Policy Research Working Paper 9831

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

WORLD BANK GROUP
Poverty and Equity Global Practice
Social Protection and Jobs Global Practice
&
Education Global Practice
November 2021

Policy Research Working Paper 9831

Abstract

This paper studies whether a particular socio-emotional skill —grit (the ability to sustain effort and interest toward long-term goals)—can be cultivated and how this affects student learning. The paper implements, as a randomized controlled trial, a nationwide 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). Students exposed to the intervention report improvements in self-regulation, in particular the perseverance-of-effort facet of grit, relative to students in a control condition. The impacts on students are larger when both students and teachers are exposed to the curriculum than when only students are treated. Among disadvantaged students, the study also finds positive impacts on grade point averages, with gains of up to 28 percent of a standard deviation one year post-treatment. However, the findings also point toward a potential downside: although the intervention made students more perseverant and industrious, there is some evidence that it may have reduced consistency in their interests over time.

This paper is a product of the Poverty and Equity, Social Protection and Jobs, and Education Global Practices. It is part of a larger effort by the World Bank to provide open access to its research and make a contribution to development policy discussions around the world. Policy Research Working Papers are also posted on the Web at http://www.worldbank.org/prwp. The authors may be contacted at amunozboudet@worldbank.org.

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 [1]

JEL codes: C93, D91, I20, I24

Keywords: education, disadvantaged students, grit, metacognition, RCT

---

* Corresponding authors. The symbol ⓡ indicates authors' orders were randomized using the American Economic Association Author Randomization Tool (Confirmation Code: O6jB2s08FnXa).

Authors' affiliations: The World Bank; Ss. Cyril and Methodius University, Skopje, and Center for Human Rights and Conflict Resolution, Skopje; University College London, CEMMAP, IFS, FAIR-NHH; Kellogg School of Management at Northwestern University; The World Bank; CEDLAS-Universidad Nacional de la Plata; London School of Economics, CEP; The World Bank; University of Pennsylvania.

[1] The team is thankful to the Ministry of Education and Science in North Macedonia; the staff from The Center for Human Rights and Conflict Resolution; the team at PUBLIK; Bojana Naceva for her guidance, inputs, and being the main contact person to the government of North Macedonia; Jasminka Sopova for all her support during implementation; Robert Gallop for his support during the analysis; and the staff from the Character Lab for their support with the design of the intervention. The program also benefited from comments and inputs by Victoria Levin, Hillary Johnson, and Maria Davalos from The World Bank, Armin Falk, Fabian Schmidt, as well as seminar participants at the annual meeting of the American Economic Association in Atlanta in 2019 and the briq/IZA Behavioral Economics of Education workshop in Bonn in 2019. Carneiro gratefully acknowledges the financial support from the Economic and Social Research Council for the ESRC Centre for Microdata Methods and Practice through grant ES/P008909/1, and the support of 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.

1. 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 of these outcomes (Levin et al., 2016; Heckman and Kautz, 2014; Stecher and Hamilton, 2014; Naemi et al., 2013;; Tough, 2012; Willingham, 1985). Among 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 conscientiousness (Eskreis-Winkler et al., 2014, 2016; Maddi et al., 2012; Duckworth et al., 2007).

In cross-section samples, adults who score high on grit (who are 'gritty') make fewer career changes, progress further in their formal education, and earn higher GPAs (Duckworth et al., 2007). Prospective longitudinal research looking at grit and achievements finds that, for example, in education, grit has been found to improve school performance in standardized tests (Alan et al., 2019); predict school graduation (Eskreis-Winkler et al., 2014); and predict the ranking of students in high-stakes competitions such as the National Spelling Bee (Duckworth et al., 2007). These findings support a growing emphasis on integrating socio-emotional skills into formal education (OECD, 2015; Guerra et al., 2014).

Grit has a strong relationship with conscientiousness, and in a nomological network of conscientiousness, represents 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 effort and interest for long-term goals without losing focus) (Duckworth et al., 2007; Duckworth and Quinn, 2009). The former correlates strongly with the productiveness facet of conscientiousness, while the latter does not, suggesting grit adds to the construct (Schmidt, Lechner and Danner 2020).

Given its importance in determining critical human development outcomes, is it possible to cultivate grit in schools? Evidence from small-scale interventions (with limited sample sizes in single classrooms or schools) suggests that the answer is "yes" (Alan et al., 2019; Eskreis-Winkler et al., 2014). But can it be done at scale? To answer this question, we designed and implemented a nationwide intervention to teach grit to middle-school students in North Macedonia. In spring 2016, all sixth and seventh-grade students in North Macedonia were randomly allocated to receive either one of two grit-building treatments, low or high-intensity, or to make part of a control condition. In the high-intensity treatment, both students and teachers are exposed to the intervention, while the lower-intensity treatment relies on student self-paced learning. The intervention consisted of a curriculum that taught and motivated students to adopt the tenets of deliberate practice[3] (to identify stretch goals, get feedback, concentrate, and repeat until mastery), as well as that characteristics such as talent, gender, or ethnicity are not deterministic of achievement. We then measured their socio-emotional

[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] Ericsson (2008), Ericsson et al. (1993).

skills development using survey data and tracked their GPAs (up to one year post-treatment) via the official administrative school records in the country.[4]

Students exposed to the intervention report improved self-regulation relative to students in the control condition in the academic quarter immediately after the intervention. This improvement comes primarily from the perseverance-of-effort facet of grit. Impacts are larger in the high-intensity treatment than in the lower-intensity treatment. Impacts are larger for female and Roma students, a group that has been traditionally disadvantaged in North Macedonia. Among Roma students, gains in GPAs become stronger over time, reachingup to 28% of a standard deviation one year post-treatment. The stronger benefits for disadvantaged students echo similar findings in a number of other social-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]

While treated students report putting in more effort and being more industrious (i.e. the perseverance-of-effort facet of grit), they score lower on the consistency-of-interest facet of grit. Thus, a potentially unintended consequence of the intervention is that it diminished interest for long-term goals. One plausible explanation for this result is that by teaching students the value of deliberate practice applied repeatedly to the same stretch goal, the intervention reduced interest in long-term goals – a potentially unintended consequence of deliberate practice itself. This could also be reflective of the intervention not intentionally targeting a specific interest, as well as of the varied interests corresponding to the age of the children participating in the intervention.[6] It is also worth noting that separate research has shown that developing consistent interests (i.e., specializing) may, paradoxically, depend on diversification (i.e., sampling) of interests earlier in life.[7]

To cultivate grit, our intervention followed a two-pronged approach. First, drawing on the evidence on *deliberate practice* (Ericsson, 2008; Ericsson et al., 1993), students were taught to distinguish effective, evidence-based ways of studying from ineffective ones as well as strategies to implement these more effective ways of studying. Second, we aimed to motivate effort. To do this, we focused 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). In Expectancy-Value Theory, expectancy is the extent to which individuals believe they will succeed and, whereby value refers to the subjective value the individual attaches to a positive outcome. The latter increases in expected benefits from effort and decreases in expected costs. Expectancy and value are reliably associated with effort expenditure and achievement, particularly in academic settings (Nagengast et al., 2013; Eccles et al., 1993; Meece et al., 1990; Eccles et al., 1984). We expect students to be more likely to engage in sustained effort if they hold a strong expectancy that doing so would improve achievement, and if they attach a high subjective value to academic achievement. It is important to instill this expectancy – particularly among disadvantaged students – given the prevalence of false, deter-

---

[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.

[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 true 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 instance, Gopnik (2020) as well as Cote and Erickson (2015).

ministic beliefs surrounding achievement, for example stereotypical beliefs that ethnicity or gender determine academic performance and later-life outcomes. Similarly, in settings where information on returns to education and student performance may be incomplete, there may be significant scope for improving perceptions about the value of effort.

By teaching students effective ways to practice and by changing their beliefs about expectancies and values of effortful practice, our intervention aimed to raise take-up of deliberate practice, which – by repeatedly applying it – is expected to build grit (Duckworth et al., 2014; 2011). In addition, to address some of the pre-existing gaps in expectancy and values between disadvantaged and non-disadvantaged groups, our intervention covered not only the key steps of deliberate practice and how to follow them, but it was also designed to counter negative stereotypes surrounding ethnicity and gender in North Macedonia by providing positive role models and counter-stereotype examples. An intervention that undoes deterministic beliefs – by pointing out that practice, not ethnicity or gender, determines achievement – should be most helpful for students who hold these deterministic beliefs most strongly at the start of the intervention. Hence, by addressing some of the underlying causes of poor(er) school performance among disadvantaged students, our intervention aimed at reducing inequality in education outcomes, creating a more leveled academic playing field.

Our intervention, therefore, joins a wave of recent research aimed at alleviating inequalities through social-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 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, namely in Turkey. Alan et al. (2019), focusing on 52 schools in Istanbul, found that students exposed to a grit-building intervention performed better on an incentivized real effort task and in standardized tests than students in the control condition. We add to this study by making several key contributions.

First, our intervention tests a nationwide application of a grit curriculum in schools and is easily scalable within country systems. It does not introduce new technology and was rolled out within the education system as it was implemented (including teacher training) through regular channels used by the Ministry of Education in 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 program can be delivered and managed at scale within existing education systems.

Second, our intervention addresses concerns of external validity that often arise in the literature. In addition to overcoming external validity issues associated with the fact that most pilots do not use country service delivery systems, our design helps address concerns related to the scale of the program. Social-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 piloted and adjusted to ensure comprehension of content and materials (i.e. that students understood the materials and that examples resonated broadly), our intervention was targeted at the general student population in North Macedonia,

[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. Trainings took place at the scheduled times for teacher trainings in the country.

for example by catering to the two 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 the program and were randomly assigned to one of the two treatment groups or the control group, again strengthening external validity. Within each school, there was no bias in the selection of teachers, classrooms, or students. Therefore, any positive effects of the intervention cannot be attributed to students', teachers', or even administrators' pre-intervention interest in the curriculum. In the field of social-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 our intervention, including the different treatment arms and the key mechanism of behavior change. Section 3 gives an overview of our survey data on socio-emotional skills and administrative data on GPAs from official records in the country. Our empirical model is outlined in Section 4. The impacts of our intervention on socio-emotional skills and GPAs, on average and by different student sub-groups, as well as robustness checks are presented in the Section 5. Section 6 discusses them against findings in the literature, including cost-effectiveness, and then concludes.

2. The Intervention

The objective of our intervention was to cultivate grit among sixth and seventh-grade students 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-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). Hence, we examined whether grit can be built in a critical developmental window that has deep, enduring consequences for a student's future.

Our intervention covered all public schools with Macedonian and Albanian language of instruction schools with sixth and seventh grade classrooms and at least five students in single-level classrooms.[10] This amounts to a total of 352 schools and about 1,780 classrooms in 80 municipalities in the country.[11] The intervention was delivered nationwide, starting in the semester after the 2016 Christmas holidays (February) and ending by the Easter holidays (March to early May, depending on school-specific holidays), i.e. the third quarter of the school year 2015/2016. It consisted of a curriculum of five consecutive lessons, delivered weekly, which were divided into two parts.

The first part of the curriculum taught students the tenets of *deliberate practice* (Ericsson, 2008; Ericsson et al., 1993), namely to: (1) identify stretch goals, (2) get feedback, (3) concentrate, and (4) repeat until mastery. Didactic slides were interspersed with activity prompts, engaging images, and exercises such as letter-writing

[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.

[10] Among the children in North Macedonia, 98% attend public education which is free and compulsory.

[11] The country has 84 municipalities. Four municipalities and 64 schools were excluded from the sample. Eight (12%) of these schools are "special schools" (e.g., art or music schools, schools for children with special needs), 31 (48%) are too small (fewer than five students in either sixth or seventh grade), 20 (32%) are schools that do not have single-level classrooms (e.g. schools that combine different grades in one classroom, especially in remote areas), and five (8%) are Turkish language schools. The 64 excluded schools represent 11% of all classes in each grade (sixth and seventh) and 7% of the students in each grade. All schools are primary schools and use a standard curriculum.

to another student.[12] The aim was to familiarize students with deliberate practice and explain how it differs from less effective forms of practice. The second part of the intervention aimed at motivating students to actually do deliberate practice. To address expectancies (i.e. subjective probabilities of success), the materials taught that characteristics such as talent, gender, or ethnicity do not deterministically fix one's level of achievement. Rather, effort – and particularly, effort invested in deliberate practice – was important for what people can accomplish. By familiarizing students with the tenets of deliberate practice and by changing their expectancies and values attached to doing it (i.e. their beliefs), the curriculum aimed at encouraging students to take up deliberate practice and, in doing so, to cultivate grit and increase achievement.

Each of the five lessons in our curriculum built 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, including a take-away self-evaluation for students to assess how successful they were in implementing the lesson of the week.[13] The five lessons were delivered on consecutive weeks during the Monday morning class hour with the headteacher. This is the first class of each week across all schools in the country, and it is typically used to talk about general issues as well as to deliver selected contents from the "Life Skills" curriculum, into which our intervention was integrated. The overall "Life Skills" curriculum aims at teaching general life and civic skills, and our intervention was designed in such a way as to have a similar structure and format as the rest of this curriculum.

We designed two treatment arms: the first treatment arm was self-paced and relied entirely on student self-learning, that is, with minimal intervention of teachers. The second added a teacher training module and relied on teachers to deliver the intervention. The control condition received the existing "Life Skills" curriculum or did other activities at the discretion of the headteacher. Table 1 provides a summary of the intervention by experimental condition.

2.1 Treatment 1: "Student Self-Learning"

The first treatment arm consisted of the five lessons organized in weekly booklets that were distributed to students to work on their own. The lessons were paper-based and self-contained, each organized to take up to one school hour (about 45 minutes) to go over them. The treatment had minimal teacher involvement: teachers 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 the school administration, and they were given only generic information about the materials. Each week's materials 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, and space to report any unusual issue affecting the class during that hour, if any).

2.2 Treatment 2: "Teacher Delivery"

The second treatment arm had the same content as Treatment 1 but relied on the headteachers to deliver the lessons. It involved a one-day teacher training session prior to the start of the intervention on its contents. During this training, the teachers received materials to familiarize themselves with the relevant concepts included in the materials and a detailed class lesson package for the five weeks of sessions they were

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

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

expected to deliver. The package also included activities booklets for the students, which were the same as in the first treatment arm but without the self-paced content elements.

For all treatment arms and the control condition, the responsible teacher was the headteacher, who was assigned by the school at the start of the school year to each class.

Table 1: Summary 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 class hour on Monday	1 lesson per week: First class hour on Monday	
Role of Teachers	No teacher involvement	1 day teacher training	
Lessons	5 lessons: (1) "Introduction" (2) "Choose Challenge" (3) "Focus 100%" (4) "Seek Feedback" (5) "Reflect, Refine, Repeat"		
Data Collection	3 sessions (baseline, endline, and additional data collection)		

3. Data

We collected data on two categories of outcomes:

(1) Socio-emotional skills, which include *(i)* deliberate practice beliefs (as a manipulation check variable), *(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)* and locus of control (Skinner et al., 1990), and *(iv)* present bias. All outcomes were measured using tested and validated self-report scales, which were adapted to children from North Macedonia and translated (back-and-forth) to Macedonian and Albanian languages.[14]

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

Data on the different socio-emotional skills and students' socio-economic characteristics (which are otherwise not available in official records) were collected through baseline and endline surveys. To analyze whether the intervention shifted socio-emotional skills generally, and to reduce issues around dimensionality and multiple hypotheses testing, we construct a socio-emotional skills index (S/E skills index) that summarizes the different measures of self-reported socio-emotional skills. It combines the measures of deliberate practice beliefs, grit, frustration reaction, motivational frameworks, locus of control, and present bias. Following Anderson (2008), we construct 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 less weight while variables that are uncorrelated and hence represent new information receive more; and then *(iv)* standardizing the resulting index.

Data on GPAs come from official, administrative records held by the Ministry of Education in North Macedonia. We exploit all grades given to the students during the school years 2015-2016 and 2016/2017, the

[14] Survey questions used to collect baseline and endline data are available upon request.

year in which the intervention took place and the year after. As the intervention targeted both sixth and seventh-grade students, we focus on the set of core subjects that are common to both years: math, English, and the student's first language (which can be either Macedonian or Albanian). Arguably, these are also the subjects that are most decisive for later educational transitions, for example the transition from primary to secondary school, which occurs in North Macedonia after grade nine.

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 21 to August 31, 2016). Besides these short-run GPAs, we also calculate medium-run (first and second quarter of the school year 2016/2017) and long-run 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 start of the intervention. When calculating GPAs, we are agnostic and treat both written and oral grades – both of which are recorded – as equally relevant.[15] Furthermore, students, depending on the subject and school year (i.e. sixth or seventh grade), may differ in the number of exams or tests they take, and accordingly, in the number of grades they receive, as well as in the share of these that comes from written or oral exams.[16] Unfortunately, standardized tests were not available in the country at the time of the intervention. As with our socio-emotional skills measures, we standardize GPAs to have mean zero and standard deviation one (i.e. z-scores).

It should be noted that our 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 teachers. Our approach to construct GPAs is, therefore, agnostic, in that we exploit sharp cut-offs based on official school year and quarter beginning and end dates to determine whether a specific grade is pre or post treatment. This is a conservative approach, as pre-treatment grades based on exams or tests taken just before the start of the treatment quarter may only be recorded in the treatment quarter, and may thus be intermingled with post-treatment grades, deflating these and resulting in lower-bound estimates.[17] While this type of bias works in favor of positive results, 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 winter and summer holidays), and only a small fraction in-between. Our intervention, however, was implemented in the third quarter while our post-treatment GPAs are constructed from fourth-quarter grades only. That said, there was only a small fraction of assessments occurring just before the fourth quarter which could have spilled over. Finally, note that, even if spillover would occur, it is unlikely to cause systematic bias, as such spillover is likely to be balanced across groups due to randomization, unless teacher grade-reporting behavior has changed systematically between groups. However, we find little evidence for this.

We test the latter formally by comparing 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 ("student self-learning") 31% (σ=0.46) and those in the second treatment group ("teacher delivery") 28% (σ=0.45). Calculating normalized differences between the treatment groups and the control group to

[15] The type of 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 on request.

[16] Such differences are unlikely to bias our results as these are balanced between experimental groups.

[17] 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 very similar to our baseline results using the sharp cut-off dates of the school year calendar.

adjust for large group sizes (there are 2,469,524 recorded grades in the fourth quarter), none of the normalized differences exceed the recommended threshold of 0.25 (Imbens and Wooldridge, 2009).[18] Hence, the number of grades recorded by teachers is balanced between groups, suggesting no systematic change in teacher grade-reporting behavior.

Finally, we collected data on students' age, gender, and ethnicity, in order to routinely control for them throughout our regressions. Table 2 shows summary statistics and balancing properties for socio-emotional skills, GPAs, and student socio-economic characteristics at baseline, by experimental group.

Table 2: 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: Perseverance-of-Effort Facet	16.300	16.380	16.250	-0.074	0.049	18,718
	(2.499)	*(2.471)*	*(2.518)*			
Grit: Consistency-of-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)*			

[18] 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 ("student self-learning") and our control group is -0.13, that between our second treatment group ("teacher delivery") and our control group is -0.19.

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	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)*			

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

Notes: Standard deviations in parentheses. T-tests with robust standard errors clustered at the school level. Sample of students with non-missing information in all the socio-emotional variables analyzed (i.e. the sample of students used in the regressions analysis). All figures rounded to three decimal places.

Source: Own survey data, administrative data, school year 2015/2016, own calculations.

Table 2 Panel A focuses on students' socio-emotional skills taken from surveys and GPAs taken from official administrative records, whereas Panel B focuses on student's socio-economic characteristics. The sample includes all students with non-missing information on all outcomes and controls in the regression analysis of Table 3. As can be seen, in contrast to GPAs taken from official records, we encountered a significant amount of missing information on socio-emotional skills and characteristics, mainly due to some surveys

not being returned, being unreadable, or being only partly filled out. We will turn to the issue of attrition in our robustness checks. Note that grit is reported as a whole as well as split into its two facets: perseverance of effort and consistency of interest. With few exceptions, pre-treatment outcomes are balanced between groups, as we would expect from random assignment of schools to treatment arms.[19]

4. 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 from students in the treatment groups to those in the control group. With equal probability, schools (and all sixth and seventh-grader students therein) were allocated either to any of the 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 the following value-added specification (Todd & 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 dummies that equal one if the student belongs to treatment group one or two, respectively, and zero else; X_{it} is a vector of controls, including dummies for age, gender, grade (i.e. whether the student is in sixth or seventh grade), and ethnicity (i.e. whether the student is Macedonian, Albanian, Roma, or belongs to any other ethnicity). When evaluating impacts on GPAs, we also control for a fourth order polynomial in pre-treatment GPAs to account for non-linearities. 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. 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 with each other in terms of effect size.

Our regressors of interest are T_1 and T_2. Because of randomization, full eligibility, and full compliance (schooling up to including grade nine is compulsory in North Macedonia), 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.

We also analyze whether the intervention had heterogeneous impacts across gender and ethnicity, as well as across the pre-treatment student achievement distribution. One could expect improvements in grit and academic achievement to be stronger among sub-groups that are at a relative disadvantage, namely girls or ethnic minorities such as Roma or Albanian, or lower-performing students. In previous socio-emotional skills interventions in education, 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, among those groups (see Cohen et al. (2009), Good et al. (2003), or Yeager and Dweck (2012), for example). In terms of delivery method, we also expect heterogeneous impacts. In particular, we expect the teacher-delivered lessons to show larger impacts than student self-learning given that they ensure exposure to the content, can generate a more intense experience, and reflect more regular learning practices in classrooms in the country.

[19] 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.

5. Results

5.1 Impacts on Socio-Emotional Skills

Deliberate Practice Beliefs, Grit, and Grit Facets

We begin by looking at socio-emotional skills and the variables most likely to be affected by the intervention: deliberate practice beliefs and grit, including its different facets.[20] We first look at the average effect of treatment, and then at heterogeneous effects by gender, ethnicity, and prior student achievement.[21]

Recall that the intervention aimed at cultivating grit by informing students about deliberate practice and motivating them – through changing their beliefs about expectancies and values of deliberate practice – 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 interested and perseverant towards their long-term learning outcomes.

Column 1 of Table 3 shows the average effect of the respective experimental condition on deliberate practice beliefs, which is a summary scale combining four items that ask about the specific elements of deliberate practice. In particular, students are asked how important they believe it is, while studying, 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 he or she is certain to have absorbed it. As with all outcomes that follow, this summary scale is standardized.

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

Table 3: 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***
"Student 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*

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

[20] We present here a more refined set of covariates than pre-registered in our pre-analysis plan. The results using the exact, pre-registered covariates can be found in Appendix 4. These confirm the results presented here.

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

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 are clustered at the school level in parentheses. All figures are rounded to three decimal places.
Source: Own survey data, school year 2015/2016, own calculations.

Column 2 of Table 3 shows the average effect of the respective experimental condition on grit, which is our main outcome of interest. The finding is surprising and goes in the opposite direction of the hypothesized effect: it seems as if the intervention actually reduced students' self-reported grit, particularly in case of student self-learning (by about 5% SD).

Grit is measured using the Short Grit Scale (Duckworth and Quinn, 2009), which is a standard instrument commonly used in the literature. The eight items in the scale set out to capture the two facets of grit: perseverance of effort and consistency of interest.

When looking at these two facets of grit separately (Table 3 Columns 3 and 4), we see an interesting pattern: while the intervention increased the perseverance-of-effort facet of grit (in both treatment arms, with again slightly stronger effects in the teacher-delivery treatment, about +6% SD *versus* +5% SD), it reduced the consistency-of-interest facet (again in both treatment arms, by about -10% SD or more). Compared to students in the control group, treated students report being more industrious and hard-working but also report losing interest more quickly. Both facets combined then yield an insignificant (teacher-delivery) or even negative impact of the intervention (student self-learning treatment) on grit.

The divergence between the two facets of grit is puzzling. There could be several reasons behind this finding. First, some aspects of the Short Grit Scale may be less suited to the age group target of the intervention (that is, the Short Grit Scale being used in a younger cohort than it was originally developed for). For example, the concept of 'projects' may not be much applicable to the daily realities of sixth and seventh-grade students (who are between twelve and thirteen years old). Second, there may be a potential language translation problem, either in general or with respect to specific terms being used. For example, the term 'projects' has more than one meaning in Macedonian. Third, there may be a problem with reverse-coded items: the consistency-of-interest facet of grit is entirely captured by items that are reversely coded while the consistency-of-effort facet is not. It may be that, in the given country or target group, respondents might misinterpret how to answer these items.[22]

When looking at the correlation between grit-effort and grit-interest in the pre-treatment period, we find only a weak (yet significant) 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 similarly found a rather weak, negative correlation, i.e. -0.135, significant at the 5% level, in a large sample of Latin-American respondents).[23] So, while there may be some issues with the Short Grit Scale, these issues do not seem to be unique to our country context. Importantly, though, such issues by themselves do not necessarily explain why our intervention significantly *decreased* grit-interest for treated

[22] 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 (either in general or with respect to specific terms). Moreover, the Short Grit Scale (and all other instruments) was piloted in the country and with a random set of students in the targeted age group. The pilots also included qualitative discussions where none of the issues above came up. This gives us confidence in the correct application of the scale.

[23] We used the 2017 round of the Development Bank for Latin America (CAF) annual household survey of socio-demographic 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).

students (in both treatment arms) relative to students in the control condition (rather than, say, have no significant impact at all).

Perhaps a more plausible explanation for the negative effect of our intervention on grit-interest may be the content of the intervention itself, and in particular the deliberate practice component. 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 them to take up this particular form of practice. The rationale is that, when students take up this form of practice and experience a first "academic success" (which becomes more likely the more students practice), this will, in turn, reinforce their beliefs, leading to sustained behavior change and potentially long-run benefits. Deliberate practice consists of four elements: *(i)* identifying stretch goals, *(ii)* getting feedback, *(iii)* concentrating fully, and *(iv)* repeating until mastery. It is plausible that, when applied repeatedly to the same (or similar) stretch goal, the repetitive nature of these elements may crowd out interest in long-term goals – a potentially unintended consequence of deliberate practice itself. In fact, Duckworth et al. (2011), Ericsson (2006, 2007, 2009), and Ericsson et al. (1993) point towards negative "side effects" of deliberate practice, in the sense that this particular form of practice may be perceived as more unpleasant and 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 age of the children participating in the intervention.[24] Developing consistency of interest among young children may, paradoxically, require diversification (sampling) of interests earlier in life.[25]

In sum, we find that our intervention induced students in both treatment arms to update their beliefs about deliberate practice relative to students in the control condition, with stronger effects when teachers deliver the curriculum as opposed to student self-learning. Impacts on grit are mixed: while students in either treatment arm report more perseverance-of-effort aspects relative to students in the control condition, impacts on consistency-of-interest aspects are negative, pointing towards the possibility of unintended consequences, and in particular, the possibility of crowding out of long-term goals-orientation.

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) that combines the measures of deliberate practice beliefs, grit, frustration reaction, motivational frameworks, locus of control, and present bias, to reduce dimensionality and multiple hypotheses testing.[26]

But before proceeding to this analysis, we first have a look at the *relative* importance of different sets of skills, in particular deliberate practice beliefs and grit *versus* all others. Table 4a shows the impacts of our intervention on our 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).

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

[25] See, for instance, Gopnik (2020).

[26] Appendix 1 Table A1.2 replicates Table 3 and shows findings for frustration reaction, motivational frameworks, locus of control, and present bias.

Table 4a: Impacts on S/E Skills Index and Different Sets of Skills (Z-Scores)

	S/E Skills Index		
	Including All Skills	Excluding Deliberate Practice Beliefs and Grit	Only Including Deliberate Practice Beliefs and Grit
	(1)	(2)	(3)
Treatment 1	0.055**	-0.004	0.070***
"Student Self-Learning"	*(0.022)*	*(0.022)*	*(0.020)*
Treatment 2	0.128***	0.063***	0.133***
"Teacher Delivery"	*(0.021)*	*(0.021)*	*(0.021)*
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*

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

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 are clustered at the school level. All figures are rounded to three decimal places.

Source: Own survey data, school year 2015/2016, own calculations.

Table 4a shows that our intervention had a significant, positive impact on our S/E skills index on average (Column 1), with stronger impacts in the teacher-delivery treatment (about 13% SD) compared to student self-learning (about 6% SD). Disaggregating the index into the different sets of skills that are included, in particular deliberate practice beliefs and grit *versus* all others, we find that, while all other skills do play a role in the teacher-delivery treatment (Column 2), deliberate practice beliefs and grit are clearly the driving forces, being twice as important in terms of effect size as all other skills (which turn out to be entirely insignificant when it comes to student self-learning (compare Columns 2 and 3).

Next, we turn to our main analysis in Table 4b. Panel A shows the impact of our intervention on our 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), grade (Columns 8 and 9), and tercile in the pre-treatment S/E skills distribution, whereby the first tercile is the lower and the third tercile the upper end (Columns 10 to 12).

When it comes to heterogeneous treatment effects by gender, the impact of both treatments is positive for girls and for boys, but they are larger for girls. Moreover, treatment effects are slightly larger for students of Macedonian ethnicity (the largest ethnic group in the country) than for students of Albanian ethnicity (the second largest group). They turn out especially large for Roma students (the most disadvantaged group), for whom the impact of the teacher-delivered treatment is about 42% SD (Column 6). The finding for Roma students should be taken with caution, though: in this heterogeneity analysis, sample size drops substantially. Columns 8 and 9 show similar impacts of the intervention on sixth and seventh grade students, whereas Columns 10 and 12 show that impacts are larger for students who already a had higher level of S/E skills to begin with (those in the second and third terciles of the pre-treatment S/E skills distribution). We cannot reject that there are no statistically significant differences between groups reported in Columns 2 to 12.

In sum, when looking at all socio-emotional skills taken together, we find that the intervention has significant, positive impacts, for students on average and for different student sub-groups. On average, and across student sub-groups, the teacher-delivered treatment consistently delivers better results. The pattern of impacts on sub-groups is broadly in line with the literature, which documents that students who are at relative disadvantage (i.e. girls and Roma students in our case) benefit relatively more (cf. Sisk et al., 2018).

Roma are a particularly disadvantaged group across the Western Balkans. In North Macedonia, ethnicity, which also determines language and religion, is a socially dividing line and a frequent cause of inter-ethnic conflict, both in the country and across borders with neighboring Albania. Roma (who make up only 3% of the population), are socially and economically marginalized, and subject to discrimination in many areas of 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, our intervention targeted minorities (i.e. Albanian and Roma students) in particular. However, it cannot be ruled out that in the teacher-delivered treatment in which teachers received training on the principles behind the curriculum this has led to a more effective delivery of contents to Roma students. In that sense, our intervention seems to reduce educational inequalities across ethic lines. At the same time, however, our finding that students at the upper end of the pre-treatment S/E skills distribution benefit more than those at the lower end suggests that reducing educational inequalities may be more difficult.

5.2 Impacts on GPAs

Short-term GPAs

Did improvements in middle-school students' general S/E skills translate into improvements in students´ academic achievement? Table 4b Panel B shows impacts on short-run GPAs, calculated immediately after the intervention in the fourth quarter across math, English, and first language. Note that the sample size is larger here because GPAs are obtained from official, administrative records and are therefore not subject to missing information.

Overall, we find little evidence of impacts of our intervention 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 the 10% level in the student self-learning treatment (with a similar point estimate and standard error in the teacher-delivered treatment). Likewise, there is little evidence for differential impacts by student sub-group: impacts are similar between gender, sixth and seventh-grade students, 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-delivered treatment. This is interesting because Roma students are the sub-group of students for whom we find the strongest impacts in terms of our S/E skills index (about 42% SD, cf. Table 4b Panel A). There is thus consistency between the finding for our S/E Skills index and the finding for GPAs of Roma students. The effect size for GPAs, however, is rather small, at about 6% SD.

Our estimates on GPAs are likely downward biased, for several reasons: first, as noted previously, there could be a lag between students taking exams or tests and teachers reporting their results. Although teachers are encouraged to report results to official records within two weeks after taking the assessment, we cannot exclude the case that some results from tests taken during the pre-treatment period (i.e. before the intervention had finished or even started) are reported only during the post-treatment period, deflating post-treatment GPAs and resulting in lower-bound estimates. Second, there is the possibility that teachers engage in "grading on a curve". Although this is not common practice in North Macedonia, we cannot exclude the case that single teachers do behave that way. Again, this kind of behavior would bias our estimates downwards, which, as in the previous case, works in our favor. Third, impacts on GPAs may need some time to materialize. We look at this next.

Table 4b: Impacts on S/E Skills Index and GPAs in Short-Run (Z-Scores)

		Gender		Ethnicity				Grade		Pre-Treatment Outcome		
	Average	Male	Female	Macedo-nian	Alba-nian	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: S/E Skills Index												
Treatment 1	0.055**	0.025	0.087***	0.065**	0.030	0.203	0.032	0.037	0.071***	-0.015	0.053*	0.117***
"Student Self-Learning"	*(0.022)*	*(0.029)*	*(0.024)*	*(0.026)*	*(0.050)*	*(0.177)*	*(0.087)*	*(0.029)*	*(0.027)*	*(0.031)*	*(0.030)*	*(0.029)*
Treatment 2	0.128***	0.080***	0.172***	0.140***	0.077*	0.417***	0.128	0.121***	0.136***	0.024	0.150***	0.176***
"Teacher Delivery"	*(0.021)*	*(0.028)*	*(0.025)*	*(0.026)*	*(0.044)*	*(0.144)*	*(0.096)*	*(0.029)*	*(0.026)*	*(0.030)*	*(0.030)*	*(0.031)*
N	18,718	9,077	9,641	13,020	4,494	360	844	8,944	9,774	4,745	6,511	7,368
N Control	7,286	3,592	3,694	4,867	2,021	127	271	3,482	3,804	1,831	2,577	2,861
N Treatment 1	5,424	2,622	2,802	3,919	1,116	109	280	2,598	2,826	1,473	1,873	2,060
N Treatment 2	6,008	2,863	3,145	4,234	1,357	124	293	2,864	3,144	1,441	2,061	2,447
R^2	*0.337*	*0.305*	*0.347*	*0.355*	*0.205*	*0.329*	*0.429*	*0.334*	*0.335*	*0.221*	*0.280*	*0.361*
Panel B: GPAs, Short-Run (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
"Student 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*	*0.930*	*0.942*	*0.671*	*0.592*	*0.518*

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

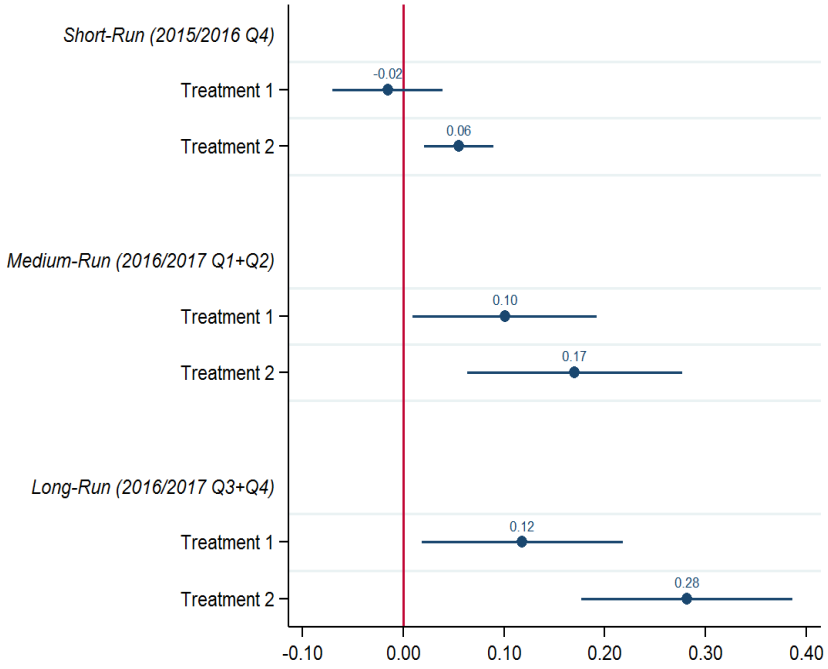
Notes: All regressions control for pre-treatment outcomes, demographic controls (including dummies for age, gender, academic year, and ethnicity), and municipality fixed effects. Robust standard errors are clustered at the school level in parentheses. All figures are rounded to three decimal places.
Source: Own survey data, administrative data, school year 2015/2016, own calculations.

Long-term GPAs

We are able to track the GPAs of students in our two treatment groups and our control group over time. Table 5 looks at medium-run GPAs (i.e. half a year later in the second quarter of the school year 2016/2017) and long-term GPAs (i.e. one year later in the fourth quarter of the school year 2016/2017).

We find evidence that impacts on GPAs become stronger over time, on average as well as for particular student sub-groups. Mostly, impacts become visible one year later, with effect sizes still being rather small, hovering between 3% SD and 5% SD depending on sub-group. It should be noted that the pattern of impacts for long-term GPAs differs from that for general S/E skills: small impacts on GPAs seem to materialize over time for males (rather than females) and students at the lower end of the pre-treatment GPA distribution (rather than at the upper end). A notable exception are again Roma students, for whom we find a consistent gradient in GPA improvement over time, which is again stronger in the teacher-delivered treatment as opposed to student self-learning: medium-run and long-run GPAs increase by about 17% and 28% SD, respectively, up from 6% SD immediately after the intervention. Figure 1 shows this improvement of GPAs for Roma students over time.

Figure 1: Impacts on GPAs of Roma Students Over Time (Z-Scores)



Notes: All regressions control for pre-treatment outcomes, 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 Table 4 for the regression results.
Source: Administrative data, school year 2016-2017, own calculations.

As our medium-term and long-term GPAs measure impacts one semester and two semesters post-treatment, students should, at this point, have passed half or the entire eighth grade, at maximum, and still be in primary education. In theory, we would not expect the materializing impacts on GPAs over time to be driven by attrition (out-of-sample selection), with only "good" students remaining in the sample and "bad" students dropping out of formal schooling. In a robustness check in which we regress the likelihood to have a record for a medium-run or long-run GPA on ethnicity, among others, we find that Roma students are between seven and ten percentage points less likely than Macedonian students to have a medium-run or

long-run GPA (Appendix Table A4.2). This is an interesting finding in itself, pointing towards either an inconsistent use of official records in the country, true dropout behavior, migration, or a combination of these. To further look into how attrition affects our findings, we re-estimate Table 5 using a balanced sample, including only students who are observable in official records during the entire period from school year 2015/2016 to school year 2016/2017 (Appendix Table A4.3). 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 materializing impacts on GPAs over time.

In sum, for Roma students, we find a consistent pattern: large impacts on socio-emotional skills are mirrored by very small impacts on GPAs immediately after the intervention, which turn into medium impacts half a year later and medium to large impacts one year after. Such a pattern of emerging impacts over time might point towards sustained behavior change and that students reap the benefits of such behavior change over time only, at least when it comes to academic achievement as reflected in school GPAs.

Which Roma students benefited the most? To answer this question exploratorily (noting that sample sizes are small at times), we dig deeper into our finding on Roma students and extend our analysis on GPAs over time by looking at heterogeneous treatment effects (Appendix 3 Table A3.1). There is some evidence that, while both male and female Roma students benefited, female students benefited more in the long-run. The difference in GPAs between male and female students, however, cannot be rejected at a conventional level of statistical significance. When it comes heterogeneous treatment effects by pre-treatment GPAs, we find that students in higher terciles clearly benefited more; for students in the upper tercile, the intervention had a positive effect on GPAs at the 1% level, increasing GPAs by about 8% SD in the short-run, 21% in the medium-run, and 46% in the long-run in the teacher-delivered treatment (effects are also significant, although lower, in the student self-learning treatment). Finally, when estimating seemingly unrelated regressions for all results on Roma students (on average, by gender, and by student achievement), we can reject the null that the coefficients of the student-teacher treatment are jointly equal to zero across models at the 1% level.[27]

5.3 Robustness

Attrition

So far, we have a different number of students in our analysis of socio-emotional skills and our analysis of GPAs, the difference arising from attrition in our survey data, mainly due to some surveys not being returned, being unreadable, or being only partly filled out. Our analysis of socio-emotional skills, therefore, includes all students with complete surveys, whereas our analysis of GPAs includes all students who are present in official, administrative records by the Ministry of Education in North Macedonia (and should have participated in the intervention according to the randomization routine). This approach is agnostic as it simply takes all observations that are available. It is also conservative, in the sense that our estimates on GPAs are again likely to be downward biased, as some, potentially non-treated, students may be allocated to either of the two treatment groups.

[27] A more complex issue arises in case that the intervention increased motivation and effort on the side of teachers such that teachers in any of the two treatment groups (and potentially differently by treatment group) recorded GPAs more swiftly post-treatment. Graphical evidence, however, suggests that the temporal distribution of when GPAs are recorded is similar between any of the two treatment groups and the control group.

Table 5: Impacts on GPAs Over Time (Z-Scores)

		Gender		Ethnicity				Academic Year		Pre-Treatment Achievement		
	Average	Male	Female	Macedo-nian	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, Medium-Run (2016/2017 Q1+Q2)												
Treatment 1	0.006	0.013	-0.002	0.015	-0.002	0.109**	0.052	0.020	-0.009	0.009	0.013	-0.005
"Student Self-Learning"	*(0.013)*	*(0.014)*	*(0.014)*	*(0.016)*	*(0.024)*	*(0.045)*	*(0.036)*	*(0.017)*	*(0.015)*	*(0.016)*	*(0.019)*	*(0.012)*
Treatment 2	0.009	0.023	-0.006	0.019	0.000	0.166***	0.014	0.022	-0.003	0.029*	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)*	*(0.011)*
N	31,310	16,154	15,156	18,568	10,348	1,045	1,349	15,697	15,613	10,437	10,552	10,321
N Control	11,600	5,992	5,608	6,533	4,190	426	451	5,881	5,719	3,759	3,970	3,871
N Treatment 1	10,166	5,265	4,901	6,310	3,107	238	511	4,996	5,170	3,417	3,438	3,311
N Treatment 2	9,544	4,897	4,647	5,725	3,051	381	387	4,820	4,724	3,261	3,144	3,139
R^2	0.878	0.872	0.870	0.876	0.850	0.807	0.907	0.873	0.887	0.502	0.436	0.364
Panel B: GPAs, Long-Run (2016/2017 Q3+Q4)												
Treatment 1	0.021	0.031	0.009	0.027	0.001	0.129**	0.045	0.032	0.010	0.010	0.040	0.005
"Student Self Learning"	*(0.018)*	*(0.020)*	*(0.018)*	*(0.017)*	*(0.038)*	*(0.052)*	*(0.034)*	*(0.021)*	*(0.023)*	*(0.024)*	*(0.029)*	*(0.009)*

Treatment 2	0.030*	0.041**	0.018	0.042**	0.006	0.279***	0.034	0.044**	0.017	0.058***	0.040	-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	10,480	10,498	10,459
N Control	11,404	5,881	5,523	6,573	4,038	360	433	5,813	5,591	3,667	3,888	3,849
N Treatment 1	10,461	5,424	5,037	6,362	3,340	238	521	5,185	5,276	3,560	3,489	3,412
N Treatment 2	9,572	4,904	4,668	5,781	3,024	387	380	4,715	4,857	3,253	3,121	3,198
R^2	*0.854*	*0.850*	*0.843*	*0.853*	*0.820*	*0.798*	*0.900*	*0.850*	*0.862*	*0.464*	*0.420*	*0.226*

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

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 are clustered at the school level in parentheses. All figures are rounded to three decimal places.
Source: Administrative data, school year 2016/2017, own calculations.

Appendix 2 Tables A2.1 to A2.3 re-estimate our previous tables by including only those students for whom we have completed surveys and who we know for sure that they did participate in the intervention. The results are quite similar to our previous findings, and the estimates obtained from using a balanced sample are indeed often larger in size than those obtained from using an unbalanced sample, in line with a lower-bound interpretation of our previous findings. Note that, for Roma students, impacts on GPAs become twice as large but significance drops from 1% to 10% only, most likely because the sample size drops substantially and the standard error becomes much larger.

Appendix 4 Table A4.1 shows the results of an attrition test for our survey data in which we regress the availability of the different survey instruments (i.e. having a baseline or endline survey, or both) on dummies for the two treatment groups, pre-treatment GPAs, and socio-demographic characteristics. Appendix 4 Table A4.2 conducts the same attrition test for our administrative data. The results show, for the survey data, that having better GPAs is associated with a higher likelihood to have completed the different survey instruments. However, the effects are rather small: increasing GPAs by 1% SD increases the likelihood to have complete either or both survey instruments by between 1% and 2%. There are notable differences in survey attrition between our first treatment group (student self-learning) and our second treatment group (teacher delivery), as well as between the first and the control group. In fact, the first group is less likely to have returned the endline survey by about 15%. This is most likely related to a printing error: the print work of the surveys for all sixth and seventh-grade students in the country was shared among three survey firms, and one of these firms printed the endline survey twice (instead of both baseline and endline survey) for about one third of the first group. This is also the reason why a large share of socio-demographic controls is missing (cf. Table 1), as these were captured in the baseline survey. While this is likely to introduce noise only due to the reduced sample size, impacts on the first group should nevertheless be interpreted with this point in mind.

With regards to the attrition test for the GPAs, things look different. Since schooling is mandatory, there should be medium-run and long-run GPAs available for each and every student. However, we can see that some students do drop out: a higher GPA reduces the likelihood to drop out by about 1%, and being in any treatment group by between 4% and 5%, respectively. This suggests that being exposed to treatment may actually *reduce* the likelihood to drop out of formal schooling relative to the control group. As described above, when looking only at students who are observable during the entire observation period using a balanced sample, our results remain robust.

Observer Effects

It could be the case that some, if not all, of the positive impacts of our intervention come about as observer (Hawthorne) effects. Teachers and students, both of whom are not blind to the experiment, may change their behavior as a result of being part of an experiment, rather than due to the actual contents being taught in the experiment. In case of students, observer effects may pertain both to self-reporting behavior (e.g. social desirability to satisfy the expectations of the experimenters) or to genuine learning behavior, motivated by feelings of being cared about and taken seriously. In the case of teachers, observer effects may pertain mostly to teachers treating students differently or to changing grading behavior. In both cases, impacts come about as artifacts of the experiment rather than genuine impacts of the contents learned.

Although we cannot exclude observer effects for either teachers or students with certainty, we argue that they are unlikely to be the main driver behind our findings. When it comes to students, note that the intervention was embedded into an existing "Life Skills" curriculum. The students were familiar with this curriculum being taught during Monday morning class hours (yet without expectation of upcoming content) and we made sure that our intervention resembled its basic structure and appeal (in fact, our local psychologist designed the "Life Skills" curriculum some years ago). Moreover, our curriculum was rather light, in the sense that it consisted only of five content sessions. Finally, baseline and endline data were collected before and after these sessions, with a timely spacing, respectively. There were never experimenters or intervention

facilitators present. Students were not monitored either, neither within sessions nor outside. Our intervention should thus not have been particularly salient among students, hence minimizing observer effects.

When it comes to teachers, observer effects are particularly relevant in the teacher-delivered treatment and impacts on GPAs, as the role of teachers in the student self-learning treatment was minimal. This begs the question of whether the teacher-delivered treatment was indeed more effective for students, either because *(a)* students simply absorbed more contents, *(b)* teachers themselves reflected on these contents and treated students in improved ways, or *(c)* teachers wary of being part of an experiment – simply changed their grading behavior. While *(a)* and *(b)* are arguably part of the genuine treatment effects of our intervention, *(c)* is an experimental artefact. Again, although we cannot exclude *(c)* with certainty, we argue that it is unlikely to be the main driver behind our findings. Recall that the "treated" teachers are the headteachers of the respective class, typically teaching one subject (which can vary depending on the subject focus of the headteacher). Our GPAs are, however, constructed across several subjects (math, English, and first language, i.e. Macedonian and Albanian), reducing the relative importance of the headteacher's subject when calculating average GPAs and hence, technically, minimizing observer effects. When recalculating our impacts for GPAs taken across all subjects (not only math, English, or first language), we obtain similar results as in our main findings.[28]

Replication Using Pre-Registered Controls Only

For all results, we run two sets of additional sensitivity analyses. The first modifies the original models using the pre-registered controls only (Appendix A4.2). The second applies a parametric bootstrap modeling of the same data using, likewise, the pre-registered controls. For the short-term results the sensitivity analysis using the pre-registered controls only confirms our findings obtained with the extended set. The parametric bootstrap modeling behaves well on all outcomes, except GPAs, with the bootstrap estimates hovering around null effects. 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 largely confirms our previous findings.

Multiple Hypotheses Testing

We examined the robustness of our results regarding multiple hypotheses testing using the stepwise p-value correction described in Romano and Wolf (2005, 2016). Accounting for 24 hypotheses in our socio-emotional skills analysis and 28 hypotheses in our GPA analysis (which include the medium-run and long-run impacts, which we only have for GPAs), we find that, for socio-emotional skills, the following impact estimates remain statistically different from zero with a significance level of 10%: the average impact of the teacher-delivered treatment, the impact of the same treatment for females, the impact of both the student self-learning and teacher-delivered treatments for seventh-grade students (the latter also statistically significant at the 5% level), and the impact of teacher-delivered treatment for students in the second tercile of the pre-treatment S/E skills distribution. For GPAs, the long-term impact on Roma students one year post-treatment also remains statistically significant at the 5% level.

[28] The results are available on request.

6. Discussion and Conclusion

To the best of our knowledge, this paper is the first rigorous examination of a nationwide school-delivered program aimed at improving socio-emotional skills – in particular grit (the ability to sustain effort and interest towards long-term goals) – among middle-school students in sixth and seventh grades. We implemented a 5-week grit intervention in North Macedonia, delivered through the regular school curriculum during the second semester of the 2015-2016 academic year. Implemented as 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. The lessons focused on deliberate practice and – by changing students' beliefs about expectancies and values (both are antecedents of effortful behavior) of engaging in deliberate practice for learning outcomes – motivated students to take up this particular form of practice. This, in turn, raises the chances of attaining these learning outcomes, which then reinforces beliefs and leads to sustained behavior change and positive long-run benefits. The intervention materials were gender and ethnicity neutral, hence counteracting prevailing stereotypes around gender and ethnicity.

We find that the intervention significantly increased socio-emotional skills and, to some extent, academic achievement among treated students, particularly the more disadvantaged Roma students. In terms of socio-emotional skills, both treatments had significant positive impacts, with higher effects found when content was delivered by teachers. In this latter case, the average student saw an improvement in socio-emotional skills of 13% SD, while Roma students saw an increase of up to 42% SD. In terms of grades, effects were measured at three stages: short-term (same semester and school year of the intervention), medium-term (first semester of the 2016-2017 school year), and long-term (second semester of the 2016-2017 school year, that is, one year after completing the program). Across all three periods disadvantaged minority students – particularly Roma students under the teacher-delivered treatment – saw the most gains in terms of grades (up to 28% SD in the long-term). For Roma students, changes in GPAs are significant and positive in all observation periods, with the impact doubling every semester. Back-of-the-envelope, these achievement gains observed are equivalent to the gains normally associated with three weeks of extra school learning.

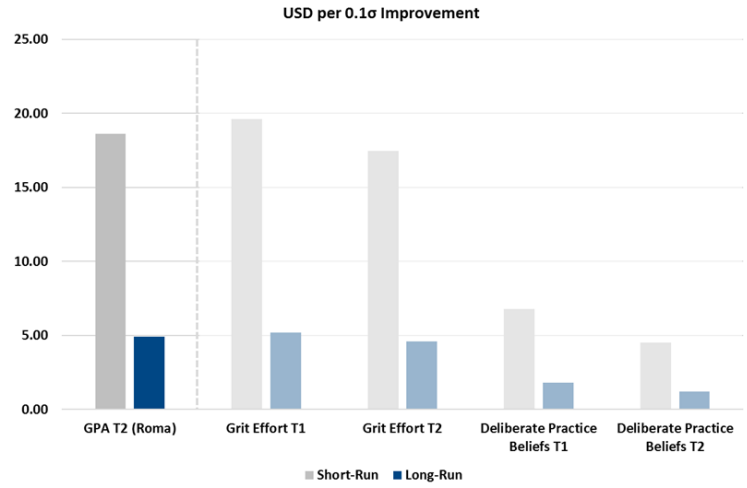
However, we also found that while the intervention increased the perseverance-of-effort facet of grit, it reduced the consistency-of-interest facet. Compared to students in the control group, treated students report being more industrious and hard-working but also report losing interest more quickly. Both facets combined then yield an insignificant (teacher-delivery) or even negative impact of the intervention (student self-learning treatment) on grit. One plausible explanation is that by teaching students the value of deliberate practice applied repeatedly to the same stretch goal, the intervention reduced interest in long-term goals, also considering that the intervention did not target nurturing a specific interest. Developing consistency of interest among young children may, paradoxically, require diversification (sampling) of interests earlier in life. This is an area that deserves further psychological and empirical research.

The impacts of the intervention confirm, thus, not only that it is possible to teach and improve self-regulation and grit among primary school students across the education system, but also that doing so can have positive impacts on GPAs, with impacts possibly increasing over time. That the impacts are particularly high, on average, among the Roma, further indicates the potential for this type of intervention to support school learning in ways that may improve equity in educational outcomes, although heterogeneous effects based on pre-treatment academic achievement (where already higher-performing students may profit more from such an intervention) suggest that improving equity in education by means of intervention may not be as straightforward as one may think. The magnitude of impacts found in this intervention compare favorably to other education interventions focused on improving socio-emotional skills, and are consistent with mounting evidence that grit and growth mindset interventions often benefit disproportionally (and sometimes only) disadvantaged students (Sisk et al., 2018). In terms of academic achievement, our results for

Roma students are comparable with those found in a metanalysis of impacts of socio-emotional skills interventions among disadvantaged groups (Sisk et al., 2018; 34% SD) and are higher than impacts among disadvantaged students in other programs: 18% SD in a math standardized test in Indonesia, although an expanded program showed no impacts on grades (Johnson et al., 2020; World Bank, 2019); and 10% SD in Peru's Expande Tu Mente program (Outes et al 2020). Apart from these studies, Alan and Ertac (2019) find average effects of 23% SD in math tests 2.5 years after a similar intervention that targets grit in participating schools in Turkey.

The intervention was also relatively cost-effective. When looking at GPAs for Roma only and allocating costs accordingly (the common cost-effectiveness ratios in educational economics are GPAs per USD spent in the intervention), we find that, in the short-run (including all cost categories), the intervention cost 18.6 USD for a 0.1 SD increase in GPAs for Roma students. After one year, this translates into 11 USD per SD improvement. Excluding design and impact evaluation costs (Figure 2), the numbers are even lower at 5 USD per 0.1 SD improvement in the short-run. Compared to the existing literature, this is favorable. Glewwe and Muralidharan (2016) find that incentive schemes (both for students and for teachers) cost between 1 USD and 3 USD per 0.1 SD improvement in test scores, CCTs between 77 USD and 138 USD for a comparable improvement, and IT in classroom (to support pedagogy) about 30 USD per 0.1 SD. However, it is difficult to compare costs between countries (for example, the bulk of the costs in the case of our intervention was on printing and distribution of materials), as some costs will vary depending on geography, and so on. Importantly, cost structures and especially opportunity costs of education will vary substantially between countries.

Figure 2: Intervention Cost-effectiveness in Long-Run (Excluding Costs of Design and Evaluation)



While the results from this intervention are promising, impacts and their heterogeneity across treatments and students suggest that important questions remain open in terms of how to foster grit among students. In particular, there are remaining questions in terms of the exact mechanisms that drive our results. In this paper, albeit not conclusively, we examine some potential avenues for impact. Gritty students do more deliberate practice and work harder for a longer period 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, which would also be consistent with the higher impacts found among Roma students. Our paper also highlights the importance of delivery mechanisms, and suggest that more intense methods – particularly involving teachers or other individuals to deliver the content – may be more impactful and still cost-effective at scale (consistent with Alan and Ertac, 2019).

In terms of grit in particular, we find impacts on the perseverance elements of grit but not on the element related to consistency of interest, suggesting – consistent with recent literature – that targeting consistency of interest may not be an effective way of cultivating grit among younger students. In fact, we show that the intervention may even crowd out interest – a potentially unintended consequence of the particular behavior our intervention tries to engrain. While this may be a point that is unique to deliberate practice and our intervention more generally, it does point towards the importance of measuring outcomes of social-psychological interventions (or literally all interventions) more broadly to detect potentially negative behavioral spillovers. Finally, and more generally, while our paper contributes to better understanding the potential of cultivating grit among primary school students at scale, further work is still needed to distill what is the most effective combination of socio-emotional skills for the needs of different students.

References

Acosta, P. and 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), pp. 1121-1162.

Allensworth, E. and 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. and Aronson, J., 2011. Stereotype threat impairs ability building: Effects on test preparation among women in science and technology. *European Journal of Social Psychology*, 41(7), pp.904-913.

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

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

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

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

Beilock, S.L. and 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), p.983.

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

Benner, A.D. and 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), pp.508-519.

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

Borghans, Lex and Duckworth, Angela and Heckman, James J. and ter Weel, Bas. 2008b. The economics and psychology of personality traits. *Journal of Human Resources* 43(4), pp. 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*, pp. 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), pp.1163-1224.

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

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

Crandall, V. C. 1969. Sex differences in expectancy of intellectual and academic reinforcement. In Smith, C. P. (ed.), *Achievement-Related Motives in Children*, Russell Sage Foundation, New York.

Credé, M., Tynan, M.C., and Harms, P.D. 2016. Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, 113(3), pp. 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. and Inzlicht, M., 2003. Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, 24(6), pp.645-662.

Guerra, N.; Modecki, K.; Cunningham, W. 2014. Developing social-emotional skills for the labor market: The PRACTICE model. *Policy Research Working Paper*; no. WPS 7123. World Bank, Washington, DC.

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), pp. 1087-101.

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

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

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

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

Eskreis-Winkler, L., Duckworth, A. L., Shulman, E. P., & Beal, S. 2014. The grit effect: predicting retention in the military, the workplace, school and marriage. *Frontiers in Psychology*, pp. 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), pp. 728.

Ericsson, K.A. 2006. The influence of experience and deliberate practice on the development of superior expert performance. In Ericsson, K.A., Feltovich, P.J., and Hoffman, R.R. (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, pp. 4-34.

Ericsson, K. A. 2008. Deliberate practice and acquisition of expert performance: a general overview. *Academic Emergency Medicine*, *15*(11), pp. 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), pp. 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), pp.54-85.

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

Gatti,R.; Karacsony,S.; Anan,K.; Ferre,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. and Muralidharan, K., 2016. Improving education outcomes in developing countries: Evidence, knowledge gaps, and policy implications. *In Handbook of the Economics of Education.* Vol. 5, pp. 653-743. Elsevier.

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

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

Gutman, L.M. and Schoon, I., 2013. *The Impact of Non-Cognitive Skills on Outcomes for Young People: Literature Review*. November, 21. Education Endowment Foundation 59 (22.2), p.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.J., Moon, S.H., Pinto, R., Savelyev, P.A., Yavitz, A. 2010. Analyzing social experiments as implemented: A reexamination of the evidence from the HighScope Perry Preschool Program. *Quantitative Economics* 1 (1), pp. 1–46.

Heckman, J. and T. Kautz. 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*. University of Chicago Press. pp. 341-430.

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

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

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

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

Johnson, H., Pinzón, D., Trzesniewski, K, Indrakesuma, T., Vakis, R. Perova, E., Muller, N., De Martino, S., and 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 Socio-emotional Skills in Adolescence: Evidence from Chicago Public Schools and the Onegoal Program*. University of Chicago.

Kautz, T., Heckman, J. J., Diris, R., Ter Weel, B., & Borghans, L. 2014. Fostering and measuring skills: Improving cognitive and socio-emotional 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, D.C.: World Bank Group.

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

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

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

Nguyen, H.H.D. and 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), p.1314.

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

Outes-León, I., A. Sánchez, and R. Vakis. 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. and Dweck, C.S., 2015. Mind-set interventions are a scalable treatment for academic underachievement. *Psychological Science*, 26(6), pp.784-793.

Pintrich, P.R., 2003. Motivation and classroom learning. *Handbook of Psychology*, pp.103-122.

Robayo-Abril, Monica; Millan, Natalia. 2019. *Breaking the Cycle of Roma Exclusion in the Western Balkans*. World Bank, Washington, DC

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

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

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

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

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

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

Schmidt, F.T., Lechner, C.M. and 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), p.e0228969.

Skinner, E.A., Wellborn, J.G. and 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), p.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, pp. 1213-1227

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

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

Steele, C.M., Spencer, S.J. and Aronson, J., 2002. Contending with group image: The psychology of stereotype and social identity threat. In *Advances in Experimental Social Psychology* Vol. 34, pp. 379-440.

Stecher, B.M. and Hamilton, L.S., 2014. Measuring hard-to-measure student competencies: A research and development plan. *Research Report*. RAND Corporation.

Sturman, E.D. and 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, pp.1-10.

Sisk, V. F., Burgoyne, A. P., Sun, J., Butler, J. L. and Macnamara, B. N. 2018. To what extent and under which circumstances are growth mindsets important to academic achievement? Two Meta-Analyses. *Psychological Science*, 29 (4), pp. 549–571.

Todd, P. E., and K. I. Wolpin. 2003. On the specification and estimation of the production function for cognitive achievement. *Economic Journal*, 113, pp. 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), pp. 877–902.

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

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

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

Wigfield, A. and Eccles, J.S., 2002. The development of competence beliefs, expectancies for success, and achievement values from childhood through adolescence. In *Development of Achievement Motivation,* pp. 91-120

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

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. and Linville, P.W., 1982. Improving the academic performance of college freshmen: Attribution therapy revisited. *Journal of Personality and Social Psychology*, 42(2), p.367.

Wilson, T. D. 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, pp. 364-369.

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

Appendix 1: Additional Outcomes

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

	Frustration Reaction	Motivation Frame-work	Locus of Control	Present Bias
	(1)	(2)	(3)	(4)
Treatment 1	-0.034	-0.009	-0.040**	-0.023
"Student 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*

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

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 are clustered at the school level. All figures are rounded to three decimal places.

Source: Own survey data, school year 2015/2016, own calculations.

Appendix 2: Balanced Sample

Table A2.1: Balancing Properties, Balanced Sample

	Group Control (1)	Treatment 1 (2)	Treatment 2 (3)	T-Test Difference (1) - (2)	Difference (1) - (3)	N
Panel A: Outcomes, Baseline						
Deliberate Practice Beliefs	16.600	16.630	16.760	-0.032	-0.156**	16,575
	(2.410)	*(2.417)*	*(2.368)*			
Grit	29.600	29.600	29.420	-0.002	0.175	16,575
	(4.067)	*(4.039)*	*(4.071)*			
Grit: Perseverance-of-Effort Facet	16.310	16.370	16.220	-0.067	0.089	16,575
	(2.473)	*(2.464)*	*(2.517)*			
Grit: Consistency-of-Interest Facet	13.290	13.230	13.200	0.064	0.086	16,575
	(3.073)	*(3.057*	*(3.089)*			
Frustration Reaction	11.033	11.124	11.062	-0.091	-0.029	16,575
	(2.535)	*(2.528)*	*(2.526)*			
Motivation Framework	20.862	20.921	20.869	-0.059	-0.007	16,575
	(2.873)	*(2.865)*	*(2.835)*			
Locus of Control	17.665	17.648	17.646	0.017	0.019	16,575
	(2.299)	*(2.329)*	*(2.330)*			
Present Bias	4.203	4.216	4.237	-0.013	-0.034	16,575
	(0.991)	*(1.006)*	*(0.988)*			
S/E Skills Index	0.053	0.073	0.084	-0.020	-0.031	16,575
	(0.993	*(1.001)*	*(1.009)*			
GPAs	3.559	3.490	3.548	0.069	0.010	16,575
	(1.078)	*(1.087)*	*(1.096)*			
Panel B: Controls, Baseline						
Age	12.490	12.480	12.500	0.002	-0.018	16,575
	(0.557)	*(0.553)*	*(0.554)*			
Female	0.503	0.519	0.522	-0.016*	-0.019**	16,575
	(0.500)	*(0.500)*	*(0.500)*			
Sixth Grader	0.483	0.486	0.467	-0.003	0.016	16,575
	(0.500	*(0.500)*	*(0.499)*			
Macedonia	0.680	0.724	0.694	-0.044	-0.014	16,575
	(0.466)	*(0.447)*	*(0.461)*			
Albanian	0.266	0.208	0.250	0.057	0.016	16,575
	(0.442)	*(0.406)*	*(0.433)*			
Roma	0.017	0.019	0.022	-0.002	-0.005	16,575
	(0.130)	*(0.135)*	*(0.148)*			
Other	0.037	0.049	0.033	-0.012	0.004	16,575
	(0.189)	*(0.216)*	*(0.179)*			

TV at Home	0.957	0.945	0.012	9,025
	(0.202)	*(0.229)*		
PC at Home	0.962	0.954	0.009	9,025
	(0.190)	*(0.210)*		
Car at Home	0.869	0.873	-0.004	9,025
	(0.337)	*(0.333)*		
Family Goes on Vacation	0.715	0.687	0.029	9,025
	(0.451)	*(0.464)*		
Mother Lives at Home	0.976	0.967	0.008**	9,025
	(0.154)	*(0.178)*		
Father Lives at Home	0.951	0.942	0.010*	9,025
	(0.215)	*(0.235)*		
Mother College Educated	0.310	0.303	0.006	9,025
	(0.462)	*(0.460)*		
Father College Educated	0.296	0.285	0.010	9,025
	(0.456)	*(0.452)*		

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

Notes: Standard deviations in parentheses. T-tests with robust standard errors clustered at the school level. All figures are rounder to three decimal places.

Source: Own survey data, administrative data, school year 2015/2016, own calculations.

Table A2.2: Impacts on Deliberate Practice Beliefs and Grit, Balanced Sample (Z-Scores)

	Deliberate Practice Beliefs	Grit	Grit: Effort	Grit: Interest
	(1)	(2)	(3)	(4)
Treatment 1	0.162***	-0.064***	0.051**	-0.127***
"Student 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*

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

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 are clustered at the school level in parentheses. All figures are rounded to three decimal places.

Source: Own survey data, school year 2015/2016, own calculations.

Table A2.3: Impacts on S/E Skills Index and GPAs in Short-Run, Balanced Sample (Z-Scores)

		Gender		Ethnicity				Academic Year		Pre-Treatment Achievement		
	Average	Male	Female	Macedonian	Albanian	Roma	Other	Sixth Grade	Seventh Grade	Tercile 1	Tercile 2	Tercile 3
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: S/E Skills Index												
Treatment 1 "Student 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												
Treatment 1 "Student 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*

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

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 are 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.

Appendix 3: Long-Term Effects and Exclusion of Non-Reverse Items in S/E Skills Index

Table A3.1: Effects on GPAs of Different Roma Students in Long-Run (Z-Scores)

		Gender		Academic Year		Pre-Treatment Achievement		
	Average	Male	Female	Sixth Grade	Seventh Grade	First Tercile	Second Tercile	Third Tercile
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Panel A: GPAs, Short-Run (2015/2016 Q4)								
Treatment 1	-0.015	-0.009	-0.029	0.022	-0.034	-0.063*	0.037	-0.048
"Student 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-Run (2016/2017 Q1+Q2)								
Treatment 1	0.109**	0.085	0.162***	0.162**	0.065	0.108*	0.079	0.174*
"Student 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-Run (2016/2017 Q3+Q4)								
Treatment 1	0.129**	0.136**	0.126	0.196**	0.077	0.104**	0.127	0.178*
"Student 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*

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$. *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 are clustered at the school level in parentheses. All figures are rounded to three decimal places.
Source: Own survey data, school year 2015/2016, own calculations.

Table A3.3: Impacts on S/E Skills Index, Exclusion of Non-Reverse Items (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)
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***
"Student 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*

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

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 are clustered at the school level in parentheses. All figures are rounded to three decimal places.
Source: Administrative data, school year 2016-2017, own calculations.

Appendix 4: Robustness Checks

Attrition Tests

Table A4.1. Attrition in Survey Data

	Has Baseline Survey (1)	Has Endline Survey (2)	Has Both Surveys (3)
Pre-Treatment GPA	0.017***	0.010***	0.016***
	(0.003)	*(0.004)*	*(0.004)*
Treatment 1 "Student Self-Learning"	-0.001	-0.149***	-0.120**
	(0.034)	*(0.045)*	*(0.047)*
Treatment 2 "Teacher Delivery"	-0.010	-0.063*	-0.068*
	(0.029)	*(0.036)*	*(0.039)*
Age 12	0.002	0.018	0.026
	(0.013)	*(0.017)*	*(0.019)*
Age 13	-0.017	-0.033	-0.020
	(0.018)	*(0.023)*	*(0.025)*
Age 14	-0.052*	-0.059	-0.044
	(0.028)	*(0.037)*	*(0.038)*
Female	0.004	0.005	0.009*
	(0.004)	*(0.005)*	*(0.005)*
Sixth Grader	-0.022	-0.071***	-0.067***
	(0.015)	*(0.018)*	*(0.019)*
Albanian	-0.118***	-0.291***	-0.289***
	(0.042)	*(0.059)*	*(0.060)*
Roma	-0.141***	-0.242***	-0.271***
	(0.052)	*(0.062)*	*(0.062)*
Other	-0.056*	-0.090**	-0.079**
	(0.032)	*(0.036)*	*(0.037)*
N	33,454	33,454	33,454
N Control	12,426	12,426	12,426
N Treatment 1	10,995	10,995	10,995
N Treatment 2	10,033	10,033	10,033
R^2	0.082	0.146	0.140

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

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 are 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.

Table A4.2. Attrition in Administrative Data

	Has GPA, Medium-Run (2016/2017 Q1+Q2) (1)	Has GPA, Long-Run (2016/2017 Q3+Q4) (2)
Pre-Treatment GPA	0.004	0.009***
	(0.003)	*(0.002)*
Treatment 1 "Student Self-Learning"	-0.016	0.035**
	(0.017)	*(0.017)*
Treatment 2 "Teacher Delivery"	0.004	0.053**
	(0.018)	*(0.021)*
Age 12	0.010	-0.019**
	(0.010)	*(0.008)*
Age 13	0.000	-0.018
	(0.013)	*(0.014)*
Age 14	-0.060***	-0.107***
	(0.022)	*(0.022)*
Female	-0.004	-0.004
	(0.003)	*(0.003)*
Sixth Grader	0.010	0.013
	(0.014)	*(0.018)*
Albanian	-0.081**	-0.042***
	(0.038)	*(0.012)*
Roma	-0.073***	-0.099**
	(0.021)	*(0.039)*
Other	-0.001	-0.013
	(0.014)	*(0.016)*
N	33,454	33,454
N Control	12,426	12,426
N Treatment 1	10,995	10,995
N Treatment 2	10,033	10,033
R^2	0.088	0.106

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

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 are clustered at the school level in parentheses. All figures are rounded to three decimal places.
Source: Administrative data, school year 2016-2017, own calculations.

Analysis Using Pre-Registered Covariates

We deviated from the pre-registered covariates including age and controlling for pre-GPA as a 4th degree polynomial term. Following on the pre-analysis plan, the registered covariates included: prior GPA, gender, ethnicity, initial cognitive ability, quality of school and geographical location. Initial cognitive ability was intended to be derived from the national standardized test, which was not possible as the country has no reliable national standardized test (confirmed by the Ministry of Education), similarly for quality of schools. For the initial ability, at baseline, the team conducted two measures (Baseline Math and Baseline Reading Comprehension), however due to a printing error of the baseline survey, only 20% of the original sample took the Math and Reading baseline skills. For the replication of baseline covariates, prior GPA was used as surrogate for Cognitive Ability. Quality of schools was originally planned to be based on the national standardized test. Given the absence of such test, adjusting for school as a source of clustering was used as surrogate for quality. For geographical location we use the municipality indicator (79-level categorical measures) as the surrogate effect of the location.

Table A4.3: Impacts on Deliberate Practice Beliefs and Grit, 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***
"Student 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 A4.4: Effects on S/E Skills Index and GPAs in Short-Run, Pre-Registered Covariates (Z-Scores)

		Gender		Ethnicity				Academic Year		Pre-Treatment Achievement		
	Average	Male	Female	Macedonian	Albanian	Roma	Other	6th Grade	7th Grade	First Tercile	Second Tercile	Third Tercile
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Panel A: S/E Skills Index												
Treatment 1 "Student Self-Learning"	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 "Teacher Delivery"	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-Run (2015/2016 Q4)												
Treatment 1 "Student Self-Learning"	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 "Teacher Delivery"	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 A4.5: Effects on GPAs Over Time, Balanced Sample (Z-Scores)

		Gender		Ethnicity				Academic Year		Pre-Treatment Achievement		
	Average	Male	Female	Macedo-nian	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, Medium-Run (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
"Student 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 B: GPAs, Long-Run (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
"Student 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 "Teacher Delivery"	0.025 *(0.017)*	0.034* *(0.020)*	0.015 *(0.018)*	0.041* *(0.021)*	0.000 *(0.032)*	0.280*** *(0.054)*	0.037 *(0.033)*	0.038* *(0.022)*	0.011 *(0.022)*	0.053** *(0.022)*	0.032 *(0.028)*	-0.010 *(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

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

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 are clustered at the school level in parentheses. All figures are rounded to three decimal places.
Source: Administrative data, school year 2015/2016, own calculations.