results by parental well-being and conflict intensity. Both figures report point estimates and 95% confidence intervals for four main outcome indices.

Two caveats are important to note. First, due to power constraints, we pool data across experiments for the heterogeneity analysis. Given the consistency of average treatment effects across waves, this pooling provides meaningful insights for potential scale-up. Second, this heterogeneity analysis was not pre-registered. Nevertheless, given that understanding if effects vary by student or contextual characteristics is relevant for the design of educational intervention we include these exploratory findings in this paper.

We find that program effects on learning are larger among older students and those underperforming at baseline. Specifically, students above age 12 benefited more—gaining 0.40 SD in math and 0.37 SD in Ukrainian language—likely because they required less adult supervision and were more independent learners. Similarly, students with lower baseline performance gained 0.36 SD in math and 0.34 SD in Ukrainian language, consistent with evidence that lower-performing students benefit disproportionately from targeted tutoring (Carlana and La Ferrara, 2021). In contrast, we find no evidence that these student characteristics drive differences in mental health outcomes observed in Section 6.

We also find meaningful differences in program impacts based on parental stress and conflict intensity (Figure 5). Children of parents with high stress levels experienced larger academic gains—0.36 SD in math and 0.33 SD in Ukrainian language—compared to 0.27 SD and 0.21 SD, respectively, for children of parents with low stress levels. This sug-

gests that external educational support may help offset disadvantages associated with caregiver stress. In contrast, children exposed to higher levels of conflict benefited less from the program. Students residing in regions with above-median conflict intensity gained 0.26 SD in math and 0.21 SD in Ukrainian, whereas those in lower-intensity regions achieved 0.37 SD in both subjects. One possible explanation is that high-conflict environments introduce greater distractions or connectivity disruptions, which may limit students' ability to fully engage with the tutoring sessions.

We also explore whether tutor characteristics and baseline perceptions influenced the effectiveness of the tutoring program. This analysis is restricted to the treatment group, as only treated students were assigned to tutors. Specifically, we examine heterogeneity by tutors' teaching experience, number of tutoring groups, stress levels, growth mindset, and bias toward more resourceful students. The results are presented in Figure A8.

We find meaningful patterns. Students assigned to tutors with above-median teaching experience achieved higher math scores, while those assigned to tutors handling a greater number of tutoring groups performed better in Ukrainian language. In addition, tutor beliefs appear to influence academic and mental health outcomes through distinct pathways. Students assigned to tutors with a more fixed mindset performed better in math, but those assigned to tutors with a less fixed mindset experienced larger reductions in stress. Similarly, students assigned to tutors with a stronger bias toward more resourceful students had greater academic gains in both math and Ukrainian language but reported smaller improvements in stress. These results suggest that tutor selection and training—particularly around growth mindset and equitable instructional practices—may be important levers to enhance both academic and psychosocial benefits of online tutoring interventions.

# 8 Robustness Checks

**Exploring potential bias due to differential attrition.** As described in Section 5.1, we collected follow-up data through two separate surveys to reduce respondent fatigue. One survey included the math assessment and additional outcome modules, while the other contained the Ukrainian language assessment. Table A11 reports differences in survey completion rates by treatment status for each round.

We find no statistically significant differences in survey completion between treatment and control groups in the first and third experiments after adjusting for multiple hypoth-

esis testing.[30] However, in the second experiment, treatment group students were 8 pp more likely to complete the math and other modules survey and 10.8 pp more likely to complete the Ukrainian language survey. This differential attrition raises concerns that our estimates could be biased—particularly if students in the treatment group who completed the surveys had systematically better academic outcomes or lower distress levels than those who did not.

To assess the sensitivity of our findings, we follow Fairlie et al. (2015) and conduct a bounds analysis under conservative assumptions about the outcomes of attritors. Columns (1), (4), and (7) of Table A12 replicate the ITT estimates from Figures 2 to 3 and Tables A4 to A6. For each experiment, we impute upper and lower bounds by adding or subtracting 5% of the SD of the observed outcome distribution. Specifically, for attritors in the treatment group, we impute the mean minus 5% of the SD; for those in the control group, we impute the mean plus 5% of the SD—and vice versa for the upper bound.[31]

Across all three experiments, the estimated effects remain robust under these bounds. The conclusions presented in Figures 2 and 3, and in Tables A4 to A6, hold under both upper- and lower-bound assumptions.

**Assessing experimenter demand in self-reported mental health outcomes.** Given the constraints of wartime data collection, the scale of our study, and our limited budget, it was unfeasible to send specialists to assess mental health status of participants. For this reason, we relied on self-administered online surveys to measure students' mental health. While self-reported outcomes can be susceptible to experimenter demand effects, several features of our design reduce this concern.

First, the DASS-Y instrument includes items that assess physiological and behavioral symptoms (e.g., "I was easily annoyed," "My hands felt shaky"), rather than directly asking students to report feeling "stressed" or "anxious." This indirect approach is less prone to demand effects because it does not require explicit self-diagnosis.

Second, while psychosocial activities were introduced in the third experiment, the program was primarily academic in nature. Mental health content was delivered briefly and uniformly, making it unlikely that students would infer a strong experimental focus

[30] There was a small difference in the completion of the Ukrainian language assessment—students in treatment group were 3.5 pp less likely to complete the assessment—in the first experiment. However, this difference is no statiscally significant at the conventional levels after adjusting for multiple hypothesis testing.

[31] This approach provides a conservative test, as 5% of a SD in our sample represents a substantial adjustment. For instance, reducing the mean math score in the second experiment by 5% of its SD lowers the treatment effect by nearly 60%.

on well-being or feel pressure to report improvement.

Third, we triangulate student self-reports with tutor-reported well-being, collected in session journals. Tutors recorded whether students "seemed happy, relaxed, or calm" during the final week of the program. Using pooled data, we find statistically significant negative correlations between students' self-reported stress and anxiety (via DASS-Y) and tutors' external assessments. Specifically, stress is negatively correlated with tutor-reported calmness at $r = -0.26$ ($p < 0.1$), and anxiety at $r = -0.43$ ($p < 0.01$). These correlations suggest that the self-reported mental health measures align with external observations and lend credibility to the main findings.

# 9 Scalability

We assess the scalability of the tutoring program along three dimensions: (i) cost-effectiveness; (ii) external validity, or how representative the sample is relative to the broader Ukrainian student population; and (iii) operational feasibility and geographic reach.

## 9.1 Cost-Effectiveness

To assess cost-effectiveness, we compare the costs of the program to projected long-term economic benefits. First, for our baseline scenario, we follow standard projections from the Ukrainian government and assumptions commonly used in the global literature on benefit-cost analysis. Table 7 summarizes these parameters. Specifically, we assume a 56% labor force participation rate, average annual earnings of $6,124 in 2024, real wage growth of 2.9%, a working life of 43 years, and a 5% annual discount rate.[32]

Second, to estimate the expected increase in future earnings resulting from the tutoring program, we rely on existing literature linking earnings to both cognitive and mental health. For cognitive skills, we draw on Hanushek et al. (2015), who use data from the OECD's Programme for the International Assessment of Adult Competencies (PIAAC) covering 23 countries. They find that a 1 SD increase in numeracy skills is associated with an 18% increase in earnings among prime-age workers, or 10% after controlling for years of schooling. In Eastern European countries—including Poland, the Slovak Republic, and Czechia—returns are estimated at 7.1% to 8.6%. Based on these findings, we assume a

[32] The sources for these assumptions are provided in Table 7. While many education projects use a 3% discount rate, we adopt a higher rate to reflect greater uncertainty about future benefits in the current context.

conservative 8% return on earnings per 1 SD increase in learning in our baseline scenario for Ukraine.

For mental health, we rely on two meta-analyses (Cabus et al., 2021; Vella, 2024) that examine the relationship between emotional stability (e.g., fear, worry, paranoia, and stress) and labor market outcomes. Across 15 and 52 studies, respectively, the estimated earnings returns to a 1 SD improvement in emotional stability range from 1.6% to 1.8%. Accordingly, we assume a 1.7% return on earnings per 1 SD improvement in mental health. 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).

Third, to estimate the costs of the program, we use a costing tool developed by World Bank's Strategic Impact Evaluation Fund.[33] 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 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). Thus, the cost per standard deviation (SD) of learning improvement was $204.19 in the first experiment, $384.66 in the second, and $394.38 in the third (Panel C, Table 7).

In the baseline scenario, comparing program costs to the expected increase in earnings (expressed in 2024 prices) shows that the tutoring program is highly cost-effective. The benefit-to-cost ratio is 56.4 for the first experiment, 31.4 for the second, and 32.9 for the third (Panel D, Table 7). These estimates indicate that the benefits of the program substantially outweigh its costs in all three experiments.

However, given evidence on fade-out effects in tutoring programs, we explore two more conservative scenarios into our cost-effectiveness analysis. Meta-analyses and follow-up studies suggest that achievement gains from tutoring often diminish by 40–60% within a few years (Bailey et al., 2020; Fryer, 2017; Kraft, Matthew A. et al., 2024; Nickow et al., 2023). For example, Finseraas, Henning et al. (2024) find that in Norway, approximately 60% of initial test score gains persisted 3.5 years post-intervention. To reflect this possibility, we introduce two alternative cost-effectiveness scenarios. In the first alternative scenario, we assume 50% of the original effect size endures over time, yielding benefit-to-

[33] This tool is designed to capture detailed (disaggregated) listing and valuing of all resources and efforts required to implement a remote instruction program.

cost ratios of 29.3 (Experiment 1), 16.38 (Experiment 2), and 17.19 (Experiment 3). In the second, more conservative scenario, assuming only 20% of the treatment effect persists, ratios fall to 11.6, 6.5, and 6.9, respectively. Even under this lower-bound assumption, the program remains cost-effective. Moreover, research suggests that tutoring may produce long-run gains in educational attainment and labor market outcomes, even if short-term academic improvements fade (Cortes and Goodman, 2014; Fryer, 2017).

We also assess the sensitivity of the estimated benefit-cost ratios by varying key parameters—namely, wage growth, discount rates, and earnings returns to improvements in learning and mental health—following the approach in Ganimian et al. (2024). Specifically, we allow discount rates to range from 3% to 7%, wage growth rates from 1% to 5%, earnings gains from improved learning from 5% to 11%, and earnings gains from improved mental health from 0.5% to 3%. The sensitivity analysis shows that, even under the most conservative scenarios, the tutoring program yields benefit-to-cost ratios of at least 26 (Figure A9).

Finally, we estimate the marginal value of public funds (MVPF), following Hendren and Sprung-Keyser (2020).[34] Although the program was implemented using donor funds, we simulate a public financing scenario. The MVPF would be 1.8 for the first experiment (112.1/4,380.0), 3.2 for the second, and 3.0 for the third. In other words, government costs would be fully recouped through future tax revenues, provided participants pay a net tax rate of at least 1.8%–3.2% on their additional earnings. These results indicate that the program would yield a strong fiscal return if publicly financed.

## 9.2 External Validity, Representativeness, and Operational Feasibility

To evaluate whether our experimental sample is representative of the broader student population in Ukraine, we compare key baseline characteristics of our sample to data from the 2022 round of the Programme for International Student Assessment (PISA).[35] For comparability, we restrict both samples to 15-year-olds residing in regions where PISA was conducted.

Table A14 shows that students in our sample are, on average, 0.4 years younger and 11 percentage points more likely to be female than their PISA counterparts. However, across other core dimensions relevant to scalability, the samples are strikingly similar.

[34] The MVPF measures the after-tax benefits that participants receive for each dollar of government expenditure.

[35] We focus on four characteristics that are measured consistently across both datasets: students' grade level, age, gender, whether their guardian has completed tertiary education, and whether electronic devices are available in the household.

For example, 68%–69% of guardians had completed tertiary education in both samples, and nearly all students (99%) reported access to internet-enabled devices at home. These comparisons suggest that our sample is broadly representative along the key dimensions that matter for delivery and uptake of online tutoring.

In terms of operational feasibility and geographic reach, the program's implementation across three waves demonstrates its operational resilience in a conflict-affected setting. Recruitment, training, and tutoring were successfully carried out amid infrastructure challenges and repeated disruptions. Most importantly, the program achieved broad national coverage. As shown in Figure A10, students from all regions of Ukraine enrolled in the program, regardless of conflict intensity, demonstrating both nationwide demand and potential feasibility of delivering virtual instruction at scale.

Enrollment patterns closely mirrored national population distributions across oblasts. We compare the regional distribution of students in our sample to the 2022 population estimates across Ukrainian oblasts. Overall, sample representation closely tracks national population distribution, with only a few exceptions.[36]

Lastly, the intervention is designed to complement the existing education infrastructure in Ukraine, further supporting its potential for scalability. The tutoring program relies on a pool of trained teachers—many of whom had spare capacity due to school closures—who were selected and onboarded through a rigorous process implemented by TFU. The model is therefore not only cost-effective but also institutionally compatible with Ukraine's current education workforce.

# 10 Concluding remarks

In this paper, we present new insights on investments in human capital during wartime. Drawing on three randomized experiments, we show that investing in education *during wartime* can have a meaningful impact, is feasible, and scalable. Across all experiments, demand for the online tutoring program is high, and take-up and attendance exceeded expectations. We consistently find that the online tutoring program led to substantial improvements in math scores across all experiments and to improvements in Ukrainian language and reductions in student stress in two of the three experiments. The interventions also enhance peer support, fostered positive learning attitudes and student invest-

[36] Our sample slightly overrepresents students from Kyiv and Dnipropetrovsk oblasts and underrepresents those from Donetsk and Luhansk.

ments, and developed social-emotional skills, thereby contributing to both academic and psychological resilience.

From a policy perspective, the estimated benefit-to-cost ratios are substantial, reinforcing the case that human capital investments should remain a priority—even during conflict. While most research in conflict settings focuses on postwar recovery or relies on observational data, our findings show that it is possible to rigorously assess, adapt, and deliver learning support even amid widespread disruption. Deferring action until after conflict risks widening existing inequalities and slowing recovery. Our results underscore that education should not be viewed solely as a post-conflict priority, but rather as integral to resilience-building and a central pillar of both humanitarian response and long-term development planning.

One key insight is that non-governmental organizations can complement public education by delivering cost-effective, technology-enabled solutions that reach students at scale. As Ukraine rebuilds its education system, such programs can support students at relatively low cost, helping maintain continuity under challenging conditions. This education model also offers lessons for other fragile and conflict-affected settings, where education systems face prolonged disruption. The scalability of the program, its alignment with existing infrastructure, and its measurable impacts on both learning and well-being suggest that similar approaches could be adapted and deployed elsewhere.

Future research should continue partnering with organizations on the ground to test innovative delivery models, evaluate scalability, assess both short- and long-term impacts on learning and well-being, and refine interventions to better meet the needs of children in crisis. Sustaining human capital development amid adversity is not only possible—it is essential for protecting the next generation and ensuring that recovery efforts rest on a strong educational foundation.

# 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 5)

**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 6)

**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 24)

**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 5)

**_ , 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 5)

**_ , _ , 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 5)

**_ , 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 5)

**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 5)

**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 54)

**Bailey, Drew H., Greg J. Duncan, Flávio Cunha, Barbara R. Foorman, and David S. Yeager**, "Persistence and Fade-Out of Educational-Intervention Effects: Mechanisms and Potential Solutions," *Psychological Science in the Public Interest*, October 2020, *21* (2), 55–97. (Cited on 29)

**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 6)

_ , _ , _ , _ , _ , **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 6)

_ , _ , _ , _ , _ , **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 6)

**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 6)

_ , **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 6)

**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 6)

**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 6)

**Bertoni, Eleonora, Michele Di Maio, Vasco Molini, and Roberto Nistico**, "Education is forbidden: The effect of the Boko Haram conflict on education in North-East Nigeria," *Journal of Development Economics*, 2019, *141*, 102249. (Cited on 12)

**Betancourt, Theresa S., Sarah E. Meyers-Ohki, Alexandra P. Charrow, and Wietse A. Tol**, "Interventions for Children Affected by War: An Ecological Perspective on Psychosocial Support and Mental Health Care," *Harvard Review of Psychiatry*, 2013, *21* (2), 70–91. (Cited on 12)

**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 5)

**Blattman, Christopher and Edward Miguel**, "Civil war," *Journal of Economic literature*, 2010, *48* (1), 3–57. (Cited on 2, 12)

**Brück, Tilman, Michele Di Maio, and Sami H Miaari**, "Learning the hard way: The effect of violent conflict on student academic achievement," *Journal of the European Economic Association*, 2019, *17* (5), 1502–1537. (Cited on 12)

**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 18)

**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 29)

**Calzolari, Giacomo and Mattia Nardotto**, "Effective reminders," *Management Science*, 2017, *63* (9), 2915–2932. (Cited on 24)

**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 5, 11, 16, 20, 23, 25)

**Carrell, Scott E., Mark Hoekstra, and Elira Kuka**, "The Long-Run Effects of Disruptive Peers," *American Economic Review*, November 2018, *108* (11), 3377–3415. (Cited on 11)

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

**Cortes, Kalena E. and Joshua S. Goodman**, "Ability-Tracking, Instructional Time, and Better Pedagogy: The Effect of Double-Dose Algebra on Student Achievement," *American Economic Review*, May 2014, *104* (5), 400–405. (Cited on 30)

**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 11)

**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, 50, 58)

**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 5)

**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 21)

**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 15, 27, 85)

**Finseraas, Henning, Nyhus, Ole Henning, Salvanes, Kari Vea, and Sandsør, Astrid Marie Jorde**, "Sustained Effects of Small-Group Instruction in Mathematics," Technical Report 24-931, EdWorking Paper 2024. (Cited on 29)

**Fryer, R.G.**, "The Production of Human Capital in Developed Countries," in "Handbook of Economic Field Experiments," Elsevier, 2017, pp. 95–322. (Cited on 29, 30)

**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 30, 53, 71)

**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 24)

**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 11)

**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 5, 20)

**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 9)

**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 6)

**Hanushek, Eric A, 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 28)

**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 5)

**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 30)

**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 5)

**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 53)

**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 5)

**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 6)

_ **and Tamara Wilder**, "Educational expectations and attainment," Technical Report, National Bureau of Economic Research 2010. (Cited on 11)

**Justino, Patricia**, "Violent conflict and human capital accumulation," *IDS Working Papers*, 2011, *2011* (379), 1–17. (Cited on 11, 12)

**_ , Marinella Leone, and Paola Salardi**, "Short-and long-term impact of violence on education: The case of Timor Leste," *The World Bank Economic Review*, 2014, *28* (2), 320–353. (Cited on 12)

**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 24)

**Khattab, Nabil**, "Students' aspirations, expectations and school achievement: What really matters?," *British educational research journal*, 2015, *41* (5), 731–748. (Cited on 11)

**Kraft, Matthew A., Schueler, Beth E., and Falken, Grace**, "What Impacts Should We Expect from Tutoring at Scale? Exploring Meta-Analytic Generalizability," Technical Report 24-1031, EdWorking Paper 2024. (Cited on 29)

**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 5)

**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 6)

**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 6)

**_ , 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 6)

**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 5)

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

**Lovibond, Sydney H and Peter F Lovibond**, *Manual for the depression anxiety stress scales*, Psychology Foundation of Australia, 1996. (Cited on 17, 46)

**Mobarak, Ahmed Mushfiq**, "Assessing social aid: the scale-up process needs evidence, too," *Nature*, 2022, *609* (7929), 892–894. (Cited on 6)

**Muralidharan, Karthik and Paul Niehaus**, "Experimentation at Scale," *Journal of Economic Perspectives*, November 2017, *31* (4), 103–24. (Cited on 6)

**National Bank of Ukraine**, "Inflation Report: July 2024," Technical Report, National Bank of Ukraine 2024. (Cited on 53)

**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 2, 4, 10, 11, 20, 29)

**OECD**, *PISA 2022 Results (Volume I): The State of Learning and Equity in Education*, Paris: OECD Publishing, 2023. (Cited on 7)

**Patrinos, Harry Anthony, Emiliana Vegas, and Rohan Carter-Rau**, "An analysis of COVID-19 student learning loss," Technical Report 2022. (Cited on 5)

**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 24)

**Rodriguez, Catherine and Fabio Sanchez**, "Armed conflict exposure, human capital investments, and child labor: evidence from Colombia," *Defence and peace economics*, 2012, *23* (2), 161–184. (Cited on 12)

**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 11)

**Sabarwal, Shwetlena, Malek Abu-Jawdeh, and Radhika Kapoor**, "Teacher beliefs: Why they matter and what they are," *The World Bank Research Observer*, 2022, *37* (1), 73–106. (Cited on 17)

**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, 50, 59)

**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 5)

**Shonkoff, Jack P., Andrew S. Garner, Committee on Psychosocial Aspects of Child and Family Health, and Committee on Early Childhood, Adoption, and Dependent Care and Section on Developmental and Behavioral Pediatrics**, "The lifelong effects of early childhood adversity and toxic stress," *Pediatrics*, 2012, *129* (1). (Cited on 12)

**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 11)

**Swee, Eik Leong**, "On war intensity and schooling attainment: The case of Bosnia and Herzegovina," *European Journal of Political Economy*, 2015, *40*, 158–172. (Cited on 12)

**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, 42, 44, 52)

**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 17, 46, 73)

**Valente, Christine**, "Education and civil conflict in Nepal," *The World Bank Economic Review*, 2014, *28* (2), 354–383. (Cited on 12)

**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 29)

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

**Ward, Zachary J and Sue J Goldie**, "Global Burden of Disease Study 2021 estimates: implications for health policy and research," *The Lancet*, 2024, *403* (10440), 1958–1959. (Cited on 17)

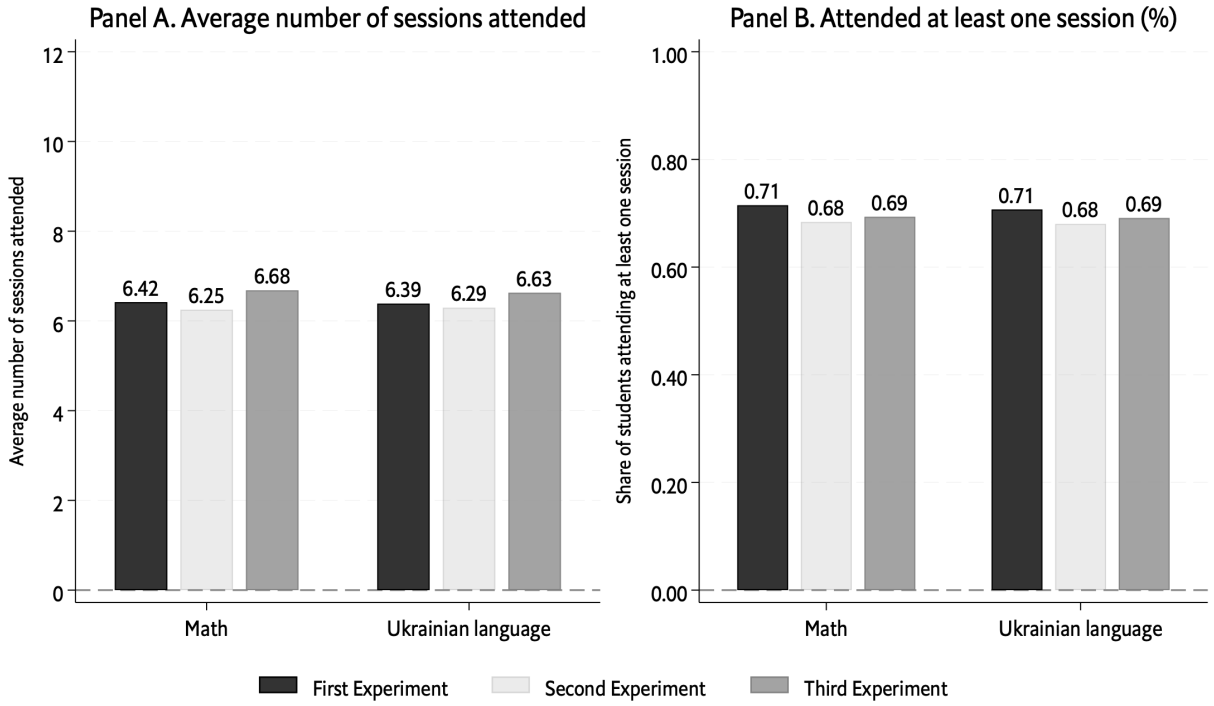
**Westfall, Peter H and S Stanley Young**, *Resampling-based multiple testing: Examples and methods for p-value adjustment*, Vol. 279, John Wiley & Sons, 1993. (Cited on 19, 48, 49, 50, 52, 77, 78, 79, 80, 81, 83, 84)

**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 7)

**_ , 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 7)

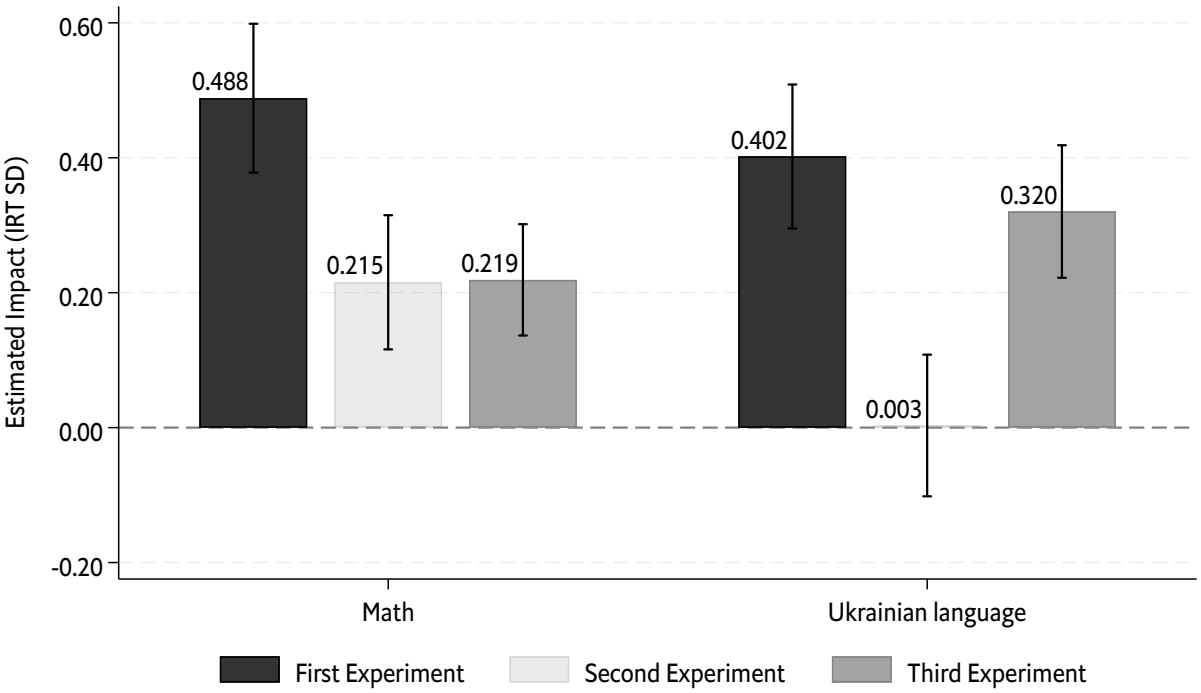
# Tables and Figures

**Figure 1:** Attendance to Tutoring Sessions



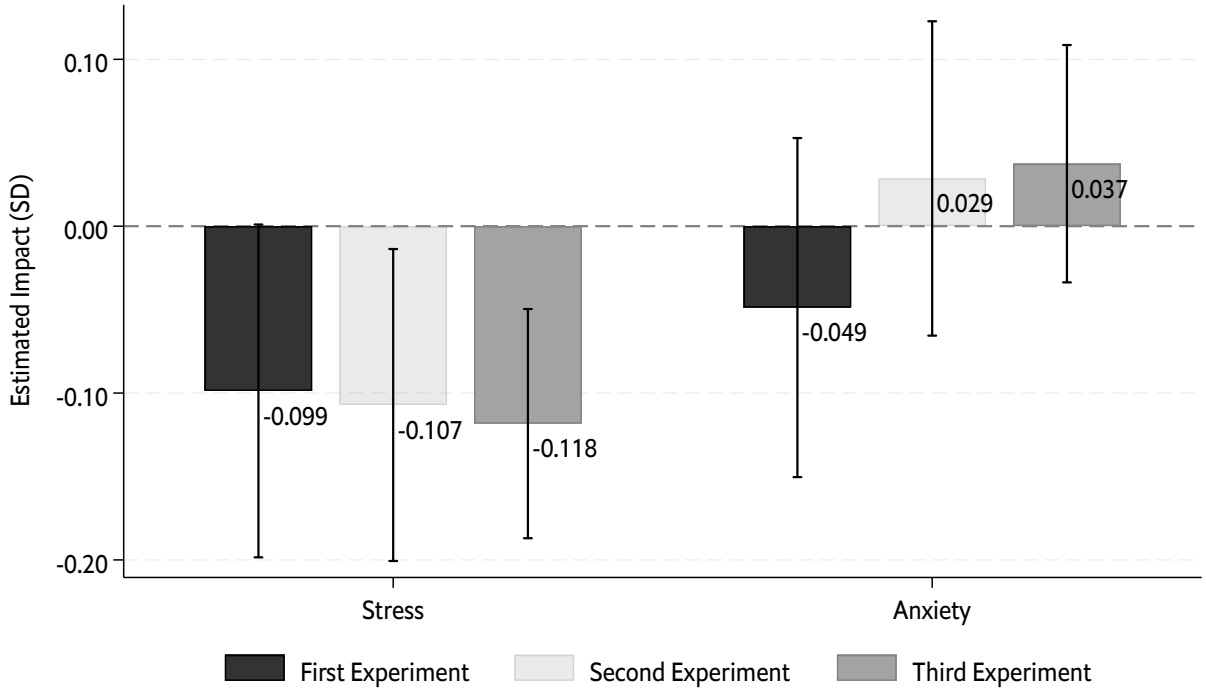

*Notes*: This figure shows the average attendance to the tutoring program (panel A) and take up of the intervention (panel B), by experiment. The average attendance is estimated as the average number of sessions of attended, separated by subject (math or Ukrainian language). As a reference, the total 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 $\beta_1$ from equation (3) 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 $\beta_1$ from equation (3) 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

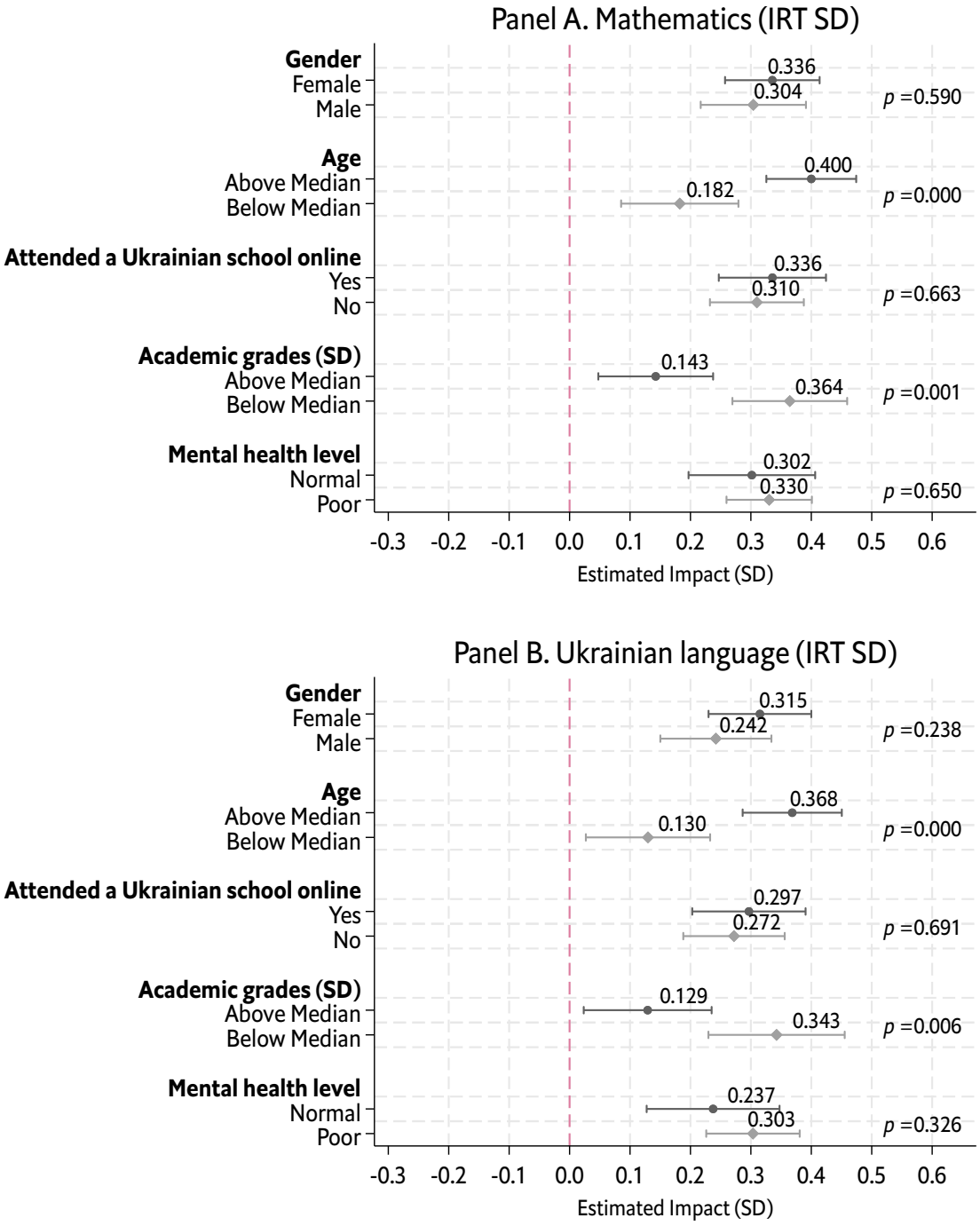
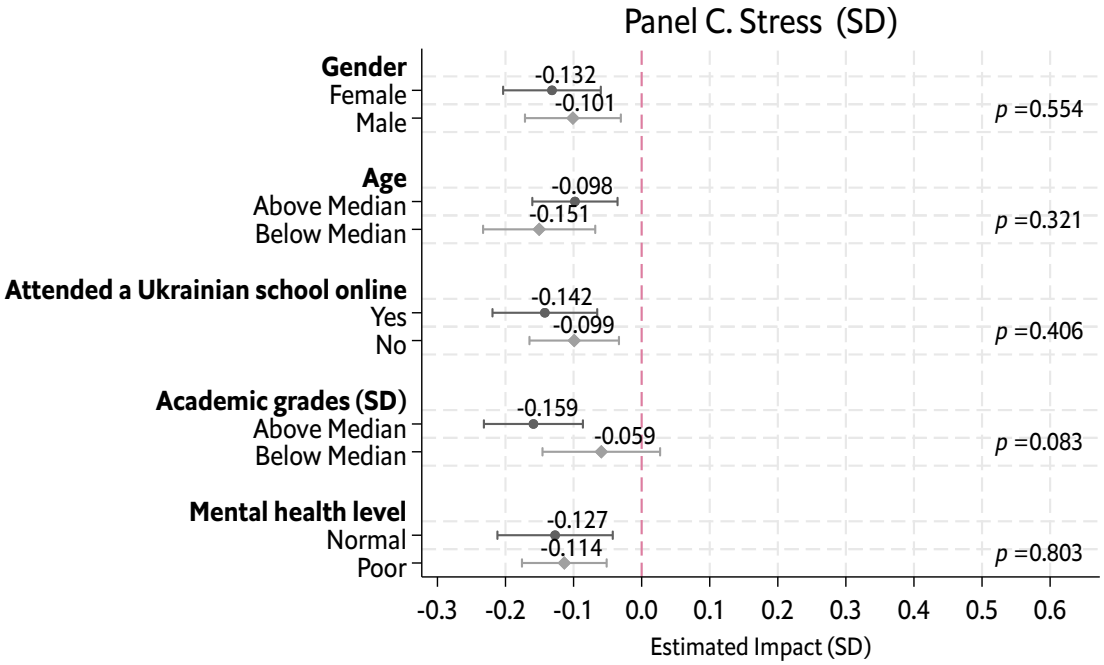
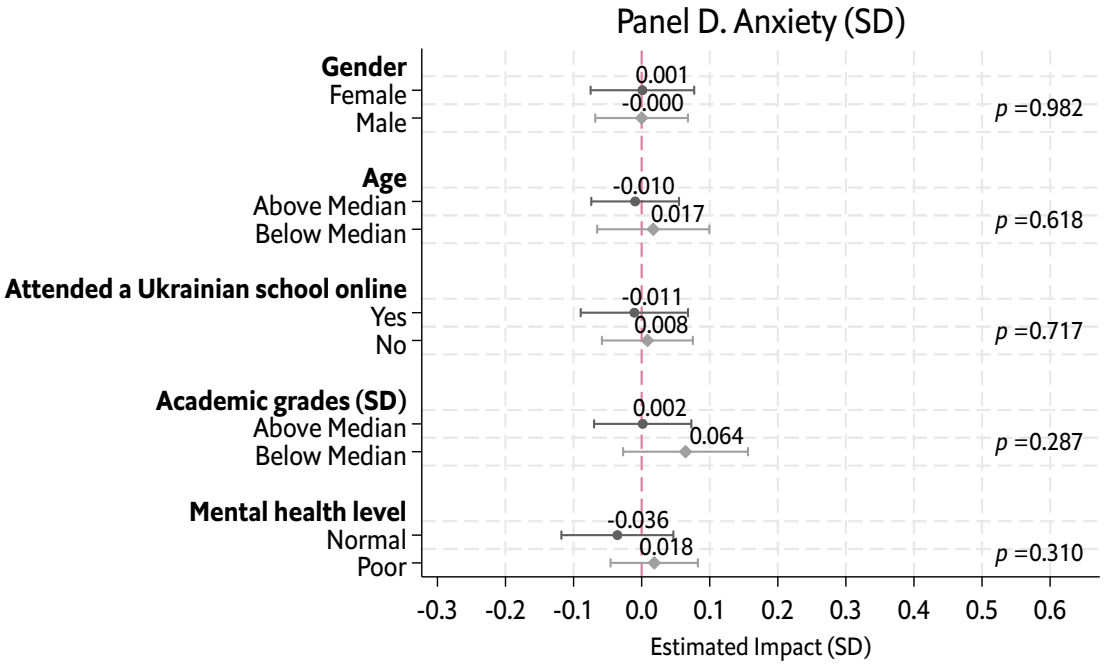


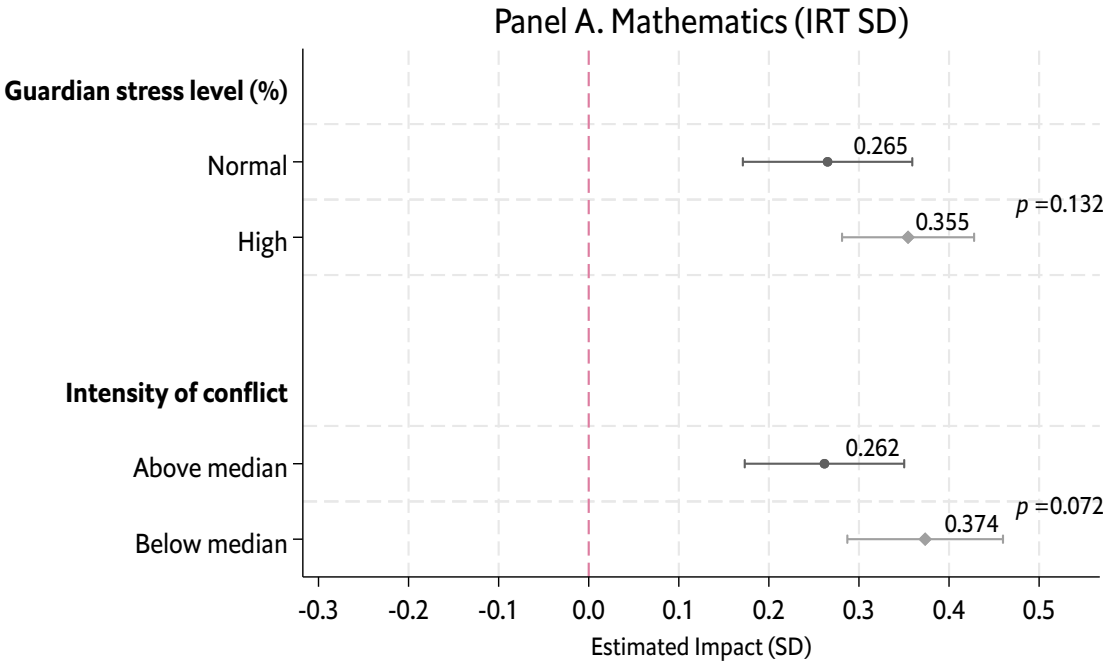
**Figure 4**: *Continued*–Heterogeneity of results by Student Characteristics




*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 (3), 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



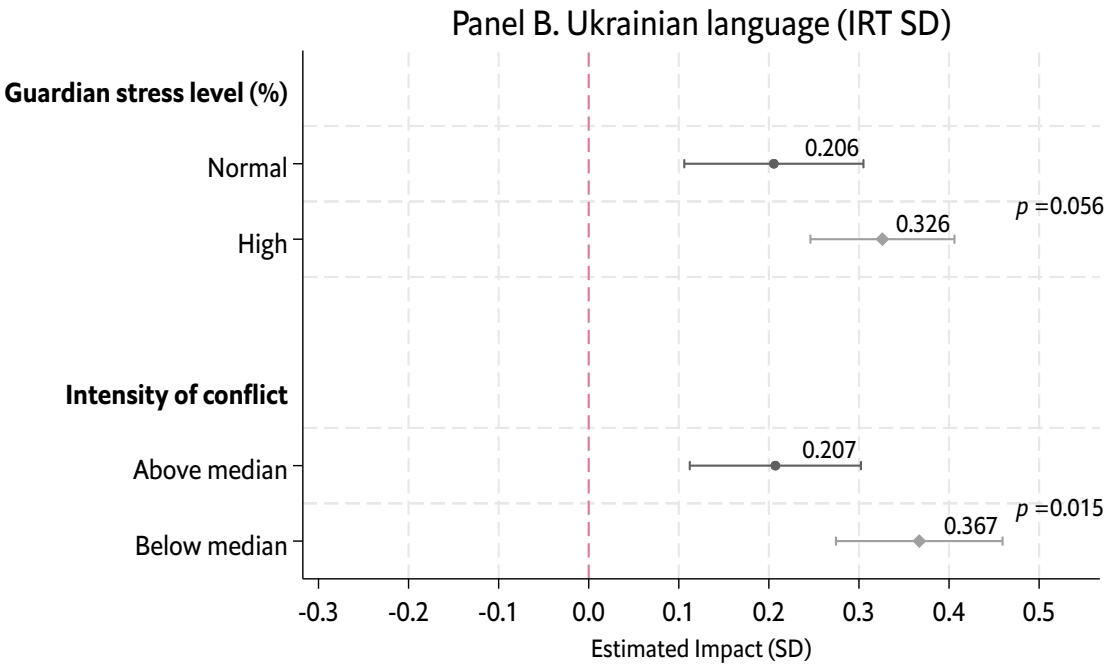
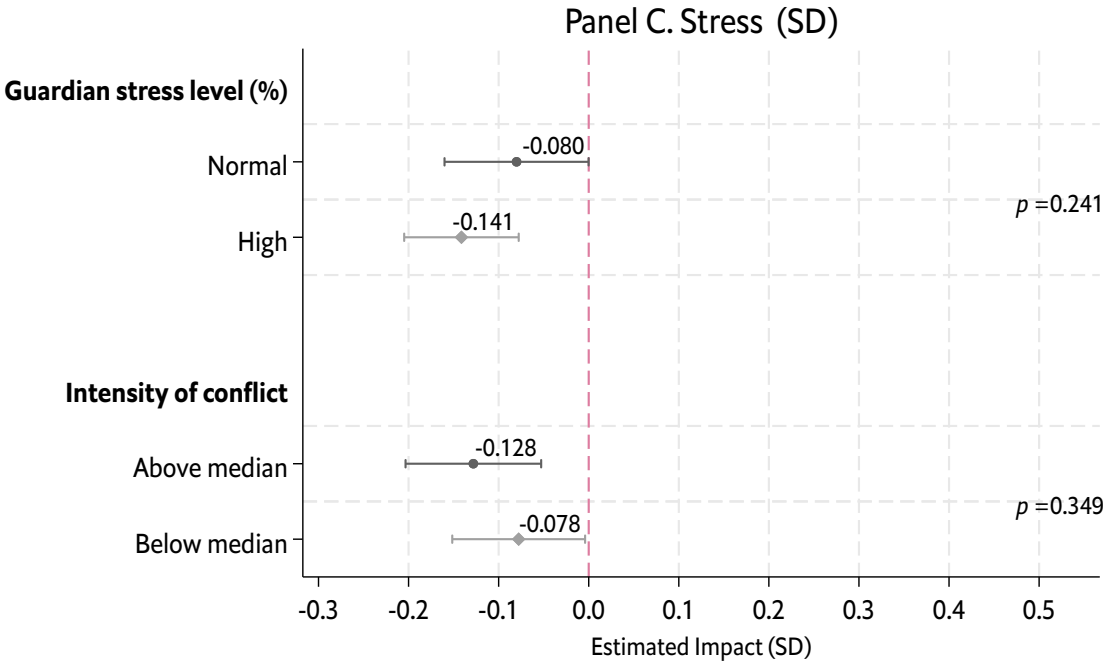
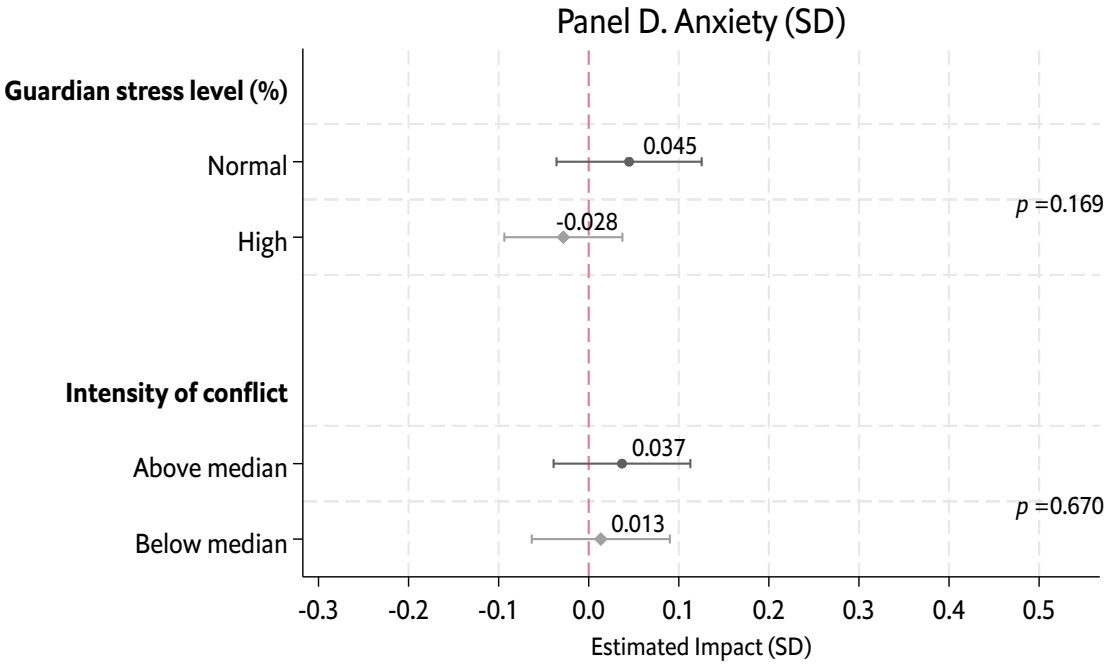



**Figure 5**: *Continued*–Heterogeneity of Results by External Factors




*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 (3), 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 $\gamma_s$ and $\varepsilon_{iswt}$ 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 $\beta_1$ from equation (3) 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 $\beta_1$ from equation (3) 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 $\beta_1$ from equation (3) 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

| | | Average daily: time on (Hours≥1): | |
|---|---|---|---|
| | Participated in additional tutoring | Online Resources | Homework |
| | (1) | (2) | (3) |
| **Panel A. First Experiment** | | | |
| Treatment | 0.211*** | 0.136*** | 0.020 |
| | (0.025) | (0.023) | (0.025) |
| | [0.000] | [0.000] | [0.410] |
| Control group outcome mean | 0.302 | 0.214 | 0.578 |
| # of control variables selected | 0 | 0 | 0 |
| Obs. | 1,562 | 1,563 | 1,562 |
| **Panel B. Second Experiment** | | | |
| Treatment | 0.306*** | 0.266*** | -0.022 |
| | (0.024) | (0.025) | (0.027) |
| | [0.000] | [0.000] | [0.371] |
| Control group outcome mean | 0.229 | 0.225 | 0.559 |
| # of control variables selected | 0 | 0 | 1 |
| Obs. | 1,368 | 1,368 | 1,368 |
| **Panel C. Third Experiment** | | | |
| Treatment | 0.330*** | 0.238*** | -0.028 |
| | (0.019) | (0.018) | (0.020) |
| | [0.000] | [0.000] | [0.191] |
| Control group outcome mean | 0.287 | 0.225 | 0.595 |
| # of control variables selected | 0 | 1 | 2 |
| Obs. | 2,456 | 2,456 | 2,456 |

*Notes*: This table shows estimates of $\beta_1$ from equation (3) on measures of: (column 1) participation in additional tutoring or subject-specific support during the past weeks; (column 2) whether the average daily time spent on online resources and asynchronous videos on the All-Ukrainian Online School platform exceeds one hour; (column 3) whether the average daily time spent on homework during the period from January to March 2024 exceeds one hour. The outcome was measured using data from the self-reported endline survey. All estimations 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. 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 $\beta_1$ from equation (4) 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 National Bank of Ukraine (2024).

# 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, war-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. Pro-

tocols 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, 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**

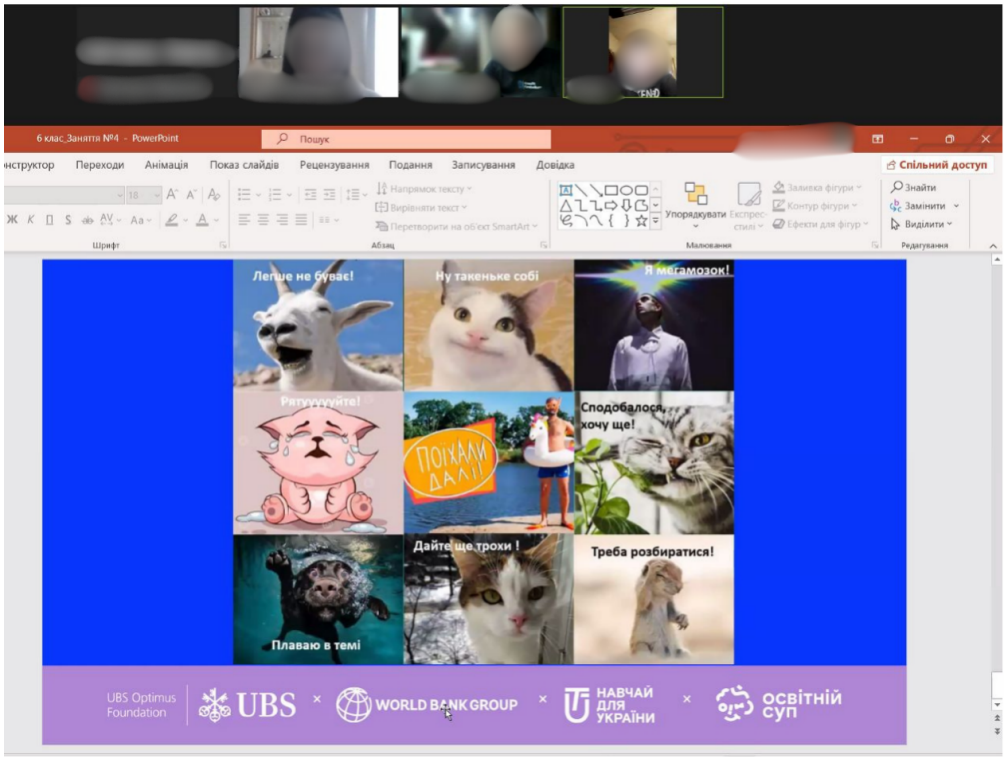
None.

# 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?

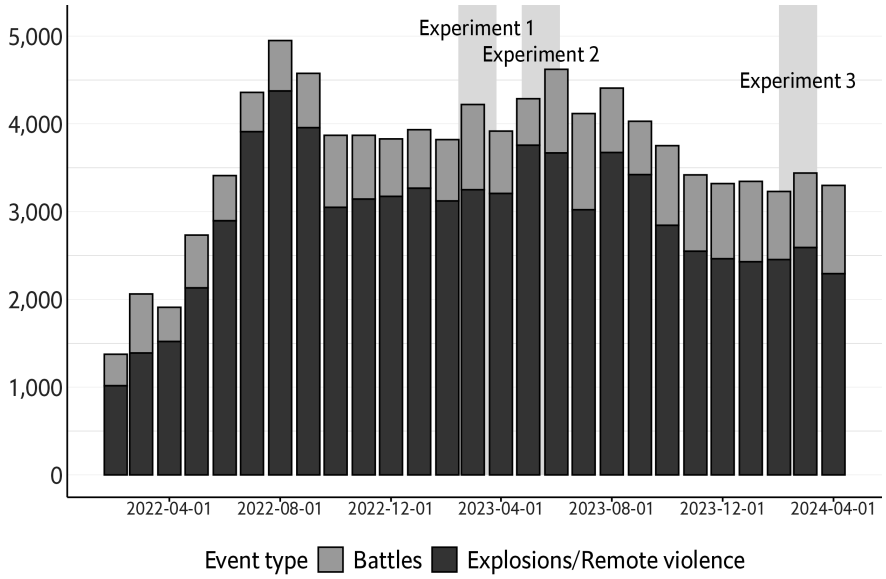
3) How many interactions did you have with them during the past 6 weeks?

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

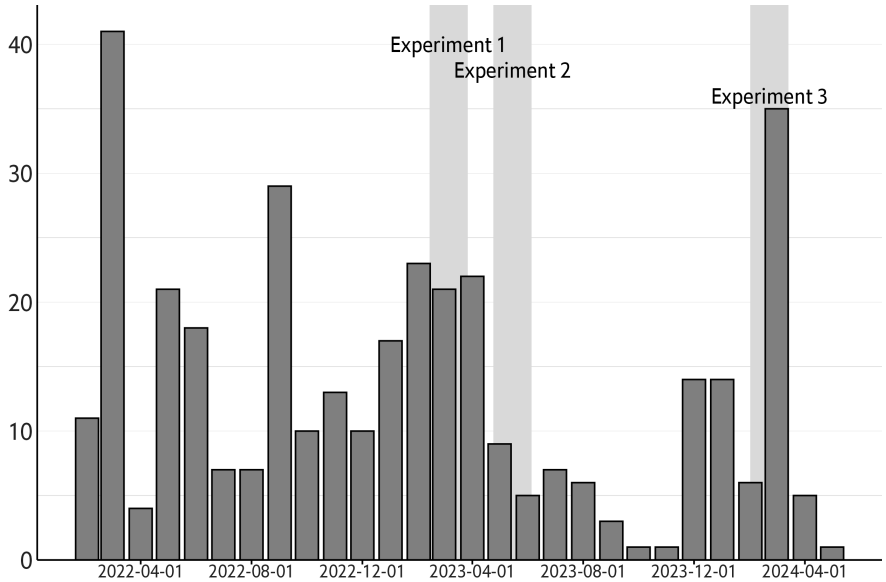
# Appendix Figures

**Figure A1:** Attacks and Violent Events in Ukraine

**(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:** 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.**

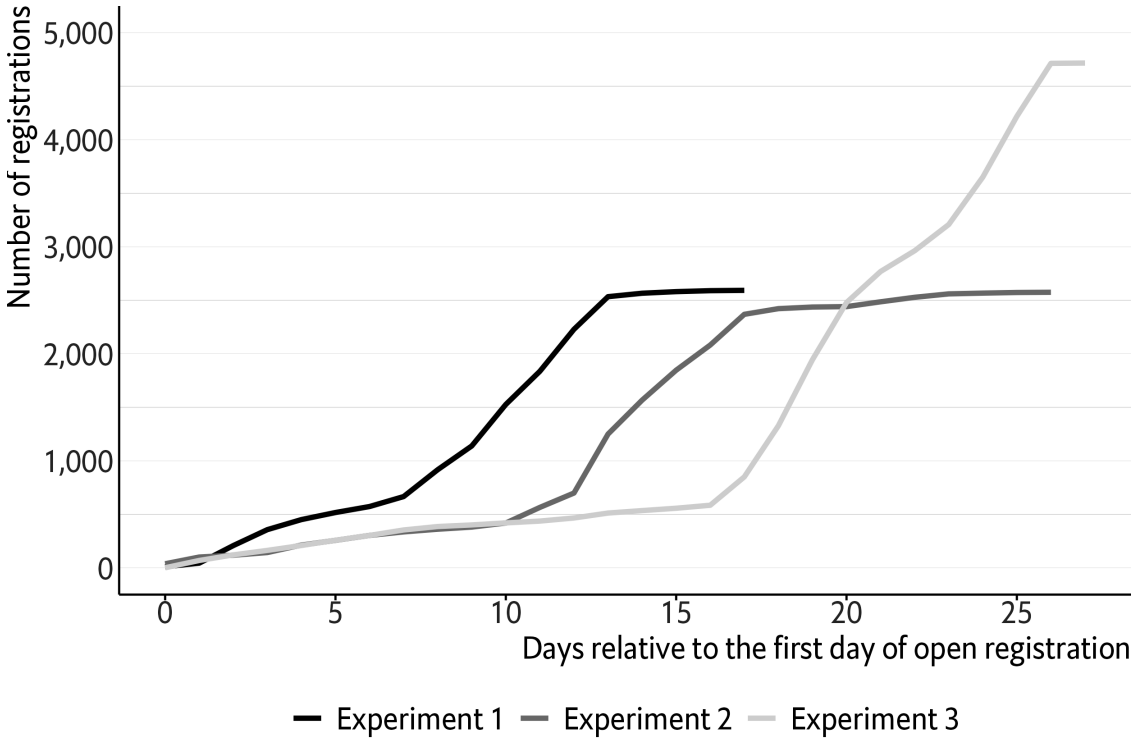
У таблиці нижче наведено середній бал групи **5_106** і всіх учнів цього класу за темами та очікуваними навчальними результатами, що, сподіваємось, допоможе Вам у індивідуалізації роботи з цими учасниками і учасницями.

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 |

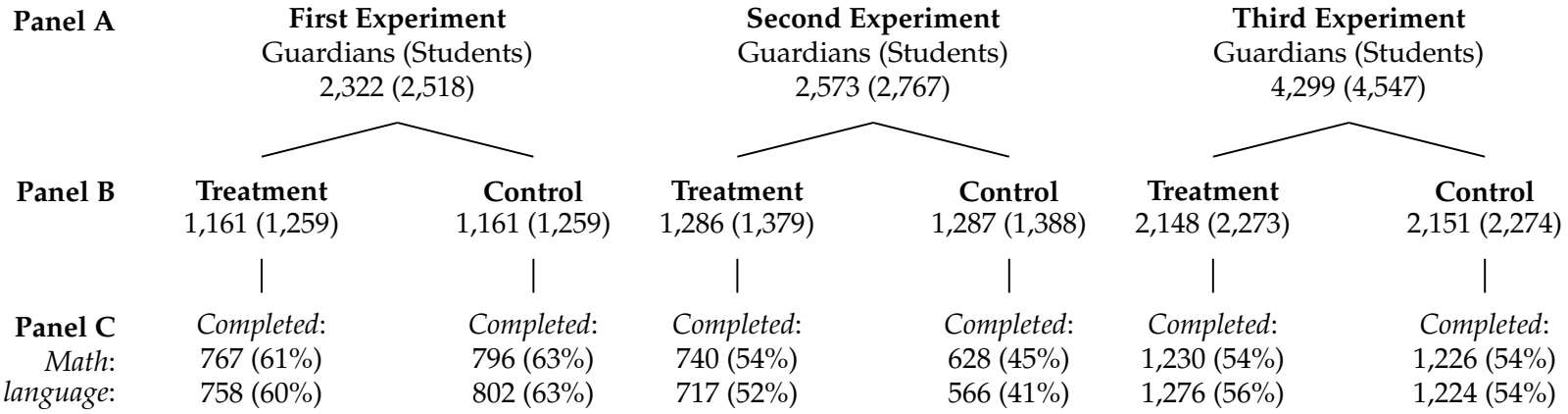
*Notes:* This figure shows examples of diagnostic reports sent to the tutors of the group 5-106 using baseline math scores. Reports for Ukrainian language had the same structure and format.

**Figure A3:** 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 A4:** Sample Sizes, Treatment Assignments, and Endline Survey Completion Rates




*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 A5:** Relationship between Tutoring Attendance and Endline Test Scores in Math and Ukrainian language

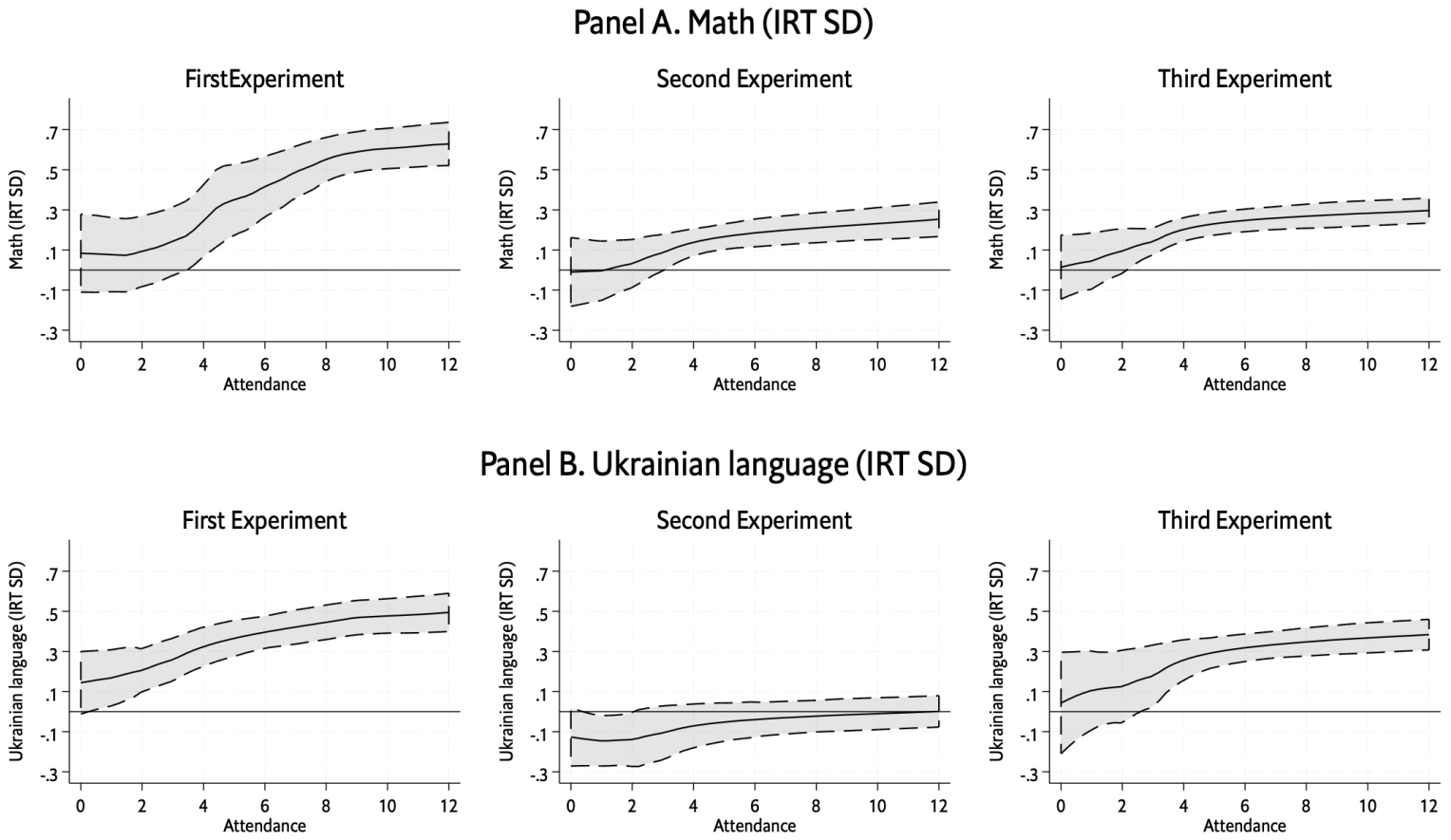
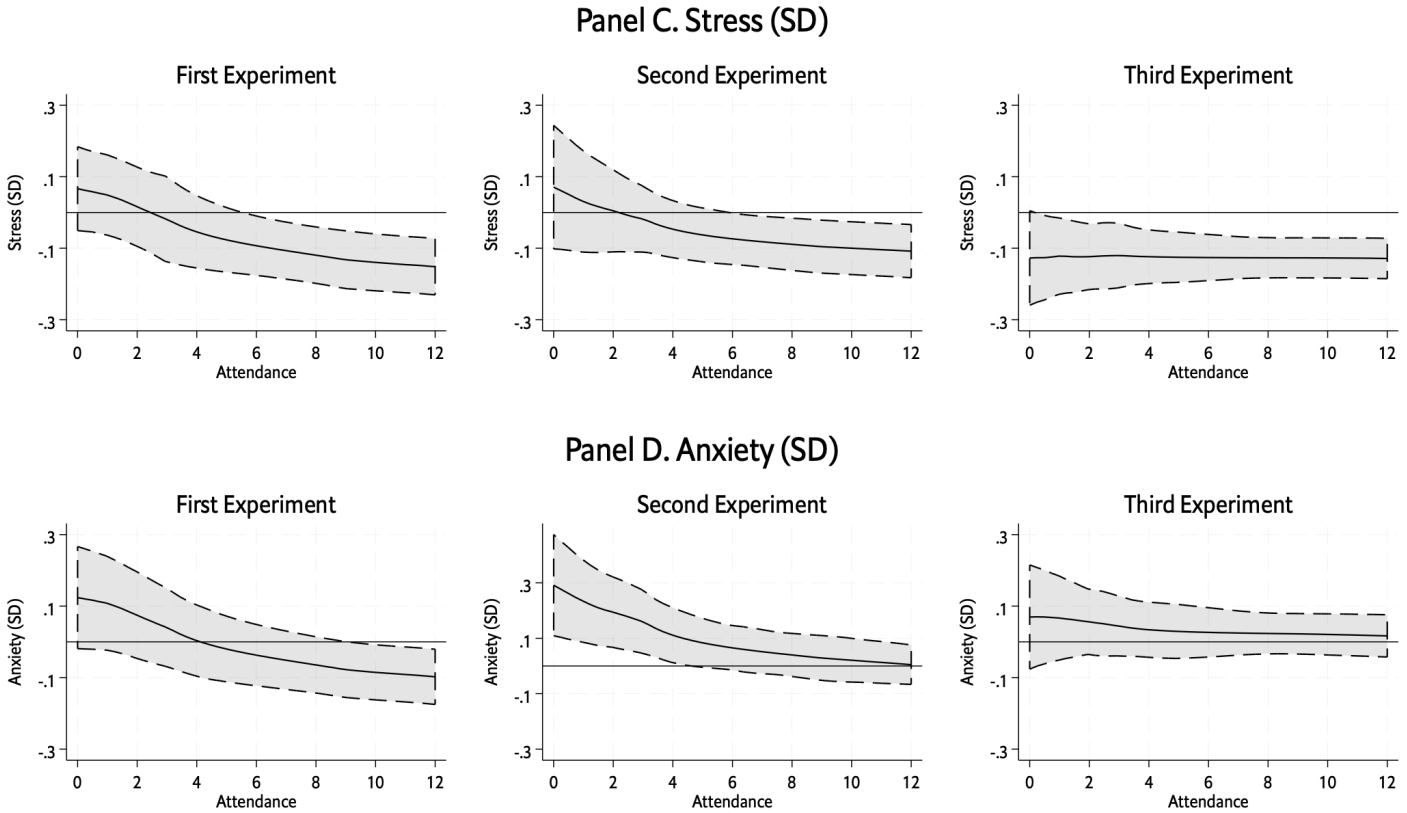
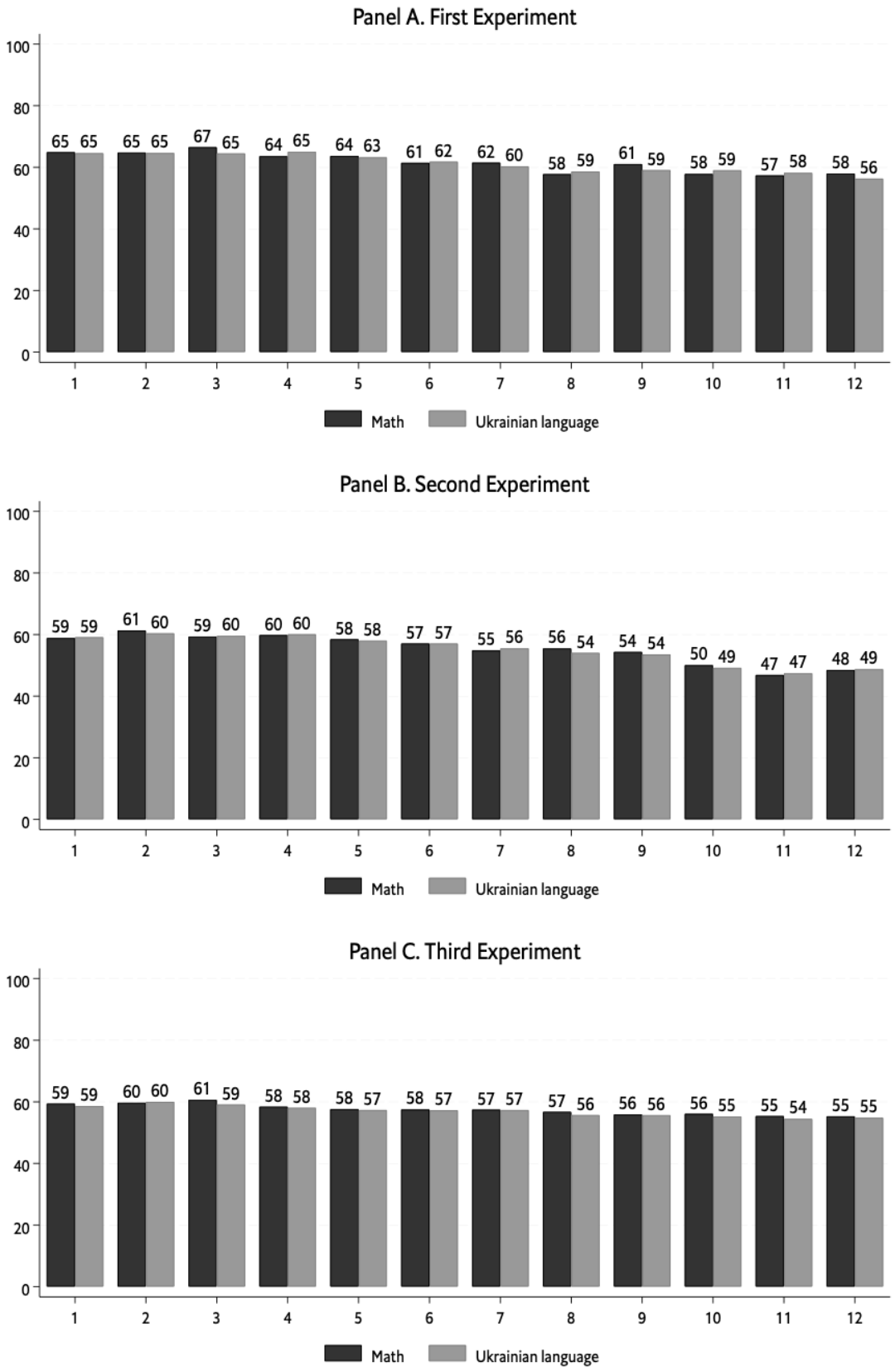


**Figure A5**: *Continued*–Relationship between Tutoring Attendance and Stress and Anxiety



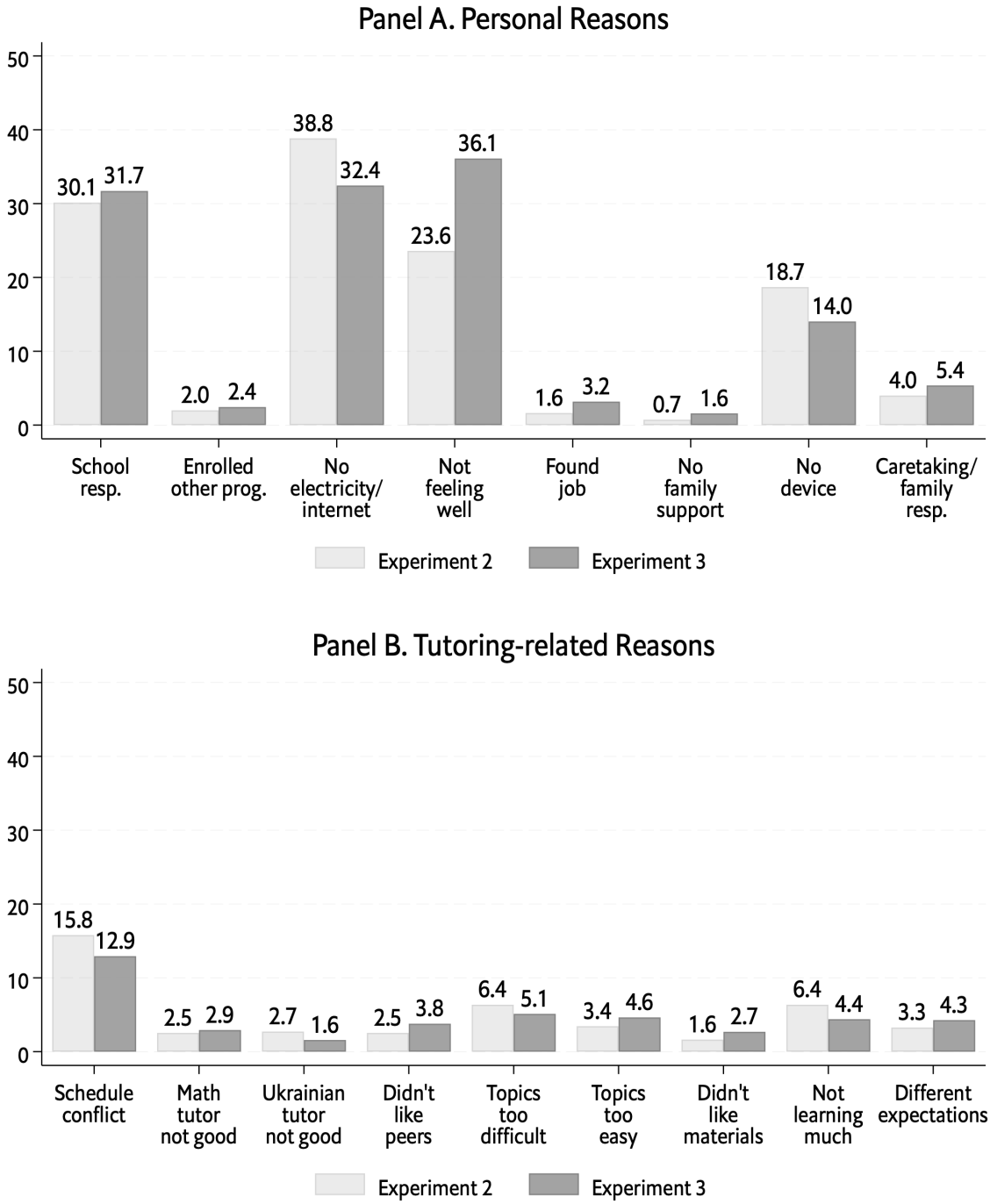

*Notes*: These figures present kernel-weighted local polynomial regressions showing the estimated conditional mean of standardized endline outcomes as a function of the number of sessions attended, separated by experiment. Panel A shows the association between attendance at math tutoring sessions and 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 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:** Heterogeneity by tutor characteristics

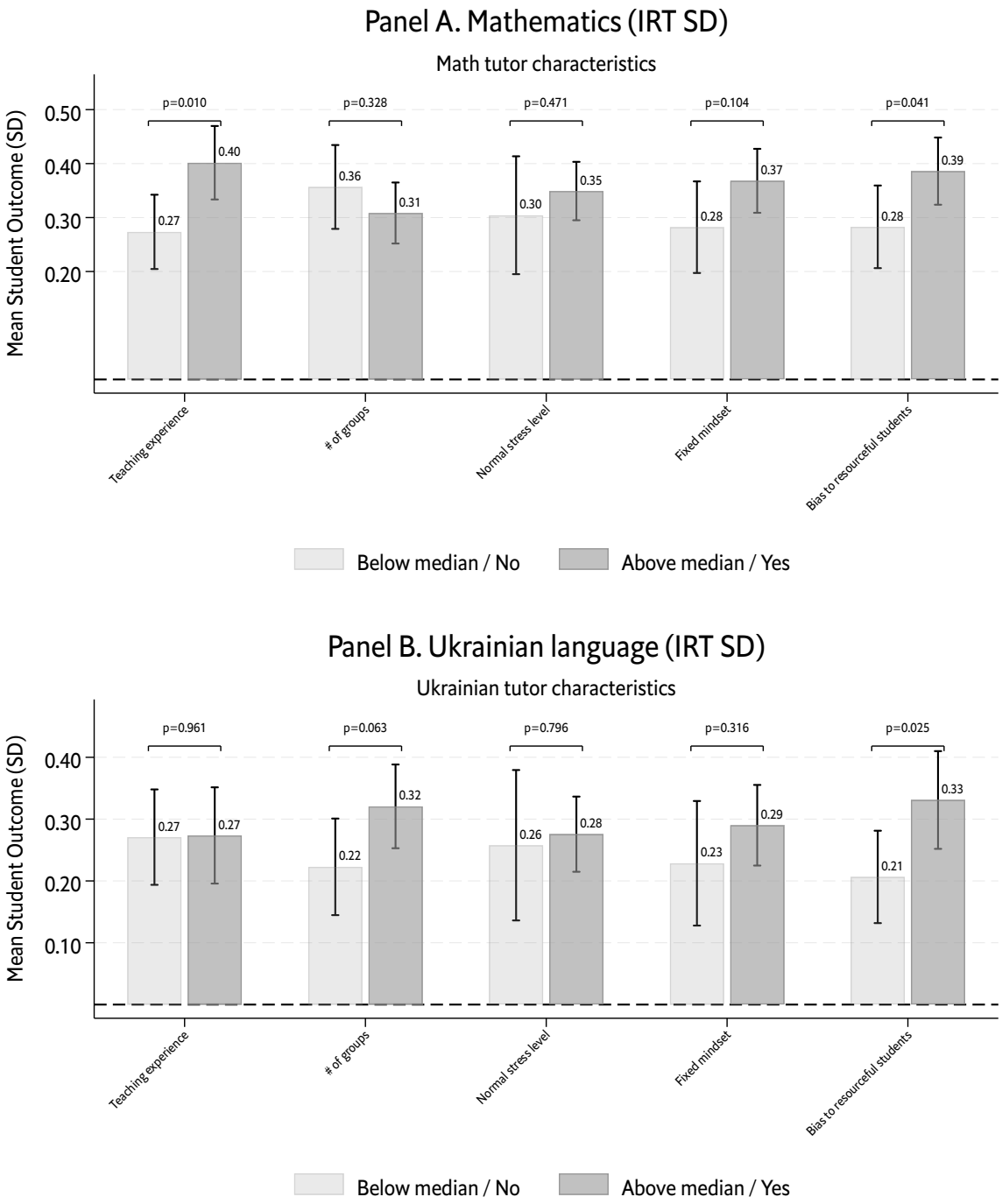
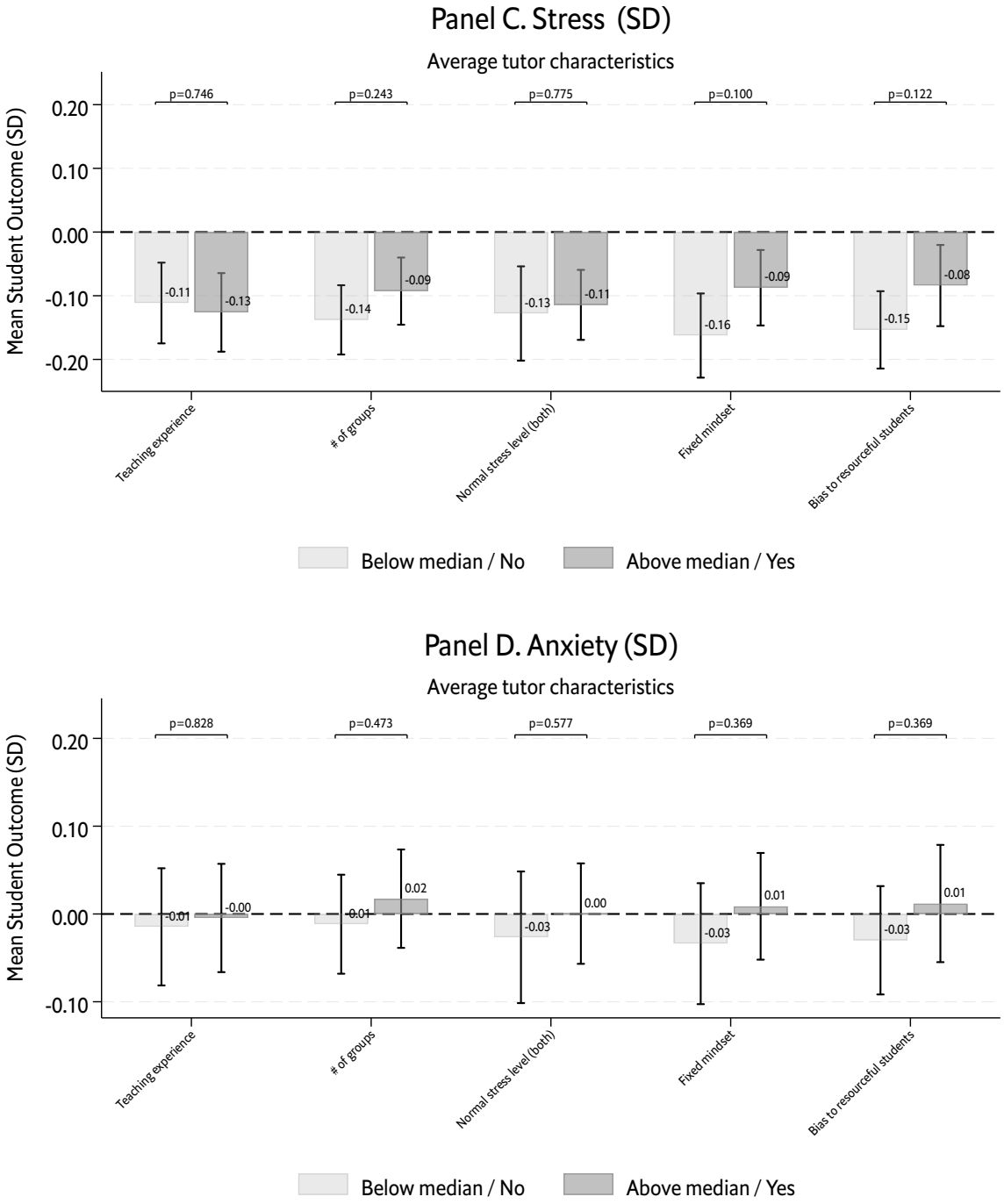
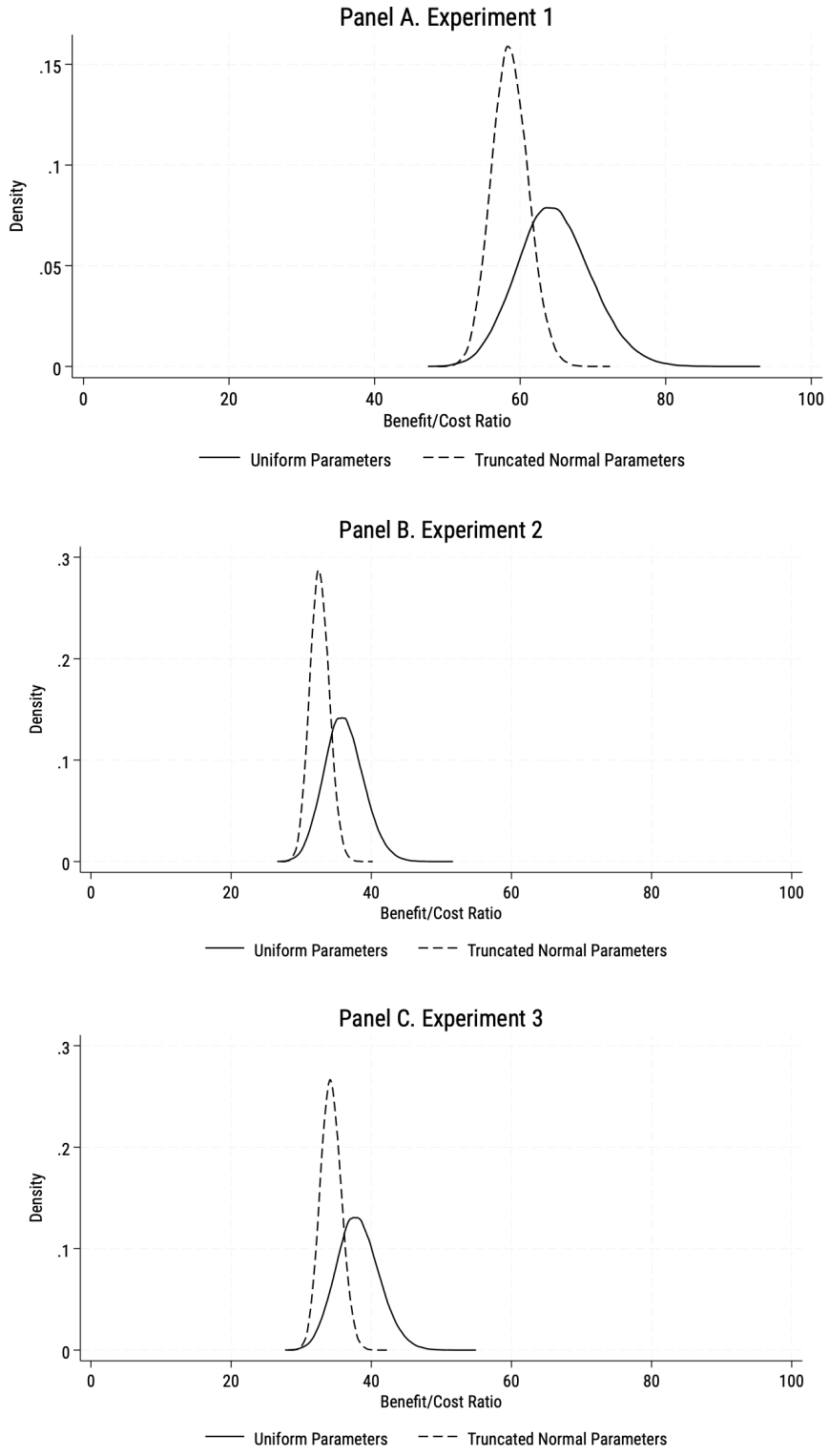


**Figure A8**: *Continued*–Heterogeneity by tutor characteristics




*Notes*: The bars in this figure show the average standardized outcome for treated students, disaggregated by tutor characteristics. Panel A presents math scores, and Panel B shows Ukrainian language scores. Panels C and D display stress and anxiety scores, respectively. Subgroups are defined based on whether the tutor's measure value on a given characteristic is above or below the sample median. These characteristics include years of teaching experience, number of tutoring groups assigned, fixed mindset score, and bias toward resourceful students score. For the "normal stress level" characteristic, the comparison is between tutors who reported normal levels of stress and those who did not. For academic outcomes (Panels A and B), subgroup assignments are based on the subject-specific tutor (math or Ukrainian language). For mental health outcomes (Panels C and D), subgroup assignments are based on the average characteristic across the two tutors assigned to each student. In the case of "normal stress level," the student is considered to have been assigned to "normal-stress" tutors only if both tutors reported normal levels of stress; otherwise, the student is classified as having been assigned to "high-stress" tutors. Error bars represent 95% confidence intervals, and *p*-values shown above each bar indicate the statistical significance of the differences between subgroups.

**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).

**Figure A10:** Intensity of Conflict Across Different Experiments

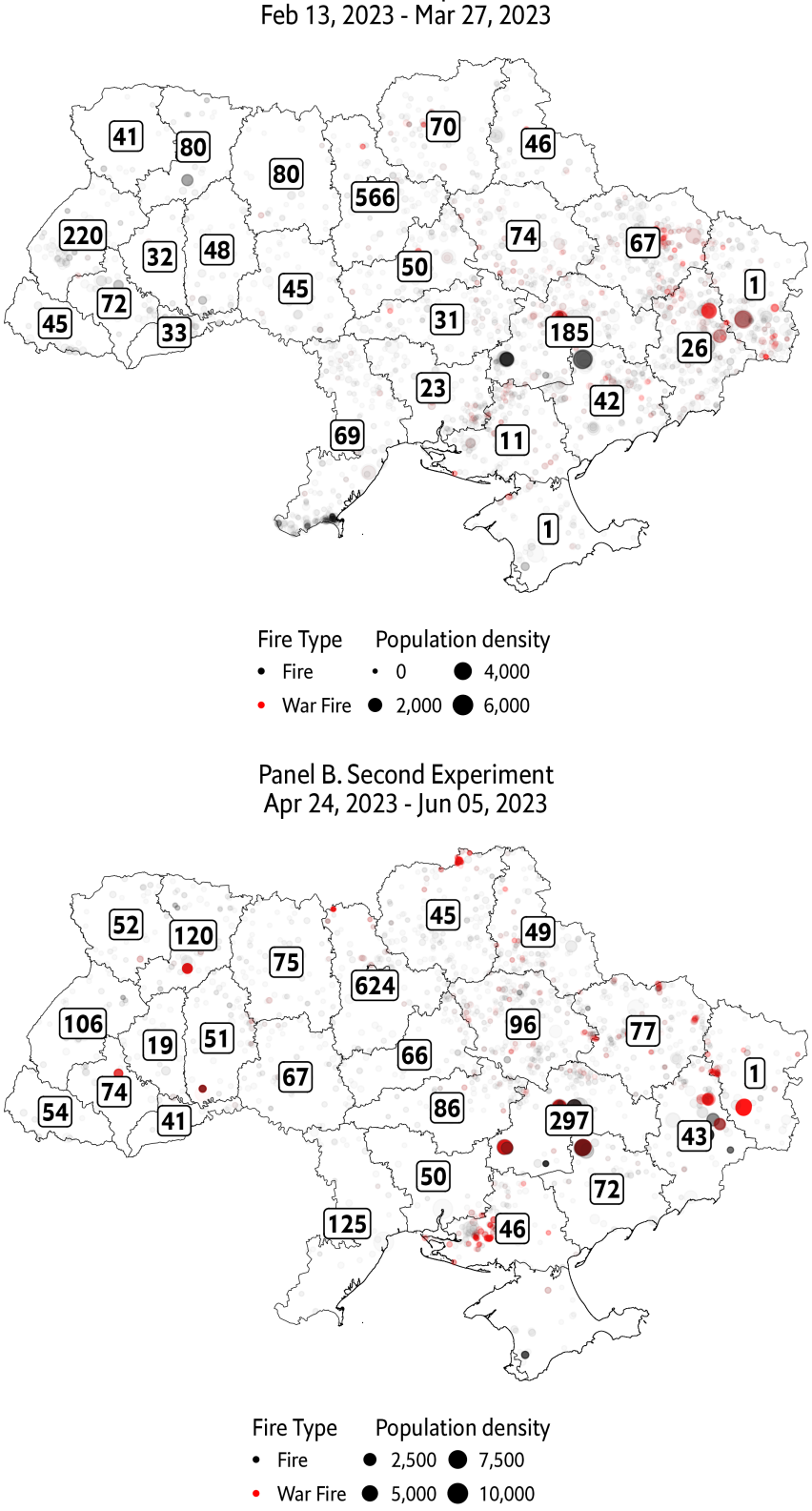
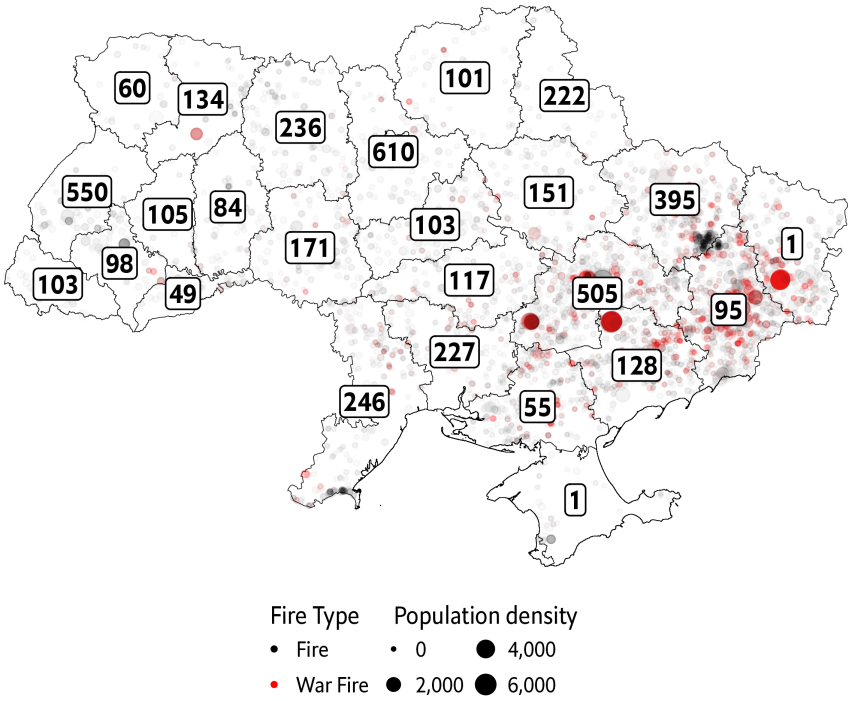


**Figure A10**: *Continued*–Intensity of Conflict Across Different Experiments




*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.

# Appendix Tables

**Table A1:** Overview of Experimental Design Across Experiments

| | **First Experiment** | **Second Experiment** | **Third Experiment** |
|---|---|---|---|
| **Registration** | Dec 2022–Jan 2023 | Mar–Apr 2023 | Dec 2023–Jan 2024 |
| **Enrollees** | 2,322 guardians<br>2,518 students | 2,573 guardians<br>2,767 students | 4,299 guardians<br>4,547 students |
| **Additional eligibility criteria** | – | First time enrolling | First time enrolling and living in Ukraine |
| **Assignment to Treatment or Control** | Random<br>*Strata:* parental education and living in/out of Ukraine | Random<br>*Strata:* parental education and region of residency | Random<br>*Strata:* parental education and region of residency |
| **Assignment to tutoring groups** | Random<br><br>*Strata:* treatment status, grade, preferred schedule | Ranking based on short baseline assessment<br>*Strata:* treatment status, grade, preferred schedule | Ranking based on short baseline assessment<br>*Strata:* treatment status, grade, preferred schedule |

*Note:* This table summarizes each of the three experimental designs included in this study. "Additional eligibility criteria" refers to other criteria in addition to being enrolled in grades 5 to 10 and providing parental consent and student assent.

**Table A2:** 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 A3:** 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 | 41.31 | 41.72 | 41.51 |
| 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 |
| *Other variables* | | | | |
| Teaching Experience (Years) | 17.09 | 15.85 | 17.53 | 17.65 |
| Normal stress level | 0.81 | 0.82 | 0.80 | 0.80 |
| Number of groups | 4.35 | 4.11 | 4.56 | 4.83 |
| Fixed mindset (index) | 0.25 | 0.25 | 0.26 | 0.25 |
| Bias towards resourceful students (index) | 0.44 | 0.42 | 0.45 | 0.44 |
| 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 tutor 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.

**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 $\beta_1$ from equation (3) 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 $\beta_1$ from equation (3) 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 $\beta_1$ from equation (3) 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 | ITT | TOT | # of control variables selected |
|---|---|---|---|---|
| | (1) | (2) | (3) | (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 (3). 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:** 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 A10:** Endline Completion in the Parental Investment Experiment

| | Completed Endline | |
|---|---|---|
| | Math | Ukrainian language |
| | (1) | (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 A11:** Endline Survey Completion, by Experiment

| | Completed Endline Academic/Mental Health Assessment | | |
|---|---|---|---|
| | First Experiment (1) | Second Experiment (2) | Third Experiment (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 5.1, 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). 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). In the Parental Investment Experiment, the control group refers to students that only receive the tutoring program. Statistical significance at the 1%, 5%, and 10% levels is indicated by ***, **, and *, respectively.

**Table A12:** Fairlie Bounds

| | Experiment 1 | | | Experiment 2 | | | Experiment 3 | | |
|---|---|---|---|---|---|---|---|---|---|
| | Main | Fairlie 5% | | Main | Fairlie 5% | | Main | Fairlie 5% | |
| Outcome | Estimate (1) | Lower (2) | Upper (3) | Estimate (4) | Lower (5) | Upper (6) | Estimate (7) | Lower (8) | Upper (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.

**Table A13:** Deviations from Pre-Analysis plan in AEA Registry

| **AEA Registry** | **Deviations from the registry or additional information and results included in the manuscript** |
|---|---|
| **General Information**<br>Title: The Effects of a Tech-Based Tutoring Program in Ukraine<br>RCT ID: AEARCTR-0010634<br>Initial registration date: December 13, 2022<br>First published: December 16, 2022, 4:03 PM EST<br>Last updated: December 28, 2023, 1:45 PM EST<br>**Locations**<br><br>Country: Ukraine<br>**Primary Investigator**<br><br>Name: Lelys Dinarte<br>Affiliation: The World Bank<br>Email: ldinartediaz@worldbank.org | *Explanation:* The last update was made on December 28, 2023, before the third experiment began and prior to the team having access to any outcome data from the first two experiments. As documented in the RCT registry history, the main changes include: updating the trial status from "in development" to "ongoing"; adding more detailed information about the intervention; updating PI Rodriguez's institutional affiliation; revising the end dates to reflect the additional time required to complete the intervention and experimental design of the third experiment; providing a more accurate estimation of sample size and statistical power; and including information on IRB approval. |
| **Other Primary Investigator(s)**<br>Name: Renata Lemos<br>Affiliation: The World Bank<br>Email: rlemos@worldbank.org<br>Name: James Gresham<br>Affiliation: The World Bank<br>Email: jgresham@worldbank.org<br>Name: Harry Patrinos<br>Affiliation: The World Bank<br>Email: hpatrinos@worldbank.org<br>Name: Rony Rodriguez<br>PI Affiliation: Harvard University<br>PI Email: rrodriguezramirez@g.harvard.edu | **Other Primary Investigator(s)**<br>Name: Harry Patrinos<br>Affiliation: University of Arkansas<br>Email: patrinos@uark.edu<br>*Explanation:* Change in the affiliation and email of PI Patrinos from World Bank to University of Arkansas. |

*Continued on next page*

Table A13–*continued from previous page*

| **AEA Registry** | **Deviations from the registry or additional information and results included in the manuscript** |
|---|---|
| **Additional Trial Information**<br>Status: On going<br>Start date: 2023-01-02<br>End date: 2024-08-31<br>Keywords: Crime, Violence, & Conflict, Education<br>Prior work: This trial does not extend or rely on any prior RCTs. | **Additional Trial Information**<br>Status: Completed<br>Start date: 2022-12-02<br><br>*Explanation:* The registration started in December 2022 and the last data collection activity was conducted in August 2024. |
| Abstract: Evidence suggests that traumatic events, such as pandemics and wars, can impact children's learning, socio-emotional development, and sense of protection (Quintana-Domeque and Ródenas-Serrano, 2017; Almond et al. 2018). Ukraine's education system faces critical constraints in providing high-quality education to its students on their path toward recovery after disruptions to schooling and learning due to years of pandemic-related school closures. While children in many countries have gone back to school, the return to in-person education has been hindered by a lack of security, significant student and teacher displacement, and school damages posed by 6 months of Russia's invasion. Currently, learning losses in Ukraine are estimated to be over one year (Angrist et al, 2022), with learning outcomes falling below the lowest-performing countries in Europe which will have substantial impacts on human capital development in the country. To mitigate the impact of these traumatic events on children, the Ukrainian education system must find new strategies for supporting learning recovery and increasing learning equity while children are not able to return to in-person schooling.<br><br>High dosage small group instructional tutoring can be a powerful strategy to improve learning outcomes and cognitive and socioemotional skills, as it offers students a massive increase in personalized instruction, enabling teaching at the right level (Banerjee et al., 2015). To test the effectiveness of these programs in a conflict-affected setting, we study a tutoring program offering supplemental learning in math and Ukrainian language and psychosocial support. | *Explanation:* There are two main differences between the Abstract reported in the registry and the manuscript content. First, due to logistical challenges and the NGO's limited implementation capacity, we included only two variations in the core structure of the tutoring program across waves: (i) light-touch tools for academic support and (ii) enhanced psychosocial support activities. These components were implemented only in the second and third experiments. We do not test the effectiveness of one component relative to the other; instead, we assess the effectiveness of each component relative to a control group with no tutoring program. These changes were reflected in the Intervention Description section (see below) during the last update on December 28, 2023, but not in the Abstract section.<br><br>Second, following careful discussions with the NGO regarding the program's structure, the primary outcomes reported are limited to math and Ukrainian language test scores, as well as mental health (stress and anxiety). The main reason is that the intervention was primarily academic in nature, with some components designed to improve students' mental health. As such, these are the outcomes most directly targeted by the intervention. Social-emotional skills, while potentially affected by the intervention, are considered to be influenced only indirectly. Therefore, they are now categorized as secondary outcomes (or mechanisms), alongside structured peer interactions, complementary investments, attitudes toward learning, and aspirations. More details are provided below. |

*Continued on next page*

Table A13–*continued from previous page*

| AEA Registry | Deviations from the registry or additional information and results included in the manuscript |
|---|---|
| The target population of the tutoring program is Ukrainian students in grades 5 to 10 who are seeking supplemental support beyond the standard online schooling schedule. Initially, students are placed in groups of 3 and will receive 3 hours of tutoring per week for 6 weeks by paid-for tutors through an online platform. The implementing partner is Teach for Ukraine (https://teachforukraine.org/en/).<br><br>The impact evaluation will include three waves, through which we will test for the effectiveness of varying certain attributes, including program length, group size, content allocation, group composition by ability, and tutors as role models. Students will be recruited in each wave. All waves will have one control and one treatment group.<br><br>The effects of the tutoring program will be measured in math and Ukrainian language test scores, socioemotional skills, and mental health. As secondary outcomes, we will measure the effects of the intervention on expectations, time use, attendance to the tutoring activities, and attitudes towards tutoring.<br><br>**Registration Citation**<br>Citation: Dinarte, Lelys et al. 2023. "The Effects of a Tech-Based Tutoring Program in Ukraine." AEA RCT Registry. December 28. https://doi.org/10.1257/rct.10634-1.1<br><br>**Sponsors**<br>Sponsor name: The World Bank<br>Sponsor location<br>Sponsor URL: https://www.worldbank.org/en/home<br>Sponsor name: UBS Foundation<br>Sponsor location<br>Sponsor URL: https://www.ubs.com/global/en/ubs-society/philanthropy/optimus-foundation.html<br>**Partner**<br>Name: Teach for Ukraine<br>Type: NGO<br>URL: https://teachforukraine.org/en/ | |

*Continued on next page*

Table A13–*continued from previous page*

| AEA Registry | Deviations from the registry or additional information and results included in the manuscript |
|---|---|
| **Interventions**<br>The intervention under evaluation consists of a tutoring program offering supplemental learning in math and Ukrainian language and psychosocial support. The learning component includes subject-specific academic content in accordance with Ukrainian educational programs. The psychosocial support component includes activities to enhance student's social and emotional well-being. The target population of the tutoring program is Ukrainian students in grades 5 to 10 who seek supplemental support beyond the standard online schooling schedule.<br><br>The structure of the intervention will vary in each wave. In wave 1 of the intervention, we implement a "Barebones Program" where the group size is three students, the number of hours per week is 3, and the number of subjects is 2 (math and Ukrainian language).<br><br>In wave 2, we keep the same structure of the program as in wave 1, but we implement a variation–that we call the "enhanced program"– with three additional features. First, we conduct a short assessment during the enrolment stage to have a proxy for student knowledge prior to the beginning of the tutoring program and used this measure to rank students and sort them in groups based on ability levels. Second, we use the information in this short assessment to prepare and share a diagnostic report with tutors for each group assigned to them, containing aggregated information on how well each group performed relative to the average for all students. Third, we developed short curriculum-based formative assessments and shared with tutors during the second week of the program to be used as a guide for tutors to assess and learn how well students were doing on several learning outcomes.<br><br>In the third wave, we use the same structure of the program as in wave 2 but we work with a team of psychologists and specialists from the Harvard Program in Refugee Trauma (HPRT) to enhance the psychosocial support component. In the updated curriculum for this component, we include activities related to breathing exercises, mental check ins, parables, and other activities. | *Explanation:* As mentioned earlier, the intervention description was already updated in the registry during the last update on December 28, 2023. In Experiment (Wave) 1, we implemented the core tutoring program. In the second experiment, we added light-touch tools for academic support to the core program. Finally, in the third experiment, we further expanded the intervention by including enhanced psychosocial support activities, building on the structure of the second experiment. |

*Continued on next page*

Table A13–*continued from previous page*

| **AEA Registry** | **Deviations from the registry or additional information and results included in the manuscript** |
|---|---|
| Intervention Start Date<br>2023-01-23<br>Intervention End Date<br>2024-04-30 | |
| **Primary Outcomes**<br>Math test score<br>Ukrainian language test score<br>Mental health | **Primary Outcomes**<br>Math test score<br>Ukrainian language test score<br>Mental health (stress and anxiety)<br>*Explanation:* Mental health is proxied using measures of stress and anxiety. |
| **Secondary Outcomes**<br>Social emotional skills<br>Personal and Academic Expectations<br>Time use<br>Attendance of the tutoring activities<br>Attitudes towards tutoring | **Secondary Outcomes**<br>Social emotional skills (GRIT and self-efficacy)<br>Attitudes towards learning and educational aspirations<br>Student self-driven learning investments<br>Attendance to and engagement during tutoring activities<br>Structured peer interactions (enrollment and interaction in the platform)<br><br>*Explanation:* The main changes in the secondary outcomes, which we called "Mechanisms" are the following. First, in addition to tracking attendance at the tutoring sessions, we used data from tutor journals to measure students' engagement during the sessions (e.g., whether they had their camera on, actively participated, or came prepared). Although this was not included in the original registration, we believe it is an important outcome that captures the quality of students' participation in the intervention.<br><br>Second, to reduce the number of hypotheses tested—and given the high correlation between "attitudes toward tutoring" and "personal and academic expectations"—we combined these into a single outcome: "Attitudes toward learning and educational aspirations." |

*Continued on next page*

Table A13–*continued from previous page*

| AEA Registry | Deviations from the registry or additional information and results included in the manuscript |
|---|---|
| | Third, the original objective of measuring "time use" was to assess whether students were exerting additional effort beyond the tutoring sessions. To better reflect this intent, we renamed the outcome to "Student self-driven learning investment." |
| **Experimental design**<br>The impact evaluation will be conducted in three waves of tutoring. In each wave, we will test for the effectiveness of varying certain attributes and following an agile experimenting approach. The exact attribute to be varied and tested in each wave will depend on the results from the previous wave.<br><br>Students will be recruited in each wave. The tutoring program will be advertised on social network platforms and disseminated by the Ministry to schools around the country, influencers in the Education Sector, and TFU. Students interested in taking part in the tutoring program will register online, provide responses to a questionnaire, and provide their own consent as well as parental consent. Students who submit this information will become part of the study sample and, for all the 3 waves, will be randomly assigned to the treatment or the control group.<br><br>The randomization procedure for each wave will consist of a two-stage stratified household-level randomization. In each wave, at the first stage, all enrolled students are randomly assigned to treatment or control groups. The stratification variables are whether the parent has completed higher education and if they were living in Ukraine for wave 1 and the interaction between region (Central, Eastern, Southern, Western, Out of Ukraine) and if the parent has completed higher education in waves 2 and 3. Those in the treatment groups receive the tutoring activities, and the ones in the control groups enrolled in the online platform and were allowed to interact among each other but did not receive the tutoring sessions. | **Experimental design**<br>The deviations from the initial experimental design are the following:<br><br>Agile experimenting approach and effectiveness of varying program attributes. As previously explained, in the first experiment we evaluated the effectiveness of the core tutoring program—two 1.5-hour weekly sessions in math and Ukrainian language, delivered over six weeks in small groups of three students, each with a dedicated subject tutor. Sessions were held online. In the second experiment, we added light-touch tools for academic support to the core program. In the third experiment, we further included enhanced psychosocial support activities in addition to the previous components.<br><br>Due to logistical constraints and the NGO's limited capacity to implement a fully agile experimentation approach, treatment and control assignment was conducted separately within each experiment. As a result, we are unable to estimate the relative effectiveness of program components across experiments. Instead, we can assess the effectiveness of each version of the tutoring model relative to a control group (i.e., no intervention) within each experiment, but not across different program variations.<br><br>Assignment to groups: Assignment to treatment status (T or C) was done through a stratified, household-level randomization conducted in a single stage. |

*Continued on next page*

Table A13–*continued from previous page*

| **AEA Registry** | **Deviations from the registry or additional information and results included in the manuscript** |
|---|---|
| For the second stage, we assign students to different tutoring groups stratifying by grade and preferred schedule in wave 1. The stratification variables in waves 2 and 3 included grade and preferred schedule and they were also ranked by ability using a short knowledge test in math and Ukrainian language collected during the registration period. Students are then sorted into small groups based on their ranking position.<br><br>Randomization Method: Randomization will be done in the office using a code developed by the PIs.<br><br>Randomization Unit: The unit of randomization will be the household. Yet, based on pilot data, the average number of students enrolled per household is 1.2. Therefore, we expect our randomization to be almost at the individual level.<br><br>Was the treatment clustered? Yes | Tutoring group formation followed this structure (deviations from the registry are italicized):<br>- First experiment: Students were *randomly assigned* to groups of three, stratifying by *treatment status* (T or C), grade level, and preferred schedule for tutoring sessions.<br>- Second and third experiments: After stratifying by *treatment status*, grade level, and preferred schedule for tutoring sessions, students were ranked based on their scores on a short baseline assessment and then grouped into groups of *three*.<br><br>*Parental investment experiment.* The impact evaluation included four experiments: three main experiments (waves) to test the impacts of variations of the tutoring program, and one additional small experiment to test the parental investment mechanism. This last experiment was not pre-registered because our initial power calculations were very conservative, assuming higher attrition and lower take-up than what we observed during implementation. However, since the results are highly informative and offer valuable insights into the challenges of engaging parents in supporting their children's learning during wartime, we believe it is important to include them in the manuscript.<br><br>*Heterogeneity results.* During the dissemination of preliminary findings, we received frequent requests regarding whether the intervention had differential impacts based on participant characteristics or baseline conditions. In the current manuscript, we present heterogeneity analyses along the following dimensions: (i) Student characteristics – gender, age, academic performance at baseline, and mental health; (ii) External factors – parental well-being (stress) and local conflict intensity. Since these heterogeneity analyses were not pre-registered, we recommend interpreting them as exploratory and complementary to the main results. |

*Continued on next page*

Table A13–*continued from previous page*

| **AEA Registry** | **Deviations from the registry or additional information and results included in the manuscript** |
|---|---|
| **Experiment Characteristics**<br>Sample size: *planned number of clusters*<br>At least 8,000 households<br><br>Sample size: *planned number of observations*<br>At least 9,600 students<br><br>Sample size (or number of clusters) by treatment arms<br>Treatment → 4,000 households; 4,800 students<br>Control → 4,000 households; 4,800 students<br><br>*Minimum detectable effect size for main outcomes (accounting for sample design and clustering):* Assuming attrition of 30%, take up of 52%, a power of 80%, and a type I error rate of 0.05, for a sample size of 9,600, we estimate an MDE of 0.08SD.<br><br>**Institutional Review Boards (IRBs)**<br>IRB Name: Human Subjects Committee for Innovations for Poverty Action IRB-USA<br>IRB Approval Date: 2023-03-30<br>IRB Approval Number: 16680 | **Experiment Characteristics**<br>The exact number of clusters and students across the three waves are the following:<br>Total:<br><br>Clusters: 9,194 households<br>Individuals: 9,832 students<br><br>*By treatment status:*<br>Treatment: 4,595 households, 4,911 students<br>Control: 4,599 households, 4,921 students<br><br>Given an attrition rate of between 40% to 55%, take up of 70%, a power of 80%, and a type I error rate of 0.05, for a sample size of 9,832 students and 9,194 households, we estimate an MDE of 0.08SD across waves. |

**Table A14:** Student and Guardian Characteristics comparison with PISA sample

| | Experiments | | | | PISA | |
|---|---|---|---|---|---|---|
| | All | 15-Year-Old Students | PISA Regions | 15-Year-Old and PISA Regions | | |
| | Mean/(SD) (1) | Mean/(SD) (2) | Mean/(SD) (3) | Mean/(SD) (4) | Obs. (5) | Mean/(SD) (6) |
| Grade | 7.07 (1.67) | 9.53 (0.66) | 7.09 (1.68) | 9.54 (0.69) | 2,417 | 9.95 (0.23) |
| Age | 12.48 (1.69) | 15.00 (0.00) | 12.48 (1.69) | 15.00 (0.00) | 2,417 | 15.36 (0.48) |
| Girl | 0.55 | 0.62 | 0.55 | 0.64 | 2,417 | 0.53 |
| Guardian's tertiary education | 0.79 | 0.68 | 0.80 | 0.68 | 1,919 | 0.69 |
| Electronics at home | 0.98 | 0.99 | 0.98 | 0.98 | 2,277 | 0.99 |
| Obs. | 9,832 | 1,669 | 6,492 | 1,115 | | |

*Notes*: This table shows unweighted mean and standard deviations (in parentheses) of student and guardian characteristics for our experimental samples and the PISA 2022 sample. Columns (1)–(4) report statistics for all experiment (col 1), students who are 15-year-old (col 2), by the subset of PISA regions that overlap with our study (col 3), and the intersection of age and PISA regions (col 4); columns (5)–(6) report the PISA comparison. "PISA regions" refers to the 18 out of 27 subnational regions that administered both the full PISA 2022 assessment and its parent modules in order to match the samples in columns (5) and (6). Note that guardian's education is observed for only 1,919 PISA respondents because the parental-background questionnaire was optional.