Santos, Indhira et al.

Working Paper

Can Grit Be Taught? Lessons from a Nationwide Field Experiment with Middle-School Students

IZA Discussion Papers, No. 15588

Provided in Cooperation with:
IZA – Institute of Labor Economics

Suggested Citation: Santos, Indhira et al. (2022) : Can Grit Be Taught? Lessons from a Nationwide Field Experiment with Middle-School Students, IZA Discussion Papers, No. 15588, Institute of Labor Economics (IZA), Bonn

This Version is available at:
https://hdl.handle.net/10419/265809

I Z A Institute
of Labor Economics
Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15588

Can Grit Be Taught? Lessons from a Nationwide Field Experiment with Middle-School Students

Indhira Santos ⓡ
Violeta Petroska-Beska ⓡ
Pedro Carneiro ⓡ
Lauren Eskreis-Winkler ⓡ
Ana Maria Munoz Boudet ⓡ
Ines Berniell ⓡ
Christian Krekel ⓡ
Omar Arias
Angela Duckworth

SEPTEMBER 2022

I Z A Institute
of Labor Economics

Initiated by Deutsche Post Foundation

DISCUSSION PAPER SERIES

IZA DP No. 15588

Can Grit Be Taught? Lessons from a Nationwide Field Experiment with Middle-School Students

Indhira Santos
World Bank

Violeta Petroska-Beska
Cyril and Methodius University in Skopje

Pedro Carneiro
University College London and IZA

Lauren Eskreis-Winkler
Northwestern University: Kellogg School of Management

Ana Maria Munoz Boudet
World Bank

Ines Berniell
Universidad Nacional de La Plata and CEDLAS

Christian Krekel
London School of Economics and IZA

Omar Arias
World Bank

Angela Duckworth
University of Pennsylvania

SEPTEMBER 2022

ISSN: 2365-9793

IZA – Institute of Labor Economics

Schaumburg-Lippe-Straße 5–9
53113 Bonn, Germany

Phone: +49-228-3894-0
Email: publications@iza.org

www.iza.org

ABSTRACT

Can Grit Be Taught? Lessons from a Nationwide Field Experiment with Middle-School Students*

We study whether a particular socio-emotional skill – grit (the ability to sustain effort and interest towards long-term goals) – can be cultivated through a large-scale program, and how this affects student learning. Using a randomized control trial, we evaluate the first nationwide implementation of a low-cost intervention designed to foster grit and self-regulation among sixth and seventh-grade students in primary schools in North Macedonia (about 33,000 students across 350 schools). The results of this interventions are mixed. Exposed students report improvements in self-regulation, in particular the perseverance-of-effort facet of grit, relative to students in a control condition. Impacts on students are larger when both students and teachers are exposed to the curriculum than when only students are treated. For disadvantaged students, we also find positive impacts on grade point averages, with gains of up to 28 percent of a standard deviation one year post-treatment. However, while this intervention made students more perseverant and industrious, it reduced the consistency-of-interest facet of grit. This means that exposed students are less able to maintain consistent interests for long periods.

JEL Classification: C93, D91, I20, I24

Keywords: socio-emotional skills, grit, gpas, middle-school students, field experiment, RCT

Corresponding author:
Christian Krekel
London School of Economics and Political Science
Houghton Street
London, WC2A 2AE
United Kingdom
E-mail: C.Krekel@lse.ac.uk

* We thank the Ministry of Education and Science of North Macedonia; the staff from The Center for Human Rights and Conflict Resolution; the team at PUBLIK DOO; Bojana Naceva and Jasminka Sopova from the World Bank; Robert Gallop, and the staff from the Character Lab for their support. Victoria Levin, Hillary Johnson, Maria Davalos, Armin Falk, Fabian Schmidt, as well as seminar participants at the annual meeting of the American Economic Association in Atlanta in 2019, the briq/IZA Behavioral Economics of Education workshop in Bonn in 2019 provided valuable comments at different stages of the design and analysis. Carneiro gratefully acknowledges the financial support from the European Research Council through grant ERC-2015-CoG-682349. The research was funded by a grant from the Umbrella Facility for Gender Equality at The World Bank and a grant by the Research Department. Russell Sage and the Walton Family Foundation supported the Character Lab. The RCT evaluated in this article is registered with the AEA Trial Registry (RCT ID: AEARCTR-0002094), and a University IRB approval by the University of Pennsylvania. The use of Ⓕ signals that the authors' names have been randomized using a uniform distribution.

I. INTRODUCTION

A growing literature in psychology and economics shows that socio-emotional skills play a key role in predicting education and labor market outcomes (Alan et al., 2019; Acosta and Muller, 2018; Kautz et al., 2014; Borghans et al., 2008a, 2008b; Heckman et al., 2006). Attributes related to self-regulation, in particular, have been found to be strong predictors (Levin et al., 2016; Heckman and Kautz, 2014; Stecher and Hamilton, 2014; Naemi et al., 2013; Tough, 2012; Willingham, 1985). Amongst these, grit – the ability to sustain effort and interest towards long-term goals (Duckworth et al., 2007) – has been found to predict various outcomes at levels comparable to IQ and the personality trait of conscientiousness (Eskreis-Winkler et al., 2014, 2016; Maddi et al., 2012; Duckworth et al., 2007).

Grit has a strong relationship with conscientiousness, representing its proactive aspects centering around industriousness and self-control (Roberts et al., 2005).[2] It has two distinct, albeit related, facets: *(i)* perseverance of effort (i.e. working strenuously towards challenging goals over long periods of time despite failure, adversity, or plateaus in progress); and *(ii)* consistency of interest (i.e. maintaining interest for long-term goals without losing focus; see Duckworth et al., 2007; Duckworth and Quinn, 2009).[3]

In cross-sectional samples, adults who score high on grit make fewer career changes, progress further in their formal education, and obtain higher GPAs (Duckworth et al., 2007). Prospective longitudinal research looking at grit and education outcomes finds that grit predicts, for example, school graduation (Eskreis-Winkler et al., 2014) and the ranking of students in high-stakes competitions such as the National Spelling Bee (Duckworth et al., 2007). Interventions that improve grit have also been found to improve performance in standardized tests (Alan et al., 2019). These findings support a growing

[2] Grit has a pairwise correlation coefficient with conscientiousness of about 0.8, leading some authors to argue that grit and conscientiousness are in fact the same (Schmidt et al., 2017; Credé et al., 2016; Eskreis-Winkler et al., 2014; Ivcevic and Brackett, 2014). While this debate is ongoing and beyond the scope of this paper, it should be noted that the literature reports a high variance in the strength of correlations (between 0.4 and 0.7) and measures used to assess conscientiousness vary substantially.

[3] The former correlates strongly with the productiveness facet of conscientiousness, while the latter does not, suggesting that grit is indeed a different construct from conscientiousness (Schmidt et al., 2020).

emphasis in education policy on integrating socio-emotional skills into formal education (OECD, 2015; Guerra et al., 2014).

Given its importance in influencing important individual outcomes, we ask whether it is possible to cultivate grit through large-scale interventions in schools. Evidence from smaller interventions suggests that the answer is "yes" (Alan et al., 2019; Eskreis-Winkler et al., 2014), at least in the short run. The question is whether it is possible to cultivate grit at scale in schools, with a nationwide program.

To answer this question, we designed and evaluated the first nationwide intervention to foster grit, implemented amongst middle-school students in North Macedonia. In the Spring of 2016, all sixth and seventh-grade students in the country were randomly allocated to receive either one of two grit-building treatments (i.e. low or high-intensity) or to be part of a control condition. In the high-intensity treatment, both students and teachers were exposed to our intervention with teachers being trained to deliver a grit curriculum to their students. The low-intensity treatment relied on student self-paced learning, from a set of materials that were developed for this purpose, without interference from teachers. Control students had no exposure to the grit curriculum.

This intervention consists of a curriculum that teaches and motivates students to adopt the tenets of *deliberate practice* – to identify stretch goals, get feedback, concentrate, and repeat until mastery (Ericsson, 2008; Ericsson et al., 1993), which has been shown to cultivate grit (Duckworth et al., 2014; 2011). It also stresses that achievement is not determined solely by immutable characteristics such as talent, gender, or ethnicity. To evaluate the impacts of this intervention, we measured students' socio-emotional skills development using survey data and tracked their GPAs (up to one year post-treatment) in official, administrative school records.[4]

The results of this intervention are mixed. Relative to the control group, students exposed to the grit curriculum show improvements in deliberate practice beliefs and the perseverance-of-effort facet of grit, as well as in socio-emotional skills more generally (which include, besides deliberate practice beliefs

[4] The pre-registered analysis plan in the AEA RCT Registry is: Arias et al. (2017). "Can Grit be Taught? Learning from a field experiment with middle school students in FYR Macedonia." https://doi.org/10.1257/rct.2094-1.0.

and grit, measures of motivation, frustration reaction, locus of control, and present bias). Impacts are larger in the high-intensity treatment, as well as for girls and for Roma students – a group that has been traditionally disadvantaged in North Macedonia. Amongst Roma students, we observe gains in GPAs that become stronger over time, reaching up to 28% SD one year post-treatment. These stronger benefits for disadvantaged students echo similar findings from other psychological interventions in education (Yeager et al., 2019; Sisk et al., 2018; Paunesku et al., 2015; Cohen et al., 2009; Hulleman and Harackiewicz, 2009; Wilson and Linville, 1982).[5]

However, we also find that, relative to students in the control group, exposed students score lower on the consistency-of-interest facet of grit, suggesting that they have a lower disposition to maintain goal-interest for longer periods of time – a potential unintended consequence of our and similar interventions. One possible explanation for this result is that, by teaching students the value of deliberate practice applied repeatedly to a stretch goal, without specifying the goal itself, this intervention shifted students focus from long-term goals to more present ones (i.e. academic performance in the next test). At the same time, it could be reflective of this intervention not intentionally targeting a specific interest or long-term goal, as well as the fact that children in our targeted age group (who are between 11 and 14 years old) have a varied set of interests.[6] Recent research has shown that developing consistent interests (i.e. specializing) later in life may, paradoxically, depend on diversification (i.e. sampling) of interests earlier in life.[7] It is possible this negative impact is attenuated or eliminated with a small change in the curriculum, fostering the development of consistent interests and goal setting.

To foster grit, this intervention follows a two-pronged approach. First, drawing on the evidence on deliberate practice (Ericsson, 2008; Ericsson et al., 1993), students are taught to distinguish effective, evidence-based ways of studying from ineffective ones, as well as strategies to implement these more

[5] Disadvantaged students, whether by income, gender, race, or ethnicity, often experience higher-than-average challenges and stress in academic settings compared to their peers (Schmader, 2010; Beilock et al., 2007; Murphy et al., 2007; Ben-Zeev et al., 2005; Steele and Aronson, 1995). As a result, the decline in grades that is generally found for all students in transition periods (in our case, the start of sixth grade in middle school) is more pronounced for students from disadvantaged backgrounds (Gutman et al., 2003).

[6] See, for example, Sturman and Zappala-Piemme (2017).

[7] See, for example, Gopnik (2020) and Cote and Erickson (2015).

effective ways of studying. Second, the intervention aims to motivate effort. To do this, it focuses on changing students' beliefs about practice, by raising expectancies and values, two psychological antecedents of motivated, effortful behavior (Pintrich, 2003; Eccles and Wigfield, 2002; Wigfield and Eccles, 2000; Feather, 1982; Crandall, 1969; Battle, 1965; Atkinson, 1957).

According to Expectancy-Value Theory, 'expectancy' is the extent to which individuals believe they will succeed, whereby 'value' refers to the subjective value individuals attach to a positive outcome, which increases in expected benefits from effort and decreases in expected costs. In academic settings, expectancy and value are reliably associated with effort expenditure and achievement (Nagengast et al., 2013; Eccles et al., 1993; Meece et al., 1990; Eccles et al., 1984). We predict that students will be more likely to engage in sustained effort if they hold a strong expectancy that doing so would improve academic achievement, and if they attach a high value to academic achievement. It is important to instill such an expectancy and valuation – particularly amongst disadvantaged students – given the prevalence of false, deterministic beliefs surrounding achievement, for example stereotypical beliefs that talent, gender, or ethnicity alone determine achievement and later-life outcomes. Similarly, in settings where students may have imperfect information on the returns to education and performance, there may be significant scope for improving perceptions about the true value of effort.

By teaching students effective ways to practice and by changing their beliefs about expectancies and values of effortful practice, this intervention aims to raise the take-up of deliberate practice, which – by repeated application – is expected to cultivate grit (Duckworth et al., 2014; 2011). In addition, to address some of the pre-existing gaps in expectancies and values between disadvantaged and non-disadvantaged groups, this intervention covers not only the key steps of deliberate practice but is also designed to counter negative stereotypes surrounding gender and ethnicity in North Macedonia. It does so by providing positive role models and counter-stereotypical examples throughout the intervention materials. An intervention that undoes deterministic beliefs – by pointing out that practice, not gender or ethnicity, or a birth-given level of intelligence, determines achievement – should be most helpful for students who hold such deterministic beliefs most strongly. Hence, by addressing some of the underlying causes of poor(er) school performance amongst disadvantaged students, this intervention also aims at reducing inequality in education outcomes, creating a more levelled academic playing field.

This intervention joins a wave of recent research aimed at alleviating educational inequalities through psychological interventions (Outes-Leon et al., 2020; Broda et al., 2018; Walton and Wilson, 2018; Inzlicht and Schmader, 2008). As opposed to changing structural variables or surrounding policies, these interventions reduce inequality by empowering disadvantaged students to make the best of the imperfect environments which they find themselves in. Our findings point towards a promising, cost-effective approach that, in combination with other policies, can contribute to closing equity gaps in educational attainment.

To the best of our knowledge, there is only one other intervention focused on grit and firmly rooted in psychological research that has been tested and rigorously evaluated in a developing country context. Alan et al. (2019) implemented a grit-building intervention in 52 schools in Istanbul, Turkey, and found that treated students performed better on an incentivized real effort task and in standardized tests than students in the control condition. We contribute to this literature by designing, implementing, and evaluating a grit curriculum at a national scale, allowing us to make several key contributions.

We are the first to study a nationwide implementation of a grit curriculum in schools which is easily scalable within already existing education infrastructure. It does not introduce new technology and is designed to be delivered and rolled out within the existing education system and processes (including the teacher training component), using the regular channels set up by the Ministry of Education and Science of North Macedonia.[8] While making the fidelity of implementation more challenging than in smaller-scale, proof-of-concept studies, reliance on regular channels shows the extent to which this intervention can be delivered and managed at scale within existing education systems.

Second, in addition to overcoming external validity issues associated with real world fidelity of implementation, our design also helps address concerns related to the targeting of the program. Psychological interventions are often "tailored" to the contexts in which they are delivered to ensure "fit" with the population of interest. This customization comes at the cost of generalization. While

[8] The intervention is paper-based, following on an assessment of the technology availability across all schools in North Macedonia, which was limited and/or had maintenance, connection, or other issues. Where teachers were involved, trainings took place at the scheduled locations and times for teacher trainings during the school year set by the Ministry of Education and Science.

piloted and adjusted to ensure comprehension of content and materials (i.e. that students understood the materials and that examples resonated broadly), this intervention is designed for the general student population in North Macedonia, for example by catering to the two main languages of instruction in the country.[9] Adaptations to language and details of featured characters (e.g. names and ethnicities) could be easily tailored and transferred to other contexts.

Third, our results are not affected by selective school buy-in. Almost all schools in the country were included in this intervention and were randomly assigned to either one of our two treatment groups or the control group, again strengthening the external validity of our results. Within each school, there was no bias in the selection of teachers, classrooms, or students who participated in this intervention. Hence, any positive effects of this intervention cannot be attributed to students', teachers', or even administrators' pre-intervention interest in our curriculum. In the field of psychological interventions, nearly all randomized controlled experiments in school contexts require school-level buy-in, and impacts could thus be specific to those schools who want or choose to participate, and not necessarily representative of all schools in a country.

This paper is organized as follows. In Section 2, we describe the intervention, including the different treatments and the key mechanism of behavior change. Section 3 gives an overview of our survey and administrative data. Our empirical model is outlined in Section 4. The impacts of the intervention on socio-emotional skills and GPAs over time, on average and by different student sub-groups, as well as robustness checks are presented in Section 5. Section 6 then discusses our results against findings in the wider literature, including the cost-effectiveness of the intervention, and concludes.

II. THE INTERVENTION

The objective of the intervention is to foster grit amongst sixth and seventh-grade students (who are between 11 and 14 years old) in North Macedonia. Past research highlights the long-term benefits of working with this age group: in early adolescence, motivated behaviors have been shown to have long-

[9] The two languages are Macedonian and Albanian. According to the Macedonian State Statistical Office, in the school year of the intervention, 65% of students were in Macedonian language classes or schools, 32% in Albanian, and 3% in Turkish.

term effects on outcomes such as high school retention, college enrollment, or workforce earnings (Allensworth and Easton, 2005; Benner and Graham, 2011; Crosnoe, 2011; Heckman et al., 2014). Moreover, the beginning of the sixth grade is an important transition point in the education system in North Macedonia, as it constitutes the beginning of middle school.[10] Hence, we examine whether grit can be built during a critical developmental window that has enduring consequences for a student's future.

The intervention covered all public schools in the country with Macedonian and Albanian language of instruction, with sixth and seventh-grade classrooms, and at least five students in each single-level classroom.[11] This amounts to a total of 35,340 students in 1,780 classrooms, 352 schools, and 80 municipalities across the country.[12] The intervention was delivered nationwide, starting in the semester following the 2016 Christmas holidays (at the beginning of February) and ending by the Easter holidays (at the end of March), i.e. the third quarter of the school year 2015/2016.

The intervention consists of a curriculum of five, hour-long consecutive lessons, delivered weekly, always in the same time slot, which are divided into two parts. The first part teaches students the tenets of *deliberate practice* (Ericsson, 2008; Ericsson et al., 1993), namely to: *(i)* identify stretch goals, *(ii)* seek feedback, *(iii)* concentrate, and *(iv)* repeat until mastery. The aim is to familiarize students with deliberate practice and explain how it differs from less effective forms of practice. The second part of the curriculum aims at motivating students to actually implement deliberate practice.

To address students' expectancies (i.e. subjective probabilities of success), the materials teach that characteristics such as talent, gender, or ethnicity do not deterministically fix one's level of academic

---

[10] In North Macedonia, primary schooling is compulsory and goes from grade one to nine. Middle schooling is considered to begin with grade six. After primary school, there are four years of secondary school after which students can enter tertiary education, or alternatively, four years of vocational or technical school.

[11] 98% of children in North Macedonia attend public education which is compulsory and free.

[12] The country has 84 municipalities, with four being excluded due to lack of schools. Moreover, 64 schools (15%) out of a total of 416 schools (100%) were excluded: eight (2%) are "special schools" (e.g. art or music schools, schools for children with special needs), 31 (7%) are too small (less than five students in either sixth or seventh grade), 20 (5%) do not have single-level classrooms (e.g. schools that combine different grades in one classroom, especially in remote areas), and five (1%) are Turkish language schools. The 64 excluded schools represent 11% of all classes in each grade (sixth and seventh) and 7% of the total student population in each grade. All schools included in our intervention are primary schools that follow a standard curriculum.

achievement. Rather, effort – and particularly, effort invested in deliberate practice – is important for what people can accomplish. To address students' values (i.e. subjective values attached to success), the materials stressed the important role that academic achievement plays for later-life outcomes, emphasizing the returns to education. By familiarizing students with the tenets of deliberate practice and by changing their expectancies and values attached to doing it (i.e. their beliefs), our curriculum aimed to encourage students to take up deliberate practice and, in doing so, to cultivate grit and increase achievement. When experiencing an initial "success", this then reinforces students' beliefs, leading to a virtuous learning cycle.

Each of the five lessons builds on the previous one, starting by recapping what had been learned in the previous lesson, followed by the introduction of new concepts, and ending with a practical, hands-on activity. [13] The five lessons were delivered on consecutive weeks during the Monday morning class hour with the headteacher of the respective class.[14] This is the first class of each week across all schools in the country, and it is typically spent with the headteacher and used to talk about general issues as well as to deliver selected contents from a "Life Skills" curriculum, into which the intervention was integrated. This "Life Skills" curriculum aims at teaching general life and civic skills, and the intervention was designed in such a way as to have a similar structure and format as the rest of this curriculum.

The intervention had two treatment arms. In the first arm, the delivery of the curriculum was self-paced and relied entirely on student self-learning, with minimal interference of headteachers. The second arm added a teacher-training module and relied on headteachers to deliver the intervention. The control group received the existing "Life Skills" curriculum or did other activities at the discretion of the headteacher. Table I provides an overview of the intervention by experimental condition.

---

[13] Three additional sessions took place: one week before and one week after the intervention to collect baseline and endline surveys, and another two weeks after to collect additional behavioral outcomes. The latter also included a pilot measure to capture 'objectively' three (out of the four) dimensions of deliberate practice. The results on these additional behavioral outcomes, however, are not included in this paper, given concerns about attrition during the post-intervention data collection.

[14] The headteacher can be a teacher in any subject, and is assigned by the school at the start of the school year as the main responsible teacher for the class, for both students and parents. This teacher only teaches one subject to the students, in addition to the Monday morning class hour.

[TABLE I ABOUT HERE]

II.A. Treatment 1: "Self-Learning"

The first treatment arm consisted of the five lessons organized in weekly booklets that were distributed to students each week to work through on their own. The lessons were paper-based and self-contained, each organized so that students would take up to one school hour (about 45 minutes) to go over the materials. Each lesson had a lesson-specific student workbook that included didactic slides that are interspersed with activity prompts, engaging images, and exercises to internalize contents. Workbooks including a take-away self-evaluation for students to assess how successful they were in implementing the lesson of the week (which served as a reinforcement).[15] The treatment had minimal teacher involvement: headteachers were only responsible for distributing the materials, answering questions for clarification, and upholding discipline. They were notified of the intervention and their expected role in it by the Ministry of Education and Science and the school administration, and were given only generic information about the materials. Each week's material came in prepared packages for the classroom, including a one-page guide for the teacher regarding the basic instructions for the hour (e.g. distribution and collection of materials, as well as space to report any unusual issue affecting the class during that hour, if any).

II.B. Treatment 2: "Teacher Delivery"

The second treatment arm had the same content as the first but relied on headteachers to deliver the lessons. It involved a one-day teacher training session about a month prior to the start of the intervention. During this training, teachers received teacher-specific materials to familiarize themselves with the relevant concepts included in the lessons and a detailed lesson plan with instructions for the five weeks of lessons they were expected to deliver. Students received weekly activity booklets, which were the same as in the first treatment arm but without the self-paced content elements, as these were outsourced to be delivered by the teachers.

[15] The entire set of materials is available upon request.

III. DATA

We collected data on two categories of outcomes:

(1) Socio-emotional skills, which included *(i)* deliberate practice beliefs, *(ii*) the Short Grit Scale (Duckworth and Quinn, 2009), *(iii)* a measure of frustration reaction, *(iv)* the Motivational Frameworks Questionnaire (Gunderson et al., 2013), *(v)* locus of control (Skinner et al., 1990), and *(iv)* present bias. All skills outcomes were measured using tested and validated self-report scales, which were adapted to children from North Macedonia by translating them (back-and-forth) to Macedonian and Albanian languages.[16]

(2) GPAs at different points in time as a measure of academic achievement. These are available in the short-term, i.e. immediately after the intervention in the fourth quarter of the school year 2015/2016; in the medium-term, i.e. half a year later in the first semester of the school year 2016/2017; and in the longer-term, i.e. one year later in the second semester of the school year 2016/2017. In North Macedonia, grades are recorded on a one-to-five scale, whereby one is the lowest and five is the highest attainable grade.

Data on the different socio-emotional skills were collected through baseline and endline surveys filled out by students. To analyze whether the intervention shifted socio-emotional skills more generally, and to mitigate concerns about multiple hypotheses testing, we constructed a socio-emotional skills index ('S/E Skills Index') that summarizes the different measures of skills. In particular, it combines the measures of deliberate practice beliefs, grit, frustration reaction, motivational frameworks, locus of control, and present bias. Following Anderson (2008), we constructed this index by *(i)* switching the sign of the variables included in the index (if needed), so that a positive direction always indicates a "better" outcome; *(ii)* standardizing each variable (to have mean zero and standard deviation one, i.e. z-scores); *(iii)* averaging the standardized variables using appropriate weights (i.e. the inverse of the variance-covariance matrix of the standardized variables) to ensure that highly correlated items receive

[16] The survey instruments are available upon request.

less weight while variables that are uncorrelated and hence represent new information receive more; and then *(iv)* standardizing the resulting index.

Data on GPAs comes from official, administrative records held by the Ministry of Education and Science of North Macedonia. We use all grades given to the students during the school years 2015/2016 and 2016/2017, the year in which the intervention took place and the year after. Since the intervention targeted both sixth and seventh-grade students, we focus on the set of core subjects that are common to both groups: math, English, and language (which can be either Macedonian or Albanian, depending on the school's language of instruction).

Access to administrative data is important for two main reasons. First, it lets us study the impact of the intervention of student learning. Second, it is available for all students in North Macedonia, even the ones for whom for one reason or another we were not able to obtain survey data.

After calculating GPAs from math, English, and language, we classify them into either pre-treatment or post-treatment, depending on the date when they were recorded. As the intervention was implemented in the third quarter of the school year 2015/2016 (to be precise, between February 15 and March 21, 2016), the post-treatment GPA is calculated over the period of the fourth quarter (March 22 to August 31, 2016). Besides these *short-term GPAs*, we also calculate *medium-term* (first and second quarter of the school year 2016/2017) and *long-term GPAs* (third and fourth quarter of the school year 2016/2017). The pre-treatment GPA is calculated over the entire academic year 2015/2016, right up to the beginning of the intervention. When calculating GPAs, we use all recorded grades, including both written and oral tests grades.[17] Furthermore, depending on the subject and level (i.e. sixth or seventh grade), students may differ in the number of exams or tests they take, in the number of grades they receive, as well as in the share of these that comes from written or oral tests.[18] Unfortunately, standardized tests were not

[17] The type of exam or test has equal weight for students' GPA at the end of the school year. We conducted sensitivity analyses with respect to oral and written grades separately but these did not result in qualitatively different findings. The results are available upon request.

[18] Such differences are unlikely to cause systematic bias, as they are likely to be balanced between groups due to randomization.

available in the country at the time of the intervention. As with our skills measures, we standardize GPAs to have mean zero and standard deviation one (i.e. z-scores).

The administrative data do not include the precise dates when exams or tests were taken, but only the (precise) dates when the resulting grades were recorded by subject teachers. While this may induce some bias in our estimates, we believe that it is quantitatively rather minor, for several reasons. First, teachers are strongly encouraged to record grades as soon as possible, ideally within two weeks after the respective exam or test was taken. Second, the vast majority of assessments are taken at the end of the second and fourth quarter of the school year (i.e. after the first and second semester, before the winter and summer holidays), and only a small fraction in-between. [19] The intervention was implemented in the third quarter while our post-treatment GPAs are constructed from fourth-quarter grades only. We observe only a small fraction of assessments being taken just before the fourth quarter, for which the recording of grades could have spilled over from the intervention to the post-intervention period. Even if such spillovers would occur, these are unlikely to cause systematic bias, as they are probably balanced between groups due to randomization, unless teacher test-taking and grade-reporting behavior would have changed systematically between groups due to the intervention.[20] This is rather unlikely: the number and timing of exams or tests is announced at the start of the semester and cannot be changed abruptly.[21] As noted previously, only few exams or tests are taken during the third quarter of a school year.

To test formally for changes in teacher test-taking and grade-reporting behavior, we compared the number of grades recorded in the fourth quarter between groups. Teachers in the control group reported about 35% of all grades (σ=0.48), those in the first treatment group ("self-learning") 31% (σ=0.46), and those in the second ("teacher delivery") 28% (σ=0.45). Calculating normalized differences between

[19] We experimented with constructing GPAs using various lags (e.g. using a delay of two weeks at the beginning of the fourth quarter rather than using its sharp start date. However, our findings remained qualitatively similar to our baseline results using the sharp cut-off dates of the school year calendar.

[20] We turn to observer and experimenter-demand effects in our robustness checks in Section V.C.

[21] Apart from changes in test-taking and grade-reporting behaviour of teachers, there may also be changes in actual grading behaviour. We turn to this issue in our robustness checks in Section V.C, where we discuss observer (Hawthorne) and experimenter-demand effects.

groups to adjust for large group sizes (there were 2,469,524 recorded grades in the fourth quarter of the school year 2015/2015), none of the normalized differences exceeds the recommended threshold of 0.25, which would indicate covariate imbalance (Imbens and Wooldridge, 2009).[22] Hence, the number of grades recorded by teachers is balanced between groups, suggesting no systematic changes in teacher test-taking and grade-reporting behavior as a result of our experiment.

Finally, we obtained data on students' demographic characteristics (i.e. age, gender, and ethnicity) from official, administrative data to routinely include them as controls in our regressions. We obtained additional data on students' socio-economic characteristics (i.e. proxies for parental household wealth, whether both parents live in the household, and parental education) from our own surveys. Table 2 shows summary statistics and balancing properties for socio-emotional skills, GPAs, as well as students' demographic and socio-economic characteristics at baseline, by experimental group, for all students with non-missing information on all outcomes and controls.

[TABLE II ABOUT HERE]

Table II Panel A shows students' socio-emotional skills taken from our surveys and GPAs taken from official, administrative data. Panel B shows students' demographic and socio-economic characteristics, taken from administrative and survey data, respectively. This is the sample we use to obtain our benchmark results, and it includes all students with non-missing information on all outcomes and controls.

As seen, the number of observations varies substantially across variables: it is highest for GPAs, and lowest for demographic and socio-economic characteristics. In contrast to GPAs, which are taken from official, administrative data and which we have for most students, there is a significant amount of missing information in the survey data, mainly due to some surveys not being returned, unreadable, or

[22] Normalized differences are calculated as $\Delta x = (\bar{x}_t - \bar{x}_c) / \sqrt{(\sigma_t^2 + \sigma_c^2)}$, where $\bar{x}_t$ and $\bar{x}_c$ is the sample mean of the covariate for the treatment and control group, respectively. σ^2 denotes the respective variance. The normalized difference between our first treatment group ("self-learning") and our control group is -0.13, that between our second treatment group ("teacher delivery") and our control group is -0.19.

only partly filled out.[23] We present a detailed discussion of survey non-response and attrition in our robustness checks in Section V.C, including formal tests for (differential) attrition by group, using a balanced sample of survey and administrative data, and using multiple imputation. Our results remain robust to accounting for attrition in these ways. Note that grit is reported as a single measure, as well as split into its two facets: *(i)* perseverance of effort and *(ii)* consistency of interest. With few exceptions, pre-treatment outcomes and controls are balanced between groups, as we would expect from random assignment of schools to treatment arms.

IV. EMPIRICAL MODEL

The intervention was implemented as a cluster-stratified randomized controlled trial. The unit of randomization was the school, to lessen the probability of contamination (e.g. via information spillovers) from students in either of our two treatment groups to those in the control group. With equal probability, all schools in the country (and all sixth and seventh-grade students therein) were allocated to either one of our two treatment groups or to the control group. Moreover, to achieve a balance of groups at a regional level and hence national representativeness, we stratified the randomization by municipality. With few exceptions, this ensured that there was an equal number of treatment and control schools within each municipality.

We estimate a value-added educational production function (Todd and Wolpin, 2003):

$$y_{it} = \beta_0 + \beta_1 T_1 + \beta_2 T_2 + \beta_3' X_{it} + \sum_{k=1}^{4} \delta_k y_{it-1}^k + \mu_m + \varepsilon_{it} \qquad (1)$$

where y_{it} is the outcome of student i at time $t=\{0, 1\}$; T_1 and T_2 are indicator variables that equal one if the student belongs to treatment group one or two, respectively, and zero if they belong to the control group; X_{it} is a vector of controls, including indicators for age, gender, academic year (i.e. whether the student is in sixth or seventh grade), and ethnicity (i.e. whether the student is Macedonian, Albanian,

[23] Note that, for treatment group one, some controls are missing due to a printing error in the endline survey. We will return to this issue in our robustness checks. Unfortunately, these controls cannot be used in our analysis but are reported here for illustrative purposes.

Roma, or belongs to any other ethnicity). When evaluating impacts on GPAs, we also control for a fourth-order polynomial in pre-treatment GPAs, which yields a better model fit compared to a linear term.[24] Finally, to reflect the impact evaluation design as a cluster-stratified randomized controlled trial, we use robust standard errors clustered at the school level while controlling for a full set of municipality fixed effects, μ_m (as the randomization was stratified by municipality). All pre-treatment and post-treatment outcomes are standardized with mean zero and standard deviation one, using the control group's mean and standard deviation, to make outcomes comparable in terms of effect size.

Our regressors of interest are T_1 and T_2. Because of randomization, full eligibility, and full compliance (recall that schooling up to including grade nine is compulsory), the coefficients of these variables can be interpreted as the average treatment effect of the respective experimental condition in the population of schools in our universe.

Finally, we analyze whether the intervention had heterogeneous impacts by gender and ethnicity, as well as across the pre-treatment outcome distribution. One might expect improvements in grit and academic achievement to be stronger amongst student sub-groups that are at a relative disadvantage, namely girls or ethnic-minority students such as Roma, or students who lack skills and are lower-performing at the outset. In some psychological interventions previously studied in the literature, larger impacts have been found for students that are at a relative disadvantage or at risk. This is likely due to the higher salience of psychological barriers, such as stereotype threat, amongst those groups (see, for example, Cohen et al., 2009; Good et al., 2003; and Yeager and Dweck, 2012). In terms of delivery method, one might also expect heterogeneous impacts. In particular, we expect the teacher-delivered lessons to show larger impacts than student self-learning, given that teachers ensure exposure to the contents, can generate a more intense experience, and reflect more regular (i.e. teacher-centered) learning practices in classrooms in the country.

[24] Our results are robust to not including these controls. Moreover, they are, with few exceptions, robust to the exclusion of pre-treatment outcomes as controls; if anything, effect sizes become slightly stronger in this case.

V. RESULTS

V.A. Impacts on Socio-Emotional Skills

Deliberate Practice Beliefs, Grit, and Grit Facets

We begin by looking at socio-emotional skills and the outcomes most likely to be affected by the intervention: deliberate practice beliefs and grit, including its different facets.[25] We first look at the average effect of treatment, and then at heterogeneous effects by gender, ethnicity, and pre-treatment outcomes.[26]

Recall that the intervention aimed at fostering grit by informing students about deliberate practice and motivating them to take up this particular form of practice. By experiencing success, students are, in theory, motivated to keep practicing which, in turn, makes them grittier, in the sense that they become more perseverant and interested towards their long-term learning outcomes.

Table III Column 1 shows the average effect of the respective treatment on deliberate practice beliefs, which is a summary scale combining four items referring to the specific elements of deliberate practice. In particular, students are asked how important (while studying) it is to: *(i)* put greater effort into yet unknown material; *(ii)* concentrate solely on studying; *(iii)* seek feedback from parents and teachers; and *(iv)* repeat the material several times until they are certain to have absorbed it.

[TABLE III ABOUT HERE]

We find that the intervention successfully changed students' beliefs about the importance of deliberate practice while studying. In particular, both treatment arms significantly increased students' beliefs, but the point estimates are larger when teachers delivered these contents (about +23% SD) compared to student self-learning (about +15% SD).

---

[25] We present here a more refined set of covariates than pre-registered in our pre-analysis plan. The analysis using the exact, pre-registered covariates can be found in Appendix C Section 2. Its results largely confirm the findings presented here.

[26] As discussed below, the results presented in this section hold when accounting for multiple hypotheses testing using the stepwise p-value correction by Romano and Wolf (2005, 2016).

Column 2 shows the average effect of the respective treatment on grit, which is our main outcome of interest. The finding is surprising, as it goes into the opposite direction of the hypothesized impact: the intervention actually *reduced* students' grit, and this negative impact is even statistically significant in the student self-learning treatment.

We measure grit using the Short Grit Scale (Duckworth and Quinn, 2009), which is a standard instrument in the literature. The eight items in the scale set out to capture the two facets of grit: *(i)* perseverance of effort and *(ii)* consistency of interest. When looking at these two facets of grit separately in Columns 3 and 4, we detect an interesting pattern: while the intervention increased the perseverance-of-effort facet (in both treatments, with slightly stronger effects when teachers deliver the contents, about +6% SD *versus* +5% SD), it actually reduced the consistency-of-interest facet (again in both treatments, about -10% SD). Compared to students in the control group, treated students report being more industrious and hard-working, but also having less consistent attachment to interests for longer periods of time. Both facets combined then yield a statistically insignificant (teacher-delivery) or even significantly negative (student self-learning) impact of the intervention on grit as a whole.

The divergence between the two facets of grit is puzzling. There could be several reasons behind this finding related to measurement. Some elements of the Short Grit Scale may be less suited for the age group of the intervention (i.e. the Short Grit Scale is used in a younger cohort than it was originally developed for); there may be a problem with how items are phrased: the consistency-of-interest facet of grit is entirely captured by items that are negatively phrased (i.e. are reverse coded, e.g. "My interests change from year to year") while the perseverance-of-effort facet-items are positively phrased (e.g. "I am a hard worker"); or terms used were not understood correctly. [27]

However, we do not believe that measurement issues are driving our results. When we look at the correlation between grit-effort and grit-interest at baseline, we find only a weak (yet significant)

[27] We cannot conclusively dispel the role of these factors. However, when translating the Short Grit Scale into Macedonian or Albanian language, we applied back-and-forth translation to make sure that there are no language translation problems. Moreover, the Short Grit Scale (and all other instruments) was piloted in the country with a random set of students in the targeted age group. These pilots also included qualitative focus group discussions where no issues came up. This gives us confidence in the correct application of the scale.

correlation on average and for both Macedonian and Albanian language versions of the Short Grit Scale (+0.055 on average, +0.114 for Macedonian students, and -0.121 for Albanian students, all significant at the 1% level). Typically, the two facets of grit have been found to be more strongly correlated, although we also found a rather weak, negative correlation, -0.135, significant at the 5% level, in a large sample of Latin-American urban respondents.[28] So, while there may be some measurement issues with the Short Grit Scale, these do not seem to be unique to our country context. Importantly, though, such issues by themselves do not explain why the intervention significantly *reduced* the consistency-of-interest facet of grit for treated students (in both treatments) relative to students in the control condition.

A more plausible explanation for the negative effect of the intervention on grit-interest may be the content of the intervention itself. Recall that the intervention induces students to update their beliefs about the expectancies and values of deliberate practice while studying, and in doing so, to motivate students to take up this particular form of practice. The rationale is that, when students take up this form of practice and experience a first "success" (which becomes more likely the more students practice), this will, in turn, reinforce their beliefs, leading to sustained behavior change and potentially long-term benefits. It is plausible that, when applied repeatedly to the same (or similar) immediate goal (i.e. studying for the next test), it crowds out interests in longer-term goals – a potentially negative side effect of deliberate practice itself. In fact, Duckworth et al. (2011), Ericsson (2009, 2007, 2006), and Ericsson et al. (1993) all point towards negative side effects of deliberate practice, in the sense that this particular form of practice may be perceived as more unpleasant and more exhaustive than other forms. This could also be reflective of the intervention not intentionally targeting a specific interest, as well as of the varied interests corresponding to the young age of children in the intervention.[29] Developing consistency of interest amongst young children may, paradoxically, require diversification (sampling) of interests earlier in life.[30]

---

[28] We used the 2017 round of the Development Bank for Latin America (CAF) annual household survey of sociodemographic information (N=10,687), which is representative of the adult population of major cities in Latin America. The 2017 round added the Short Grit Scale to the questionnaire (CAF, 2017).

[29] See, for example, Sturman and Zappala-Piemme (2017).

[30] See, for example, Gopnik (2020).

In sum, we find that the intervention induced students in both treatment groups to update their beliefs about deliberate practice, relative to students in the control group, with stronger effects when teachers delivered our curriculum as opposed to student self-paced learning. Impacts on grit were mixed: while students in either treatment group showed more perseverance-of-effort aspects relative to students in the control group, impacts on consistency-of-interest aspects were negative, pointing towards a potential unintended consequence, in particular the possibility of crowding out longer-term goal-orientation, of ours and other similar interventions.

Index of Socio-Emotional Skills

As discussed above, we collected survey data on various other measures of socio-emotional skills. As all of them focus on some element of self-regulation, in our main analysis, we look at a single index ('S/E Skills Index') that combines the measures of deliberate practice beliefs, grit, frustration reaction, motivational frameworks, locus of control, and present bias.[31] Note that our S/E Skills Index reduces our sample size considerably, as it requires non-missing values for all its constituent elements (i.e. scales and sub-scales). In Section V.C, we show that the results presented are robust to survey non-response and attrition, by conducting formal tests for (differential) attrition by experimental group and by using multiple imputation techniques.

Table IV shows the impacts of the intervention on the S/E Skills Index on average (Column 1), when excluding deliberate practice beliefs and grit (but including all other skills) (Column 2), and when including only deliberate practice beliefs and grit (excluding all others) (Column 3).

[TABLE IV ABOUT HERE]

The intervention had a significant, positive impact on the S/E Skills Index on average (Column 1), with stronger impacts in the teacher-delivery treatment (about 13% SD) compared to student self-paced learning (about 6% SD). Disaggregating the index into its different sets of skills we find that, while all

[31] Appendix A Table A1.1 replicates Table 3 for frustration reaction, motivational frameworks, locus of control, and present bias as separate socio-emotional skills outcomes. Table A1.2 replicates Table 3 for the sample used for our S/E Skills Index. Our previous findings continue to hold for this sample.

skills do play some role in the teacher-delivery treatment (Column 2), deliberate practice beliefs and grit are clearly the driving force behind the impacts shown in the first column.

Next, we look at heterogeneous impacts of the intervention. Table V Panel A shows the impact of the intervention on the S/E Skills Index on average (Column 1) and for different sub-groups defined by gender (Columns 2 and 3), ethnicity (Columns 4 to 7), academic year (Columns 8 and 9), and tercile in the pre-treatment S/E Skills Index distribution, whereby the first tercile is the lower and the third tercile the upper tail (Columns 10 to 12)There are many hypotheses being tested in this table. We show below that our main results are robust to accounting for multiple hypothesis testing.

[TABLE V ABOUT HERE]

When it comes to heterogeneous impacts by gender, the intervention had positive impacts for girls and for boys (especially when teachers delivered the contents), but they were clearly much stronger for girls. Moreover, impacts were stronger for Macedonian students (the largest ethnic group in the country) than for Albanian students (the second largest group). They turn out particularly strong for Roma students (the most disadvantaged group), for whom the impact of the teacher-delivery arm is about 42% SD (Column 6). Although the sub-sample of Roma students is much smaller than those of the other groups, and the resulting standard errors are larger (so point estimates should be taken with caution), the impact is still statistically significant. Columns 8 and 9 show that the intervention had similar impacts on sixth and seventh-grade students (especially in the teacher-delivery arm), whereas Columns 10 and 12 show that impacts were somewhat larger for students who already a had higher level of S/E skills to begin with (i.e. those in the third tercile of the pre-treatment skills distribution). We cannot reject that there are no statistically significant differences between the different groups reported in Columns 2 to 12.

In sum, when looking at all socio-emotional skills together, we find that the intervention had positive impacts, for all students on average, and for particular sub-groups. Estimated impacts tend to be stronger

for students who are at a relative disadvantage (i.e. girls and Roma students, in our case). Finally, the teacher-delivery arm consistently produced better results.[32]

The Roma are a particularly disadvantaged group in the Western Balkans. In North Macedonia, ethnicity, which also determines language and religion, is a socially dividing line and a frequent cause of social conflict. The Roma (who make up only 3% of the population) are socially and economically marginalized, and subject to discrimination in many areas of everyday life (Robayo-Abril and Millan, 2019; Gatti et al., 2016). Their children typically lag far behind Macedonian and Albanian children in terms of academic achievement. By providing positive role models and using an inclusive language directly aimed at providing counter-stereotypical examples, with a distinct focus on ethnic minorities, we targeted Roma students in particular. We cannot however rule out the possibility that the teacher-delivery arm, in which teachers received training on the principles behind our curriculum, has led to a more effective delivery of contents particularly to Roma students. Whatever the exact mechanism, the intervention seems to have reduced educational inequalities across ethnic lines. At the same time, our finding that students at the upper end of the pre-treatment skills distribution benefit somewhat more suggests that reducing educational inequalities more broadly may require a multi-faceted intervention.

V.B. Impacts on GPAs

Short-Term GPAs

Did improvements in students' socio-emotional skills translate into improvements in their academic achievement? Table V Panel B shows impacts on short-term GPAs, calculated immediately after the intervention in the fourth quarter across math, English, and language. Note that our sample size is much larger for GPAs than for our socio-emotional skills outcomes reported so far, as GPAs are obtained from official, administrative records and are therefore much less subject to missing information.

Overall, we find little evidence for impacts on short-term GPAs. If anything, there is an improvement in short-term GPAs of about 2% SD on average, which is only marginally significant at

[32] In an exploratory analysis in Appendix A Table A1.3, we replicate our results in Table 5 Panel A using a modified S/E Skills Index that excludes all reverse-coded items in each scale. Our findings continue to hold and, if anything, become more pronounced in terms of effect size.

the 10% level, for the student self-learning treatment (yet with a similar point estimate and standard error in the teacher-delivery treatment). Likewise, there is little evidence for differential impacts by student sub-group: impacts are similar between gender, academic year, and students in different terciles of the pre-treatment GPA distribution. Interestingly though, we do find a larger impact on Roma students in the teacher-delivery treatment (about 6% SD). Roma students are also the sub-group of students for whom we find the strongest impacts on the S/E Skills Index (about 42% SD, cf. Table 5 Panel A Column 6). There is thus consistency between our finding for the S/E Skills Index and our finding for GPAs of Roma students. The effect size for GPAs, however, is rather small.

Note that the finding of no short-term impacts on GPAs for almost all groups could also be explained by teachers grading on a curve. However, according to official sources (and personal communication with local teachers and educators), grading on a curve is not a common practice in North Macedonia.

Longer-Term GPAs

Since schooling in North Macedonia is compulsory up to (and including) grade nine, we are able to track the GPAs of students in our two treatment groups and our control group over time. Table VI looks at medium-term GPAs (i.e. half a year later, in the first semester of the school year 2016/2017) and long-term GPAs (i.e. one year later, in the second semester).

We find evidence that impacts of the intervention on GPAs become stronger over time, on average as well as for different student sub-groups. More specifically, impacts become statistically significant one year after the intervention, with effect sizes still being small, hovering between 3% SD and 6% SD depending on sub-group. It should be noted that the pattern of impacts for longer-term GPAs differs from that for the S/E Skills Index: small impacts on GPAs seem to materialize over time for males (rather than females) and for students at the lower end of the pre-treatment GPA distribution (rather than at the upper end). A notable exception are Roma students, for whom we find a consistent gradient in GPA improvement over time, which is again stronger in the teacher-delivery than in the student self-learning treatment. Here, medium-term and long-term GPAs increase by about 17% and 28% SD, respectively, up from about 6% SD immediately after the intervention. Figure I shows this improvement in GPAs for Roma students over time.

[TABLE VI ABOUT HERE]

[FIGURE I ABOUT HERE]

In theory, we would not expect materializing impacts on GPAs over time to be driven by attrition (out-of-sample selection), with only high-achieving students remaining in the sample and lower-achieving ones dropping out of formal schooling. This is because, as we have just discussed, schooling in North Macedonia is compulsory up to (and including) grade nine. However, in a robustness check in which we regress the likelihood of having a medium or long-term GPA record in the official, administrative data on ethnicity, amongst others, we find that Roma students are between seven and ten percentage points *less* likely than Macedonian students to be recorded half a year and one year after the intervention had ended (Appendix C Table A3.2), although this does not differ much between the control and any of the two treatment arms (if anything, Roma students are *more* likely to have a record for the long-term GPA in the teacher-delivery arm).

To formally check whether our findings for Roma students continue to hold when accounting for attrition, we re-estimate Table V using a balanced panel, including only those students who are observable in the official, administrative data during the entire observation period from school year 2015/2016 to school year 2016/2017 (Appendix B Table A2.1). We find that our results (especially those for Roma students) remain robust using this specification, suggesting that attrition and resulting changes in sample composition – although a real phenomenon – are unlikely to drive the impacts on GPAs over time. We turn to a more detailed discussion on survey non-response and attrition, including sensitivity analyses and robustness checks, in Section V.C.

In an exploratory analysis (Appendix B Table A2.2), we also document that, while both male and female Roma students benefited from the intervention, females benefited more in the long-term. However, we cannot reject that the difference in impacts on the GPAs of male and female Roma students is equal to zero. When it comes heterogeneous impacts by pre-treatment GPAs, we find that Roma students in higher terciles of the pre-treatment GPA distribution benefited more; in fact, for students in the upper tercile, the intervention had a positive effect on GPAs, increasing these scores by about 8% SD in the short-term, 21% in the medium-term, and 46% in the long-term when teachers delivered the

intervention (effects are also significant, although lower, for student self-paced learning). Finally, when estimating seemingly unrelated regressions (SURs) for all subgroups Roma students (on average, by gender and by pre-treatment GPAs), we can reject the null that the coefficients of the teacher-delivery treatment are jointly equal to zero across all models.[33]

In sum, we find a consistent pattern for Roma Students: large impacts on socio-emotional skills are mirrored by very small impacts on GPAs immediately after the intervention, which become larger half a year later and substantial one year after. Such a pattern of emerging impacts over time might point towards sustained behavior change and suggests that students reap the benefits of such behavior change over time only, at least when it comes to academic achievement as reflected in GPAs. This pattern echoes findings from other social-psychological interventions in education that show stronger benefits for more disadvantaged students (Yeager et al., 2019; Sisk et al., 2018; Paunesku et al., 2015; Cohen et al., 2009; Hulleman and Harackiewicz, 2009; Wilson and Linville, 1982).

V.C. Robustness

Survey Non-Response and Attrition

So far, our analysis of socio-emotional skills includes all students with completed surveys, whereas our analysis of GPAs includes all students who are present in official records (and should thus have participated in the intervention according to the randomization routine). This approach uses all observations that are available for each outcome, in order to maximize sample size.

Yet, this approach results in a different number of students in our analysis of socio-emotional skills that in our analysis of GPAs, the difference arising from survey non-response and attrition. There are various reasons for these. In what follows, we discuss survey non-response and attrition in both administrative and survey data and show that our results are robust to corrections for attrition.

[33] A more complex issue arises in case that our intervention increased motivation and effort on side of teachers such that teachers in any of the two treatment arms (and potentially differently by treatment arm) recorded GPAs more swiftly post-treatment. Graphical evidence, however, suggests that the temporal distribution of when GPAs were recorded is similar between any of the two treatment groups and the control group. For a detailed discussion, see Section 2.

Administrative Data. All teachers in North Macedonia, from primary to secondary school, are required to record electronically all grades of students, across all subjects (in a so-called "Electronic Gradebook"). Grades should be entered within two weeks of the marking of the assessment. Therefore, it is not surprising that there is little attrition in our administrative data.

Officially, there were 35,340 sixth and seventh-grade students who were randomised to be part of our experiment in the school year 2015/2016. There are 33,454 students (95% of the total) for whom we can study short-term impacts (on GPAs in the academic quarter right after the intervention). Moreover, 31,310 students (88% of the total) have data available to study medium-term impacts (on GPAs in the first semester of the following school year). Finally, 31,437 students (89% of the total) have data available to study long-term impacts (on GPAs in the second semester of the following school year). The small differences in the number of schools and students between different aggregation periods suggest that not all teachers strictly abide to the formal requirement to record all grades within two weeks, although the vast majority certainly complies with it. There may also be some natural fluctuation of students genuinely entering and leaving the formal education system in North Macedonia (e.g. migration).

Survey Data. In contrast to our administrative data, attrition in our survey data is much more pronounced. The intervention was nationwide, directly scaled-up to cover all primary schools in the entire country. Due to missing IT infrastructure in most schools and classrooms, our surveys were paper-based and printed ahead of data collection. They were delivered to schools together with the intervention materials ahead of the intervention start date.

About 700,000 pages of intervention materials had to be printed in a time span of only a few weeks. To accomplish this, printing was divided between three large local printing companies, each of which printed complete packages (i.e. baseline surveys, endline surveys, workbooks, name lists of students, and stickers with student IDs) for a subset of schools defined by treatment arm. Printing and packaging were supervised to ensure sufficient materials were printed, and that they were appropriately packaged and labelled. Materials were then delivered to schools two weeks prior to the intervention start date in packages corresponding to each class. At about the same time, teachers received written instructions in

a separate letter posted to them (and e-mailed to the principal), explaining how to distribute surveys and how to have them completed by students. Instructions were also included in print in the survey packages.

When teachers had students complete surveys (baseline survey: one week before the intervention start, endline survey: one week after the intervention had ended), they were instructed to use stickers with student IDs to be attached to surveys. Student IDs are used as unique person identifiers to merge baseline with endline surveys, as well as with the administrative data. Once completed, surveys were stored by the school in the same boxes they were originally delivered in and kept for collection. This process was the same for all intervention materials. Both baseline and endline surveys were collected from schools by our local survey company starting from two weeks after the intervention had ended. All boxes were stored at a central warehouse owned by the survey company, and then successively digitalised over the next few months by the company's data entry team.

Given the nationwide scale and the highly complex logistical nature of our survey data collection, attrition could have occurred at several points of the field work, due to several reasons. We identify the most important sources to be: *(i)* boxes of baseline and/or endline surveys went missing within schools prior to baseline and/or endline survey completion or by the time of boxes collection; *(ii)* boxes of baseline and/or endline surveys were returned empty; *(iii)* some baseline and/or endline surveys were filled out but no stickers with ID were attached to them, in which case it is impossible to uniquely identify a student. Apart from these, there was attrition of a more typical nature, in the sense that some responses were invalid or illegible.

Our analysis of socio-emotional skills uses the sample of students for whom we have both completed baseline and endline surveys. Students can thus drop out of our analysis if they have either a missing endline survey, a missing baseline survey, or both. Starting with the latter, about 2,000 students or 6% (out of a total of 35,340 students) have neither a baseline nor an endline survey, and are thus omitted from our analysis. About 31,000 (88% of the total) have a valid baseline survey. There are some differences in completion rates between scales, e.g. 31,544 observations for deliberate practice beliefs compared to 29,487 for grit at baseline. About 27,000 students (76% of the total) have a valid endline survey (again, with slight differences between items, e.g. 27,161 for deliberate practice beliefs compared

to 25,815 for grit at endline). This implies an attrition from baseline to endline of about 4,000 students or 11%. Finally, there are some students who have completed an endline survey without having completed a baseline survey: about 3,000 students or 8%.

Our analysis of socio-emotional skills, therefore, uses a sample of about 24,000 students (68% of the total) for whom we have completed both baseline and endline surveys, so about 32% are missing. What is more, our sample reduces to 18,718 students (53% of the total) when using our S/E Skills Index, which combines all socio-emotional skills outcomes. Note that outcomes are multi-item summed scales that become missing if (at least) one of their items becomes missing (even if the others do not); generally, the more items, the greater the chance that the entire scale has a missing value. The S/E Skills Index is particularly susceptive to this. In what follows, we test for (differential) attrition by experimental group, while replicating our baseline results using a sample that is balanced between our survey and administrative data, as well as multiple imputation techniques to correct for attrition.

Appendix C Table A3.1 shows the results of a regression of a binary indicator for the availability of the different survey instruments (i.e. having a baseline or an endline survey, or both) on indicators for being in one of our two treatment groups, pre-treatment GPAs, and demographic characteristics (Columns 1, 3, and 5). Moreover, Columns 2, 4, and 6 additionally interact the indicators for the two treatment groups with these other covariates. Table A3.2 runs the same regression for the availability of GPAs over time (i.e. having a medium-term or long-term GPA, or both). Our aim is to study which patterns of missing data are correlated with our two treatment groups and with these other covariates (i.e. attrition), and whether these patterns differ by experimental group (i.e. differential attrition).

When it comes to the different survey instruments (Table A3.1), having better pre-treatment GPAs is associated with a higher likelihood of completion of the survey. However, effects are small: increasing pre-treatment GPAs by 1% SD increases the likelihood to have a completed baseline or endline survey, or both, by between one and two percentage points. Moreover, there are notable differences in attrition between our first (self-learning) and second (teacher delivery) treatment group, as well as between our first treatment group and our control group.

In particular, our first treatment group is less likely to have a valid endline survey by about 15%. We have however determined this to be related to a printing error: one of the printing firms printed the endline survey twice (instead of both baseline and endline survey) for about one third of our first treatment group, most likely due to the first page of both surveys looking the same. This is also the reason why students' socio-economic characteristics are missing (cf. Table I), as these were captured at the end of the baseline survey. This is likely to cause primarily random noise (as opposed to systematic bias) due to a reduced sample size, since the schools which received the wrongly printed surveys were a random set of schools.

Finally, we find that, compared to being Macedonian, being Albanian or Roma is also associated with a (significantly) lower likelihood of having a completed baseline or endline survey, or both. However, when looking at the interactions between the dummies for our two treatment groups and these other covariates, we find little evidence for differential attrition by experimental group (if anything, Roma students are significantly *more* likely to have a valid endline survey in our first treatment group). When it comes to the availability of GPAs over time (Table A3.2), we obtain broadly similar results.[34]

To examine the sensitivity of our previous results for attrition, in Appendix C Table A3.3, we re-estimate Table V by including only those students for whom we have both completed baseline and endline surveys and for whom we know for sure that they participated in the intervention, i.e. a balanced sample between survey and administrative data. The results are very similar to our previous results, and the estimates obtained from using such a balanced sample are indeed even larger in size than those obtained from using an unbalanced sample. Note that, for Roma students, impacts on short-term GPAs become twice as large but also more imprecise, most likely because the sample size drops substantially. We have already shown in Appendix B Table A2.1 that materializing impacts on GPAs for Roma students remain the same when using balanced panel, i.e. including only those students for whom we have short-term, medium-term, and long-term GPAs, to account for attrition.

[34] Note that, relative to our control group, being in any treatment group is associated with a higher likelihood of having a long-term GPA by between four and five percentage points. This may suggest that being exposed to treatment may actually reduce the likelihood to drop out of formal schooling.

Finally, we use a multiple imputation procedure to correct for attrition in both survey and administrative data. More specifically, we impute each post-treatment outcome (i.e. deliberate practice beliefs, grit and grit facets, the S/E Skills Index, as well as short-term, medium-term, and long-term GPAs) using students' pre-treatment GPAs, age, gender, ethnicity, and academic year. We use a multivariate normal regression and an iterative Markov chain Monte Carlo (MCMC) approach to simulate our data. Appendix C Tables A3.4 to A3.9 replicate our previous results from Tables III to VI using this imputation procedure. As seen, our results remain qualitatively the same, the only exception being the S/E Skills Index of Roma students in the teacher-delivery treatment. Note, however, that this result was based only on a very small number of observations. The impacts on Roma students' short-term, medium-term, and long-term GPAs remain the same.

Observer (Hawthorne) and Experimenter-Demand Effects

It could be that some, if not all, of the positive impacts of the intervention were due to observer (Hawthorne) or experimenter-demand effects. Teachers and students, both of whom were not blind to the experiment, may have changed their behavior as a result of being part of the intervention, rather than due to the actual contents being taught in it.

Although we cannot fully exclude that observer effects may have played a role for either teachers or students, we argue that they are unlikely to be the main driver behind our findings. When it comes to students, the intervention was embedded into an existing "Life Skills" curriculum which was already implemented in schools. Students were familiar with lessons from this curriculum being taught during Monday morning class hours (yet without any expectation of upcoming content), and we made sure that the intervention resembled its basic structure and appeal (in fact, the local psychologist working on the adaptation of the intervention to the local context also designed this "Life Skills" curriculum). Moreover, our curriculum was rather light, in the sense that it consisted only of five content sessions. Finally, baseline and endline surveys were collected before and after these sessions, with a timely spacing (i.e. one week before intervention start and one week after intervention end), to avoid any bias arising from the exposure to the material itself. There were never experimenters nor intervention facilitators present at any point during the intervention. Students were not monitored either, neither within sessions nor

outside. Hence, the intervention should not have been particularly salient amongst students, thereby minimizing potential observer or experimenter-demand effects.

When it comes to teachers, such effects are particularly relevant in the teacher-delivery treatment and for impacts on GPAs, as the role of teachers in the self-learning treatment was minimal. This begs the question of whether the teacher-delivery treatment was indeed more effective, either because *(i)* students simply absorbed contents more or more effectively, *(ii)* teachers themselves reflected on these contents and treated students in improved ways, or *(iii)* teachers – weary of being part of an experiment – simply changed their grading behavior. While *(i)* and *(ii)* are arguably part of the genuine treatment of the intervention, *(iii*) is an experimental artefact. Again, although we cannot exclude *(iii)* with certainty, we argue that it is unlikely to be the main driver behind our findings. The "treated" teachers were the headteachers of the respective class, and they typically teach only one subject, not necessarily corresponding to the subjects we used to construct GPAs. Moreover, we constructed GPAs across several subjects taught by multiple teachers (i.e. math, English, and language), thereby reducing the relative importance of a single subject and hence of experimenter-demand effects. When re-calculating impacts for GPAs taken across *all* subjects (not only math, English, or language), we obtain similar results as our previous ones.[35]

Replication Using Pre-Registered Controls Only, Parametric Bootstrap

For our baseline results, we ran two sets of additional sensitivity analyses. The first re-estimated our previous models using pre-registered controls only (Appendix C Tables A3.10 and A3.12). The second fitted these models with a mixed-effects model structure incorporating parametric bootstrap estimates and using, likewise, pre-registered controls only (Tables A3.11 and A3.13). This mixed-effects model is described in detail in Appendix C.

For the short-term impacts, the sensitivity analysis using the pre-registered controls only confirms our previous findings obtained with our extended set of controls. The parametric bootstrap modeling shows similar results for all outcomes, except GPAs, where bootstrap estimates are hovering around

[35] These results are available upon request.

zero impacts. Note that the parametric bootstrap modeling uses the residual variance and the predicted estimate from a mixed-effects model. As GPAs do not have much variance, this reduces the likelihood of the iterations of the mixed model to *not* produce zero-variance estimates or to converge, which is a likely reason behind the null effects. Taken together, however, our re-analysis using pre-registered controls only and using a different modeling approach with bootstraps largely confirms our previous findings.

Multiple Hypotheses Testing

We examined the robustness of our previous results regarding multiple hypotheses testing using the stepwise p-value correction by Romano and Wolf (2005, 2016). Accounting for the 24 hypotheses in our socio-emotional skills analysis and the 28 hypotheses in our GPA analysis (which include medium-term and long-term impacts, which we only have for GPAs), we find that, for our socio-emotional skills, the following estimates remain statistically significantly different from zero (at least) at the 10% level: *(i)* the average impact of the teacher-delivery treatment, *(ii)* the impact of the same treatment for females, *(iii)* the impact of both the self-learning and the teacher-delivery treatments on seventh-grade students, and *(iv)* the impact of the teacher-delivery treatment on students in the second tercile of the pre-treatment S/E Skills Index distribution. Importantly, when it comes to GPAs, the long-term impact on Roma students one year post-treatment remains significant at the 5% level.

VI. DISCUSSION AND CONCLUSION

To the best of our knowledge, we are the first to rigorously evaluate a nationwide, school-delivered intervention aimed at fostering grit (the ability to sustain effort and interest towards long-term goals). The intervention is implemented amongst middle-school students in North Macedonia. A five-week program was delivered through regular channels and the regular school curriculum during the second semester of the 2015-2016 school year.

The evaluation of this intervention's impacts relies on a multi-arm cluster-randomized controlled trial. In one treatment group of schools, lessons were delivered through student self-paced learning, while in another, lessons were teacher-delivered. There was also a control group where students did not

receive the grit curriculum. The lessons focused on deliberate practice and – by changing students' beliefs about the expectancies and values of engaging in practice – motivated students to take up this particular type of practice. This, in turn, raises the chances of attaining better learning outcomes, which then reinforces beliefs and, thereby, leads to sustained behavior change, with positive long-term benefits. The intervention also countered negative stereotypes surrounding gender and ethnicity in North Macedonia and the region as a whole, by providing positive role models and counter-stereotypical examples throughout the intervention materials.

We found that the intervention significantly increased socio-emotional skills and, to some extent, academic achievement amongst exposed students, particularly the more disadvantaged Roma students. In terms of socio-emotional skills, both treatments had significant positive impacts, with higher impacts found when contents were delivered by teachers. In this latter case, the average student experienced an improvement in socio-emotional skills of about 13% SD, while Roma students saw an improvement of up to 42% SD. In terms of GPAs, impacts were measured at three stages: short-term (in the quarter right after the intervention), medium-term (in the first semester of the following academic year), and long-term (in the second semester of the following academic year, i.e. one year after completing the program). Across all three periods, disadvantaged Roma students in the teacher-delivery treatment experienced most gains in terms of GPAs (up to 28% SD in the long-term). Changes in GPAs were significant and positive across all time periods, with impacts roughly doubling every semester. Back-of-the-envelope, these achievement gains are roughly equivalent to the gains normally associated with three weeks of additional instruction time. Impacts on GPAs of Roma students mirror those on socio-emotional skills for this group, and they remain robust to using a balanced panel, using multiple imputation techniques, and when accounting for multiple hypotheses testing.

However, we also find that, while the intervention increased the perseverance-of-effort facet of grit, it reduced its consistency-of-interest facet. Compared to students in the control group, treated students reported to be more perseverant and hard-working, but also to more quickly lose their interest for longer-term goals. Both facets combined then yielded an insignificant (teacher-delivery) or even significant negative impact (self-learning) of the intervention on grit as a whole. One plausible explanation is that, by teaching students the value of deliberate practice applied to a more immediate stretch goal, our

intervention reduced their interest in longer-term goals, considering that our intervention was not aimed at nurturing a specific interest. Developing consistency of interest amongst young children may, paradoxically, require diversification (sampling) of interests earlier in life (Gopnik, 2020; Cote and Erickson, 2015). This is an important area that requires further empirical research.

As a whole, though, our findings confirm not only that it is possible to effectively teach grit amongst students across the education system but also that, by doing so, this can have positive impacts on academic achievement, possibly with impacts increasing over time. These impacts are particularly high amongst disadvantaged students, which further indicates the potential for this type of intervention to support school learning in ways that may improve equity in educational outcomes, though heterogeneous impacts by pre-treatment academic achievement (where higher-performing students benefited relatively more from our intervention) suggest that improving equity in education may require a more multi-faceted intervention. The magnitude of impacts found in our intervention compares favorably to other educational interventions focused on improving socio-emotional skills and are consistent with a mounting body of evidence that grit and growth mindset interventions often benefit disproportionally (and sometimes only) disadvantaged students. In terms of GPAs, our impacts on Roma students are comparable with those found in a recent meta-analysis of socio-emotional skills interventions amongst disadvantaged groups (34% SD, cf. Sisk et al., 2018), and are higher than impacts amongst disadvantaged students in other programs: 18% SD in a standardized test in math in Indonesia, where an expanded program showed no impacts on GPAs (Johnson et al., 2020; World Bank, 2019); 10% SD in Peru's "Expande Tu Mente" program (Outes et al., 2020). Alan and Ertac (2019) find average effects of 23% SD in a standardized test in math 2.5 years after a similar intervention that targeted grit in participating schools in Istanbul, Turkey.

Our intervention was very cost-effective. When looking only at Roma students for whom we find significant positive impacts on GPAs one year post-treatment (the common cost-effectiveness ratios in educational economics are annual GPAs per USD spent) and allocating costs accordingly, we find that, when including all cost categories, our intervention cost about 3.7 USD for a 0.1 SD increase in annual GPAs of Roma students. Excluding costs of design and evaluation, this translates into about 1 USD per

0.1 SD increase.[36] Compared to the literature, these are very favourable ratios. Glewwe and Muralidharan (2016) find that incentive schemes (for both students and teachers) cost between 1 and 3 USD per 0.1 SD improvement in test scores, CCTs between 77 USD and 138 USD for a comparable improvement, and pedagogy-supporting classroom-IT about 30 USD per 0.1 SD improvement. These results are illustrative only, as implementation costs are likely to vary considerably by country, as do opportunity costs of education.

While the results from our intervention are promising, our identified impacts and their heterogeneity across treatments and students suggest that important questions remain in terms of how to foster grit amongst students. In particular, we were not able to identify the exact mechanisms that drive our results. Albeit not conclusively, we examined some potential avenues for impact. Gritty students do more deliberate practice and work harder for longer periods of time in order to achieve their goals, and both treatments worked positively in this regard. However, our intervention also addressed issues of self-efficacy and stereotype threat by providing positive role models and counter-stereotypical examples, which would also be consistent with the higher impacts found amongst Roma students. Arguably, however, such role models and examples (i.e. success despite adversity) could also serve as an inspiration for the average student who may be of a different ethnicity. Our paper also highlighted the importance of delivery mechanisms, and suggests that more intense methods – particularly those involving teachers or other individuals to deliver the contents – may be more impactful while still cost-effective at scale (consistent with Alan and Ertac, 2019).

In terms of grit in particular, we identified impacts on its perseverance-of-effort facet but none or even negative impacts on its consistency-of-interest facet, suggesting that – in line with recent research (cf. Gopnik, 2020; Cote and Erickson, 2015) – targeting consistency of interest may not be an effective way of fostering grit amongst younger students. In fact, we show that our intervention may have even crowded out interest – a potentially unintended consequence of the particular behaviors our and similar

[36] The total costs of our intervention were USD 343,616. With a total of 34,454 students in our analysis of GPAs, this yields a cost per student of about USD 10.3 (including all costs) or USD 2.7 when excluding costs of design and evaluation. There are 1,161 Roma students in total. Hence, we obtain a cost-effectiveness ratio of (10.3 x 0.1) / 0.28 = 3.7 USD per 0.1 SD improvement in annual GPAs of Roma students when including all costs or (2.7.3 x 0.1) / 0.28 = 1 USD per 0.1 SD improvement when excluding costs of design and evaluation.

interventions try to engrain. While this may be a point that is unique to deliberate practice and our intervention more generally, it does stress the importance of measuring outcomes of interventions more broadly to detect potentially negative behavioral spillovers. Finally, and more generally, while our paper contributes to better understanding the potential of fostering grit amongst students at scale, further work is still needed to distill what is the most effective combination of socio-emotional skills for the educational needs of different students.

THE WORLD BANK, UNITED STATES

SS. CYRIL AND METHODIUS UNIVERSITY, SKOPJE, AND CENTER FOR HUMAN RIGHTS AND CONFLICT RESOLUTION, SKOPJE, NORTH MACEDONIA

UNIVERSITY COLLEGE LONDON, CEMMAP, IFS, FAIR-NHH, UNITED KINGDOM

KELLOGG SCHOOL OF MANAGEMENT AT NORTHWESTERN UNIVERSITY, UNITED STATES

THE WORLD BANK, UNITED STATES

CEDLAS-UNIVERSIDAD NACIONAL DE LA PLATA, ARGENTINA

LONDON SCHOOL OF ECONOMICS, CEP, UNITED KINGDOM

THE WORLD BANK, UNITED STATES

UNIVERSITY OF PENNSYLVANIA, UNITED STATES

SUPPLEMENTARY MATERIAL

Please see Appendix in a separate file

APPENDIX

Appendix A: Other Socio-Emotional Skills Outcomes

Appendix B: Impacts on GPAs Over Time

Appendix C: Robustness Checks

Appendix A: Other Socio-Emotional Skills Outcomes

Table A1.1: Impacts on Other Socio-Emotional Skills (Z-Scores)

	Frustration Reaction	Motivational Frameworks	External Locus of Control	Present Bias (Reverse Coded)
	(1)	(2)	(3)	(4)
Treatment 1	-0.034	-0.009	-0.040**	-0.023
"Self-Learning"	*(0.023)*	*(0.021)*	*(0.019)*	*(0.020)*
Treatment 2	0.003	0.063***	-0.001	0.040**
"Teacher Delivery"	*(0.021)*	*(0.020)*	*(0.017)*	*(0.020)*
N	23,832	23,909	23,724	23,335
N Control	9,277	9,378	9,238	9,174
N Treatment 1	6,953	6,890	6,921	6,745
N Treatment 2	7,602	7,641	7,565	7,416
R^2	*0.210*	*0.366*	*0.292*	*0.200*

Notes: All regressions control for pre-treatment outcomes, demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.2: Impacts on Deliberate Practice Beliefs, Grit, and Grit Facets,

Sample Used for S/E Skills Index (Z-Scores)

	Deliberate Practice Beliefs	Grit	Grit: Effort	Grit: Interest
	(1)	(2)	(3)	(4)
Treatment 1	0.162***	-0.064***	0.051**	-0.127***
"Self-Learning"	*(0.019)*	*(0.020)*	*(0.021)*	*(0.019)*
Treatment 2	0.236***	-0.032	0.062***	-0.094***
"Teacher Delivery"	*(0.018)*	*(0.022)*	*(0.019)*	*(0.022)*
N	18,718	18,718	18,718	18,718
N Control	7,286	7,286	7,286	7,286
N Treatment 1	5,424	5,424	5,424	5,424
N Treatment 2	6,008	6,008	6,008	6,008
R^2	*0.324*	*0.344*	*0.335*	*0.238*

Notes: All regressions control for pre-treatment outcomes, demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A1.3: Exclusion of Non-Reverse Items in S/E Skills Index (Z-Scores)

Outcome: Modified S/E Skills Index

		Gender		Ethnicity				Academic Year		Pre-Treatment Outcome		
	Average	Male	Female	Macedonian	Albanian	Roma	Other	Sixth Grade	Seventh Grade	Tercile 1 (Low)	Tercile 2 (Middle)	Tercile 3 (High)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Treatment 1	0.117***	0.107***	0.129***	0.135***	0.106***	-0.100	0.022	0.108***	0.120***	0.080**	0.120***	0.148***
"Self-Learning"	*(0.020)*	*(0.027)*	*(0.022)*	*(0.021)*	*(0.039)*	*(0.156)*	*(0.103)*	*(0.026)*	*(0.026)*	*(0.034)*	*(0.028)*	*(0.025)*
Treatment 2	0.171***	0.164***	0.178***	0.189***	0.098***	0.245*	0.220*	0.188***	0.154***	0.120***	0.196***	0.180***
"Teacher Delivery"	*(0.019)*	*(0.026)*	*(0.023)*	*(0.023)*	*(0.033)*	*(0.146)*	*(0.117)*	*(0.027)*	*(0.024)*	*(0.035)*	*(0.028)*	*(0.025)*
N	20,059	9,801	10,258	13,835	4,917	389	918	9,636	10,423	5,228	6,988	7,739
N Control	9,451	3,876	3,940	5,175	2,207	134	300	3,767	4,049	2,018	2,767	3,012
N Treatment 1	7,068	2,817	2,978	4,161	1,216	121	297	2,779	3,016	1,611	2,007	2,157
N Treatment 2	7,757	3,108	3,340	4,499	1,494	134	321	3,090	3,358	1,599	2,214	2,570
R^2	*0.373*	*0.354*	*0.365*	*0.392*	*0.308*	*0.389*	*0.472*	*0.369*	*0.384*	*0.305*	*0.335*	*0.391*

Notes: All regressions control for pre-treatment outcomes, demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Appendix B: Impacts on GPAs Over Time

Table A2.1: Impacts on GPAs Over Time, Balanced Panel (Z-Scores)

		Gender		Ethnicity				Academic Year		Pre-Treatment Achievement		
	Average	Male	Female	Macedonian	Albanian	Roma	Other	Sixth Grade	Seventh Grade	First Tercile	Second Tercile	Third Tercile
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: GPAs, Short-Term (2015/2016 Q4)												
Treatment 1	0.015	0.016	0.014	0.019	0.010	-0.008	0.006	0.016	0.013	0.006	0.026*	0.010
"Self-Learning"	*(0.011)*	*(0.011)*	*(0.012)*	*(0.013)*	*(0.019)*	*(0.025)*	*(0.020)*	*(0.016)*	*(0.013)*	*(0.013)*	*(0.015)*	*(0.008)*
Treatment 2	0.009	0.005	0.012	0.006	0.006	0.039**	-0.008	0.011	0.006	0.011	0.005	0.008
"Teacher Delivery"	*(0.012)*	*(0.013)*	*(0.012)*	*(0.015)*	*(0.019)*	*(0.016)*	*(0.024)*	*(0.015)*	*(0.015)*	*(0.014)*	*(0.018)*	*(0.008)*
N	29,303	15,135	14,168	17,851	9,253	928	1,271	14,757	14,546	9,723	9,782	9,798
N Control	10,607	5,480	5,127	6,304	3,554	340	409	5,497	5,110	3,350	3,622	3,635
N Treatment 1	9,716	5,034	4,682	6,072	2,932	225	487	4,793	4,923	3,305	3,230	3,181
N Treatment 2	8,980	4,621	4,359	5,475	2,767	363	375	4,467	4,513	3,068	2,930	2,982
R^2	*0.936*	*0.934*	*0.930*	*0.930*	*0.927*	*0.919*	*0.954*	*0.931*	*0.943*	*0.675*	*0.597*	*0.508*
Panel B: GPAs, Medium-Term (2016/2017 Q1+Q2)												
Treatment 1	0.004	0.011	-0.004	0.016	-0.008	0.103**	0.059	0.018	-0.014	0.004	0.013	-0.007
"Self-Learning"	*(0.013)*	*(0.014)*	*(0.015)*	*(0.016)*	*(0.027*	*(0.047)*	*(0.036)*	*(0.018)*	*(0.016)*	*(0.017)*	*(0.020)*	*(0.013)*
Treatment 2	0.008	0.020	-0.006	0.019	0.000	0.167***	0.023	0.019	-0.004	0.030*	0.013	-0.016
"Teacher Delivery"	*(0.013)*	*(0.015)*	*(0.014)*	*(0.017)*	*(0.024)*	*(0.061)*	*(0.043)*	*(0.018)*	*(0.017)*	*(0.018)*	*(0.020)*	*(0.012)*

N	29,303	15,135	14,168	17,851	9,253	928	1,271	14,757	14,546	9,723	9,782	9,798
N Control	10,607	5,480	5,127	6,304	3,554	340	409	5,497	5,110	3,350	3,622	3,635
N Treatment 1	9,716	5,034	4,682	6,072	2,932	225	487	4,793	4,923	3,305	3,230	3,181
N Treatment 2	8,980	4,621	4,359	5,475	2,767	363	375	4,467	4,513	3,068	2,930	2,982
R^2	*0.878*	*0.872*	*0.869*	*0.875*	*0.851*	*0.805*	*0.907*	*0.873*	*0.887*	*0.508*	*0.432*	*0.364*
Panel C: GPAs, Long-Term (2016/2017 Q3+Q4)												
Treatment 1	0.016	0.027	0.004	0.027	-0.015	0.116**	0.043	0.024	0.007	0.004	0.035	0.004
"Self-Learning"	*(0.018)*	*(0.020)*	*(0.018)*	*(0.018)*	*(0.037)*	*(0.051)*	*(0.034)*	*(0.021)*	*(0.024)*	*(0.024)*	*(0.030)*	*(0.009)*
Treatment 2	0.025	0.034*	0.015	0.041*	0.000	0.280***	0.037	0.038*	0.011	0.053**	0.032	-0.010
"Teacher Delivery"	*(0.017)*	*(0.020)*	*(0.018)*	*(0.021)*	*(0.032)*	*(0.054)*	*(0.033)*	*(0.022)*	*(0.022)*	*(0.022)*	*(0.028)*	*(0.009)*
N	29,303	15,135	14,168	17,851	9,253	928	1,271	14,757	14,546	9,723	9,782	9,798
N Control	10,607	5,480	5,127	6,304	3,554	340	409	5,497	5,110	3,350	3,622	3,635
N Treatment 1	9,716	5,034	4,682	6,072	2,932	225	487	4,793	4,923	3,305	3,230	3,181
N Treatment 2	8,980	4,621	4,359	5,475	2,767	363	375	4,467	4,513	3,068	2,930	2,982
R^2	*0.855*	*0.852*	*0.843*	*0.853*	*0.822*	*0.797*	*0.900*	*0.851*	*0.864*	*0.468*	*0.424*	*0.227*

Notes: All regressions control for pre-treatment outcomes (up to a fourth-order polynomial), demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. *$* p < 0.05$, **$ p < 0.01$, ***$ p < 0.001$*

Table A2.2: Impacts on GPAs of Roma Students Over Time, Unbalanced Panel (Z-Scores)

		Gender		Academic Year		Pre-Treatment Outcome		
	Average	Male	Female	Sixth Grade	Seventh Grade	Tercile 1 (Low)	Tercile 2 (Middle)	Tercile 3 (High)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: GPAs, Short-Term (2015/2016 Q4)								
Treatment 1	-0.015	-0.009	-0.029	0.022	-0.034	-0.063*	0.037	-0.048
"Self-Learning"	*(0.028)*	*(0.034)*	*(0.036)*	*(0.034)*	*(0.035)*	*(0.036)*	*(0.035)*	*(0.055)*
Treatment 2	0.055***	0.056***	0.063***	0.003	0.108**	0.006	0.076***	0.083***
"Teacher Delivery"	*(0.017)*	*(0.020)*	*(0.023)*	*(0.026)*	*(0.045)*	*(0.043)*	*(0.028)*	*(0.024)*
N	1,161	611	550	573	588	388	391	382
N Control	476	254	222	228	248	189	146	141
N Treatment 1	262	142	120	136	126	84	105	73
N Treatment 2	423	215	208	209	214	115	140	168
R^2	*0.913*	*0.908*	*0.923*	*0.920*	*0.917*	*0.521*	*0.403*	*0.862*
Panel B: GPAs, Medium-Term (2016/2017 Q1+Q2)								
Treatment 1	0.109**	0.085	0.162***	0.162**	0.065	0.108*	0.079	0.174*
"Self-Learning"	*(0.045)*	*(0.053)*	*(0.060)*	*(0.074)*	*(0.042)*	*(0.060)*	*(0.053)*	*(0.090)*
Treatment 2	0.166***	0.179***	0.170***	0.232**	0.113***	0.126*	0.177***	0.210***
"Teacher Delivery"	*(0.056)*	*(0.060)*	*(0.054)*	*(0.093)*	*(0.028)*	*(0.065)*	*(0.048)*	*(0.057)*
N	1,045	551	494	521	524	349	352	344
N Control	426	229	197	207	219	170	132	124
N Treatment 1	238	128	110	125	113	76	94	68
N Treatment 2	381	194	187	189	192	103	126	152

R^2	*0.807*	*0.772*	*0.844*	*0.809*	*0.825*	*0.440*	*0.231*	*0.703*
Panel C: GPAs, Long-Term (2016/2017 Q3+Q4)								
Treatment 1	0.129**	0.136**	0.126	0.196**	0.077	0.104**	0.127	0.178*
"Self-Learning"	*(0.052)*	*(0.062)*	*(0.086)*	*(0.077)*	*(0.057)*	*(0.040)*	*(0.082)*	*(0.102)*
Treatment 2	0.279***	0.240***	0.342***	0.327***	0.228***	0.170***	0.232***	0.460***
"Teacher Delivery"	*(0.055)*	*(0.035)*	*(0.094)*	*(0.071)*	*(0.038)*	*(0.057)*	*(0.054)*	*(0.042)*
N	985	522	463	512	473	330	328	327
N Control	360	193	167	199	161	150	108	102
N Treatment 1	238	129	109	125	113	76	93	69
N Treatment 2	387	200	187	188	199	104	127	156
R^2	*0.798*	*0.786*	*0.821*	*0.806*	*0.815*	*0.257*	*0.316*	*0.691*

Notes: All regressions control for pre-treatment outcomes (up to a fourth-order polynomial), demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. *$* p < 0.05$, **$ p < 0.01$, ***$ p < 0.001$*

Appendix C: Robustness Checks

1. Survey Non-Response and Attrition

Table A3.1: Predicting Attrition in Survey Data

	Has Baseline Survey		Has Endline Survey		Has Both Surveys	
	(1)	(2)	(3)	(4)	(5)	(6)
Treatment 1	-0.001	-0.010	-0.149***	-0.106	-0.120**	-0.120
	(0.034)	*(0.061)*	*(0.045)*	*(0.083)*	*(0.047)*	*(0.084)*
Treatment 1 x Pre-Treatment GPA		-0.008		-0.016		-0.016
		(0.009)		*(0.012)*		*(0.013)*
Treatment 1 x Age 12		0.011		0.003		0.007
		(0.028)		*(0.046)*		*(0.047)*
Treatment 1 x Age 13		0.016		-0.010		-0.001
		(0.041)		*(0.058)*		*(0.060)*
Treatment 1 x Age 14		0.037		0.076		0.086
		(0.066)		*(0.084)*		*(0.086)*
Treatment 1 x Female		0.023**		0.020*		0.022*
		(0.009)		*(0.011)*		*(0.012)*
Treatment 1 x Sixth Grader		0.001		0.005		0.032
		(0.035)		*(0.044)*		*(0.045)*
Treatment 1 x Albanian		0.023		-0.025		0.039
		(0.080)		*(0.097)*		*(0.102)*
Treatment 1 x Roma		0.066		0.246***		0.243**
		(0.127)		*(0.088)*		*(0.105)*
Treatment 1 x Other Ethnicity		-0.051		-0.013		0.013
		(0.062)		*(0.077)*		*(0.081)*
Treatment 2	-0.010	-0.016	-0.063*	-0.193**	-0.068*	-0.156*
	(0.029)	*(0.063)*	*(0.036)*	*(0.082)*	*(0.039)*	*(0.087)*
Treatment 2 x Pre-Treatment GPA		0.004		0.020*		0.012

		(0.011)		*(0.011)*		*(0.012)*
Treatment 2 x Age 12		-0.027		-0.013		-0.031
		(0.027)		*(0.038)*		*(0.041)*
Treatment 2 x Age 13		-0.020		0.042		0.006
		(0.038)		*(0.053)*		*(0.056)*
Treatment 2 x Age 14		-0.082		0.042		-0.024
		(0.064)		*(0.085)*		*(0.088)*
Treatment 2 x Female		0.003		-0.006		-0.004
		(0.010)		*(0.011)*		*(0.012)*
Treatment 2 x Sixth Grader		-0.003		0.024		0.022
		(0.038)		*(0.044)*		*(0.045)*
Treatment 2 x Albanian		0.038		0.122		0.147*
		(0.061)		*(0.077)*		*(0.087)*
Treatment 2 Roma		0.142		0.081		0.137
		(0.117)		*(0.096)*		*(0.091)*
Treatment 2 x Other Ethnicity		-0.087		-0.126		-0.089
		(0.068)		*(0.079)*		*(0.086)*
Pre-Treatment GPA	0.017***	0.019***	0.010***	0.010	0.016***	0.0184**
	(0.003)	*(0.006)*	*(0.004)*	*(0.007)*	*(0.004)*	*(0.008)*
Age 12	0.002	0.006	0.018	0.018	0.026	0.030
	(0.013)	*(0.018)*	*(0.017)*	*(0.027)*	*(0.019)*	*(0.029)*
Age 13	-0.017	-0.017	-0.033	-0.047	-0.020	-0.026
	(0.018)	*(0.025)*	*(0.023)*	*(0.034)*	*(0.025)*	*(0.038)*
Age 14	-0.052*	-0.042	-0.059	-0.108**	-0.044	-0.075
	(0.028)	*(0.044)*	*(0.037)*	*(0.052)*	*(0.038)*	*(0.057)*
Female	0.004	-0.005	0.005	0.000	0.009*	0.003
	(0.004)	*(0.007)*	*(0.005)*	*(0.007)*	*(0.005)*	*(0.007)*
Sixth Grader	-0.022	-0.023	-0.071***	-0.083***	-0.067***	-0.087***
	(0.015)	*(0.024)*	*(0.018)*	*(0.024)*	*(0.019)*	*(0.027)*
Albanian	-0.118***	-0.134**	-0.291***	-0.314***	-0.289***	-0.337***
	(0.042)	*(0.055)*	*(0.059)*	*(0.074)*	*(0.060)*	*(0.081)*

Roma	-0.141***	-0.203*	-0.242***	-0.329***	-0.271***	-0.378***
	(0.052)	*(0.114)*	*(0.062)*	*(0.079)*	*(0.062)*	*(0.092)*
Other Ethnicity	-0.056*	-0.013	-0.090**	-0.051	-0.079**	-0.061
	(0.032)	*(0.034)*	*(0.036)*	*(0.053)*	*(0.037)*	*(0.055)*
N	33,454	33,454	33,454	33,454	33,454	33,454
N Control	12,426	12,426	12,426	12,426	12,426	12,426
N Treatment 1	10,995	10,995	10,995	10,995	10,995	10,995
N Treatment 2	10,033	10,033	10,033	10,033	10,033	10,033
F	*9.62*	*4.40*	*9.34*	*4.70*	*12.75*	*6.28*
R^2	*0.082*	*0.084*	*0.146*	*0.153*	*0.140*	*0.146*

Notes: All regressions control for municipality fixed effects. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places.
Source: Own survey data, administrative data, school year 2015/2016, own calculations. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A3.2: Predicting Attrition in Administrative Data

	Has GPA, Medium-Run (2016/2017 Q1+Q2)		Has GPA, Long-Run (2016/2017 Q3+Q4)	
	(1)	(2)	(3)	(4)
Treatment 1	-0.016	0.059	0.035**	0.124***
	(0.017)	*(0.039)*	*(0.017)*	*(0.044)*
Treatment 1 x Pre-Treatment GPA		-0.010*		-0.009
		(0.006)		*(0.006)*
Treatment 1 x Age 12		0.000		-0.002
		(0.026)		*(0.021)*
Treatment 1 x Age 13		-0.012		-0.084**
		(0.033)		*(0.034)*
Treatment 1 x Age 14		0.033		-0.025
		(0.056)		*(0.055)*
Treatment 1 x Female		0.012*		0.009
		(0.007)		*(0.007)*
Treatment 1 x Sixth Grader		-0.056		-0.105**
		(0.035)		*(0.043)*
Treatment 1 x Albanian		-0.043		0.066*
		(0.043)		*(0.036)*
Treatment 1 x Roma		0.004		0.148**
		(0.034)		*(0.069)*
Treatment 1 x Other Ethnicity		-0.015		0.063
		(0.038)		*(0.047)*
Treatment 2	0.004	0.081**	0.053**	0.172***
	(0.018)	*(0.038)*	*(0.021)*	*(0.051)*
Treatment 2 x Pre-Treatment GPA		-0.012*		-0.009
		(0.007)		*(0.006)*
Treatment 2 x Age 12		-0.022		-0.010
		(0.020)		*(0.023)*
Treatment 2 x Age 13		-0.054**		-0.104***
		(0.027)		*(0.037)*
Treatment 2 x Age 14		-0.040		-0.097*
		(0.046)		*(0.057)*
Treatment 2 x Female		0.006		0.004
		(0.006)		*(0.007)*
Treatment 2 x Sixth Grader		-0.013		-0.129***
		(0.035)		*(0.042)*
Treatment 2 x Albanian		0.018		0.064
		(0.055)		*(0.050)*
Treatment 2 Roma		-0.049		0.109*
		(0.034)		*(0.058)*
Treatment 2 x Other Ethnicity		0.018		0.035
		(0.028)		*(0.046)*
Pre-Treatment GPA	0.004	0.011***	0.009***	0.014***
	(0.003)	*(0.003)*	*(0.002)*	*(0.004)*

Age 12	0.010	0.016	-0.019**	-0.016
	(0.010)	*(0.016)*	*(0.008)*	*(0.016)*
Age 13	0.000	0.020	-0.018	0.041
	(0.013)	*(0.020)*	*(0.014)*	*(0.028)*
Age 14	-0.060***	-0.059	-0.107***	-0.072*
	(0.022)	*(0.037)*	*(0.022)*	*(0.043)*
Female	-0.004	-0.009**	-0.004	-0.008
	(0.003)	*(0.004)*	*(0.003)*	*(0.005)*
Sixth Grader	0.010	0.032	0.013	0.086**
	(0.014)	*(0.022)*	*(0.018)*	*(0.038)*
Albanian	-0.081**	-0.072**	-0.042***	-0.078***
	(0.038)	*(0.034)*	*(0.012)*	*(0.024)*
Roma	-0.073***	-0.055**	-0.099**	-0.174***
	(0.021)	*(0.026)*	*(0.039)*	*(0.064)*
Other Ethnicity	-0.001	0.002	-0.013	-0.048
	(0.014)	*(0.017)*	*(0.016)*	*(0.043)*
N	33,454	33,454	33,454	33,454
N Control	12,426	12,426	12,426	12,426
N Treatment 1	10,995	10,995	10,995	10,995
N Treatment 2	10,033	10,033	10,033	10,033
F	*4.45*	*2.84*	*6.50*	*4.83*
R^2	*0.088*	*0.093*	*0.106*	*0.114*

Notes: All regressions control for municipality fixed effects. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$

Table A3.3: Impacts on S/E Skills Index and Short-Run GPAs, Balanced Sample of Survey and Administrative Data (Z-Scores)

		Gender		Ethnicity				Academic Year		Pre-Treatment Outcome		
	Average (1)	Male (2)	Female (3)	Macedonian (4)	Albanian (5)	Roma (6)	Other (7)	Sixth Grade (8)	Seventh Grade (9)	Tercile 1 (Low) (10)	Tercile 2 (Middle) (11)	Tercile 3 (High) (12)
Panel A: S/E Skills Index												
Treatment 1 "Self-Learning"	0.060***	0.034	0.087***	0.069**	0.029	0.225	0.104	0.057*	0.121***	-0.007	0.047	0.122***
	(0.023)	*(0.030)*	*(0.026)*	*(0.027)*	*(0.054)*	*(0.182)*	*(0.104)*	*(0.031)*	*(0.029)*	*(0.033)*	*(0.033)*	*(0.030)*
Treatment 2 "Teacher Delivery"	0.149***	0.105***	0.192***	0.168***	0.068	0.525***	0.249**	0.138***	0.182***	0.017	0.168***	0.210***
	(0.023)	*(0.031)*	*(0.027)*	*(0.028)*	*(0.048)*	*(0.145)*	*(0.110)*	*(0.032)*	*(0.027)*	*(0.034)*	*(0.031)*	*(0.033)*
N	16,575	8,059	8,516	11,562	4,038	318	657	7,943	8,632	4,167	5,808	6,600
N Control	6,606	3,282	3,324	4,493	1,754	113	246	3,192	3,414	1,615	2,366	2,625
N Treatment 1	4,979	2,394	2,585	3,604	1,037	93	245	2,419	2,560	1,329	1,741	1,909
N Treatment 2	4,990	2,383	2,607	3,465	1,247	112	166	2,332	2,658	1,223	1,701	2,066
R^2	*0.336*	*0.303*	*0.348*	*0.353*	*0.205*	*0.360*	*0.439*	*0.331*	*0.385*	*0.228*	*0.282*	*0.356*
Panel B: GPAs, Short-Run (2015/2016 Q4)												
Treatment 1 "Self-Learning"	0.023*	0.030**	0.016	0.028*	0.014	-0.025	-0.022	0.019	0.027*	0.030*	0.029*	0.006
	(0.012)	*(0.013)*	*(0.014)*	*(0.015)*	*(0.018)*	*(0.049)*	*(0.029)*	*(0.018)*	*(0.015)*	*(0.017)*	*(0.017)*	*(0.010)*
Treatment 2 "Teacher Delivery"	0.019	0.024*	0.013	0.017	0.011	0.113*	-0.019	0.007	0.035**	0.027	0.013	0.014
	(0.014)	*(0.015)*	*(0.015)*	*(0.018)*	*(0.018)*	*(0.060)*	*(0.030)*	*(0.019)*	*(0.017)*	*(0.018)*	*(0.020)*	*(0.009)*
N	16,575	8,059	8,516	11,562	4,038	318	657	7,943	8,632	5,550	5,569	5,456
N Control	6,606	3,282	3,324	4,493	1,754	113	246	3,192	3,414	2,152	2,242	2,212
N Treatment 1	4,979	2394	2,585	3,604	1,037	93	245	2,419	2,560	1,742	1,684	1,553
N Treatment 2	4,990	2383	2,607	3,465	1,247	112	166	2,332	2,658	1,656	1,643	1,691

R^2	*0.929*	*0.928*	*0.923*	*0.923*	*0.926*	*0.916*	*0.954*	*0.922*	*0.939*	*0.706*	*0.556*	*0.444*

Notes: All regressions control for pre-treatment outcomes (in case of GPAs, up to a fourth-order polynomial), demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A3.4: Replication of Table 3 Using Imputed Outcomes

	Deliberate Practice Beliefs	Grit	Grit: Effort	Grit: Interest
	(1)	(2)	(3)	(4)
Original S/E Skills Index				
Treatment 1	0.151***	-0.052***	0.052***	-0.116***
"Self-Learning"	*(0.018)*	*(0.019)*	*(0.020)*	*(0.019)*
Treatment 2	0.227***	-0.029	0.059***	-0.096***
"Teacher Delivery"	*(0.017)*	*(0.021)*	*(0.019)*	*(0.021)*
N	24,276	21,925	23,049	23,267
N Control	9,451	8,528	8,953	9,077
N Treatment 1	7,068	6,365	6,714	6,745
N Treatment 2	7,757	7,032	7,382	7,445
R^2	*0.320*	*0.334*	*0.331*	*0.223*
Imputed S/E Skills Index				
Treatment 1	0.115***	-0.038**	0.038**	-0.084***
"Self-Learning"	*(0.016)*	*(0.016)*	*(0.016)*	*(0.017)*
Treatment 2	0.181***	-0.022	0.053***	-0.077***
"Teacher Delivery"	*(0.016)*	*(0.017)*	*(0.017)*	*(0.018)*
N	35,340	35,340	35,340	35,340
N Control	13,226	13,226	13,226	13,226
N Treatment 1	10,970	10,970	10,970	10,970
N Treatment 2	11,144	11,144	11,144	11,144
R^2	-	-	-	-

Notes: All regressions control for pre-treatment outcomes, demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. We impute outcomes using multiple imputation from students' pre-treatment GPAs, age, gender, ethnicity, and academic year. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. *$p < 0.05$, **$p < 0.01$, ***$p < 0.001$*

Table A3.5: Replication of Table 4 Using Imputed Outcomes

	Including All Skills (1)	Excluding Deliberate Practice Beliefs and Grit (2)	Including Only Deliberate Practice Beliefs and Grit (3)
Original S/E Skills Index			
Treatment 1	0.055**	-0.004	0.070***
"Self-Learning"	*(0.022)*	*(0.022)*	*(0.020)*
Treatment 2	0.128***	0.063***	0.133***
"Teacher Delivery"	*(0.021)*	*(0.021)*	*(0.020)*
N	18,718	18,718	18,718
N Control	7,286	7,286	7,286
N Treatment 1	5,424	5,424	5,424
N Treatment 2	6,008	6,008	6,008
R^2	*0.337*	*0.345*	*0.308*
Imputed S/E Skills Index			
Treatment 1	0.042**	0.001	0.053***
"Self-Learning"	*(0.017)*	*(0.016)*	*(0.017)*
Treatment 2	0.100***	0.051***	0.102***
"Teacher Delivery"	*(0.019)*	*(0.018)*	*(0.018)*
N	35,340	35,340	35,340
N Control	13,226	13,226	13,226
N Treatment 1	10,970	10,970	10,970
N Treatment 2	11,144	11,144	11,144
R^2	-	-	-

Notes: All regressions control for pre-treatment outcomes, demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. We impute outcomes using multiple imputation from students' pre-treatment GPAs, age, gender, ethnicity, and academic year. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A3.6: Replication of Table 5 Panel A Using Imputed Outcomes

		Gender		Ethnicity				Academic Year		Pre-Treatment Outcome		
	Average	Male	Female	Macedonian	Albanian	Roma	Other	Sixth Grade	Seventh Grade	Tercile 1 (Low)	Tercile 2 (Middle)	Tercile 3 (High)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Original S/E Skills Index												
Treatment 1	0.055**	0.025	0.087***	0.065**	0.030	0.203	0.032	0.037	0.071***	0.043	0.040	0.078**
"Self-Learning"	*(0.022)*	*(0.029)*	*(0.024)*	*(0.026)*	*(0.050)*	*(0.177)*	*(0.087)*	*(0.029)*	*(0.027)*	*(0.030)*	*(0.032)*	*(0.032)*
Treatment 2	0.128***	0.080***	0.172***	0.140***	0.077*	0.417***	0.128	0.121***	0.136***	0.128***	0.110***	0.150***
"Teacher Delivery"	*(0.021)*	*(0.028)*	*(0.025)*	*(0.026)*	*(0.044)*	*(0.144)*	*(0.096)*	*(0.029)*	*(0.026)*	*(0.032)*	*(0.029)*	*(0.034)*
N	18,718	9,077	9,641	13,020	4,494	360	844	8,944	9,774	5,871	6,236	6,611
N Control	7,286	3,592	3,694	4,867	2,021	127	271	3,482	3,804	2,301	2,469	2,516
N Treatment 1	5,424	2,622	2,802	3,919	1,116	109	280	2,598	2,826	1,708	1,776	1,940
N Treatment 2	6,008	2,863	3,145	4,234	1,357	124	293	2,864	3,144	1,862	1,991	2,155
R^2	*0.337*	*0.305*	*0.347*	*0.355*	*0.205*	*0.329*	*0.429*	*0.334*	*0.350*	*0.122*	*0.072*	*0.184*
Imputed S/E Skills Index												
Treatment 1	0.042**	0.023	0.063***	0.053**	0.032	0.074	0.038	0.031	0.055***	0.035	0.038	0.057**
"Self-Learning"	*(0.017)*	*(0.022)*	*(0.020)*	*(0.021)*	*(0.031)*	*(0.120)*	*(0.077)*	*(0.022)*	*(0.020)*	*(0.028)*	*(0.023)*	*(0.028)*
Treatment 2	0.100***	0.069***	0.132***	0.110***	0.080**	0.078	0.144	0.099***	0.102***	0.097***	0.078***	0.125***
"Teacher Delivery"	*(0.019)*	*(0.023)*	*(0.024)*	*(0.023)*	*(0.033)*	*(0.101)*	*(0.087)*	*(0.024)*	*(0.023)*	*(0.028)*	*(0.025)*	*(0.029)*
N	35,340	18,068	17,062	21,078	11,139	1,238	1,683	17,414	17,740	11,411	11,410	11,410
N Control	13,226	6,738	6,362	7,389	4,676	523	520	6,560	6,556	4,293	4,348	4,119
N Treatment 1	10,970	5,681	5,289	6,835	3,215	322	598	5,401	5,569	3,626	3,457	3,617
N Treatment 2	11,144	5,649	5,411	6,854	3,248	393	565	5,453	5,615	3,492	3,605	3,674

R^2	-		-		-		-		-		-		-		-		-		-		-		-		-

Notes: All regressions control for pre-treatment outcomes, demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. We impute outcomes using multiple imputation from students' pre-treatment GPAs, age, gender, ethnicity, and academic year. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A3.7: Replication of Table 5 Panel B Using Imputed Outcomes

		Gender		Ethnicity				Academic Year		Pre-Treatment Outcome		
	Average	Male	Female	Macedonian	Albanian	Roma	Other	Sixth Grade	Seventh Grade	Tercile 1 (Low)	Tercile 2 (Middle)	Tercile 3 (High)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Original GPAs, Short-Term (2015/2016 Q4)												
Treatment 1	0.018*	0.019*	0.017	0.019	0.017	-0.015	0.012	0.017	0.020	0.010	0.028*	0.014
"Self-Learning"	*(0.011)*	*(0.011)*	*(0.012)*	*(0.012)*	*(0.019)*	*(0.028)*	*(0.020)*	*(0.015)*	*(0.014)*	*(0.013)*	*(0.015)*	*(0.009)*
Treatment 2	0.016	0.013	0.019	0.007	0.020	0.055***	-0.007	0.014	0.019	0.020	0.015	0.011
"Teacher Delivery"	*(0.012)*	*(0.012)*	*(0.013)*	*(0.015)*	*(0.018)*	*(0.017)*	*(0.023)*	*(0.014)*	*(0.016)*	*(0.014)*	*(0.018)*	*(0.008)*
N	33,454	17,270	16,184	19,460	11,423	1,161	1,410	16,563	16,891	11,181	11,127	11,146
N Control	12,426	6,415	6,011	6,880	4,605	476	465	6,228	6,198	4,078	4,187	4,161
N Treatment 1	10,995	5,711	5,284	6,657	3,527	262	549	5,442	5,553	3,683	3,688	3,624
N Treatment 2	10,033	5,144	4,889	5,923	3,291	423	396	4,893	5,140	3,420	3,252	3,361
R^2	*0.935*	*0.934*	*0.930*	*0.930*	*0.925*	*0.913*	*0.953*	*0.931*	*0.942*	*0.671*	*0.592*	*0.518*
Imputed GPAs, Short-Term (2015/2016 Q4)												
Treatment 1	0.016	0.018*	0.013	0.016	0.018	-0.014	0.009	0.014	0.018	0.011	0.024*	0.010
"Self-Learning"	*(0.010)*	*(0.010)*	*(0.011)*	*(0.011)*	*(0.017)*	*(0.036)*	*(0.019)*	*(0.014)*	*(0.013)*	*(0.012)*	*(0.014)*	*(0.008)*
Treatment 2	0.015	0.014	0.016	0.006	0.022	0.054**	-0.015	0.012	0.017	0.021*	0.012	0.012
"Teacher Delivery"	*(0.010)*	*(0.011)*	*(0.011)*	*(0.013)*	*(0.016)*	*(0.023)*	*(0.021)*	*(0.012)*	*(0.014)*	*(0.012)*	*(0.016)*	*(0.008)*
N	35,340	18,068	17,062	21,078	11,139	1,238	1,683	17,414	17,740	11,583	11,793	11,754
N Control	13,226	6,738	6,362	7,389	4,676	523	520	6,560	6,556	4,257	4,454	4,389
N Treatment 1	10,970	5,681	5,289	6,835	3,215	322	598	5,401	5,569	3,743	3,661	3,566

N Treatment 2	11,144	5,649	5,411	6,854	3,248	393	565	5,453	5,615	3,583	3,678	3,799
R^2	-	-	-	-	-	-	-	-	-	-	-	-

Notes: All regressions control for pre-treatment outcomes (up to a fourth-order polynomial), demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. We impute outcomes using multiple imputation from students' pre-treatment GPAs, age, gender, ethnicity, and academic year. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

Table A3.8: Replication of Table 6 Panel A Using Imputed Outcomes

		Gender		Ethnicity				Academic Year		Pre-Treatment Outcome		
	Average	Male	Female	Macedonian	Albanian	Roma	Other	Sixth Grade	Seventh Grade	Tercile 1 (Low)	Tercile 2 (Middle)	Tercile 3 (High)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Original GPAs, Medium-Term (2016/2017 Q1+Q2)												
Treatment 1	0.006	0.013	-0.002	0.015	-0.002	0.109**	0.052	0.020	-0.009	0.008	0.015	-0.006
"Self-Learning"	*(0.013)*	*(0.014)*	*(0.014)*	*(0.016)*	*(0.024)*	*(0.045)*	*(0.036)*	*(0.017)*	*(0.015)*	*(0.016)*	*(0.019)*	*(0.013)*
Treatment 2	0.009	0.023	-0.006	0.019	0.000	0.166***	0.014	(0.022)	-0.003	0.028*	0.015	-0.014
"Teacher Delivery"	*(0.013)*	*(0.014)*	*(0.014)*	*(0.017)*	*(0.020)*	*(0.056)*	*(0.046)*	*(0.018)*	*(0.015)*	*(0.017)*	*(0.019)*	*(0.012)*
N	31,310	16,154	15,156	18,568	10,348	1,045	1,349	15,697	15,613	10,338	10,423	10,549
N Control	11,600	5,992	5,608	6,533	4,190	426	451	5,881	5,719	3,716	3,935	3,949
N Treatment 1	10,166	5,265	4,901	6,310	3,107	238	511	4,996	5,170	3,391	3,391	3,384
N Treatment 2	9,544	4,897	4,647	5,725	3,051	381	387	4,820	4,724	3,231	3,097	3,216
R^2	*0.878*	*0.872*	*0.870*	*0.876*	*0.850*	*0.807*	*0.907*	*0.873*	*0.887*	*0.503*	*0.434*	*0.377*
Imputed GPAs, Medium-Term (2016/2017 Q1+Q2)												
Treatment 1	0.004	0.009	-0.003	0.010	0.002	0.079*	0.040	0.015	-0.009	0.008	0.012	-0.008
"Self-Learning"	*(0.012)*	*(0.013)*	*(0.013)*	*(0.015)*	*(0.021)*	*(0.046)*	*(0.034)*	*(0.016)*	*(0.014)*	*(0.014)*	*(0.018)*	*(0.012)*
Treatment 2	0.011	0.022*	-0.001	0.022	0.001	0.159***	0.002	0.020	0.002	0.026*	0.018	-0.006
"Teacher Delivery"	*(0.012)*	*(0.013)*	*(0.013)*	*(0.016)*	*(0.018)*	*(0.053)*	*(0.039)*	*(0.016)*	*(0.014)*	*(0.015)*	*(0.017)*	*(0.012)*
N	35,340	18,068	17,062	21,078	11,139	1,238	1,683	17,414	17,740	11,583	11,793	11,754
N Control	13,226	6,738	6,362	7,389	4,676	523	520	6,560	6,556	4,257	4,454	4,389
N Treatment 1	10,970	5,681	5,289	6,835	3,215	322	598	5,401	5,569	3,743	3,661	3,566

N Treatment 2	11,144	5,649	5,411	6,854	3,248	393	565	5,453	5,615	3,583	3,678	3,799
R^2	-	-	-	-	-	-	-	-	-	-	-	-

Notes: All regressions control for pre-treatment outcomes (up to a fourth-order polynomial), demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. We impute outcomes using multiple imputation from students' pre-treatment GPAs, age, gender, ethnicity, and academic year. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. $^{*}p < 0.05$, $^{**}p < 0.01$, $^{***}p < 0.001$

Table A3.9: Replication of Table 6 Panel B Using Imputed Outcomes

		Gender		Ethnicity				Academic Year		Pre-Treatment Outcome		
	Average	Male	Female	Macedonian	Albanian	Roma	Other	Sixth Grade	Seventh Grade	Tercile 1 (Low)	Tercile 2 (Middle)	Tercile 3 (High)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Original GPAs, Long-Term (2016/2017 Q3+Q4)												
Treatment 1	0.021	0.030	0.009	0.027	0.001	0.129**	0.045	0.032	0.010	0.007	0.043	0.004
"Self-Learning"	*(0.018)*	*(0.020)*	*(0.018)*	*(0.017)*	*(0.038)*	*(0.052)*	*(0.034)*	*(0.021)*	*(0.023)*	*(0.024)*	*(0.029)*	*(0.010)*
Treatment 2	0.030*	0.041**	0.018	0.042**	0.006	0.279***	0.034	0.043	0.017	0.056**	0.043	-0.008
"Teacher Delivery"	*(0.017)*	*(0.019)*	*(0.017)*	*(0.021)*	*(0.031)*	*(0.055)*	*(0.033)*	*(0.022)*	*(0.021)*	*(0.022)*	*(0.028)*	*(0.009)*
N	31,437	16,209	15,228	18,716	10,402	985	1,334	15,713	15,724	10,247	10,502	10,688
N Control	11,404	5,881	5,523	6,573	4,038	360	433	5,813	5,591	3,585	3,893	3,926
N Treatment 1	10,461	5,424	5,037	6,362	3,340	238	521	5,185	5,276	3,479	3,497	3,485
N Treatment 2	9,572	4,904	4,668	5,781	3,024	387	380	4,715	4,857	3,183	3,112	3,277
R^2	*0.854*	*0.850*	*0.843*	*0.853*	*0.820*	*0.798*	*0.900*	*0.850*	*0.862*	*0.459*	*0.426*	*0.240*
Imputed GPAs, Long-Term (2016/2017 Q3+Q4)												
Treatment 1	0.015	0.022	0.006	0.022	0.004	0.104*	0.031	0.027	0.004	0.004	0.033	0.005
"Self-Learning"	*(0.016)*	*(0.018)*	*(0.017)*	*(0.016)*	*(0.032)*	*(0.055)*	*(0.031)*	*(0.020)*	*(0.021)*	*(0.021)*	*(0.026)*	*(0.010)*
Treatment 2	0.027*	0.035**	0.017	0.044**	0.002	0.255***	0.014	0.036*	0.017	0.045**	0.039	-0.000
"Teacher Delivery"	*(0.015)*	*(0.017)*	*(0.016)*	*(0.019)*	*(0.026)*	*(0.056)*	*(0.029)*	*(0.019)*	*(0.019)*	*(0.019)*	*(0.024)*	*(0.010)*
N	35,340	18,068	17,062	21,078	11,139	1,238	1,683	17,414	17,740	11,583	11,793	11,754
N Control	13,226	6,738	6,362	7,389	4,676	523	520	6,560	6,556	4,257	4,454	4,389
N Treatment 1	10,970	5,681	5,289	6,835	3,215	322	598	5,401	5,569	3,743	3,661	3,566

N Treatment 2	11,144	5,649	5,411	6,854	3,248	393	565	5,453	5,615	3,583	3,678	3,799
R^2	-	-	-	-	-	-	-	-	-	-	-	-

Notes: All regressions control for pre-treatment outcomes (up to a fourth-order polynomial), demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. We impute outcomes using multiple imputation from students' pre-treatment GPAs, age, gender, ethnicity, and academic year. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

2. Analysis Using Pre-Registered Covariates Only, Parametric Bootstrap

We deviated from the pre-registered covariates by including age and controlling for pre-treatment GPA as a fourth-order polynomial. In our pre-analysis plan, the pre-registered covariates were: prior GPA, gender, ethnicity, initial cognitive ability, quality of schools, and geographical location. Initial cognitive ability was intended to be derived from a national standardized test; however, this was not possible as the country had no reliable national standardized test at the time (confirmed by Ministry of Education and Science). Likewise, there was no reliable measure of quality of schools. For initial cognitive ability at baseline, we collected two skills test measures (i.e. baseline math and baseline reading comprehension); however, due to a printing error of the baseline survey, only about 20% of the original sample took these skills tests.

We first replicate our results using pre-registered covariates only, to the extent possible. To do so, pre-treatment GPA was used as a surrogate for initial cognitive ability. Moreover, adjusting for school as a source of clustering was used as a surrogate for quality of schools. For geographical location, the municipality indicator (a 80-level categorical measure) was used as a surrogate for location. As seen in Tables A3.10 and A3.12, our baseline results continue to hold under these modifications.

Next, we replicate our results using a different modelling approach. In particular, as an alternative to our linear models, we also pre-registered generalized linear mixed-effects models (GLMM) to evaluate intervention outcomes. In these models, the data have a two-level structure with students nested within schools. Metric outcomes (e.g. deliberate practice beliefs) are evaluated using a Gaussian residual distribution and the identify link function, whereas binary outcomes are evaluated using a binomial residual distribution and the logit link function. Parameter estimates are estimated using restricted maximum likelihood (REML) and Gauss-Hermite quadrature, respectively, for metric and binary outcomes. Confidence intervals for parameter estimates are estimated using the parametric percentile bootstrap method (with 5,000 random draws). These models allow schools to have their own intercepts, deviating randomly from the mean intercept. Given that intervention condition is assigned at the highest level of analysis (i.e. schools), the intervention effects, comparing each treatment to the control condition, are fixed.

Cross-level interactions between intervention conditions and pre-intervention outcome levels and pre-intervention GPA are estimated, thus allowing treatment effects to vary across levels of pre-intervention outcome variable and prior GPA. Equations A3.1 and A3.2 shows our models for metric and binary outcomes, respectively.

Equation A3.1: Metric Outcomes

$$
\begin{aligned}
y^{ij} &= \beta_0^j + \beta_1^j Pre_1^{ij} + \beta_2^j PriorGPA_2^{ij} + \beta_3^j Gender_3^{ij} + \beta_4^j Ethnicity_4^{ij} + \beta_5^j SES_5^{ij} + e^{ij} \\
\beta_0^j &= \gamma_{00} + \gamma_{01} Achieve_1^j + \gamma_{02} T1_2^j + \gamma_{03} T2_3^j + u^j \\
\beta_1^j &= \gamma_{10} + \gamma_{11} T1_1^j + \gamma_{12} T_2^j \\
\beta_2^j &= \gamma_{20} + \gamma_{21} T_1^j + \gamma_{22} T_2^j \\
\beta_3^j &= \gamma_{30} \\
\beta_4^j &= \gamma_{40} \\
\beta_5^j &= \gamma_{50} \\
e^{ij} &\sim N(0, \sigma^2) \\
u^j &\sim N(0, \tau_{00})
\end{aligned}
$$

Equation A3.2: Binary Outcomes

$$
\begin{aligned}
\eta^{ij} &= \beta_0^j + \beta_1^j Pre_1^{ij} + \beta_2^j PriorGPA_2^{ij} + \beta_3^j Gender_3^{ij} + \beta_4^j Ethnicity_4^{ij} + \beta_5^j SES_5^{ij} \\
\beta_0^j &= \gamma_{00} + \gamma_{01} Achieve_1^j + \gamma_{02} T1_2^j + \gamma_{03} T2_3^j + u^j \\
\beta_1^j &= \gamma_{10} + \gamma_{11} T1_1^j + \gamma_{12} T_2^j \\
\beta_2^j &= \gamma_{20} + \gamma_{21} T_1^j + \gamma_{22} T_2^j \\
\beta_3^j &= \gamma_{30} \\
\beta_4^j &= \gamma_{40} \\
\beta_5^j &= \gamma_{50} \\
y^{ij} \mid \pi^{ij} &\sim Bernouli(\pi^{ij}) \\
\eta^{ij} &= logit(\pi^{ij}) \\
u^j &\sim N(0, \tau_{00})
\end{aligned}
$$

where Pre = Pre-intervention level of the outcome (y^{ij}); PriorGPA = GPA prior to the intervention; Gender = Female/Male; Ethnicity = Ethnicity; SES = Socioeconomic status; Achieve = School's achievement level on standardized tests; T1 = Simple contrast of treatment 1 vs. control; T2 = Simple contrast of treatment 2 vs. control. y_{ij} = Level 1 outcome for student *i* within school *j*; βs = Level 1 regression coefficients; e_{ij} = Level 1 residuals; γs = Level 2 regression coefficients; *u*s = Level 2 residuals (individual school intercept deviations from mean intercept); σ^2 = Variance in level 1 residuals; τ_{00} = Variance in level 2 residuals (intercept variability); π = Probability of a success (i.e., probability of score equal to one); η = Link function.

Table A3.10: Impacts on Deliberate Practice Beliefs, Grit, and Grit Facets,

Pre-Registered Covariates (Z-Scores)

	Deliberate Practice Beliefs	Grit	Grit: Effort	Grit: Interest
	(1)	(2)	(3)	(4)
Treatment 1	0.157***	-0.048*	0.060**	-0.115***
"Self-Learning"	*(0.019)*	*(0.020)*	*(0.020)*	*(0.019)*
Treatment 2	0.230***	-0.029	0.060**	-0.097***
"Teacher Delivery"	*(0.018)*	*(0.021)*	*(0.019)*	*(0.021)*
N	24,151	21,815	22,929	23,153
N Control	9,429	8,507	8,930	9,056
N Treatment 1	7,041	6,346	6,692	6,723
N Treatment 2	7,681	6962	7,307	7,374
R^2	*0.335*	*0.350*	*0.348*	*0.232*

*$p<0.10$, * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$*

Table A3.11: Impacts on Deliberate Practice Beliefs, Grit, and Grit Facets,

Pre-Registered Covariates with Parametric Bootstrap (Z-Scores)

	Deliberate Practice Beliefs	Grit	Grit: Effort	Grit: Interest
	(1)	(2)	(3)	(4)
Treatment 1	0.153***	-0.059***	0.051***	-0.121***
"Self-Learning"	*(0.013)*	*(0.014)*	*(0.013)*	*(0.015)*
Lower and Upper Bound	*[0.127 – 0.178]*	*[-0.085 – -0.032]*	*[0.032 – 0.070]*	*[-0.150 – -0.092]*
Treatment 2	0.231***	-0.034*	0.062***	-0.104***
"Teacher Delivery"	*(0.013)*	*(0.014)*	*(0.013)*	*(0.015)*
Lower and Upper Bound	*[0.206 – 0.257]*	*[-0.062 – -0.007]*	*[0.036 – 0.088]*	*[-0.134 – -0.075]*
N	24,151	21,815	22,929	23,153
N Control	9,429	8,507	8,930	9,056
N Treatment 1	7,041	6,346	6,692	6,723
N Treatment 2	7,681	6962	7,307	7,374
R^2	-	-	-	-

*p<0.10, * p < 0.05, ** p < 0.01, *** p < 0.001*

Table A3.12: Impacts on S/E Skills Index and Short-Term GPAs, Pre-Registered Covariates (Z-Scores)

		Gender		Ethnicity				Academic Year		Pre-Treatment Outcome		
	Average	Male	Female	Macedonian	Albanian	Roma	Other	Sixth Grade	Seventh Grade	Tercile 1 (Low)	Tercile 2 (Middle)	Tercile 3 (High)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: S/E Skills Index												
Treatment 1	0.063**	0.034	0.092***	0.072**	0.034	0.198	0.115	0.044	0.078**	-0.011	0.053#	0.120***
	(0.022)	*(0.029)*	*(0.025)*	*(0.026)*	*(0.056)*	*(0.173)*	*(0.098)*	*(0.030)*	*(0.028)*	*(0.031)*	*(0.030)*	*(0.029)*
Treatment 2	0.131***	0.086**	0.173***	0.144***	0.088#	0.363*	0.126	0.121***	0.139***	0.023	0.150***	0.178***
	(0.027)	*(0.029)*	*(0.026)*	*(0.028)*	*(0.046)*	*(0.147)*	*(0.098)*	*(0.030)*	*(0.027)*	*(0.030)*	*(0.030)*	*(0.031)*
N	18,624	9,036	9,588	12,975	4,460	357	832	8,915	9,709	4,745	6,511	7,368
N Control	7,269	3,587	3,682	4,867	2,004	127	271	3,482	3,787	1,831	2,577	2,861
N Treatment 1	5,406	2,613	2,793	3,919	1,100	108	279	2,595	2,811	1,473	1,873	2,060
N Treatment 2	5,949	2,836	3,113	4,189	1,356	122	282	2,838	3,111	1,441	2,061	2,447
R^2	*0.352*	*0.326*	*0.358*	*0.368*	*0.229*	*0.341*	*0.440*	*0.352*	*0.361*	*0.236*	*0.281*	*0.361*
Panel B: GPAs, Short-Term (2015/2016 Q4)												
Treatment 1	0.018#	0.019#	0.017	0.019	0.017	-0.015	0.012	0.017	0.020	0.009	0.028#	0.014
	(0.011)	*(0.011)*	*(0.012)*	*(0.012)*	*(0.019)*	*(0.028)*	*(0.020)*	*(0.015)*	*(0.014)*	*(0.013)*	*(0.015)*	*(0.009)*
Treatment 2	0.016	0.013	0.018	0.007	0.019	0.056**	-0.005	0.014	0.019	0.020	0.015	0.011
	(0.012)	*(0.012)*	*(0.013)*	*(0.015)*	*(0.018)*	*(0.018)*	*(0.023)*	*(0.014)*	*(0.016)*	*(0.014)*	*(0.018)*	*(0.008)*
N	33,454	17,270	16,184	19,460	11,423	1,161	1,410	16,563	16,891	11,181	11,127	11,146
N Control	12,426	6,415	6,011	6,880	4,605	476	465	6,228	6,198	4,078	4,187	4,161
N Treatment 1	10,995	5,711	5,284	6,657	3,527	262	549	5,442	5,553	3,683	3,688	3,624
N Treatment 2	10,033	5,144	4,889	5,923	3,291	423	396	4,893	5,140	3,420	3,252	3,361
R^2	*0.935*	*0.933*	*0.930*	*0.930*	*0.925*	*0.913*	*0.953*	*0.931*	*0.942*	*0.671*	*0.592*	*0.518*

*# p<0.10, * p < 0.05, ** p < 0.01, *** p < 0.001*

Table A3.13: Impacts on S/E Skills Index and Short-Term GPAs, Pre-Registered Covariates With Parametric Bootstrap (Z-Scores)

		Gender		Ethnicity				Academic Year		Pre-Treatment Outcome		
	Average (1)	Male (2)	Female (3)	Macedonian (4)	Albanian (5)	Roma (6)	Other (7)	Sixth Grade (8)	Seventh Grade (9)	Tercile 1 (Low) (10)	Tercile 2 (Middle) (11)	Tercile 3 (High) (12)
Panel A: S/E Skills Index												
Treatment 1	0.036*	0.028	0.053**	0.038	-0.004	0.545	0.129	0.018	0.061***	0.127	0.026	-0.204
	(0.014)	*(0.020)*	*(0.021)*	*(0.019)*	*(0.029)*	*(0.187)*	*(0.082)*	*(0.020)*	*(0.019)*	*(0.069)*	*(0.024)*	*(0.105)*
Lower and Upper Bound	*[0.010 – 0.063]*	*[-0.011 – 0.066]*	*[0.013 – 0.093]*	*[-0.001 – 0.075]*	*[-0.062 – 0.050]*	*[0.186 – 0.928]*	*[-0.039 – 0.28]*	*[-0.022 – 0.056]*	*[0.022 – 0.100]*	*[-0.008 – 0.261]*	*[-0.020 – 0.075]*	*[-0.406 – 0.007]*
Treatment 2	0.108***	0.082***	0.151***	0.108***	0.082*	0.560***	0.114	0.090***	0.127***	0.101	0.139***	-0.178
	(0.014)	*(0.020)*	*(0.020)*	*(0.019)*	*(0.027)*	*(0.196)*	*(0.074)*	*(0.020)*	*(0.019)*	*(0.068)*	*(0.024)*	*(0.104)*
Lower and Upper Bound	*[0.080 – 0.135]*	*[0.043 – 0.122]*	*[0.111 – 0.192]*	*[0.070 – 0.146]*	*[0.030 – 0.137]*	*[0.188 – 0.931]*	*[-0.036 – 0.253]*	*[0.050 – 0.132]*	*[0.089 – 0.163]*	*[-0.033 – 0.234]*	*[0.093 – 0.186]*	*[-0.382 – 0.027]*
N	18,624	9,036	9,588	12,975	4,460	357	832	8,915	9,709	4,745	6,511	7,368
N Control	7,269	3,587	3,682	4,867	2,004	127	271	3,482	3,787	1,831	2,577	2,861
N Treatment 1	5,406	2,613	2,793	3,919	1,100	108	279	2,595	2,811	1,473	1,873	2,060
N Treatment 2	5,949	2,836	3,113	4,189	1,356	122	282	2,838	3,111	1,441	2,061	2,447
R^2	-	-	-	-	-	-	-	-	-	-	-	-
Panel B: GPAs, Short-Term (2015/2016 Q4)												
Treatment 1	-0.001	0.001	-0.001	-0.003	0.002	-0.027	0.035	-0.008	0.002	0.010	0.010	0.044*
	(0.004)	*(0.005)*	*(0.006)*	*(0.005)*	*(0.008)*	*(0.035)*	*(0.017)*	*(0.006)*	*(0.005)*	*(0.018)*	*(0.008)*	*(0.019)*
Lower and Upper Bound	*[-0.009 – 0.007]*	*[-0.010 – 0.011]*	*[-0.012 – 0.010]*	*[-0.014 – 0.007]*	*[-0.013 – 0.018]*	*[-0.095 – 0.042]*	*[-0.000 – 0.070]*	*[-0.019 – 0.003]*	*[-0.008 – 0.012]*	*[-0.026 – 0.046]*	*[-0.006 – 0.026]*	*[0.006 – 0.083]*
Treatment 2	0.012*	0.012*	0.017*	0.010	0.005	0.006	0.018	0.010	0.016*	0.045*	0.013	0.050*
	(0.004)	*(0.006)*	*(0.006)*	*(0.006)*	*(0.008)*	*(0.031)*	*(0.019)*	*(0.006)*	*(0.005)*	*(0.018)*	*(0.008)*	*(0.019)*
Lower and Upper Bound	*[0.004 – 0.020]*	*[0.001 – 0.023]*	*[0.006 – 0.028]*	*[-0.000 – 0.021]*	*[-0.010 – 0.020]*	*[-0.056 – 0.067]*	*[-0.018 – 0.057]*	*[-0.001 – 0.022]*	*[0.006 – 0.026]*	*[0.008 – 0.081]*	*[-0.003 – 0.029]*	*[0.012 – 0.089]*
N	33,454	17,270	16,184	19,460	11,423	1,161	1,410	16,563	16,891	11,181	11,127	11,146
N Control	12,426	6,415	6,011	6,880	4,605	476	465	6,228	6,198	4,078	4,187	4,161

N Treatment 1	10,995	5,711	5,284	6,657	3,527	262	549	5,442	5,553	3,683	3,688	3,624
N Treatment 2	10,033	5,144	4,889	5,923	3,291	423	396	4,893	5,140	3,420	3,252	3,361
R^2	-	-	-	-	-	-	-	-	-	-	-	-

*# p<0.10, * p < 0.05, ** p < 0.01, *** p < 0.0*

REFERENCES

Acosta, P. M., & Muller, N. (2018). The role of cognitive and socio-emotional skills in labor markets. *IZA World of Labor*.

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

Allensworth, E., & Easton, J. Q. (2005). *The on-track indicator as a predictor of high school graduation*. Consortium on Chicago School Research. University of Chicago.

Appel, M., Kronberger, N., & Aronson, J. (2011). Stereotype threat impairs ability building: Effects on test preparation amongst women in science and technology. *European Journal of Social Psychology*, *41*(7), 904-913.

Aronson, J., Steele, C. M., Elliot, A. J., & Dweck, C. S. (2005). Stereotypes and the fragility of academic competence, motivation, and self-concept. *Handbook of Competence and Motivation*, 436-456.

Atkinson, J. W. (1957). Motivational determinants of risk-taking behavior. *Psychological Review*, *64*(6p1), 359.

Battle, E. S. (1965). Motivational determinants of academic task persistence. *Journal of Personality and Social Psychology*, *2*(2), 209-2018.

Beilock, S. L., Gunderson, E. A., Ramirez, G., & Levine, S. C. (2010). Female teachers' math anxiety affects girls' math achievement. *Proceedings of the National Academy of Sciences*, *107*(5), 1860-1863.

Beilock, S. L., & DeCaro, M. S. (2007). From poor performance to success under stress: working memory, strategy selection, and mathematical problem solving under pressure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*(6), 983.

Ben-Zeev, T., Fein, S., & Inzlicht, M. (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology*, *41*(2), 174-181.

Benner, A. D., & Graham, S. (2011). Latino adolescents' experiences of discrimination across the first 2 years of high school: Correlates and influences on educational outcomes. *Child Development*, *82*(2), 508-519.

Borghans, L., Meijers, H., & Ter Weel, B. (2008a). The role of noncognitive skills in explaining cognitive test scores. *Economic Inquiry*, *46*(1), 2-12.

Borghans, L., Duckworth, A. L., Heckman, J. J., & Ter Weel, B. (2008b). The economics and psychology of personality traits. *Journal of Human Resources*, *43*(4), 972-1059.

Broda, M., Yun, J., Schneider, B., Yeager, D. S., Walton, G. M., & Diemer, M. (2018). Reducing inequality in academic success for incoming college students: A randomized trial of growth mindset and belonging interventions. *Journal of Research on Educational Effectiveness*, *11*, 317-338.

CAF. (2017). *Encuesta CAF 2017: Trayectorias Laborales y Productivas en América Latina.* Retrieved from http://scioteca.caf.com/handle/123456789/1400

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

Cohen, G. L., Garcia, J., Purdie-Vaughns, V., Apfel, N., & Brzustoski, P. (2009). Recursive processes in self-affirmation: Intervening to close the minority achievement gap. *Science*, *324*(5925), 400-403.

Côté, J., & Erickson, K. (2015). Diversification and deliberate play during the sampling years. *Routledge Handbook of Sport Expertise*, 305-316.

Crandall, V. C. (1969). Sex differences in expectancy of intellectual and academic reinforcement. In C. P. Smith (Ed.). *Achievement-related motives in children*. New York: Russell Sage Foundation.

Credé, M., Tynan, M. C., & Harms, P. D. (2016). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*(3), 492-511.

Crosnoe, R. (2011). *Fitting in, standing out: Navigating the social challenges of high school to get an education*. Cambridge University Press.

Good, C., Aronson, J., & Inzlicht, M. (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, *24*(6), 645-662.

Guerra, N., Modecki, K., & Cunningham, W. (2014). Developing social-emotional skills for the labor market: The PRACTICE model. *World Bank Policy Research Working Paper*, (7123), Washington, DC: World Bank.

Duckworth, A. L., Peterson, C., Matthews, M. D., & Kelly, D. R. (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087-101.

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

Duckworth, A. L., Kirby, T. A., Tsukayama, E., Berstein, H., & Ericsson, K. A. (2011). Deliberate practice spells success: Why grittier competitors triumph at the National Spelling Bee. *Social Psychological and Personality Science*, *2*(2), 174-181.

Duckworth, A. L., Gendler, T. S., & Gross, J. J. (2014). Self-control in school-age children. *Educational Psychologist*, *49*(3), 199-217.

Eccles, J. S., & Wigfield, A. (2002). Motivational beliefs, values, and goals. *Annual Review of Psychology*, *53*(1), 109-132.

Eskreis-Winkler, L., Shulman, E. P., Beal, S. A., & Duckworth, A. L. (2014). The grit effect: predicting retention in the military, the workplace, school and marriage. *Frontiers in Psychology*, 5-36.

Eskreis-Winkler, L., Shulman, E. P., Young, V., Tsukayama, E., Brunwasser, S. M., & Duckworth, A. L. (2016). Using wise interventions to motivate deliberate practice. *Journal of Personality and Social Psychology*, *111*(5), 728.

Ericsson, K. A. (2006). The influence of experience and deliberate practice on the development of superior expert performance. In K. A. Ericsson, P.J. Feltovich, & R.R. Hoffman (eds).

Cambridge Handbook of Expertise and Expert Performance. Cambridge: Cambridge University Press.

Ericsson, K. A. (2007). Deliberate practice and the modifiability of body and mind: toward a science of the structure and acquisition of expert and elite performance. *International Journal of Sport Psychology*, *38*, 4-34.

Ericsson, K., A. (2008). Deliberate practice and acquisition of expert performance: a general overview. *Academic Emergency Medicine*, *15*(11), 988-994.

Ericsson, K. A. (2009). Enhancing the development of professional performance: Implications from the study of deliberate practice. In Ericsson, K. A. (Ed.). *The Development of Professional Expertise: Toward Measurement of Expert Performance and Design of Optimal Learning Environments*. New York: Cambridge University Press.

Ericsson, K. A., Krampe, R. T., & Tesch-Römer, C. (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363.

Feather, N. T. (1982). Expectancy-value approaches: Present status and future directions. In Feather, N. T. (Ed.). *Expectations and Actions: Expectancy-Value Models in Psychology*, Erlbaum, Hillsdale, NJ.

Fordham, S. (1988). Racelessness as a factor in Black students' school success: Pragmatic strategy or pyrrhic victory? *Harvard Educational Review*, *58*(1), 54-85.

Fryer Jr, R. G., & Torelli, P. (2010). An empirical analysis of 'acting white'. *Journal of Public Economics*, *94*(5-6), 380-396.

Gatti, R., Karacsony, S., Anan, K., Ferré, C., & de Paz Nieves, C. (2016). *Being Fair, Faring Better: Promoting Equality of Opportunity for Marginalized Roma*. Directions in Development--Human Development. Washington, DC: World Bank.

Glewwe, P., & Muralidharan, K. (2016). Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. *Handbook of the Economics of Education*, *5,* 653-743.

Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B*, *375*(1803), 20190502.

Gunderson, E. A., Gripshover, S. J., Romero, C., Dweck, C. S., Goldin-Meadow, S., & Levine, S. C. (2013). Parent praise to 1-to 3-year-olds predicts children's motivational frameworks 5 years later. *Child Development*, *84*(5), 1526-1541.

Gutman, L. M., & Schoon, I. (2013). *The Impact of Non-Cognitive Skills on Outcomes for Young People: Literature Review*. Education Endowment Foundation, 2019.

Heckman, J. J., Stixrud, J., & Urzua, S. (2006). The effects of cognitive and noncognitive abilities on labor market outcomes and social behavior. *Journal of Labor Economics*, *24*(3), 411-482.

Heckman, J., Moon, S. H., Pinto, R., Savelyev, P., & Yavitz, A. (2010). Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics*, *1*(1), 1-46.

Heckman, J. J., & Kautz, T. (2014). Fostering and measuring skills: Interventions that improve character and cognition. In J. Heckman, J.E. Humphries and T. Kautz (eds.) *The Myth of*

Achievement Tests: The GED and the Role of Character in American Life (341-430). University of Chicago Press.

Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting interest and performance in high school science classes. *Science*, *326*(5958), 1410-1412.

Imbens, G. W., & Wooldridge, J. M. (2009). Recent developments in the econometrics of program evaluation. *Journal of Economic Literature*, *47*(1), 5-86.

Ivcevic, Z., & Brackett, M. (2014). Predicting school success: Comparing conscientiousness, grit, and emotion regulation ability. *Journal of Research in Personality*, *52*, 29-36.

Johns, M., Inzlicht, M., & Schmader, T. (2008). Stereotype threat and executive resource depletion: Examining the influence of emotion regulation. *Journal of Experimental Psychology: General*, *137*(4), 691.

Johnson, H., Pinzón, D., Trzesniewski, K, Indrakesuma, T., Vakis, R. Perova, E., Muller, N., De Martino, S., & Catalán, D. (2020) *Can teaching growth mindset and self-management at school shift student outcomes and teacher mindsets? Evidence from a randomized controlled trial in Indonesia*. World Bank Report.

Kautz, T., & Zanoni, W. (2014). *Measuring and fostering non-cognitive skills in adolescence: Evidence from Chicago Public Schools and the OneGoal Program*. Chicago, IL: University of Chicago.

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., & Borghans, L. (2014). Fostering and measuring skills: Improving cognitive and non-cognitive skills to promote lifetime success. *National Bureau of Economic Research Working Paper* WP20749. Cambridge, MA.

Levin, V., Guallar Artal, S., & Safir, A. (2016). *Skills for work in Bulgaria: the relationship between cognitive and socioemotional skills and labor market outcomes*. Washington, DC: World Bank.

Maddi, S. R., Matthews, M. D., Kelly, D. R., Villarreal, B., & White, M. (2012). The role of hardiness and grit in predicting performance and retention of USMA cadets. *Military Psychology*, *24*(1), 19-28.

Murphy, M. C., Steele, C. M., & Gross, J. J. (2007). Signaling threat: How situational cues affect women in math, science, and engineering settings. *Psychological Science*, *18*(10), 879-885.

Naemi, B., Gonzalez, E., Bertling, J., Betancourt, A., Burrus, J., Kyllonen, P.C., Minsky, J., Lietz, P., Klieme, E., Vieluf, S. & Lee, J., (2013). Large-scale group score assessments: Past, present, and future. *Oxford Handbook of Child Psychological Assessment*, 129-149.

Nguyen, H. H. D., & Ryan, A. M. (2008). Does stereotype threat affect test performance of minorities and women? A meta-analysis of experimental evidence. *Journal of Applied Psychology*, *93*(6), 1314.

OECD. (2015) *Fostering and Measuring Skills: Improving Cognitive and Socio-emotional Skills to Promote Lifetime Success*. OECD: Paris.

Outes-Leon, I., Sánchez, A., & Vakis, R. (2020). The power of believing you can get smarter: The impact of a growth-mindset intervention on academic achievement in Peru. *World Bank Policy Research Working Paper*, 9141. Washington, DC: World Bank.

Paunesku, D., Walton, G. M., Romero, C., Smith, E. N., Yeager, D. S., & Dweck, C. S. (2015). Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science*, *26*(6), 784-793.

Pintrich, P. R. (2003). Motivation and Classroom Learning. *Handbook of Educational Psychology*, 7, 103-122.

Robayo-Abril, M., & Millan, N. (2019). *Breaking the cycle of Roma exclusion in the Western Balkans*. Washington, DC: World Bank.

Roberts, B. W., Chernyshenko, O. S., Stark, S., & Goldberg, L. R. (2005). The structure of conscientiousness: An empirical investigation based on seven major personality questionnaires. *Personnel Psychology*, *58*(1), 103-139.

Robertson-Kraft, C., & Duckworth, A. L. (2014). True grit: Trait-level perseverance and passion for long-term goals predicts effectiveness and retention amongst novice teachers. *Teachers College Record*, *116*(3).

Romano, J. P., & Wolf, M. (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, *100*(469), 94-108.

Romano, J. P., & Wolf, M. (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters*, *113*, 38-40.

Rubin, D. B. (1974). Estimating causal effects of treatments in randomized and nonrandomized studies. *Journal of Educational Psychology*, *66*(5), 688-701.

Schmader, T. (2010). Stereotype threat deconstructed. *Current Directions in Psychological Science*, *19*(1), 14-18.

Schmader, T. (2012). *Stereotype threat: Theory, process, and application*. Oxford University Press.

Schmidt, F. T., Fleckenstein, J., Retelsdorf, J., Eskreis-Winkler, L., & Möller, J. (2019). Measuring grit: A German validation and a domain-specific approach to grit. *European Journal of Psychological Assessment*, *35*(3), 436-447.

Schmidt, F. T., Lechner, C. M., & Danner, D. (2020). New wine in an old bottle? A facet-level perspective on the added value of Grit over BFI–2 Conscientiousness. *PloS One*, *15*(2), e0228969.

Skinner, E. A., Wellborn, J. G., & Connell, J. P. (1990). What it takes to do well in school and whether I've got it: A process model of perceived control and children's engagement and achievement in school. *Journal of Educational Psychology*, *82*(1), 22.

Stone, J., Lynch, C. I., Sjomeling, M., & Darley, J. M. (1999). Stereotype threat effects on Black and White athletic performance. *Journal of Personality and Social Psychology*, *77*(6), 1213-1227.

Steele, C. M., & Aronson, J. (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, *69*(5), 797.

Steele, C. M. (1997). A threat in the air: How stereotypes shape intellectual identity and performance. *American Psychologist*, *52*(6), 613.

Steele, C. M., Spencer, S. J., & Aronson, J. (2002). Contending with group image: The psychology of stereotype and social identity threat. *Advances in Experimental Social Psychology, 34*, 370-440.

Stecher, B. M., & Hamilton, L. S. (2014). Measuring Hard-to-Measure Student Competencies: A Research and Development Plan. *Research Report*. RAND Corporation.

Sturman, E. D., & Zappala-Piemme, K. (2017). Development of the grit scale for children and adults and its relation to student efficacy, test anxiety, and academic performance. *Learning and Individual Differences*, *59*, 1-10.

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L., & Macnamara, B. N. (2018). To what extent and under which circumstances are growth mind-sets important to academic achievement? Two meta-analyses. *Psychological science*, *29*(4), 549-571.

Todd, P. E., & Wolpin, K. I. (2003). On the specification and estimation of the production function for cognitive achievement. *The Economic Journal*, *113*, F3-F33.

Tough, P. (2012). *How children succeed: Grit, curiosity, and the hidden power of character*. Houghton Mifflin Harcourt.

Usher, E. L., Li, C. R., Butz, A. R., & Rojas, J. P. (2019). Perseverant grit and self-efficacy: Are both essential for children's academic success? *Journal of Educational Psychology*, *111*(5), 877-902.

Walton, G.M. and Cohen, G.L. (2003) Stereotype lift. *Journal of Experimental Social Psychology*, 39(5), 456-467.

Walton, G. M. & Wilson, T. D. (2018) Wise interventions: Psychological remedies for social and personal problems. *Psychological Review, 125*, 617-655.

Wigfield, A. (1994). Expectancy-value theory of achievement motivation: A developmental perspective. *Educational Psychology Review*, 6(1), 49-78.

Wigfield, A., & Eccles, J. S. (2002). The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. *Development of achievement motivation*, 91-120.

Willingham, W. W. (1985). *Success in college: The role of personal qualities and academic ability*. New York: College Board Publications.

Willingham, D.T. (2009). *Why don't students like school? A cognitive scientist answers questions about how the mind works and what it means for the classroom*. John Wiley & Sons.

Wilson, T. D., & Linville, P. W. (1982). Improving the academic performance of college freshmen: Attribution therapy revisited. *Journal of personality and social psychology*, *42*(2), 367.

Wilson, T. (2011). *Redirect: The surprising new science of psychological change*. New York, NY: Little, Brown.

World Bank. (2019). Instilling a growth mindset in Indonesia. *eMBeD brief*. Washington, DC: World Bank.

Yeager, D. S., Hanselman, P., Walton, G. M., Murray, J. S., Crosnoe, J., Muller, C., & Tipton, E. (2019). A national experiment reveals where a growth mindset improves achievement. *Nature* 573, no. 7774, 364-369.

Yeager, D.S. & Dweck, C.S. (2012). Mindsets that promote resilience: When students believe that personal characteristics can be developed. *Educational Psychologist*, *47*(4), 302-314.

MANUSCRIPT TABLES AND FIGURES

TABLE I: OVERVIEW OF INTERVENTION

	Experimental Condition		
	Treatment 1	**Treatment 2**	**Control**
Target	Students	Students and teachers	No intervention
Delivery	Student self-learning	Teacher-delivered lessons	
Timing	1 lesson per week: First hour on Monday morning spent with headteacher	1 lesson per week: First hour on Monday morning spent with headteacher	
Teachers	Supervision of self-paced activities, minimal teacher involvement	Delivery of intervention contents by teacher following one-day teacher training	
Lessons	5 lessons: (1) "Introduction" (2) "Choose Challenge" (3) "Focus 100%" (4) "Seek Feedback" (5) "Reflect, Refine, Repeat"		
Data Collection	3 points (baseline, i.e. one week before start of intervention; endline, i.e. one week after end of intervention; additional data collection, i.e. two weeks after end of intervention), same timing as lessons		

TABLE II: SUMMARY STATISTICS AND BALANCING PROPERTIES

	Group			T-Test		
	Control	Treatment 1	Treatment 2	Difference	Difference	
	(1)	(2)	(3)	(1) - (2)	(1) - (3)	N
Panel A: Outcomes, Baseline						
Deliberate Practice Beliefs	16.580	16.640	16.740	-0.059	-0.156**	18,718
	(2.435)	(2.439)	(2.375)			
Grit	29.590	29.600	29.470	-0.006	0.121	18,718
	(4.072)	(4.050)	(4.066)			
Grit: Effort Facet	16.300	16.380	16.250	-0.074	0.049	18,718
	(2.499)	(2.471)	(2.518)			
Grit: Interest Facet	13.290	13.220	13.220	0.069	0.072	18,718
	(3.077)	(3.060)	(3.073)			
Frustration Reaction	11.042	11.138	11.062	-0.096	-0.020	18,718
	(2.533)	(2.510)	(2.539)			
Motivational Frameworks	20.836	20.895	20.912	-0.059	-0.076	18,718
	(2.900)	(2.886)	(2.817)			
Locus of Control	17.638	17.628	17.667	0.010	-0.029	18,718
	(2.314)	(2.332)	(2.317)			
Present Bias	4.208	4.207	4.221	0.001	-0.013	18,718
	(0.987)	(1.015)	(0.999)			
S/E Skills Index	0.0529	0.072	0.083	-0.019	0.031	18,718
	(0.997)	(1.005)	(1.010)			
GPAs	3.324	3.304	3.311	0.019	0.012	33,454
	(1.150)	(1.141)	(1.157)			
Panel B: Controls, Baseline						
Age	12.490	12.490	12.490	-0.001	-0.001	18,718
	(0.557)	(0.555)	(0.553)			
Female	0.507	0.517	0.523	-0.010	-0.016*	18,718
	(0.500)	(0.500)	(0.499)			
Sixth Grader	0.478	0.479	0.477	-0.001	0.001	18,718
	(0.500)	(0.500)	(0.499)			
Macedonian	0.668	0.723	0.705	-0.055	-0.037	18,718
	(0.471)	(0.448)	(0.456)			
Albanian	0.277	0.206	0.226	0.072	0.052	18,718
	(0.448)	(0.404)	(0.418)			
Roma	0.017	0.020	0.021	-0.003	-0.003	18,718
	(0.131)	(0.140)	(0.142)			
Other Ethnicity	0.037	0.052	0.049	-0.014	-0.012	18,718
	(0.189)	(0.221)	(0.215)			
TV at Home	0.957		0.945		0.012	10,355
	(0.204)		(0.228)			
PC at Home	0.959		0.955		0.005	10,355
	(0.197)		(0.208)			

Car at Home	0.868	0.870	-0.002	10,355
	(0.339)	(0.337)		
Family Goes on Vacation	0.707	0.684	0.023	10,355
	(0.455)	(0.465)		
Mother Lives at Home	0.975	0.969	0.007*	10,355
	(0.155)	(0.174)		
Father Lives at Home	0.952	0.943	0.009*	10,355
	(0.214)	(0.233)		
Mother College Educated	0.303	0.304	-0.001	10,355
	(0.460)	(0.460)		
Father College Educated	0.288	0.284	0.005	10,355
	(0.453)	(0.451)		

Notes: Standard deviations in parentheses. T-tests with robust standard errors clustered at the school level. Sample of students with non-missing information on all outcomes and controls (i.e. the sample we use to obtain our baseline results). All figures rounded to three decimal places. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE III: IMPACTS ON DELIBERATE PRACTICE BELIEFS, GRIT, AND GRIT FACETS (Z-SCORES)

	Deliberate Practice Beliefs	Grit	Grit: Effort	Grit: Interest
	(1)	(2)	(3)	(4)
Treatment 1	0.151***	-0.052***	0.052***	-0.116***
"Self-Learning"	*(0.018)*	*(0.019)*	*(0.020)*	*(0.019)*
Treatment 2	0.227***	-0.029	0.059***	-0.096***
"Teacher Delivery"	*(0.017)*	*(0.021)*	*(0.019)*	*(0.021)*
N	24,276	21,925	23,049	23,267
N Control	9,451	8,528	8,953	9,077
N Treatment 1	7,068	6,365	6,714	6,745
N Treatment 2	7,757	7,032	7,382	7,445
R^2	*0.320*	*0.334*	*0.331*	*0.223*

Notes: All regressions control for pre-treatment outcomes, demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE IV: IMPACTS ON S/E SKILLS INDEX AND DIFFERENT SETS OF SKILLS (Z-SCORES)

	Including All Skills	Excluding Deliberate Practice Beliefs and Grit	Including Only Deliberate Practice Beliefs and Grit
	(1)	(2)	(3)
Treatment 1	0.055**	-0.004	0.070***
"Self-Learning"	*(0.022)*	*(0.022)*	*(0.020)*
Treatment 2	0.128***	0.063***	0.133***
"Teacher Delivery"	*(0.021)*	*(0.021)*	*(0.020)*
N	18,718	18,718	18,718
N Control	7,286	7,286	7,286
N Treatment 1	5,424	5,424	5,424
N Treatment 2	6,008	6,008	6,008
R^2	*0.337*	*0.345*	*0.308*

Notes: All regressions control for pre-treatment outcomes, demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

TABLE V: IMPACTS ON S/E SKILLS INDEX AND SHORT-TERM GPAS (Z-SCORES)

		Gender		Ethnicity				Academic Year		Pre-Treatment Outcome		
	Average	Male	Female	Macedonian	Albanian	Roma	Other	Sixth Grade	Seventh Grade	Tercile 1 (Low)	Tercile 2 (Middle)	Tercile 3 (High)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: S/E Skills Index												
Treatment 1	0.055**	0.025	0.087***	0.065**	0.030	0.203	0.032	0.037	0.071***	0.043	0.040	0.078**
"Self-Learning"	*(0.022)*	*(0.029)*	*(0.024)*	*(0.026)*	*(0.050)*	*(0.177)*	*(0.087)*	*(0.029)*	*(0.027)*	*(0.030)*	*(0.032)*	*(0.032)*
Treatment 2	0.128***	0.080***	0.172***	0.140***	0.077*	0.417***	0.128	0.121***	0.136***	0.128***	0.110***	0.150***
"Teacher Delivery"	*(0.021)*	*(0.028)*	*(0.025)*	*(0.026)*	*(0.044)*	*(0.144)*	*(0.096)*	*(0.029)*	*(0.026)*	*(0.032)*	*(0.029)*	*(0.034)*
N	18,718	9,077	9,641	13,020	4,494	360	844	8,944	9,774	5,871	6,236	6,611
N Control	7,286	3,592	3,694	4,867	2,021	127	271	3,482	3,804	2,301	2,469	2,516
N Treatment 1	5,424	2,622	2,802	3,919	1,116	109	280	2,598	2,826	1,708	1,776	1,940
N Treatment 2	6,008	2,863	3,145	4,234	1,357	124	293	2,864	3,144	1,862	1,991	2,155
R^2	*0.337*	*0.305*	*0.347*	*0.355*	*0.205*	*0.329*	*0.429*	*0.334*	*0.350*	*0.122*	*0.072*	*0.184*
Panel B: GPAs, Short-Term (2015/2016 Q4)												
Treatment 1	0.018*	0.019*	0.017	0.019	0.017	-0.015	0.012	0.017	0.020	0.010	0.028*	0.014
"Self-Learning"	*(0.011)*	*(0.011)*	*(0.012)*	*(0.012)*	*(0.019)*	*(0.028)*	*(0.020)*	*(0.015)*	*(0.014)*	*(0.013)*	*(0.015)*	*(0.009)*
Treatment 2	0.016	0.013	0.019	0.007	0.020	0.055***	-0.007	0.014	0.019	0.020	0.015	0.011
"Teacher Delivery"	*(0.012)*	*(0.012)*	*(0.013)*	*(0.015)*	*(0.018)*	*(0.017)*	*(0.023)*	*(0.014)*	*(0.016)*	*(0.014)*	*(0.018)*	*(0.008)*
N	33,454	17,270	16,184	19,460	11,423	1,161	1,410	16,563	16,891	11,181	11,127	11,146

N Control	12,426	6,415	6,011	6,880	4,605	476	465	6,228	6,198	4,078	4,187	4,161
N Treatment 1	10,995	5,711	5,284	6,657	3,527	262	549	5,442	5,553	3,683	3,688	3,624
N Treatment 2	10,033	5,144	4,889	5,923	3,291	423	396	4,893	5,140	3,420	3,252	3,361
R^2	*0.935*	*0.934*	*0.930*	*0.930*	*0.925*	*0.913*	*0.953*	*0.931*	*0.942*	*0.671*	*0.592*	*0.518*

Notes: All regressions control for pre-treatment outcomes (in case of GPAs, up to a fourth-order polynomial), demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

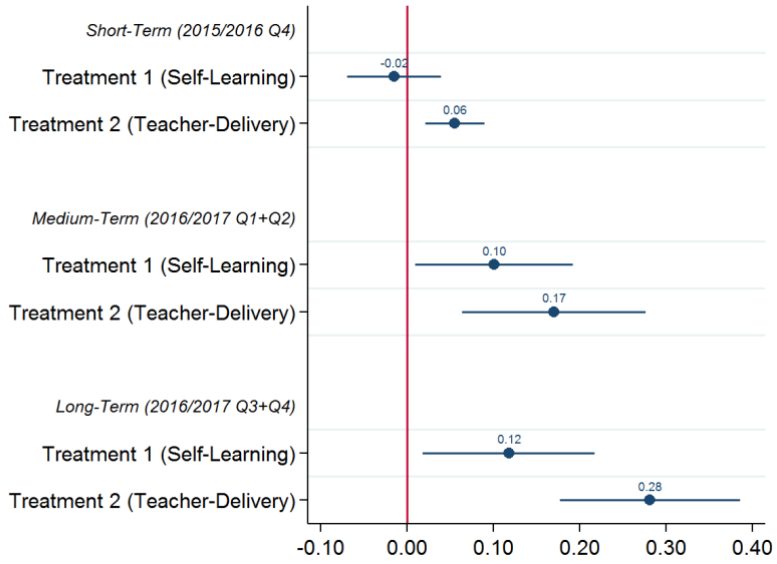
TABLE VI: IMPACTS ON LONGER-TERM GPAS (Z-SCORES)

		Gender		Ethnicity				Academic Year		Pre-Treatment Outcome		
	Average	Male	Female	Macedonian	Albanian	Roma	Other	Sixth Grade	Seventh Grade	Tercile 1 (Low)	Tercile 2 (Middle)	Tercile 3 (High)
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: GPAs, Medium-Term (2016/2017 Q1+Q2)												
Treatment 1	0.006	0.013	-0.002	0.015	-0.002	0.109**	0.052	0.020	-0.009	0.008	0.015	-0.006
"Self-Learning"	*(0.013)*	*(0.014)*	*(0.014)*	*(0.016)*	*(0.024)*	*(0.045)*	*(0.036)*	*(0.017)*	*(0.015)*	*(0.016)*	*(0.019)*	*(0.013)*
Treatment 2	0.009	0.023	-0.006	0.019	0.000	0.166***	0.014	0.022	-0.003	0.028*	0.015	-0.014
"Teacher Delivery"	*(0.013)*	*(0.014)*	*(0.014)*	*(0.017)*	*(0.020)*	*(0.056)*	*(0.046)*	*(0.018)*	*(0.015)*	*(0.017)*	*(0.019)*	*(0.012)*
N	31,310	16,154	15,156	18,568	10,348	1,045	1,349	15,697	15,613	10,338	10,423	10,549
N Control	11,600	5,992	5,608	6,533	4,190	426	451	5,881	5,719	3,716	3,395	3,949
N Treatment 1	10,166	5,265	4,901	6,310	3,107	238	511	4,996	5,170	3,391	3,391	3,384
N Treatment 2	9,544	4,897	4,647	5,725	3,051	381	387	4,820	4,724	3,231	3,097	3,216
R^2	*0.878*	*0.872*	*0.870*	*0.876*	*0.850*	*0.807*	*0.907*	*0.873*	*0.887*	*0.503*	*0.434*	*0.377*
Panel B: GPAs, Long-Term (2016/2017 Q3+Q4)												
Treatment 1	0.021	0.030	0.009	0.027	0.001	0.129**	0.045	0.032	0.010	0.007	0.043	0.004
"Self-Learning"	*(0.018)*	*(0.020)*	*(0.018)*	*(0.017)*	*(0.038)*	*(0.052)*	*(0.034)*	*(0.021)*	*(0.023)*	*(0.024)*	*(0.029)*	*(0.010)*
Treatment 2	0.030*	0.041**	0.018	0.042**	0.006	0.279***	0.034	0.043**	0.017	0.056**	0.043	-0.008
"Teacher	*(0.017)*	*(0.019)*	*(0.017)*	*(0.021)*	*(0.031)*	*(0.055)*	*(0.033)*	*(0.022)*	*(0.021)*	*(0.022)*	*(0.028)*	*(0.009)*

Delivery"												
N	31,437	16,209	15,228	18,716	10,402	985	1,334	15,713	15,724	10,247	10,502	10,688
N Control	11,404	5,881	5,523	6,573	4,038	360	433	5,813	5,591	3,585	3,893	3,926
N Treatment 1	10,461	5,424	5,037	6,362	3,340	238	521	5,185	5,276	3,479	3,497	3,485
N Treatment 2	9,572	4,904	4,668	5,781	3,024	387	380	4,715	4,857	3,183	3,112	3,277
R^2	*0.854*	*0.850*	*0.843*	*0.853*	*0.820*	*0.798*	*0.900*	*0.850*	*0.862*	*0.459*	*0.426*	*0.240*

Notes: All regressions control for pre-treatment outcomes (up to a fourth-order polynomial), demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places. * $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

FIGURE I: IMPACTS ON GPAS OF ROMA STUDENTS OVER TIME (Z-SCORES)



Notes: All regressions control for pre-treatment outcomes (up to a fourth-order polynomial), demographic controls (including dummies for age, gender, school year, and ethnicity), and municipality fixed effects. Robust standard errors are clustered at the school level. Confidence bands are 95%. See Tables V and VI for the respective regressions.