









CAN GRIT BE TAUGHT? LESSONS FROM A NATIONWIDE FIELD EXPERIMENT WITH MIDDLE-SCHOOL STUDENTS¹

INDHIRA SANTOS  VIOLETA PETROSKA-BESKA  PEDRO CARNEIRO  LAUREN ESKREIS-WINKLER  ANA MARIA MUNOZ BOUDET  INES BERNIELL  CHRISTIAN KREKEL ,
OMAR ARIAS, ANGELA L. DUCKWORTH [^]

We study the impacts of a large-scale grit intervention on students' grit, related socio-emotional skills, and academic achievement. We evaluate an RCT implemented nationwide amongst all middle school students in North Macedonia. Relative to controls, treated students score higher on deliberate practice and academic motivation. While also scoring better on the perseverance facet of grit, exposed students display lower overall grit scores, although the latter result could be driven by systematic errors in survey response. We find small average impacts on GPAs one year post-treatment, but large impacts can be observed for disadvantaged minority (Roma) students, which grow over time. (JEL codes: C93, D91, I20, I24)

¹ We thank the Ministry of Education and Science of North Macedonia; the staff at The Center for Human Rights and Conflict Resolution; the team at PUBLIK DOO; Bojana Naceva and Jasminka Sopova at The World Bank; Robert Gallop; and the staff at the Character Lab for their support. Victoria Levin, Hillary Johnson, Maria Davalos, and Fabian Schmidt as well as participants at the 2019 meeting of the American Economic Association in Atlanta and the briq/IZA Behavioural Economics of Education workshop in Bonn in 2019 provided valuable comments and suggestions at different stages of this project. The research was funded by grants from the Umbrella Facility for Gender Equality and the Research Department at The World Bank. Carneiro gratefully acknowledges the financial support from the European Research Council through grant ERC-2015-CoG-682349. Russell Sage and the Walton Family Foundation supported the Character Lab.

[^]The symbol  indicates that the author order (Santos, Petroska-Beska, Carneiro, Eskreis-Winkler, Munoz Boudet, Berniell, and Krekel) was randomized using the American Economic Association Author Randomization Tool (Confirmation Code: O6jB2s08FnXa). Corresponding Author: Christian Krekel, c.krekel@lse.ac.uk, +44(0)20 7107 5317, London School of Economics (LSE), Houghton Street, London WC2A 2AE, UK.

Affiliations: Santos: The World Bank, United States; Petroska: Ss. Cyril and Methodius University, Skopje, and The Center for Human Rights and Conflict Resolution, Skopje, North Macedonia; Carneiro: University College London, CEMMAP, IFS, United Kingdom; Eskreis-Winkler: Kellogg School of Management at Northwestern University, United States; Munoz Boudet: The World Bank, United States; Berniell: CED-LAS-Universidad Nacional de la Plata, Argentina; Krekel: Centre for Economic Performance (CEP), London School of Economics (LSE); Department of Psychological and Behavioural Science, LSE, United Kingdom; Arias: The World Bank, United States; Duckworth: University of Pennsylvania, United States.

I. INTRODUCTION

A growing literature in psychology and economics shows that socio-emotional (or “non-cognitive”) skills play a key role in predicting education and labor market outcomes (Cipriano et al., 2023; Alan et al., 2019; Acosta and Muller, 2018; Kautz et al., 2015; Almlund et al., 2011; Borghans et al., 2008a, 2008b; Heckman et al., 2006). Attributes related to the personality trait of conscientiousness, in particular, are strong predictors of such outcomes (Roberts et al., 2007; Levin et al., 2016; Heckman and Kautz, 2014; Stecher and Hamilton, 2014; Naemi et al., 2013; Tough, 2013; Willingham, 1985). Amongst these, grit – defined as the ability to sustain effort and interest towards long-term goals – has been found to be especially relevant for education and high achievement (Alan et al., 2019; Duckworth et al., 2007).

Grit has a strong relationship with the Big Five personality domain of conscientiousness (Roberts et al., 2005) but goes further.² It has two distinct, albeit related, facets: *(i)* perseverance of effort (i.e. working towards challenging goals despite failure, adversity, or plateaus in progress), dubbed grit-perseverance; and *(ii)* consistency of interest (i.e. maintaining interest for long-term goals without losing focus), dubbed grit-consistency-of-interest (Duckworth and Quinn, 2009; Duckworth et al., 2007, 2021). While conscientiousness leads a person to do what she is supposed to do, exercising self-control to avoid impulses and distractions (driven by the expectation of a short-term benefit), grit inclines an individual to pursue a challenging goal that requires sustained effort and enthusiasm over an extended period of time.

In cross-sectional samples, grit is associated with various social and economic outcomes (Eskreis-Winkler et al., 2016, 2014; Kautz et al., 2015; Maddi et al., 2012; Duckworth et al., 2007). Adults who score high on grit make fewer career changes, progress further in their formal education, and obtain higher GPAs, are more likely to graduate from high school and perform better in high-stakes student competitions (Eskreis-Winkler et al., 2014; Duckworth et al., 2007).

Recent evidence suggests that grit is malleable through school-based interventions, as discussed in Hwang & Nam (2021), Alan et al. (2019), and Eskreis-Winkler et al. (2014). For example, Alan et al. (2019) evaluate a comprehensive intervention that cultivates grit amongst two independent

² This has led some authors to question whether grit is a distinctive skill from conscientiousness (e.g. Datu, 2021, Credé et al., 2016; Rimfeld et al., 2016; Ivcevic and Brackett, 2014). See also Duckworth et al. (2021) for a recent contribution on the factor structure of grit and a discussion. We discuss the relationship between conscientiousness, grit, and its two facets in Section III.

samples of over 3,200 fourth-grade students across 52 schools (roughly half of which were treated) in Istanbul, Turkey. They find that treated students performed better on an incentivized real effort task that measures grit attributes, as well as in standardized tests of math and Turkish language, compared to students in a control condition. In other words, Alan et al. (2019) find that it is possible to change grit through school interventions, and changing grit can lead to changes in learning.³

Given the importance of grit for individual success and its apparent malleability through school-based interventions, it is natural to consider it as a promising target of education policy. Nevertheless, not all interventions scale up well. Several interventions are effective when implemented in smaller groups under strict quality control, but ineffective when they are offered universally under potentially less favorable conditions, so evaluations of scaled up version of programs that were effective at a smaller scale are of great academic and policy importance. Furthermore, evaluations of interventions at a large scale allow to learn about impacts across a wide range of people, in a way that is not possible with studies which do not cover the entire population.

In this paper we examine the impacts of a grit intervention implemented experimentally in two thirds of (nearly) all middle schools of the Republic of North Macedonia, with the remaining third being used as the control group. To our knowledge, this is the first experimental evaluation of a nationwide school-based program (with nearly universal coverage) targeting socio-emotional skills of children. We ask whether improvements in students' behaviors and academic achievement can be obtained through simple, inexpensive, easily implementable at-scale school-based grit interventions and the extent to which these programs can help to reduce achievement gaps of disadvantaged students. Our study covers a total of 35,340 students in 352 schools across the entire country.⁴ This allows us not only to examine program impacts on a large scale, but also whether they differ across groups in the population, even ethnic minorities such as Roma students. Our study addresses recent

³ A recent meta-analysis also found that grit overall and its two facets are correlated with academic achievement across different cultures (Lam and Zhou, 2022). There is an ongoing debate about focusing on cultivating grit in education, with some observers and practitioners arguing that it distracts from a focus on policies to reduce inequality in school resources which may lead to disadvantaged students receiving lower education quality. This debate is not inherent to grit and pertains more broadly to the question of how much focus schools should place on socio-emotional skills, and as such is beyond the scope of this paper.

⁴ In contrast, nearly all RCTs in school contexts are based on sub-samples of the universe of schools and require school-level buy-in to participate. Assessed impacts could thus be specific to those schools who want or choose to participate, and not necessarily scale to all schools in a country. This raises questions about the external validity or scalability of results, although these concerns can be alleviated with proper designs (such as in the study by Alan et al., 2019).

calls in the economics literature to learn from evidence-based interventions that operate at a large scale (Gupta et al., 2021; List, 2022, 2024).

Interventions focusing on non-academic skills, such as the grit intervention we study, are of particular interest since they may be especially effective at reaching the most disadvantaged students such as those from Roma origin. The achievement gap of these disadvantaged students may result in part from gaps in socio-emotional skills like in perseverance, patience, focus or aspirations (due, for example, to the absence of role models). There is hardly any evidence of effective interventions focusing primarily on cognitive skills of adolescents, while in contrast, several recent studies have shown that a focus on non-cognitive skills can have important impacts in the lives of disadvantaged teenagers (e.g., Grossman and Tierney, 1998, Blattman et al., 2017, Heller et al., 2017, Oreopoulos et al., 2017).

We assess and contrast the effectiveness of two different modes of delivery of our intervention: (i) a modality with student self-paced learning (low intensity); and (ii) a modality where teachers are directly responsible for delivering the curriculum in a classroom setting (higher intensity). Both modalities are easily transportable to other settings, even in low resource contexts. One relevant difference between these two modalities is that the self-paced modality does not require teacher training, and thus is cheaper and easier to implement, but at the same time could be potentially less effective considering the key role of teacher-student interactions in learning new skills.

We also innovate by developing and delivering a curriculum that, besides influencing student beliefs about the malleability of their abilities (as in a standard grit intervention), motivates them to develop them by adopting the tenets of *deliberate practice* (Ericsson, 2008; Ericsson et al., 1993), which has been shown to be a mediator between grit and performance (Duckworth et al., 2014, 2011). In other words, our intervention not only teaches students about grit, but also demonstrates practical strategies to cultivate it. We expect this two-pronged approach to be more impactful to cultivate grit.

Finally, our nationwide data is large enough to allow us to investigate how impacts vary across different groups, including among minorities such as Roma students. The Roma are a subject of intense social policy attention in many countries and development organizations. They are the largest ethnic minority in Europe (10-12 million), and they often live in extreme poverty and are socially excluded. Roma students have poorer educational outcomes than non-Roma students (Gatti et al., 2016, Kertesi and Kezdi, 2011). An intervention that is able to remedy achievement gaps of such an extremely disadvantaged group would be remarkable. To date, it has been difficult to find large

scale interventions that have important positive impacts on the most socially excluded populations, such as the Roma.

The intervention we study took place in the Spring of 2016, when all middle schools in North Macedonia were randomly allocated to receive one of the two grit-building treatments (i.e. low or high intensity) or to be part of a control condition. In the higher-intensity treatment, both students and teachers were exposed to the intervention and teachers were specifically trained to deliver the curriculum to their students. All sixth and seventh grade students in treated schools in North Macedonia received the grit curriculum. Students in control schools had no exposure to the curriculum. To evaluate the impacts of the intervention, we measured students' grit and related socio-emotional skills using our own student surveys, and utilize data on student educational achievement from official school records tracking students' school grades up to one year post-treatment. The combination of a nationwide intervention and access to administrative records allows us to assess the impacts of the interventions for different ethnic groups, in particular the Roma, for which information is typically hard to collect and rarely available with larger samples.

The curriculum of the intervention is firmly rooted in psychological research on grit and related constructs, and consists of a set of lessons and materials that teaches and motivates students to: (i) adopt beliefs that their ability is malleable, can be improved through effort and practice, and stress that achievement is not determined solely by immutable characteristics such as talent, gender, or ethnicity (Yeager & Dweck, 2020; Yeager & Dweck, 2012; Dweck, 1999); and (ii) implement *deliberate practice* – to identify stretch goals, seek feedback, concentrate, and repeat until mastery (Ericsson, 2008; Ericsson et al., 1993). The intervention is also designed to counter negative stereotypes surrounding gender and ethnicity by providing positive role models and counter-stereotypical examples throughout the intervention materials.

We find that, relative to students in the control group, treated students showed much higher levels of deliberate practice and moderately higher grit-perseverance and academic motivation. Impacts were larger in the higher-intensity treatment that involves teachers directly in the delivery of the curriculum, compared to the self-learning modality of our intervention.

Moreover, while we find only small average impacts on GPAs in the high-intensity, teacher-delivered modality, large impacts can be observed for Roma students, which grow over time, reaching up to 28% SD a year post-treatment. These GPA gains represent 18% of the pre-intervention

achievement gap between Roma and Macedonian students. This is remarkable given that our intervention was not particularly intensive, consisting of only one session a week for 5 weeks, and ended over a year before some of these impacts were measured.

We also find that, relative to students in the control group, exposed students scored *lower* on the overall grit scale. This is driven by a negative impact of the intervention on the consistency-of-interest facet of grit. The impact of the program on the perseverance-of-effort facet is positive. This result might indicate that an unintended consequence of the intervention was to lower students' disposition to maintain goal-interest for longer periods of time. However, we present evidence that this result could be driven instead by systematic errors in students' responses to the grit-interest questions (which were negatively phrased) in the grit scale. Finally, there is no evidence that the intervention had significant effects on other measures of socio-emotional skills covering aspects of conscientiousness or self-regulation: frustration reaction, motivation for achievement, student perceptions of teachers' mindset, or locus of control.

While our measures of grit and related socio-emotional skills are self-reported, those related to academic achievement (GPAs) are not. The fact that we observe persistent impacts on GPAs alleviates concerns that the estimated impacts on socio-emotional skills could be driven solely by changes in reporting or measurement issues, and not changes in actual behaviors. In fact, our main conclusions about the overall impacts of the intervention on behaviors and attitudes (which are based on the entire battery of survey items) and on student achievement (which is immune to self-report biases since they are based on administrative records) hold under several robustness checks.

The stronger impacts on the academic achievement of minority disadvantaged students mirror findings from other educational interventions (cf. 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 Our intervention adds to this body of research by showing that educational interventions that cultivate socio-emotional skills such as grit could reduce educational inequalities (Liu et al., 2021; Outes-Leon et al., 2020; Broda et al., 2018; Walton and Wilson, 2018; Inzlicht and Schmader, 2008). Notwithstanding the need for policies that tackle the socio-economic root causes of inequality, such

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, for example, the decline in grades that is generally found for all students in transition periods (which, in our case, is the start of sixth grade in middle school) is more pronounced for students from disadvantaged backgrounds (Gutman et al., 2013).

interventions may empower and propel disadvantaged students to persist through the setbacks that are inherent to achieving academic success.

II. THE INTERVENTION

The intervention we study covered all public schools in North Macedonia with Macedonian and Albanian language of instruction, with sixth and seventh-grade classrooms, and with at least five students in each single-level classroom. This amounts to a total of 35,340 students in 1,780 classrooms, 352 schools, and 80 municipalities across the entire country, 93 percent of the total student population for these grades.⁶ The five-weeks intervention was delivered nationwide to this student population during the third quarter of the school year 2015/2016 (February to April 2016).

The objective of the intervention was to cultivate grit amongst sixth and seventh-grade students (11 to 14 years old). Past research highlights the long-term benefits of interventions focused on 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 (Heckman et al., 2014; Benner and Graham, 2011; Crosnoe, 2011; Allensworth and Easton, 2005).

Our intervention consists of a curriculum of five, one-hour long, consecutive lessons delivered weekly that teaches students the tenets of *deliberate practice* (Ericsson, 2008; Ericsson et al., 1993), and how these can be applied to more effective ways of studying.⁷ Each of the five lessons builds on the previous one, starting by recapping what had been learned in the previous lesson, followed by the introduction of new concepts and an example of their application, and ending with a practical hands-on activity.⁸ The lessons were delivered as a component of an existing “Life Skills” curriculum in North Macedonia, which is part of the activities headteachers implement during the

⁶ About 98% of children in North Macedonia attend public education, which is compulsory until ninth grade and provided free of charge. The country has 84 municipalities, 4 were excluded due to lack of schools. Out of the total 416 schools 64 schools (15%) were excluded before the start of the intervention for the following reasons: using a different language of instruction (5), being “special schools” (8), having less than five students in the classes targeted (31), and not having single-level classrooms (20).

⁷ Appendix A.IV includes more information about our intervention. The full set of intervention materials is available upon request.

⁸ Three additional sessions took place: one week before and one week after the intervention to collect baseline and endline surveys, and another two weeks after to collect additional behavioural measures of grit-related outcomes. The latter measures aimed to capture three out of the four dimensions of deliberate practice. The results of these additional outcomes, however, are not included in this paper, given concerns about their reliability due to systematic errors and attrition during the post-intervention data collection.

Monday morning “class hour”, i.e. the first hour they have each week with their class.⁹ The intervention’s lessons were integrated in this curriculum in such a way as to have a similar structure and format.¹⁰

Specifically, students learn how to: (i) identify stretch goals, (ii) concentrate, (iii) seek feedback, and (iv) reflect, refine, and repeat until mastery. They learn how to set stretch goals and what it means to choose a challenge, for example by focusing on problems that may be difficult at first or revising their work to constantly improve it. Students are also stimulated to find ways to focus by consciously avoiding distractions. They are encouraged to get feedback by checking how they do on tasks, finding out what they got right and wrong, and importantly, how they can do better next time. Finally, they are prompted to reflect on how their tasks went, refine what they did, and then repeat them to fix mistakes and improve. They are taught that, by repeated application of these steps, they can get better and better at school and beyond. To address students’ expectancies of success, the materials across the five lessons underscore that characteristics such as innate talent, gender, or ethnicity do not pre-determine an individual’s performance, including academic achievement. To address students’ values (i.e. subjective values attached to success), the materials stressed the important role that academic achievement plays in later-life outcomes. By raising both expectancies and values associated with deliberate practice, we hypothesized that students would take up and apply this form of practice.

Our intervention comprised two treatment arms to allow us to assess and contrast the effectiveness of different modes of delivery. These two arms were contrasted with a control group that received lessons from the existing “Life Skills” curriculum or did other school-related activities, as it was customary, at the discretion of the headteacher. Teachers in the treatment arms were notified of the intervention and their expected role in it by the Ministry of Education and Science in North Macedonia and the school administration.

Treatment 1: “Self-Learning” (Low Intensity) consisted of the five lessons organized in weekly booklets that were distributed to students each week to work through on their own during class time.

⁹ The headteacher can be a teacher in any subject and is assigned by the school at the start of the academic year as the main focal point teacher for the class, for both students and parents. This teacher only teaches one subject to the students, in addition to the Monday morning “class hour”.

¹⁰ The “Life Skills” curriculum is a multi-topic curriculum covering the topics of (i) Personal Development, (ii) Social Relationships, (iii) Relationship with the Environment, and (iv) Civic Responsibilities. These topics comprise eight to ten sub-topics each. Teachers can choose from any of them to cover during their lessons with students. All lessons in this curriculum include an introduction to the topic, a practical exercise, and a final reflection activity, similar to our intervention.

The lessons were paper-based and self-contained, each organized so that students would take up to one school hour to go over the materials. Each lesson had a specific student workbook with didactic slides, which are interspersed with activity prompts, engaging images, and exercises to practice content. Workbooks included a take-away self-evaluation for students to assess how successful they were in implementing the lesson of the week (which served as a reinforcement). Teachers were only asked to distribute the materials, answer questions for clarification, and exercise regular classroom management. They were given only generic information about the materials. Each week’s material came in prepared packages for the classroom, including a one-page teacher guide with the basic instructions for the hour (e.g. distribution and collection of materials as well as space to report any unusual issue affecting the class during that hour, if any), stickers to label the materials with individual student’s IDs, and instructions for storage. Teachers also received a form to report any class disruptions (e.g. fire alarm).

Treatment 2: “Teacher Delivery” (Higher Intensity) had the same content as the first but relied on headteachers to deliver the lessons. It included a one-day teacher training session about a month prior to the start of our intervention. During this training, teachers received teacher-specific materials to familiarize themselves with the concepts included in the lessons and a detailed lesson plan with instructions for each of the five weeks of lessons they were expected to deliver. Students received weekly activity booklets, which were the same in content as in the low-intensity treatment but without the self-paced guidance elements. As in the low-intensity treatment, prepared packages of materials and labels were given to the teachers together with a form to document how the lessons went.¹¹

III. DATA

We measured students’ grit and related socio-emotional skills using our own survey data and GPAs using administrative data from official school records.¹² We have two categories of outcomes:

- (1) *Socio-emotional skills*, which included (i) deliberate practice beliefs and (ii) the Short Grit Scale (Duckworth and Quinn, 2009), the main outcomes targeted by our intervention; (iii) a measure of frustration reaction (Peters et al., 1980); (iv) the Motivational Frameworks Questionnaire

¹¹ In reviewing these forms, we found no evidence of reported issues that could have systematically affected the implementation of our intervention in either treatment arm.

¹² The RCT was registered as AEARCTR-0002094 (Arias et al., 2017) and has IRB approval by the University of Pennsylvania.

(adapted from Gunderson et al., (2013) and Kinlaw and Kurtz-Costes (2007)); (*v*) a student assessment of teacher’s mindsets (adapted from Haimovitz and Dweck, (2016) and Friedel et al., (2007)); and (*vi*) locus of control (Skinner et al., 1990). All outcomes were measured using validated self-report scales, which were adapted to children in North Macedonia and translated (back and forth) to Macedonian and Albanian languages.

(2) *Academic achievement*, measured by GPAs in three main subjects -math, first language (either Macedonian or Albanian), and English- at different points in time. These are available in the short-term, i.e. immediately after the intervention in the fourth quarter of the school year 2015/2016; in the medium-term, i.e. half a year later in the first semester of the school year 2016/2017; and in the longer-term, i.e. one year later in the second semester of the school year 2016/2017. In North Macedonia, school grades are recorded on a one-to-five scale, where one is the lowest and five is the highest attainable grade.

A. Survey Data

We used baseline and endline surveys to collect data on student basic demographic (e.g., age, gender, ethnicity, which were also collected via administrative data) and socio-economic characteristics (e.g., parental education, household assets) and socioemotional skills. The survey questionnaires were filled out by students themselves.¹³

The questionnaire items on deliberate practice beliefs asked students about the key elements of deliberate practice targeted by our intervention, namely how important they believe it is to (while studying): (*i*) put greater effort into yet unknown material; (*ii*) concentrate solely on studying; (*iii*) seek feedback from parents and teachers; and (*iv*) review the material several times until they are certain to have absorbed it. We combine their responses into a summary score of deliberate practice beliefs.

We measured grit using the eight-item Short Grit Scale (Duckworth and Quinn, 2009). Grit scores are computed both for the overall measure and for each of its two facets (i.e. perseverance of effort and consistency of interest).¹⁴ Frustration reaction uses Peters et al., (1980) 3-item equally

¹³ The survey instruments are available upon request. The questionnaire includes also questions on aspirations and perceptions of the value of education.

¹⁴ The literature reports a correlation between grit and conscientiousness of about 0.8, leading some authors to argue that grit and conscientiousness are not distinct skills (Datu, 2021, Credé et al., 2016; Rimfeld et al., 2016; Ivcevic and Brackett, 2014). However, the literature also reports a high variance in the strength of

weighted scale, framed around homework. The motivational frameworks metric includes a battery of six items, three on children's beliefs about intelligence and academic abilities and three on preference for easy tasks (Gunderson et al., 2013), adapted to common class subjects for children in North Macedonia. For locus of control, we use the five items control beliefs from Skinner et al. (1990) that prompt children to assess the extent to which they are able to induce positive and prevent negative outcomes in the school sphere. We adapt the student assessment of teacher's mindsets (Haimovitz and Dweck, 2016; Friedel et al., 2007) to capture the perceived value that teachers assigned to effort, learning, intelligence, as well as gender bias.

All questions solicit responses following a five-point Likert scale (from 'not at all' to 'really a lot'). Students are asked whether certain attitudes and behaviors are desirable or reflect how they see themselves and how they behave. In order to minimize the possibility of acquiescence bias – that students mechanically respond to items with agreement/affirmation without reflecting on their content, some statements in the scales were negatively phrased and required reverse coding. Scales were separated into eight groups, each clearly marked and containing an introduction to be read.

We followed best practices in the adaptation of all scales to the local context. Scales were translated into Macedonian and Albanian and back into English to ensure appropriate wording of items. Moreover, qualitative cognitive interviewing about the interpretation of items using focus group discussions of students was implemented during a pilot phase. Finally, the Short Grit and Deliberate Practice beliefs scales were pre-piloted with a small random set of students in the targeted age group, none of which indicated any issues.

In addition to the scores for individual scales corresponding to different attitudes and behaviors, we also construct an overall index incorporating information from all the scales. The socio-emotional skills index ("S/E Skills Index") combines all measures discussed above. Following Anderson (2008), we constructed this index by (i) switching the sign of the variables included in the index (if needed), so that a positive direction always indicates a "better" outcome; (ii) standardizing each variable (to have mean zero and standard deviation one, i.e. z-scores); (iii) averaging the standardized variables using appropriate weights (i.e. the inverse of the variance-covariance matrix of the standardized variables) to ensure that highly correlated items receive less weight while items that are uncorrelated and hence represent new information receive more; and (iv) computing z-

correlations (between 0.4 and 0.7). Moreover, the consistency-of-interest facet of grit appears to be a distinct construct from conscientiousness (Schmidt et al., 2017, 2020; Eskreis-Winkler et al., 2014), and some authors argue that this renders grit a distinct construct overall (Jachimowicz et al., 2018).

scores of the resulting overall index. All our results are robust to using alternative methods to construct scores for the individual scales from the items, as well as for the overall index, these results are presented in Appendix III.

B. Administrative Records

Our measures of student academic achievement are Grade Point Averages (GPAs) obtained from administrative data from official school records held by the Ministry of Education and Science in North Macedonia. Unfortunately, standardized tests for the relevant academic years were not available in the country at the time of our intervention. We use all grades given to the students during the school years 2015/2016 and 2016/2017, i.e. the year in which the intervention took place and the year after, respectively. Since our intervention targeted both sixth and seventh-grade students, we focus on the set of core subjects that are common to both sixth and seventh-grade students: math, English, and first language (which can be either Macedonian or Albanian, depending on the school's language of instruction). Grades are available for all students in North Macedonia, even those for whom, as discussed below, we have missing survey data.

An often-cited concern with the use of GPAs to assess educational interventions is that these may not be comparable across students in different classrooms and schools. Student grades are often based on classroom specific assessments and may be influenced by reporting biases such as variation in teachers' judgements or standards (some teachers may be easier graders than others), grade inflation, or teachers grading on a curve. However, these factors would not lead to systematic biases in our results to the extent that they were balanced across treatment and control groups with randomization. It could be, though, that teachers exposed to our intervention (especially those in the higher-intensity treatment) may have shifted their grading standards, particularly for disadvantaged groups, as a result of having been exposed to the intervention materials but without any actual impact on students' academic achievement. Only headteachers were "treated" and they typically teach only one subject, not necessarily corresponding to the subjects we used to construct GPAs. As discussed further below, this and our robustness analysis reassures us that it is unlikely that our results are an artifact of changes in teachers' grading behavior. Finally, when calculating GPAs, we use all recorded grades, which in North Macedonia includes both written and oral tests grades.¹⁵ Depending on the subject and level (i.e. sixth or seventh), students may differ in the number of

¹⁵ The type of exam or test (i.e. whether written or oral) receives equal weight for students' GPAs at the end of the school year. We conducted sensitivity analyses with respect to oral and written grades separately, but these did not result in qualitatively different findings. The results are available upon request.

exams or tests they take, in the number of grades they receive, as well as in the share of these that comes from either written or oral tests. Such differences are unlikely to cause bias to the extent that these were balanced across treatment and control groups due to randomization.

There are also advantages of using GPAs. It is now well-documented in the literature that while standardized achievement tests provide reliable measures and normalized comparisons of student learning, they are also incomplete measures of the value added of schools (and teachers). In particular, course grades and GPAs are better at capturing the development of socio-emotional skills that are important for students' long-term education and labor market outcomes. The literature has found that the evaluation of the value added of schools, teachers, and educational programs can fruitfully use course grades and GPAs to account for these skills that are not well-captured by standardized test scores (Petek and Pope, 2023; Kraft, 2019; Jackson and Kirabo, 2018; Almlund et al., 2011). Consistent with this, in Appendix I we show that the main survey-based measures of socio-emotional skills we use exhibit modest (0.3 to 0.5) positive correlations with GPAs.

GPAs allow us to study the impacts of our intervention on student achievement, including both the academic and non-academic components, over time. We classify GPAs into either pre-treatment or post-treatment, depending on the date when they were recorded.¹⁶ The pre-treatment GPA is calculated over the entire academic year 2015/2016, right up to the beginning of the intervention. The first post-treatment GPAs (which we call *short-term GPAs*) are calculated over the fourth quarter of the school year 2015/2016, the *medium-term* over the first and second quarter of the school year 2016/2017) and *long-term GPAs* over the third and fourth quarter of the school year 2016/2017.¹⁷ As with our skills measures, we standardize GPAs to have mean zero and standard deviation one (i.e. z-scores).

C. Balance

Table I shows summary statistics and balancing properties for all socio-emotional skills, GPAs, as well as students' demographic and socio-economic characteristics at baseline and by treatment assignment group. These correspond to the subset of students with non-missing information on

¹⁶ The administrative data do not include the precise dates when exams or tests were taken, but only the dates when the resulting grades were recorded. Qualitative information obtained from school administrators during and after implementation indicated that such time gaps in grade recording are not significant. While this may induce some bias in our estimates, we believe that if at all present, it is quantitatively very small.

¹⁷ More precisely, as the intervention was implemented between February 15 and March 21, 2016, the first post-treatment GPAs are calculated over the period from March 22 to August 31, 2016, corresponding to the fourth quarter of the school year 2016/2017.

each of the respective outcomes and control variables used to obtain the main results of our analysis.

TABLE I: SUMMARY STATISTICS AND BALANCING PROPERTIES

	Group			T-Test		N
	Control (1)	Treatment 1 (2)	Treatment 2 (3)	Difference (2) - (1)	Difference (3) - (1)	
<i>Panel A: Outcomes at Baseline</i>						
Deliberate Practice Beliefs	16.478 (2.492)	16.486 (2.562)	16.623 (2.480)	0.007	0.145*	24,276
Grit	29.484 (4.070)	29.501 (4.043)	29.338 (4.090)	0.017	-0.146	21,925
Frustration Reaction	10.966 (2.530)	11.005 (2.537)	10.954 (2.543)	0.039	-0.012	23,832
Motivational Frameworks	20.640 (2.969)	20.686 (2.990)	20.747 (2.881)	0.046	0.107	23,909
Teacher Mindset	23.352 (3.489)	23.658 (3.463)	23.621 (3.450)	0.306	0.270	22,605
Locus of Control	17.457 (2.424)	17.420 (2.487)	17.499 (2.431)	-0.037	0.042	23,724
S/E Skills Index	0.007 (0.629)	0.017 (0.654)	0.028 (0.633)	0.010	0.021	25,054
GPA	3.324 (1.150)	3.304 (1.141)	3.311 (1.157)	-0.019	-0.012	33,454
<i>Panel B: Controls at Baseline</i>						
Age	12.475 (0.572)	12.481 (0.570)	12.488 (0.571)	0.006	0.014	33,454
Female	0.484 (0.500)	0.481 (0.500)	0.487 (0.500)	-0.003	0.004	33,454
Sixth Grader	0.501 (0.500)	0.495 (0.500)	0.488 (0.500)	-0.006	-0.014	33,454
Macedonian	0.554 (0.497)	0.605 (0.489)	0.590 (0.492)	0.052	0.037	33,454
Albanian	0.371 (0.483)	0.321 (0.467)	0.328 (0.470)	-0.050	-0.043	33,454
Roma	0.038 (0.192)	0.024 (0.153)	0.042 (0.201)	-0.014	0.004	33,454
Other Ethnicity	0.037 (0.190)	0.050 (0.218)	0.039 (0.195)	0.013	0.002	33,454
TV at Home	0.946 (0.227)		0.935 (0.247)		-0.011	17,671
PC at Home	0.945 (0.227)		0.942 (0.233)		-0.003	18,058
Car at Home	0.854 (0.353)		0.860 (0.347)		0.006	17,392
Family Goes on Vacation	0.693 (0.461)		0.676 (0.468)		-0.017	16,348

Mother Lives at Home	0.958 (0.201)	0.950 (0.218)	-0.008*	18,498
Father Lives at Home	0.933 (0.250)	0.924 (0.265)	-0.009*	18,302
Mother College Educated	0.264 (0.441)	0.276 (0.447)	0.012	18,318
Father College Educated	0.256 (0.437)	0.263 (0.440)	0.007	18,162

Notes: Standard deviations in parentheses. T-tests with robust standard errors clustered at the school level. Subset of students with non-missing information on all outcomes and covariates. All figures are rounded to three decimal places.

Table I Panel A shows students' socio-emotional skills derived from our own surveys and GPAs computed from official school records. Panel B shows students' demographic and socio-economic characteristics, taken from survey and administrative data. We find that pre-treatment outcomes and controls are generally well-balanced between groups; the only exception are students in our second treatment group who score slightly higher in terms of deliberate practice beliefs pre-treatment. However, this difference is small.

The number of observations varies across variables: it is highest for GPAs taken from the administrative data and lowest for demographic and socio-economic characteristics collected from our own surveys. This difference arises mainly due to some surveys not having been returned, being unreadable, or having only partially been filled out. Section V.C.1 discusses attrition issues and the robustness of our results to survey non-response and attrition. Appendix III shows the balance tables and analysis using a restricted sample of students for whom we have a complete set of data for all variables.

Our analysis places a special focus on the impact on Roma students since they are a particularly disadvantaged minority group. The Roma (who make up about 3% of the population in North Macedonia) are socially and economically marginalized, and subject to discrimination in many areas, notably in education (UNICEF, 2022; Robayo-Abril and Millan, 2019; Gatti et al., 2016). Roma children typically lag significantly behind Macedonian children in terms of academic achievement. This is corroborated in our data: Appendix Table A.I.1 replicates Table I, showing pre-treatment outcomes by ethnicity, including grit and its two facets, related socio-emotional skills, our S/E Skills Index, and GPAs. The pre-intervention gaps in our S/E skills index and GPAs between Roma and Macedonian students are about 0.50 SD and 1.30 SD, respectively.

IV. RESEARCH DESIGN AND EMPIRICAL MODEL

Our 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 either of our two treatment groups to those in our control group. With equal probability, all schools in the country (and all sixth and seventh-grade students therein) were allocated to either one of our two treatment groups or to our control group. To achieve a balance of groups at a regional level and hence national representativeness, we further 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 regress each outcome (e.g. our S/E skills or GPAs) on its pre-treatment value alongside dummies for each treatment group, covariates, and municipality fixed effects:

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

where y_{it} is the outcome of student i at time $t \in \{0, 1\}$; T_1 and T_2 are binary indicators that equal one if the student belongs to the first (self-learning) and second (teacher delivery) treatment group, respectively, and zero if they belong to the control group; X_{it} is a vector of controls, including age, gender, school year (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). For each outcome, we control for a fourth-order polynomial in pre-treatment measures of the outcome, which we found yields a better model fit compared to a linear term.¹⁸ Finally, to account for 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 (based on the control group's mean and standard deviation) to make outcomes comparable in terms of effect size.

The coefficients of interest are β_1 and β_2 for each respective outcome. Because of randomization, full eligibility, and full compliance (recall that schooling up to grade nine is compulsory in

¹⁸ Our results are robust to excluding these controls. Moreover, they are, with few exceptions, robust to the exclusion of pre-treatment outcomes as controls; if anything, effect sizes become slightly stronger in this case, though somewhat less precisely estimated. These and other robustness results are available upon request.

North Macedonia), these coefficients measure the average treatment effects of each of the experimental conditions, in a large representative sample (for GPAs nearly the entire population) of sixth and seventh grade students in North Macedonia.

We also analyze whether the intervention had heterogeneous impacts by gender and ethnicity, as well as across the pre-treatment distribution of academic achievement. As noted, some studies of psychological interventions have found larger impacts for students that are at a relative disadvantage or at risk due to psychological barriers such as stereotype threat rooted in socio-economic differences between groups (cf. Good et al., 2003; Cohen et al., 2009; and Yeager and Dweck, 2012). One might expect improvements in grit, other related socio-emotional skills, and academic achievement to be stronger amongst some student sub-groups, namely girls or ethnic-minority students such as Roma, or students with lower socio-emotional skills and who are lower-performing at the outset. We also hypothesize heterogeneous impacts across our two treatment groups. Specifically, we expect the teacher-delivered lessons to show larger impacts than student self-learning given that teachers' exposure to the content of the intervention could be reflected in their regular teaching practices in the classroom and lead to a higher intensity treatment.

V. RESULTS

A. *Impacts on Socio-Emotional Skills*

We first look at the socio-emotional skills directly targeted by our intervention: deliberate practice beliefs and grit. We then examine impacts on other socio-emotional skills that could also potentially be affected by the intervention (and which were listed as subject of analysis in the pre-registration of the trial): frustration reaction, achievement motivation, locus of control, and perceived teachers mindset. We examine average treatment effects as well as heterogeneous effects by gender, ethnicity, and the levels of pre-treatment outcomes. These impacts are measured immediately after the intervention, i.e. in the fourth quarter of the school year 2015/2016. Since we are examining multiple (eight) outcomes, we present p-values for the test of whether the coefficients are equal to zero (labelled Stepdown P-Values in the table), adjusting for multiple hypotheses testing using the procedure developed in Romano and Wolf (2004).

Table II shows impacts (effect sizes) of the two intervention arms (self-learning, and teacher delivery) on deliberate practice beliefs and grit. The largest impacts of the intervention are on deliberate practice, indicating that it successfully changed students' beliefs about the importance of exercising deliberate practice while studying. Both treatment arms improved students' beliefs but,

consistent with our initial hypothesis, the point estimates are significantly larger when teachers delivered the curriculum (about +23% SD) compared to student self-learning (about +15% SD). We can reject that both coefficients are equal to zero (adjusting for multiple hypotheses testing).

In contrast, the estimated impacts of the intervention on grit are negative, although the magnitude of the coefficients is small and, as discussed below, the impacts are divergent across grit facets. Only for T1 we reject that the average treatment effect is statistically different from zero. These divergent impacts on deliberate practice and grit are unexpected. However, as discussed below, we cannot rule out that mismeasurement is a serious factor behind the grit results. We show how, in our context, mismeasurement is likely to affect grit much more than deliberate practice scores.

Estimates of program impacts on the remaining socio-emotional outcomes in Table I are small (0.03 SD or less) and statistically indistinguishable from zero (after accounting for multiple hypotheses testing) with one exception: academic motivation. T2 is estimated to increase this measure by about 0.07 SD, and we can reject that this effect is equal to zero. Finally, we estimate that T1 has a precisely estimated zero impact on an index of all these measures (as described above), while T2 leads to an increase in this same index of 0.08 SD.

TABLE II: AVERAGE TREATMENT EFFECTS ON S/E SKILLS INDEX, DELIBERATE PRACTICE BELIEFS, GRIT, AND OTHER SOCIO-EMOTIONAL SKILLS (Z-SCORES)

	S/E Skills Index	Deliberate Practice Beliefs	Grit	Frustration Reaction	Motivational Frameworks	Teacher Mindset	Locus of Control
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
<i>Self-Learning</i> (T1)	0.007 (0.018)	0.153*** (0.018)	-0.051*** (0.018)	-0.031 (0.022)	-0.008 (0.021)	-0.031 (0.023)	-0.038** (0.019)
<i>Teacher Delivery</i> (T2)	0.079*** (0.019)	0.229*** (0.018)	-0.029 (0.020)	0.006 (0.021)	0.065*** (0.020)	0.018 (0.022)	0.001 (0.017)
Stepdown P (T1)	0.984	0.002	0.022	0.531	0.984	0.531	0.208
Stepdown P (T2)	0.002	0.002	0.531	0.984	0.006	0.862	0.984
N	25,054	24,276	21,925	23,832	23,909	22,605	23,724
R ²	0.471	0.322	0.336	0.211	0.371	0.305	0.295

Notes: All regressions control for a fourth-order polynomial of each corresponding pre-treatment outcome, demographic controls (including age, gender, school year, and ethnicity), and a full set of municipality fixed effects. Samples correspond to the subset of students without any missing responses for each corresponding socio-emotional skills scale and control variable. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places.

Next, we look at heterogeneous impacts. We present here impacts on the socio-emotional index. Estimates for all individual socio-emotional skills are shown in Appendix II. Table III Panel

A shows the impacts of the intervention on this index for the overall sample (column 1), by gender (Columns 2 and 3), ethnicity (Columns 4 to 7), and tercile in the pre-treatment distribution of GPA (Columns 8 to 10), whereby the first tercile is the lower and the third tercile the upper tail of the distribution. The estimated impacts of T2 are larger for females than males, and for individuals with higher GPA levels prior to the treatment. We can reject that these differences are statistically zero (accounting for multiple hypotheses testing). For ethnicity, the estimated impacts are of similar magnitude for Macedonian, Roma, and students of other ethnicities, and lower for Albanian students. Only the effects for Macedonian students are statistically significant, and the other effects are imprecisely estimated. However, we cannot reject equality of the coefficients for Macedonian and Roma/Albanian students in either Treatment 1 or Treatment 2. For T1, the impacts are of smaller magnitude and statistically indistinguishable from zero, except for positive impacts on students at the top of the pre-treatment GPA distribution. In Appendix II we show that these results hold and the effects are larger for deliberate practice.¹⁹

¹⁹ We obtain similar results with a socio-emotional index built from a factor model (available upon request).

TABLE III: AVERAGE AND HETEROGENEOUS TREATMENT EFFECTS ON S/E SKILLS INDEX (Z-SCORES)

	Average (1)	Gender		Ethnicity				Pre-Treatment GPA		
		Male (2)	Female (3)	Macedonian (4)	Albanian (5)	Roma (6)	Other (7)	Tercile 1 (Low) (8)	Tercile 2 (Middle) (9)	Tercile 3 (High) (10)
<i>Self-Learning</i> (T1)	0.007 (0.018)	-0.013 (0.023)	0.028 (0.021)	0.007 (0.022)	0.028 (0.035)	-0.099 (0.090)	-0.047 (0.073)	-0.066** (0.028)	0.016 (0.023)	0.077*** (0.027)
<i>Teacher Delivery</i> (T2)	0.079*** (0.019)	0.047** (0.023)	0.111*** (0.021)	0.087*** (0.024)	0.049 (0.031)	0.111 (0.081)	0.094 (0.090)	0.018 (0.028)	0.083*** (0.022)	0.130*** (0.027)
Stepdown P (T1)	0.932	0.886	0.639	0.932	0.858	0.773	0.886	0.112	0.894	0.032
Stepdown P (T2)	0.004	0.192	0.002	0.006	0.511	0.639	0.778	0.894	0.006	0.002
N	25,054	12,718	12,336	16,338	6,940	582	1,194	7,703	8,513	8,709
R ²	0.471	0.471	0.471	0.471	0.471	0.471	0.471	0.487	0.487	0.487

Notes: All regressions control for a fourth-order polynomial of pre-treatment outcomes, demographic controls (including age, gender, school year, and ethnicity), and a full set of municipality fixed effects. Samples correspond to the subset of students without any missing responses for each corresponding socio-emotional skills scale and control variable. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places.

A. Disentangling the Impacts on Grit

We now discuss the impacts of the intervention on grit in more detail. Our hypothesis was that the intervention would have positive impacts on grit. Instead, we find that the impacts are negative for grit overall and, as we now discuss, they differ across grit facets. We find suggestive evidence that this is likely due to mismeasurement of grit, in particular one of its facets.

There are several ways in which grit could be mismeasured. The Short Grit Scale (Duckworth and Quinn, 2009) could be less suited for the age group of our intervention or the Macedonian cultural context, an issue that has been documented before for self-reported scales of personality and socio-emotional skills (cf. Laaj et al., 2019). There could also be a problem with translation or interpretation of certain items depending on how these are phrased. But as noted above, we found no evidence suggesting issues with translation and adaptation of the grit scale during the intervention pilot design stage. Moreover, these problems would be common to all scales. So, what would make grit especially vulnerable to mismeasurement?

The grit scale has two components, perseverance of effort and consistency of interest. The consistency of interest facet of grit, which accounts for half of the items in the grit scale, is entirely captured by items that are negatively phrased and require reverse coding (e.g. “I often set a goal but later choose to pursue a different goal”). Meanwhile items pertaining to the perseverance of effort facet are all positively phrased (e.g. “I am a hard worker”).

Columns 1 and 2 of Table IV show the effects of the intervention on the two facets of grit separately. We find that, while the intervention increased the perseverance of effort facet (in both treatments, by about +5 to 6% SD), it reduced the consistency of interest facet (in both treatments, by about -10% to 12% SD). Compared to students in the control group, treated students report being more industrious and hard-working but also being less inclined to sustain interests for longer periods of time. Combining both facets then results in an insignificant (teacher-delivery) or significantly negative (student self-learning) impact of our intervention on overall grit.

Somewhat unexpectedly, we found noticeably different patterns of responses for positively and negatively-phrased items for all scales, which impact especially the grit scores, as we document in Table IV (also see Table A.III.9). We observe that students appear to have responded to negatively phrased statements (reverse coded items) on some basis unrelated to the content the items intend to measure. This can happen due to acquiescence bias -i.e. a tendency to concentrate on the positive side of the scale (Paulhus, 1991), or casual responding, i.e. if students failed to notice the changes in item wording (items were alternating from positive to negatively phrased) and did not adjust

their responses accordingly, either because they were distracted or were rushing to fill the questionnaires. Although negatively phrased items are commonly used in psychological scales, they can potentially induce response bias as has been documented in the psychological and education literature (Steinmann et al., 2021; Lindwall et al., 2012). As show in Appendix III, we find some evidence that answers to negatively phrased items may indeed contain more measurement error. Therefore, to investigate whether this issue could lead to the observed divergent impacts in grit we used factor models to construct indices of socio-emotional outcomes either using only regularly coded or reverse coded items. We also constructed versions of the overall S/E index excluding the grit items. These results are also reported in Table IV.

Column 3 of Table IV shows the impact of the intervention on the index of socioemotional skills using only positively phrased items, and the estimates are positive and statistically different from zero. In contrast, in column 4, when we use only negatively phrased items to construct the index the estimates are negative and statistically significant. Columns 5 and 6 show that these patterns remain when we exclude all grit items from the construction of the indices, although the negative effects are attenuated. In fact, the estimated impact of T2 on the socio-emotional index using negatively phrased items is small and statistically indistinguishable from zero.

In sum, negatively phrased items not only have different response patterns, but they also lead to reverse sign treatment effects. It is possible that many students become confused when answering these items in the questionnaire, which causes irrelevant systematic variance in their responses. Given that the grit consistency of interest facet items are all negatively phrased, the negative impact of the intervention on this grit facet could be spurious.

TABLE IV: AVERAGE TREATMENT EFFECTS ON GRIT FACETS AND DIFFERENT S/E SKILLS INDICES (Z-SCORES)

	Grit-Effort	Grit-Interest	S/E Skills Index of Positively phrased Items	S/E Skills Index of Negatively phrased Items	S/E Skills Index of Positively phrased Items Excluding Grit	S/E Skills Index of Negatively phrased Items Excluding Grit
	(1)	(2)	(3)	(4)	(5)	(6)
<i>Self-Learning</i> (T1)	0.055*** (0.020)	-0.115*** (0.018)	0.051*** (0.018)	-0.117*** (0.020)	0.046** (0.019)	-0.094*** (0.022)
<i>Teacher Delivery</i> (T2)	0.061*** (0.019)	-0.096*** (0.020)	0.087*** (0.019)	-0.041** (0.020)	0.091*** (0.019)	-0.012 (0.022)
Stepdown P Value (T1)	0.018	0.002	0.018	0.002	0.018	0.002
Stepdown P Value (T2)	0.008	0.002	0.002	0.052	0.002	0.513
N	23,049	23,267	25,142	25,111	25,140	25,064
R ²	0.333	0.225	0.491	0.443	0.461	0.457

Notes: All regressions control for a fourth-order polynomial of each corresponding pre-treatment outcome, demographic controls (including age, gender, school year, and ethnicity), and a full set of municipality fixed effects. Samples correspond to the subset of students without any missing responses for each corresponding socio-emotional skills scale and control variable. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places.

It is also possible that there was an influence of our intervention on student self-reports operating through another channel. In particular, responses may be subject to reference bias, i.e. the tendency for self-reports to be influenced by the individual's reference points and social context (Heine et al., 2008, 2002). For example, when a student is asked about the extent to which they agree or disagree with a survey prompt such as "I often set a goal but later choose to pursue a different one," they might implicitly compare themselves to a reference group that can lead to different standards of what it means to be consistent in goal-pursuing. The curriculum exposed students to role models of gritty behavior,²⁰ and may thus have led students in our treatment groups to hold themselves to a higher standard than those in our control group in terms of what a consistent interest looks like. This potential problem with self-reports is not unique to grit or our study. For instance, Dobbie and Fryer (2015) obtain a similar result when assessing the impact of winning a lottery to attend a high-performing charter school on grit and related socio-emotional skills using self-report measures. Similarly, West et al. (2016) find that students who won a lottery to attend two high-achieving charter middle schools also scored lower on self-reported measures of grit and conscientiousness (among other socio-emotional skills). They report suggestive evidence that the lower scores in self-reports may be driven by a change in reference groups and social context.

We cannot dispel that reference bias, correlated with our treatments, may be behind the insignificant or even negative impacts on overall grit. If this were the case, we would expect our average treatment effects for overall grit and its two facets to be biased downwards.²¹

An alternative to using self-reports would have been to use real effort tasks such as those employed by Sutter et al. (2022), Alaoui and Fons-Rosen (2021), and Alan et al. (2019), or teacher questionnaires such as those employed by Boon-Falleur et al. (2022), neither of which rely on student self-reports. However, task-based measures posed challenges to be feasibly and reliably implemented at the scale of our intervention and teacher reports are not immune to response biases themselves (cf. Duckworth and Yeager 2015).²²

²⁰ Some of the examples included in our intervention were internationally famous singer Beyonce, Chess world champion Magnus Carlsen, and young Macedonian high achievers like Katerina Bacheva, winner of the President's National Award for best young scientist.

²¹ See Jachimowicz et al. (2018) for a discussion of psychological considerations surrounding the definition and development of grit and its two facets, as well as other articles in the same issue of PNAS.

²² We implemented a (non-incentivized) real effort task when collecting additional behavioural measures of grit-related outcomes two weeks post-treatment. However, it exhibited high attrition and lacked both reliability and validity, which is why we refrain from using it.

Finally, it is possible that the intervention did lead children to be less committed to sustain interests over long periods, and that this was an unintended consequence. The curriculum exposed students to the value of deliberate practice applied repeatedly to a stretch goal, without specifying the goal itself. The role models used encompassed a variety of backgrounds, including sports, arts, and academia. The intervention did not intentionally target a specific interest or long-term goal, albeit being applied in the education setting it intended to cultivate grit across learning domains. It is possible that this made students less able or inclined to sustain consistent interests for longer periods because it made them more self-aware of the importance of identifying a passion (long-term goal) that they are willing to commit to, and more inclined to sample various interests before committing (cf. Sturman and Zappala-Piemme, 2017). Recent research suggests that developing consistent interests later in life may, paradoxically, depend on diversification (i.e. sampling) of interests earlier in life (cf. Gopnik, 2020; Côté and Erickson, 2015).²³

C. Impacts on Academic Achievement

Did improvements in students' deliberate practice beliefs, grit-effort and related socio-emotional skills translate into improvements in their academic achievement? Table V Panels A to C show the impacts of our intervention on GPAs averaged across math, English, and first language, in the short-term right after our intervention (fourth quarter of the school year 2015/2016), in the medium-term (first semester of the school year 2016/2017), and in the long-term (second semester of the school year 2016/2017). Note that our sample is much larger for GPAs than for socio-emotional skills, as GPAs rely on administrative data from official school records and are not subject to missing information. Appendix III shows the results for a restricted balanced sample of individuals with survey and administrative data.

First, looking at short-term GPAs (Panel A), we find only small impacts which, in most cases, do not reach significance at conventional levels. There is an improvement of 2% SD, on average, for students in the self-learning treatment, but it is only marginally significant (at the 10% level). The exception are Roma students, for whom we detect a larger, 6% SD increase in short-term

²³ When looking at the correlations between grit-effort and grit-interest at baseline (see Appendix I), we find a small, positive correlation (0.04) in the overall sample. In other studies of grit, its two facets are also found to be weakly correlated (Jachimowicz et al., 2018), and in some contexts the correlation is negative. For instance, a small negative correlation (-0.135, significant at the 5% level) was found in a large sample of Latin-American urban respondents (CAF, 2017). This corroborates that the two facets of grit are distinctive, that the measures are different, and that it is plausible for the intervention to have led to differential impacts on each of them.

GPA which is significant at the 1% level. Except for the Roma, there is little evidence of consistent heterogeneous impacts by gender, ethnicity, or tercile of the pre-treatment distribution of academic achievement.

Next, looking at medium-term and long-term GPAs (Panels B and C), we find some evidence that the impacts of our intervention on academic achievement become stronger over time, particularly for specific groups. Impacts, on average, are estimated at 3% SD (significant at the 5% level) for the teacher-delivery treatment one year post-treatment, with no significant impact for the student self-learning arm. Even if these are still small impacts, they are notable considering the short duration and relatively low intensity of our intervention and that evaluations of human capital interventions often yield fade-out effects over longer time periods (Hart et al 2023; Bailey et al 2020; Bouguen et al 2019).

The estimated average GPA gains for the teacher-delivery treatment reflect the academic improvement of students in the bottom tercile of the pre-treatment GPA distribution and are particularly stronger for Roma students. For both the Roma and the low-achieving students, we find a consistent gradient of GPA gains over time in this higher intensity treatment.²⁴ For Roma students exposed to the teacher-delivery treatment, the impact on GPAs goes from 6% SD in the short-term to 17% SD in the medium term and reaches up to 28% SD a year post treatment. The one-year GPA gains represent 18% of the pre-intervention achievement gap between Roma and Macedonian students. For Roma students exposed to the student self-learning, the impact on GPAs range between 11 to 13% SD half a year to one year post-treatment, compared to no impact in the short-term.

²⁴ The distribution of achievement is left skewed, probably because grades are truncated between 0 and 5 (a relatively narrow range), and most students score above 3. Therefore, an alternative is to use percentile ranks for the achievement analysis, instead of the standardized GPA scores used so far. In the Appendix we show that results are similar regardless of the measure we use.

TABLE V: AVERAGE AND HETEROGENEOUS TREATMENT EFFECTS ON GPAs (Z-SCORES)

	Average	Gender		Ethnicity				Pre-Treatment GPA		
	(1)	Male (2)	Female (3)	Macedonian (4)	Albanian (5)	Roma (6)	Other (7)	Tercile 1 (Low) (10)	Tercile 2 (Middle) (11)	Tercile 3 (High) (12)
<i>Panel A: GPAs, Short-Term (2015/2016 Q4)</i>										
<i>Self-Learning</i> (T1)	0.018* (0.011)	0.019* (0.011)	0.017 (0.012)	0.019 (0.012)	0.017 (0.019)	-0.015 (0.028)	0.012 (0.020)	0.010 (0.013)	0.028* (0.015)	0.013 (0.009)
<i>Teacher Delivery</i> (T2)	0.016 (0.012)	0.013 (0.012)	0.019 (0.013)	0.007 (0.015)	0.020 (0.018)	0.055*** (0.017)	-0.007 (0.023)	0.020 (0.014)	0.016 (0.018)	0.010 (0.008)
Stepdown P (T1)	0.1485	0.1386	0.3663	0.2376	0.7921	0.9703	0.9505	0.8614	0.0792	0.2475
Stepdown P (T2)	0.4158	0.6931	0.3663	0.9802	0.6535	0.0198	1.0000	0.3663	0.8020	0.4554
N	33,454	17,270	16,184	19,460	11,423	1,161	1,410	11,181	11,150	11,123
R ²	0.935	0.933	0.930	0.930	0.925	0.909	0.951	0.668	0.590	0.512
<i>Panel B: GPAs, Medium-Term (2016/2017 Q1+Q2)</i>										
<i>Self-Learning</i> (T1)	0.006 (0.013)	0.013 (0.014)	-0.002 (0.014)	0.015 (0.016)	-0.002 (0.024)	0.109** (0.045)	0.052 (0.036)	0.007 (0.016)	0.015 (0.019)	-0.006 (0.012)
<i>Teacher Delivery</i> (T2)	0.009 (0.013)	0.023 (0.014)	-0.006 (0.014)	0.019 (0.017)	0.000 (0.020)	0.166*** (0.056)	0.014 (0.046)	0.028* (0.017)	0.016 (0.019)	-0.015 (0.012)
Stepdown P (T1)	0.9703	0.7327	1.0000	0.7525	1.0000	0.0198	0.4158	0.9901	0.8317	0.9802
Stepdown P (T2)	0.8614	0.1980	0.9901	0.6040	1.0000	0.0198	1.0000	0.1683	0.8218	0.4455
N	31,310	16,154	15,156	18,568	10,348	1,045	1,349	10,338	10,444	10,528
R ²	0.878	0.871	0.869	0.875	0.850	0.797	0.902	0.498	0.430	0.371
<i>Panel C: GPAs, Long-Term (2016/2017 Q3+Q4)</i>										

<i>Self-Learning</i> (T1)	0.021 (0.018)	0.030 (0.020)	0.009 (0.018)	0.027 (0.017)	0.001 (0.038)	0.129** (0.052)	0.045 (0.034)	0.007 (0.024)	0.043 (0.029)	0.004 (0.010)
<i>Teacher Delivery</i> (T2)	0.030* (0.017)	0.041** (0.019)	0.018 (0.017)	0.042** (0.021)	0.006 (0.031)	0.279*** (0.055)	0.034 (0.033)	0.056** (0.022)	0.044 (0.028)	-0.009 (0.009)
Stepdown P (T1)	0.5743	0.2475	0.9703	0.2376	1.0000	0.0198	0.4158	1.0000	0.2475	0.9901
Stepdown P (T2)	0.1287	0.0396	0.7327	0.0495	1.0000	0.0099	0.7327	0.0198	0.1980	0.7525
N	31,437	16,209	15,228	18,716	10,402	985	1,334	10,247	10,524	10,666
R ²	0.854	0.849	0.842	0.853	0.819	0.787	0.894	0.454	0.422	0.233

Notes: All regressions control for pre-treatment GPAs (up to a fourth-order polynomial), demographic controls (including age, gender, school year, and ethnicity), and a full set of municipality fixed effects. Robust standard errors clustered at the school level in parentheses. All figures are rounded to three decimal places.

The consistent pattern of strong gains in academic achievement for Roma students is remarkable. This mirrors findings from other psychological interventions in education which show stronger benefits for more disadvantaged students, although these often fade out over time (Yeager et al., 2019; Sisk et al., 2018; Paunesku et al., 2015; Cohen et al., 2009; Hulleman and Harackiewicz, 2009; Wilson and Linville, 1982). Recall that we found positive impacts of the intervention (treatment 2) on the socio-emotional skills of Roma students, which although imprecisely measured are statistically indistinguishable from the positive and significant effects for Macedonian students. We find similar results when we analyzed the intervention’s impacts on deliberate practice and grit-perseverance of Roma students, as reported in Table A.II.3, and in our robustness checks below.

Taken together, the results indicate that our intervention seems to have cultivated grit-related socio-emotional skills as well as improved academic achievement, especially amongst disadvantaged Roma students. This suggests that our intervention’s hypothesized mechanism – that by fostering grit-related socio-emotional skills, we could boost the academic achievement of students – operated more strongly among students with larger gaps in socio-emotional skills and achievement.²⁵

C. Robustness

C.1. Survey Non-Response and Attrition

So far, our analysis of socio-emotional skills included all students who have complete surveys and used all observations that are available for each outcome, whereas our analysis of GPAs included all students who are present in official school records and, therefore, should have participated in our intervention according to the randomization protocol. While this maximizes sample size, it also results in different estimation samples. We now look at survey non-response and attrition both in survey and administrative data and show that our results remain robust to accounting for both.

Administrative Data. There is little attrition in our administrative data, which come from the official grade reporting system by the North Macedonian government, and which is mandatory for teachers. There were 35,340 sixth and seventh-grade students who were randomized to be part of our intervention in the school year 2015/2016. Out of this initial sample, for 33,454 students (95% of the total) we can study short-term impacts on GPAs (in the quarter right after our intervention);

²⁵ The correlations between GPAs and deliberate practice scores and grit-perseverance amongst Roma students are similar (even stronger) to those in the overall student population (reported in the Appendix), for the various sub-samples (control and treatments) at baseline and after treatment.

for 31,310 students (88%) we can study medium-term impacts (in the first semester of the following school year); and, finally, for 31,437 students (89%) we can study long-term impacts (in the second semester of the following school year, i.e. one year post treatment). The relatively small differences in the number of students across different periods are most likely due to a combination of natural fluctuation of students entering or leaving the education system (e.g. due to external migration, transition to private education, and school dropout).²⁶

Survey Data. In contrast to our administrative data, attrition is much more pronounced in our survey data. One source of attrition is typical survey non-response (i.e. failure of students to complete surveys either partially or completely). Another is that for a number of students some responses were invalid or illegible. A third source has to do with the logistical complexities associated with the nationwide implementation of our intervention. Due to lack of technology in many schools, our surveys were paper-based and printed and delivered to schools ahead of data collection. About 700,000 pages of intervention and data collection materials had to be printed in a time span of only a few weeks so that these could be delivered to schools ahead of the intervention start date.²⁷ Teachers received written instructions (in a letter directed to them and an e-mailed to the principal) on how to distribute surveys, have them completed, and having them labelled with unique student IDs to enable matching with official school records.²⁸ The process was the same for both the baseline (one week before the intervention started) and endline (one week after the intervention had ended) surveys.

Attrition occurred at several points during field work. We identify the most important sources to be: (i) a printing error: a double printing of the endline questionnaire (instead of printing a base-

²⁶ Primary school completion rate was estimated at 93% for 2017 (UNESCO, 2023).

²⁷ Printing was divided between three local printing companies, each of which printed complete school packages (i.e. baseline surveys, endline surveys, workbooks, name lists of students, and stickers with student IDs) defined by treatment and control groups. Printing and packaging were supervised to ensure sufficient materials were printed, appropriately packaged and labelled. Materials were then delivered to schools two weeks prior to the intervention's start date in packages corresponding to each class.

²⁸ Teacher instructions were also included in print in the survey packages. Teachers were instructed to label the completed surveys with stickers containing student IDs. The latter were used as unique identifiers to match baseline and endline surveys and administrative data. Schools were instructed to store completed surveys in the same boxes they were originally delivered until collection (starting two weeks after the intervention ended). All boxes were stored at a central warehouse owned by the survey company, and the survey responses were digitalized over the next few months by the company's data entry team. This process was the same for all intervention materials.

line *and* an endline survey) for about one third of the self-learning treatment arm, presumably because both surveys looked the same except for the socio-economic characteristics module that was at the end of the endline survey. This resulted in missing values for these characteristics, as shown in Table I; (ii) missing materials: boxes containing baseline and/or endline surveys went missing within some schools prior to baseline and/or endline survey collection or by the time of materials retrieval; (iii) compliance: baseline and/or endline surveys were returned without having been completed; (iv) labelling: some completed baseline and/or endline surveys were not properly labelled, making it impossible to uniquely identify students across surveys (and to match their survey data to official school records).

Our analysis of socio-emotional skills uses the sample of students for whom we have both completed and identifiable baseline and endline surveys. Of the initial sample of 35,340 students, about 2,000 students (6% of the total) are dropped because they are missing both the baseline and endline surveys. About 31,000 students (88%) have a valid baseline survey, and about 27,000 (76%) a valid endline survey.

Further, there was an 11% attrition from baseline to endline (about 4,000 students) due to incomplete responses on some of the items of the various socio-emotional skills measures. A total of 29,487 students (83%) have a fully complete Short Grit Scale at baseline, while 25,815 have it at endline (73%). Finally, there are about 3,000 students (8%) who have an endline survey without having a completed baseline survey. Thus, altogether our analysis of socio-emotional skills uses a sample of about 22,000 to 24,000 students (62% to 68% of the initial total student population) who have complete both baseline and endline surveys for deliberate practice beliefs and the Short Grit Scale. The sample reduces to 18,718 students (53%) when we further restrict to those with complete responses for *all* socio-emotional skills measures in our robustness analysis (Appendix I and III show the results of checking for balance in outcomes and individual characteristics and of the regressions for this subsample of students).²⁹

To assess the robustness of our results to survey non-response and attrition, we first examine whether there are systematic patterns of differential attrition between our treatment and control groups, in terms of correlations with observable characteristics. Appendix Table A.III.1 shows the

²⁹ As each scale is multi-item, if a student does not respond to at least one of the items, the score for the entire scale (or sub-scale in the case of grit) gets a missing value. In general, the more items, the greater the chance that the entire scale has a missing value. Our S/E Skills Index is particularly susceptible to this as it relies on full responses on all scales.

results of regressions of binary indicators for whether the different survey instruments (i.e. baseline survey, endline survey, or both) are available on treatment dummies, pre-treatment GPAs (standardized), and demographic and socio-economic characteristics of students (Columns 1, 3, and 5). We also interact the treatment dummies with these observable characteristics (Columns 2, 4, and 6). Table A.III.2 then estimates the same regression for the availability of GPAs over time (i.e. medium-term GPA, long-term GPA, or both).

Table A.III.1 shows that, when it comes to the different survey instruments, there are notable differences in attrition between groups with regard to having a completed endline survey. In particular, the self-learning and the teacher-delivery treatment arms are about 15% and 6% less likely, respectively, to have a completed endline survey. This difference in attrition between groups is primarily related to the printing error mentioned before. This printing error, however, is likely to cause only random measurement error (as opposed to systematic bias) in our analysis of socio-emotional skills due to a reduced sample size, since the schools that received the wrongly printed (duplicate) surveys were a random set of schools scattered across the entire country.

Turning to students' demographic and socio-economic characteristics, there are some differences in attrition with regard to observable characteristics, but for the most part these are small and not systematically different across experimental groups, except for ethnicity. For instance, pre-treatment GPA is positively associated with a higher likelihood of having completed surveys. Yet, the correlation is very weak (i.e., a one to two percentage points higher likelihood for a 1 SD higher pre-treatment GPA) and not significantly different across treatment and control groups. There is, however, a strong correlation between ethnicity and attrition. Potentially worrisome for our results, Roma students in our control group are 20% and 33% more likely than Macedonian students to lack a completed baseline and endline survey, respectively, and 40% more likely to be missing both surveys. Similar results, although with smaller effects, hold for Albanian students. On the contrary, Roma students in the self-learning treatment arm are 25% *more* likely to have a completed endline survey compared to Macedonian students or other ethnic groups. There is no differential survey-related attrition by ethnicity in the teacher-delivery treatment arm. This point towards differential attrition for Roma (and to a lesser extent Albanian) vis-à-vis Macedonian students, by experimental group, which may potentially bias our estimated treatment effects for Roma students if exposure to our intervention was somehow correlated with attrition and our outcomes of interest. This could happen, for example, if treated Roma students became more likely to have completed endline surveys because they became grittier and were more judicious in completing grit and other socio-emotional skills questionnaires.

Table A.III.2 shows that, when it comes to differential attrition related to the availability of GPAs over time, we obtain results that are broadly similar, though correlations and associated differential attrition rates are smaller. Recall that attrition in official school records is much smaller given that schooling in North Macedonia is compulsory up to (and including) grade nine. Again, the most relevant results are the correlations with ethnicity across experimental conditions. Roma students in our control group are, on average, between six to 17 percentage points *less* likely than Macedonian students to have a record for medium-term and long-term GPAs, respectively. Meanwhile, Roma in the self-learning and teacher-delivery treatment arms are eleven to 17 percentage points *more* likely than Macedonian students to have a record for long-term GPAs.

In light of these findings, we analyze the sensitivity of our results to attrition. Appendix Table A.III.3 replicates Table V using the (much) smaller sample of students for whom we have completed baseline and endline surveys as well as short-term, medium-term, and long-term GPAs, i.e. a balanced sample between survey and administrative data. This allows us to assess whether differential attrition between survey and administrative data are driving our results. The results using this balanced sample largely corroborate our baseline results; if anything, the impacts are slightly stronger. For instance, for Roma students, impacts on short-term GPAs are three times as large than those in Table V, albeit more imprecisely estimated (though still significant at 1%) as the sample sizes drops substantially.

Next, we study whether the gradient of stronger impacts on the GPAs of Roma students over time is driven by attrition. This would be a problem if higher-achieving Roma students would remain in our sample while lower-achieving ones drop out. Although we do not believe that out-of-sample selection is driving our results (as formal education in North Macedonia is compulsory for the age range targeted by our intervention and dropout rates are not particularly large),³⁰ we nevertheless re-estimate Table V using a balanced panel, including only those students who are observable in the official school records during the entire observation period from school year 2015/2016 to school year 2016/2017. Appendix Table A.III.4 shows our results. We find that the positive gradient of GPAs for Roma students over time as well as the magnitude of impacts remains largely the same. This suggests that attrition and resulting changes in sample composition are unlikely to be driving the gradient of stronger impacts on the GPAs of Roma students over time.

³⁰ For our intervention year, the State Statistical Office reports 192,715 students at the start of the school year, and 190,715 at the end of the school year in both primary and lower secondary, or a 2% attrition of students. No differences are reported by ethnicity.

Likewise, the positive gradient of GPA gains (from initially 0 to 5% SD) for students in the bottom tercile of achievement remains.

Finally, we use a multiple imputation procedure to impute missing data due to attrition in both survey and administrative data. More specifically, we use a multivariate normal regression and an iterative Markov chain Monte Carlo (MCMC) to impute each missing outcome (i.e. deliberate practice beliefs, grit and its two facets, our S/E Skills Index, as well as short-term, medium-term, and long-term GPAs) using students' age, gender, ethnicity, and school year. Appendix Tables A.III.5 through A.III.8 re-estimate Tables III and V and Tables A.II.1 and A.II.2 using the imputed missing data in the resulting much larger samples (without attrition). As seen, our results remain qualitatively the same. For the most part, our point estimates are somewhat smaller but preserve their statistical significance. Importantly, the strong positive impacts on Roma students' short-term, medium-term, and long-term GPAs remain largely the same.

C.2 Observer (Hawthorne) or Experimenter-Demand Effects

Another potential source of bias in our results is that teachers and students, both of whom were not entirely blind to the experiment, may have changed their behavior merely because of being participants in our intervention, rather than because of the actual contents of the curriculum.

Although we cannot directly analyze and fully dispel that observer (Hawthorne) or experimenter-demand effects may have played a role for either teachers or students, we argue that they are unlikely to be driving our results, for several reasons. When it comes to students, our intervention was embedded into an existing "Life Skills" curriculum which was already implemented in schools. Students were familiar with the lessons from this curriculum and were used to them being taught during Monday morning class hours (yet without any expectation of upcoming content).³¹ Moreover, there were never experimenters or intervention facilitators present at any point during our intervention. Students were not monitored either within sessions or outside, beyond their headteachers being present. Hence, our intervention should not have been particularly salient amongst students, thereby minimizing potential observer or experimenter-demand effects, particularly that could last a year post-treatment. Finally, baseline and endline surveys were collected before and after these sessions, with a timely spacing (i.e. one week before intervention start and

³¹ In fact, the local psychologist working on the adaptation of our intervention to the local context was also responsible for the design of the "Life Skills" curriculum.

one week after intervention end), to minimize any bias due to immediate priming arising from the exposure to the material itself.

When it comes to teachers, such effects are particularly relevant in the teacher-delivery treatment arm and for the impacts on GPAs. The findings that the teacher-delivery treatment arm was indeed more effective could be because *(i)* students simply absorbed the curriculum content more effectively, *(ii)* teachers themselves absorbed the content and incorporated it into their teaching and interactions with students, or *(iii)* teachers – weary of being part of an intervention – simply changed their grading behavior. While *(i)* and *(ii)* are arguably part of the genuine and intended goal of our intervention, *(iii)* is an experimental artefact. Again, although we cannot fully dispel *(iii)* with certainty, we argue that it is unlikely changes in grading behavior occur or can be a driver behind our results. The “treated” teachers were the headteachers of the respective class, and they typically teach only one subject, which does not necessarily correspond to the subjects we used to construct GPAs. Moreover, we constructed GPAs across several subjects taught by multiple teachers (i.e. math, English, and first language), thereby reducing the relative importance of a single subject in the overall GPA and hence of experimenter-demand effects. Finally, when we re-estimated our models using GPAs taken across *all* subjects (not only math, English, or first language), we obtain similar treatment effect results, which reassures that our findings are not driven by changes in teachers grading behavior.³²

VI. CONCLUSION

Our study designs, implements, and rigorously evaluates a nationwide school-based grit intervention. To our knowledge, this is the first experimental evaluation of a nationwide school-based program targeting the socio-emotional skills of children. The intervention targeted all middle-school students in North Macedonia as part of their regular school curriculum during the second semester of the 2015-2016 school year. Our RCT design comprised two randomized treatment arms with different modalities of delivery of the program embedded into an existing life-skills curriculum: a self-learning arm where students received the materials in the classroom and went over these on their own, and a teacher-delivered arm where the material was instead delivered through classroom teacher instruction. The control group of students received only the pre-existing life-skills curriculum.

³² More details on these results and the subjects taught by headteachers are available upon request.

The results of our intervention are mixed. There were strong positive impacts of the intervention on deliberate practice beliefs, which were especially large for the teacher-delivery treatment. There were also small positive average impacts on academic motivation. There were, however, very small impacts on the remaining dimensions (indistinguishable from zero), except on the case of grit, where the impact of the self-learning treatment was negative. When we disentangle the impacts on grit across its two facets—consistency of interest and perseverance of effort, we find suggestive evidence that the negative impacts could be due to measurement issues caused by the negative phrasing of the survey items for the grit-consistency of interest facet. If we restrict ourselves to positively phrased statements in the grit scale, which then narrows the focus to the scale containing only perseverance of effort items, the impact of the two program modalities is positive. Although we cannot be sure that this measurement issue explains the negative impacts of the self-learning treatment arm on grit, these results are consistent with others in the educational and psychological literature that raise concerns about response bias associated with the use of negatively phrased statements. In fact, we find a similar pattern of negative impacts when we analyze negatively phrased items in our other socio-emotional scales. We should also note that whereas consistency of interest is characteristic of high achievement in adulthood, it is typically preceded by an extended period of sampling among diverse interests during the teen and youth years (Duckworth, 2016; Güllich et al., 2021). Thus, while it is developmentally appropriate to cultivate mindsets and skill sets supporting perseverance of effort in school-age children, it could be, in our view, potentially counterproductive to encourage them to commit prematurely to specific passions.

While our measures of grit and related socio-emotional skills are self-reported, we observe impacts on GPAs (measured from official government records) in the teacher-delivery arm, which alleviate possible concerns about the estimated impacts of the intervention being spurious. Notably, while average impacts on GPA are small, the impacts are very large among Roma students, the most disadvantaged group of students in our study population in North Macedonia (and Europe's largest ethnic minority). In the teacher-delivery treatment arm, Roma students experienced gains in GPAs of up to 28% SD one year post-treatment. These gains grew larger over time, roughly doubling every semester in the period observed.

Our findings confirm previous results that at least the perseverance-of-effort facet of grit is malleable and can be taught (Alan et al., 2019), and we further show that it is possible to do this cost-effectively through a nationwide school-based relatively low intensity intervention. They also add to the evidence that socio-emotional skills interventions can have positive impacts on academic achievement, and that, these gains can be sustained or even grow over time. These impacts are

notably meaningful for disadvantaged Roma students – the intervention reduced about 18% of the pre-intervention achievement gap in terms of GPAs between Roma and Macedonian students. This underscores the potential for school-based interventions to improve equity in educational outcomes by developing grit and related socio-emotional skills in ways that empower and propel disadvantaged students to persist through the setbacks that are inherent to academic success. Of course, this does not preclude the need for other policies that improve access to quality education and tackle other root causes of socio-economic inequality.

The magnitudes of impacts found compare favorably with other educational interventions for improving socio-emotional skills and are consistent with a mounting body of evidence that grit and related socio-emotional skills often benefit disproportionately disadvantaged students. In terms of GPAs, our impacts on Roma students are comparable in size to those found in a recent meta-analysis of socio-emotional skills interventions amongst disadvantaged groups (34% SD, Sisk et al., 2018), and are larger than impacts amongst disadvantaged students in other programs, e.g. 18% SD in a standardized test in math in Indonesia, where a growth mindset program showed no impacts on GPAs (Johnson et al., 2020; World Bank, 2019); 10% SD in Peru’s “Expande Tu Mente” program (Outes et al., 2020). Notably, Alan and Ertac (2019) found average positive effects of a grit intervention of 23% SD gains in a standardized test in math 2.5 years after the intervention in participating schools in Istanbul, Turkey. Yeager et al. (2019) estimated a standardized mean difference effect size of 0.11 for core GPA courses in a nationally representative sample of students in secondary education in the US.

Moreover, our intervention was quite cost-effective. When considering only Roma students for whom we find impacts on GPAs one year post-treatment and allocating (pro-rated) direct costs of the intervention accordingly (all cost categories included), we find that our intervention cost about 3.7 USD for a 0.1 SD increase in annual GPAs of Roma students. Excluding design and evaluation costs, the cost-effectiveness ratio translates into about 1 USD per 0.1 SD increase in GPA.³³ These ratios compare very favorably with cost-effectiveness ratios in the literature. For example, Glewwe and Muralidharan (2016) find that incentive schemes (for both students and teachers) cost between

³³ The total direct costs of our intervention were USD 343,616, which mainly comprise the salaries of the team that designed and oversaw the implementation of the intervention and conducted the data analysis, the costs of developing, printing, and delivering materials, and the design and delivery of the teacher training. With a total of 34,454 students in our main analysis, this yields a cost per student of about USD 10.3 or USD 2.7 when excluding costs of design and evaluation. There are 1,161 Roma students in total. Hence, we obtain a cost-effectiveness ratio of 3.7 USD per 0.1 SD improvement in annual GPAs of Roma students when including all costs or 1 USD per 0.1 SD improvement when excluding costs of design and evaluation.

1 and 3 USD per 0.1 SD improvement in test scores, cash transfers conditional on school attendance between 77 USD and 138 USD for a comparable improvement, and pedagogy-supporting classroom-IT about 30 USD per 0.1 SD improvement. These results are illustrative only, of course, as implementation and other costs are likely to vary considerably by country.

While we cannot conclusively pinpoint the exact mechanisms that drive our results, we believe that they can be attributed to two main channels. First, the intervention positively impacted students' expectancies and values attached to sustained effortful behavior, which made them engage in deliberate practice and become grittier, in the industriousness (perseverant) sense. Both treatments had positive impacts on these dimensions, with effects larger when involving teachers directly in the delivery of the grit curriculum (a finding that is consistent with Alan and Ertac, 2019). Second, our intervention may also have addressed stereotype threat by providing positive role models and counter-stereotypical examples of pursuing challenges and achieving success despite adversity, which might also be behind the higher impacts found amongst Roma students. Considering key aspects of the design and implementation of the intervention as well as evidence discussed in this paper, it is unlikely that our results are driven by observer or experimenter-demand effects or simply teachers grading more leniently (particularly Roma students) as a result of exposure to our intervention.

The results from our intervention bring to the fore important questions for future research around how to foster grit and other socio-emotional skills amongst students. The potential measurement issues surrounding self-reported scales of socio-emotional skills highlight the need for more research to develop more robust measures that can be utilized at scale and in school settings. If the lack of positive impacts on the consistency of interest facet of grit is not entirely spurious, there remains a question of how and when this facet of grit can be nurtured. Children and young adults may need to try out a variety of interests before developing consistent ones later in life. Whether the path to becoming grittier requires experimentation and being less committed to a specific interest early in life is an important avenue for future research. More generally, while our paper contributes to a better understanding of the potential and challenges to cultivate grit in schools and at scale, further work is warranted to design, implement, and evaluate other cost-effective ways to develop this skill, including in combination with other socio-emotional skills that serve the diverse educational and socio-emotional learning needs of children.

REFERENCES

- Acosta, P. M., & N. Muller, N. (2018). *The role of cognitive and socio-emotional skills in labor markets*. IZA World of Labor.
- Alan, S., T. Boneva, & S. Ertac (2019). Ever Failed, Try Again, Succeed Better: Results from a Randomized Educational Intervention on Grit. *Quarterly Journal of Economics*, 134(3), 1121-1162.
- Alaoui, L., & C. Fons-Rosen (2021). Know when to fold'em: The flip side of grit. *European Economic Review*, 136, 103736.
- Allensworth, E., & J. Q. Easton (2005). *The on-track indicator as a predictor of high school graduation*. Consortium on Chicago School Research. University of Chicago.
- Almlund, M., A. L. Duckworth, J. J. Heckman, & T. Kautz (2011). *Personality Psychology and Economics*. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.). *Handbook of the Economics of Education* (pp. 1-181).
- Bailey, D. H., Duncan, G. J., Cunha, F., Foorman, B. R., & Yeager, D. S. (2020). Persistence and Fade-Out of Educational-Intervention Effects: Mechanisms and Potential Solutions. *Psychological Science in the Public Interest*, 21(2), 55-97.
- Beilock, S. L., & M. S. DeCaro (2007). From poor performance to success under stress: working memory, strategy selection, and mathematical problem solving under pressure. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 33(6), 983-998.
- Benner, A. D., & S. Graham (2011). Latino Adolescents' Experiences of Discrimination Across the First 2 Years of High School: Correlates and Influences on Educational Outcomes. *Child Development*, 82(2), 508-519.
- Ben-Zeev, T., S. Fein, & M. Inzlicht (2005). Arousal and stereotype threat. *Journal of Experimental Social Psychology*, 41(2), 174-181.
- Blattman, C., Jamison, J., and Sheridan, M. (2017). Reducing Crime and Violence: Experimental Evidence from Cognitive Behavioral Therapy in Liberia, *American Economic Review*, 107(4), 1165-1206.
- Boon-Falleur, M., A. Bouguen, A. Charpentier, Y. Algan, E. Huillery, & C. Chevallier (2022). Simple questionnaires outperform behavioral tasks to measure socio-emotional skills in students. *Scientific Reports*, 12, 442.
- Borghans, L., A. L. Duckworth, J. J. Heckman, & B. ter Weel (2008a). The Economics and Psychology of Personality Traits. *Journal of Human Resources*, 43(4), 972-1059.
- Borghans, L., H. Meijers, & B. ter Weel (2008b). The role of noncognitive skills in explaining cognitive test scores. *Economic Inquiry*, 46(1), 2-12.
- Bouguen, Adrien and Huang, Yue and Kremer, Michael R. and Miguel, Edward, Using Randomized Controlled Trials to Estimate Long-Run Impacts in Development Economics (August 2019). *Annual Review of Economics*, Vol. 11, pp. 523-561, 2019.
- Broda, M., J. Yun, B. Schneider, D. S. Yeager, G. M. Walton, & M. Diemer (2018). Reducing inequality in academic success for incoming college students: A randomized trial of growth mindset and belonging interventions. *Journal of Research on Educational Effectiveness*, 11, 317-338.
- CAF (2017). *Encuesta CAF 2017: Trayectorias Laborales y Productivas en América Latina*. Online: <http://scioteca.caf.com/handle/123456789/1400>.
- Cipriano C, Strambler MJ, Naples LH, Ha C, Kirk M, Wood M, Sehgal K, Zieher AK, Eveleigh A, McCarthy M, Funaro M, Ponnock A, Chow JC, Durlak J. The state of evidence for social and emotional learning: A contemporary meta-analysis of universal school-based SEL interventions. *Child Dev*. 2023 Sep-Oct;94(5):1181-1204.

- Cohen, G. L., J. Garcia, V. Purdie-Vaughns, N. Apfel, & P. Brzustoski (2009). Recursive Processes in Self-Affirmation: Intervening to Close the Minority Achievement Gap. *Science*, *324*(5925), 400-403.
- Côté, J., & K. Erickson (2015). *Diversification and deliberate play during the sampling years*. In J. Baker & D. Farrow (Eds.). Routledge Handbook of Sport Expertise (pp. 305-316).
- Credé, M., M. C. Tynan, & P. D. Harms (2016). Much ado about grit: A meta-analytic synthesis of the grit literature. *Journal of Personality and Social Psychology*, *113*(3), 492-511.
- Crosnoe, R. (2011). *Fitting in, standing out: Navigating the social challenges of high school to get an education*. Cambridge: Cambridge University Press.
- Datu, J. (2021). Beyond Passion and Perseverance: Review and Future Research Initiatives on the Science of Grit. *Frontiers in Psychology*, *11*, 545526.
- Dobbie, W., & R. G. Fryer (2015). The Medium-Term Impacts of High-Achieving Charter Schools. *Journal of Political Economy*, *123*(5), 985-1037.
- Duckworth, A. L. (2016). *Grit: The power of passion and perseverance*. Simon & Schuster.
- Duckworth, A. L., & D. S. Yeager (2015). Measurement matters: Assessing personal qualities other than cognitive ability for educational purposes. *Educational Researcher*, *44*(4), 237-251.
- Duckworth, A. L., & P. D. Quinn (2009). Development and validation of the Short Grit Scale (GRIT-S). *Journal of Personality Assessment*, *91*(2), 166-174.
- Duckworth, A. L., P. D. Quinn, & E. Tsukayama (2021). Revisiting the Factor Structure of Grit: A Commentary on Duckworth and Quinn (2009). *Journal of Personality Assessment*, *103*(5), 573-575.
- Duckworth, A. L., C. Peterson, M. D. Matthews, & D. R. Kelly (2007). Grit: Perseverance and passion for long-term goals. *Journal of Personality and Social Psychology*, *92*(6), 1087-1101.
- Duckworth, A. L., P. D. Quinn, & M. E. P. Seligman (2009). Positive Predictors of Teacher Effectiveness. *Journal of Positive Psychology*, *4*(6), 540-547.
- Duckworth, A. L., T. A. Kirby, E. Tsukayama, H. Berstein, & K. A. Ericsson (2011). Deliberate practice spells success: Why grittier competitors triumph at the National Spelling Bee. *Social Psychological and Personality Science*, *2*(2), 174-181.
- Duckworth, A. L., T. S. Gendler, & J. J. Gross (2014). Self-control in school-age children. *Educational Psychologist*, *49*(3), 199-217.
- Dweck, C. S. (1999). *Self-theories: Their role in motivation, personality, and development*. Taylor and Francis/Psychology Press.
- Ericsson, K. A., R. T. Krampe, & C. Tesch-Römer (1993). The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, *100*(3), 363-406.
- Ericsson, K. A. (2008). Deliberate practice and acquisition of expert performance: a general overview. *Academic Emergency Medicine*, *15*(11), 988-994.
- Eskreis-Winkler, L., E. P. Shulman, S. A. Beal, & A. L. Duckworth (2014). The grit effect: predicting retention in the military, the workplace, school and marriage. *Frontiers in Psychology*, *5*(36), 1-12.
- Eskreis-Winkler, L., E. P. Shulman, V. Young, E. Tsukayama, S. M. Brunwasser, & A. L. Duckworth (2016). Using Wise Interventions to Motivate Deliberate Practice. *Journal of Personality and Social Psychology*, *111*(5), 728-744.
- Frederick, S., G. Loewenstein, & T. O'Donoghue (2002). Time Discounting and Time Preference: A Critical Review. *Journal of Economic Literature*, *40*(2), 351-401.

- Friedel, J.M., Cortina, K.S., Turner, J.C. and Midgley, C., 2007. Achievement goals, efficacy beliefs and coping strategies in mathematics: The roles of perceived parent and teacher goal emphases. *Contemporary Educational Psychology*, 32(3), pp.434-458.
- Gatti, R., S. Karacsony, K. Anan, C. Ferré, & C. de Paz Nieves (2016). *Being Fair, Faring Better: Promoting Equality of Opportunity for Marginalized Roma. Directions in Human Development*. World Bank Report. Washington, DC: World Bank.
- Glewwe, P., & K. Muralidharan (2016). *Improving Education Outcomes in Developing Countries: Evidence, Knowledge Gaps, and Policy Implications*. In E. A. Hanushek, S. Machin, & L. Woessmann (Eds.). *Handbook of the Economics of Education* (pp. 653-743).
- Good, C., J. Aronson, & M. Inzlicht (2003). Improving adolescents' standardized test performance: An intervention to reduce the effects of stereotype threat. *Journal of Applied Developmental Psychology*, 24(6), 645-662.
- Gopnik, A. (2020). Childhood as a solution to explore–exploit tensions. *Philosophical Transactions of the Royal Society B*, 375(1803), 20190502.
- Grossman, J. and Tierney, J. (1998)., Does Mentoring Work? An Impact Study of the Big Brothers Big Sisters Program, *Evaluation Review*, 22(3), 403-426.
- Güllich, A., B. N. Macnamara, & D. Z. Hambrick (2021). What Makes a Champion? Early Multidisciplinary Practice, Not Early Specialization, Predicts World-Class Performance. *Perspectives on Psychological Science*, 17(1), 6-29.
- Gunderson, E. A., S. J. Gripshover, C. Romero, C. S. Dweck, S. Goldin-Meadow, & S. C. Levine (2013). Parent Praise to 1-3 Year-Olds Predicts Children's Motivational Frameworks 5 Years Later. *Child Development*, 84(5), 1526-1541.
- Gupta, S., L. H. Supplee, D. Suskind, & J. A. List (2021). *Failed to Scale: Embracing the Challenge of Scaling in Early Childhood*. Mimeo.
- Gutman, L. M., & I. Schoon (2013). *The Impact of Non-Cognitive Skills on Outcomes for Young People: Literature Review*. Education Endowment Foundation.
- Haimovitz, K. and Dweck, C.S., 2016. Parents' views of failure predict children's fixed and growth intelligence mind-sets. *Psychological Science*, 27(6), pp.859-869.
- Harris, C., & D. Laibson (2013). Instantaneous Gratification. *Quarterly Journal of Economics*, 128(1), 205-248.
- Hart, Emma R., Drew H. Bailey, Sha Luo, Pritha Sengupta, and Tyler W. Watts. (2023). Do Intervention Impacts on Social-Emotional Skills Persist at Higher Rates than Impacts on Cognitive Skills? A Meta-Analysis of Educational RCTs with Follow-Up. (EdWorkingPaper: 23-782). Retrieved from Annenberg Institute at Brown University: <https://doi.org/10.26300/7j8s-dy98>
- Heckman, J. J., & T. Kautz (2014). *Fostering and measuring skills: Interventions that improve character and cognition*. In J. J. Heckman, J. E. Humphries, & T. Kautz (Eds.) *The Myth of Achievement Tests: The GED and the Role of Character in American Life* (pp. 341-430). Chicago: University of Chicago Press.
- Heckman, J. J., J. Stixrud, & S. Urzua (2006). The Effects of Cognitive and Noncognitive Abilities on Labor Market Outcomes and Social Behavior. *Journal of Labor Economics*, 24(3), 411-482.
- Heine, S. J., E. E. Buchtel, & A. Norenzayan (2008). What Do Cross-National Comparisons of Personality Traits Tell Us?: The Case of Conscientiousness. *Psychological Science*, 19(4), 309-313.
- Heller, S., Shah, A., Guryan, J., Ludwig, J., Mullainathan, S., and Pollack, H (2017), Thinking, Fast and Slow? Some Field Experiments to Reduce Crime and Dropout in Chicago, *Quarterly Journal of Economics*, 132(1), 1-54.
- Hwang, M. H., & Nam, J. K. (2021). Enhancing grit: Possibility and intervention strategies. In L. E. Van-Zyl, C. Olckers, & L. Van-der-Vaart (Eds.), *Multidisciplinary perspectives on grit* (pp. 77–93). Springer.

- Hulleman, C. S., & Harackiewicz, J. M. (2009). Promoting Interest and Performance in High School Science Classes. *Science*, 326(5958), 1410-1412.
- Imbens, G. W., & J. M. Wooldridge (2009). Recent Developments in the Econometrics of Program Evaluation. *Journal of Economic Literature*, 47(1), 5-86.
- Ivcevic, Z., & M. Brackett (2014). Predicting school success: Comparing Conscientiousness, Grit, and Emotion Regulation Ability. *Journal of Research in Personality*, 52, 29-36.
- Jachimowicz, J. M., A. Wihler, E. R. Bailey, & A. D. Galinsky (2018). Why grit requires perseverance and passion to positively predict performance. *Proceedings of the National Academy of Sciences*, 115(40), 9980-9985.
- Jackson, C. K. (2018). What Do Test Scores Miss? The Importance of Teacher Effects on Non-Test Score Outcomes. *Journal of Political Economy*, 126(5), 2072-2107.
- Johnson, H., D. Pinzón, K. Trzesniewski, T. Indrakesuma, R. Vakis, E. Perova, N. Muller, S. De Martino, & D. 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. Washington, DC: World Bank.
- Kautz, T., J. J. Heckman, R. Diris, B. ter Weel, & L. Borghans (2015). *Fostering and Measuring Skills: Improving Cognitive and Socio-emotional Skills to Promote Lifetime Success*. OECD Policy Report. OECD: Paris.
- Kertesi, G. and Kézdi, G., (2011). The Roma/non-Roma test score gap in Hungary. *American Economic Review*, 101(3), pp.519-525.
- Kinlaw, C.R. and Kurtz-Costes, B., 2007. Children's theories of intelligence: Beliefs, goals, and motivation in the elementary years. *The Journal of General Psychology*, 134(3), pp.295-311.
- Kraft, M. A. (2019). Teacher Effects on Complex Cognitive Skills and Social-Emotional Competencies. *Journal of Human Resources*, 54(1), 1-36.
- Lam, K. K. L., & Zhou, M. (2022). Grit and academic achievement: A comparative cross-cultural meta-analysis. *Journal of Educational Psychology*, 114(3), 597–621.
- Levin, V., S. Guallar-Artal, & A. Safier (2016). *Skills for work in Bulgaria: the relationship between cognitive and socioemotional skills and labor market outcomes*. World Bank Report. Washington, DC: World Bank.
- Lindwall, M., Barkoukis, V., Grano, C., Lucidi, F., Raudsepp, L., Liukkonen, J., et al. (2012). Method effects: the problem with negatively versus positively keyed items. *Journal of Personality Assessment*. 94, 196–204.
- List, J. A. (2022). *The Voltage Effect*. London: Penguin.
- List, J. A. (2024). “Optimally Generate Policy-Based Evidence Before Scaling.” *Nature* (626): 491–499.
- Liu, S., P. Liu, M. Wang, & B. Zhang (2021). Effectiveness of stereotype threat interventions: A meta-analytic review. *Journal of Applied Psychology*, 106(6), 921-949.
- Maddi, S. R., M. D. Matthews, D. R. Kelly, V. Brandilynn, & M. White (2012). The role of hardiness and grit in predicting performance and retention of USMA cadets. *Military Psychology*, 24(1), 19-28.
- Murphy, M. C., C. M. Steele, & J. J. Gross (2007). Signaling threat: how situational cues affect women in math, science, and engineering settings. *Psychological Science*, 18(10), 879-885.
- Naemi, B., E. Gonzalez, J. Bertling, A. Betancourt, J. Burrus, P. C. Kyllonen, J. Minsky, P. Lietz, E. Klieme, S. Vieluf, J. Lee, & R. D. Roberts (2013). *Large-Scale Group Score Assessments: Past, Present, and Future*. In D. H. Saklofske, C. R. Reynolds, & V. Schwane (Eds.). *Oxford Handbook of Child Psychological Assessment* (pp. 129-149), Oxford: Oxford University Press.
- Oreopoulos, P., Brown, S., and Lavecchia, A. (2017). Pathways to Education: An Integrated Approach to Helping At-Risk High School Students, *Journal of Political Economy*, 125(4).

- Outes-Leon, I., A. Sánchez, & 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.
- Paulhus, D. L. (1991). Measurement and control of response bias. In J. P. Robinson, P. R. Shaver, & L. S. Wrightsman (Eds.), *Measures of personality and social psychological attitudes* (pp. 17–59). Academic Press.
- Paunesku, D., G. M. Walton, C. Romero, E. N. Smith, D. S. Yeager, & C. S. Dweck (2015). Mind-Set Interventions Are a Scalable Treatment for Academic Underachievement. *Psychological Science*, 26(6), 784-793.
- Petek, N., & N. G. Pope (2023). The Multidimensional Impact of Teachers on Students. *Journal of Political Economy*, 131(4), 1057-1107.
- Peters, L. H., E. J. O'Connor, & C. J. Rudolf (1980). The behavioral and affective consequences of performance-relevant situational variables. *Organizational Behavior and Human Performance*, 25(1), 79-96.
- Pintrich, P. R. (2003). *Motivation and classroom learning*. In W. M. Reynolds & G. E. Miller (Eds.), *Handbook of Psychology: Educational Psychology* (pp. 103–122). John Wiley & Sons.
- Rimfeld, K., Y. Kovas, P. S. Dale, & R. Plomin (2016). True grit and genetics: Predicting academic achievement from personality. *Journal of Personality and Social Psychology*, 111(5), 780-789.
- Robayo-Abril, M., & N. Millan (2019). *Breaking the cycle of Roma exclusion in the Western Balkans*. Washington, DC: World Bank.
- Roberts, B. W., N. R. Kuncel, R. Shiner, A. Caspi, & L. R. Goldberg (2007). The Power of Personality: The Comparative Validity of Personality Traits, Socioeconomic Status, and Cognitive Ability for Predicting Important Life Outcomes. *Perspectives on Psychological Science*, 2(4), 313-345.
- Roberts, B. W., O. S. Chernyshenko, S. Stark, & L. R. Goldberg (2005). The Structure of Conscientiousness: An Empirical Investigation Based on Seven Major Personality Questionnaires. *Personnel Psychology*, 58(1), 103-139.
- Romano, J. P., & M. Wolf (2005). Exact and approximate stepdown methods for multiple hypothesis testing. *Journal of the American Statistical Association*, 100(469), 94-108.
- Romano, J. P., & M. Wolf (2016). Efficient computation of adjusted p-values for resampling-based stepdown multiple testing. *Statistics & Probability Letters*, 113, 38-40.
- Schmader, T. (2010). Stereotype Threat Deconstructed. *Current Directions in Psychological Science*, 19(1), 14-18.
- Schmidt, F. T. C., J. Fleckenstein, J. Retelsdorf, L. Eskreis-Winkler, & J. Möller (2019). A German Validation and a Domain-Specific Approach to Grit. *European Journal of Psychological Assessment*, 35(3), 436-447.
- Schmidt, F. T. C., C. M. Lechner, & D. Danner (2020). New wine in an old bottle? A facet-level perspective on the added value of Grit over BFI–2 Conscientiousness. *PLoS One*, 15(2), e0228969.
- Sisk, V. F., A. P. Burgoyne, J. Sun, J. L. Butler, & B. N. Mcnamara (2018). To What Extent and Under Which Circumstances Are Growth Mind-Sets Important to Academic Achievement? Two Meta-Analyses. *Psychological Science*, 29(4), 549-571.
- Skinner, E. A., J. G. Wellborn, & J. P. Connell (1990). What it takes to do well in school and whether I've got it: A process model of perceived control and children's engagement and achievement in school. *Journal of Educational Psychology*, 82(1), 22-32.
- State Statistical Office (2017) Primary, lower secondary and upper secondary schools at the end of the school year 2016/2017
- Stecher, B. M., & L. S. Hamilton (2014). *Measuring Hard-to-Measure Student Competencies: A Research and Development Plan*. RAND Corporation.

- Steele, C. M., & J. Aronson (1995). Stereotype threat and the intellectual test performance of African Americans. *Journal of Personality and Social Psychology*, 69(5), 797-811.
- Steinmann, I., Sánchez, D., Van Laar, S. & J. Braeken (2022) The impact of inconsistent responders to mixed-worded scales on inferences in international large-scale assessments, *Assessment in Education: Principles, Policy & Practice*, 29:1, 5-26
- Sturman, E. D., & K. Zappala-Piemme (2017). Development of the grit scale for children and adults and its relation to student efficacy, test anxiety, and academic performance. *Learning and Individual Differences*, 59, 1-10.
- Sutter, M., A. Untertrifaller, & C. Zoller (2022). Grit increases strongly in early childhood and is related to parental background. *Scientific Reports*, 12, 3561.
- Todd, P. E., & Wolpin, K. I. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *Economic Journal*, 113(485), F3-F33.
- Tough, P. (2013). *How Children Succeed: Grit, Curiosity, and the Hidden Power of Character*. Mariner Books.
- UNESCO (2023) Institute for Statistics (UIS Stat) Bulk Data Download Service. Accessed September 19, 2023. apiportal.uis.unesco.org/bdds.
- UNICEF (2022). *Education Pathways in Roma Settlements: Understanding Inequality in Education and Learning*. UNICEF Regional Office for Europe and Central Asia.
- Walton, G. M., & T. D. Wilson, (2018). Wise interventions: Psychological remedies for social and personal problems. *Psychological Review*, 125(5), 617-655.
- West, M. R., M. A. Kraft, A. S. Finn, R. E. Martin, A. L. Duckworth, C. F. O. Gabrieli, & J. D. E. Gabrieli (2016). Promise and paradox: Measuring students' non-cognitive skills and the impact of schooling. *Educational Evaluation and Policy Analysis*, 38(1), 148-170.
- Willingham, W. W. (1985). *Success in College: The Role of Personal Qualities and Academic Ability*. New York: College Board Publications.
- Wilson, T. D., & P. W. Linville (1982). Improving the academic performance of college freshmen: Attribution therapy revisited. *Journal of Personality and Social Psychology*, 42(4), 367-376.
- World Bank (2019). *Instilling a Growth Mindset in Indonesia*. eMBEd Brief. Washington, DC: World Bank.
- Yeager, D. S., & Dweck, C. S. (2020). What can be learned from growth mindset controversies? *American Psychologist*, 75(9), 1269–1284.
- Yeager, D. S., & C. S. Dweck (2012). Mindsets That Promote Resilience: When Students Believe That Personal Characteristics Can Be Developed. *Educational Psychologist*, 47(4), 302-314.
- Yeager D. S. (2019). The National Study of Learning Mindsets, [United States], 2015-2016 (ICPSR 37353) [Data set]. Inter-university Consortium for Political and Social Research. <https://doi.org/10.3886/ICPSR37353.v1>
- Yeager, D. S., P. Hanselman, G. M. Walton, J. S. Murray, R. Crosnoe, C. Muller, E. Tipton, B. Schneider, C. S. Hulleman, C. P. Hinojosa, D. Paunesku, C. Romero, K. Flint, A. Roberts, J. Trott, R. Iachan, J. Buontempo, S. M. Yang, C. M. Carvalho, P. R. Hahn, M. Gopalan, P. Mhatre, R. Ferguson, A. L. Duckworth, & C. S. Dweck (2019). A national experiment reveals where a growth mindset improves achievement. *Nature*, 573, 364-369.